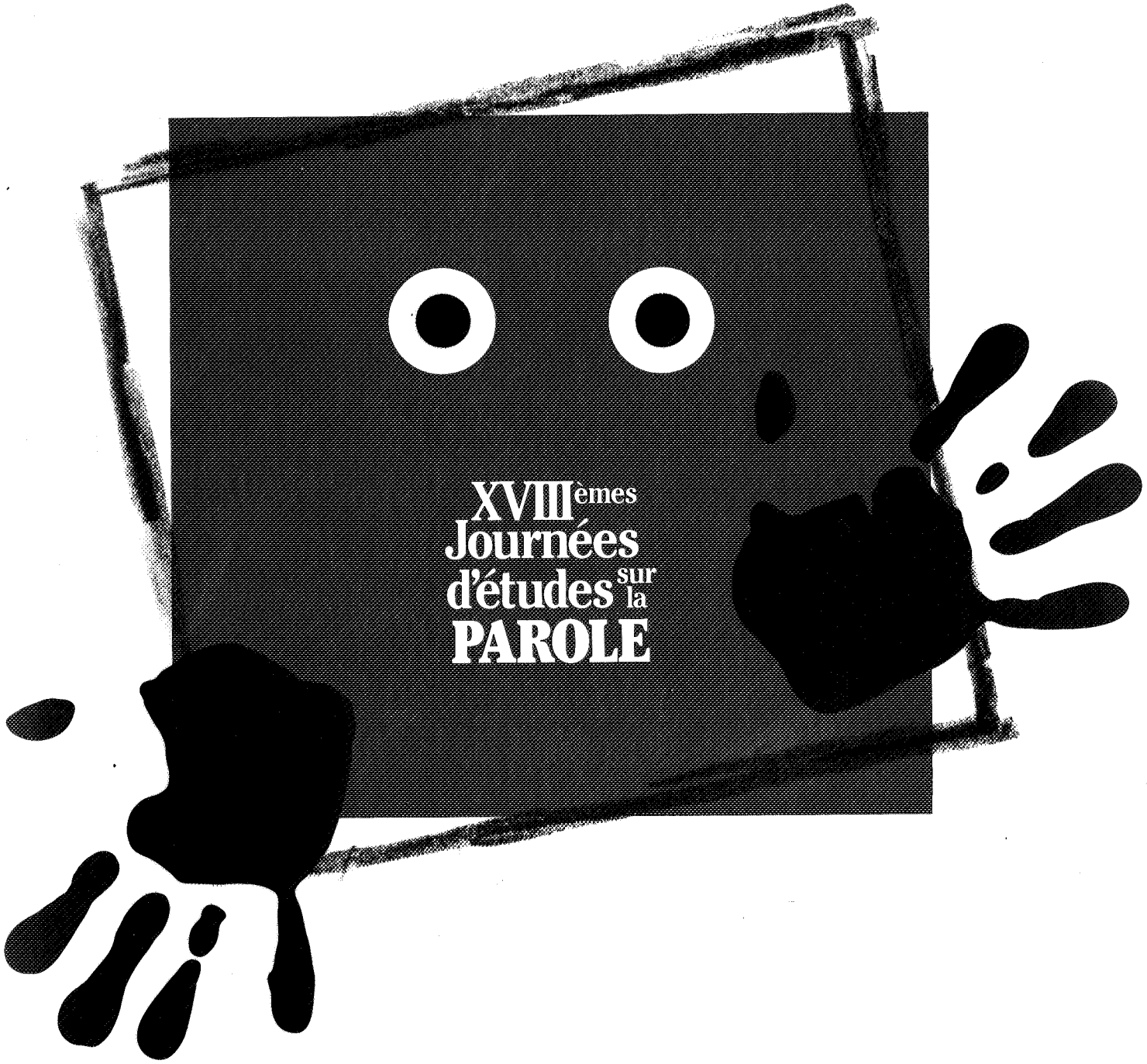
Montréal – Canada



Actes

Université de Montréal
Faculté des arts et des sciences
Département de linguistique et philologie

28-31 mai 1990

Laboratoire de phonétique
Département de Linguistique
Faculté des arts et des sciences
Université de Montréal

Christine MEUNIER
~~1 rue Lucas de Montigny~~
~~13100 Aix en Provence~~
~~Tél. 42.21.23.44~~

XVIII^{èmes} Journées d'études sur la parole

Organisées à l'Université de Montréal
avec la collaboration
de Communications Canada
et de la Société Française d'Acoustique

Du 28 au 31 mai 1990

XVIIIèmes JOURNÉES D'ÉTUDES SUR LA PAROLE

Comité d'organisation:

Danièle Archambault
Université de Montréal

Raymond Descout
Communication Canada-CCRIT-Laval

Avec la collaboration de: Julie Brousseau

Jean-François Couturier
Lise Malo
Isabelle Roy
Anne Sanfaçon
Louise Tremblay
Chantal Trépanier

Comité scientifique:

Présidents: D. Archambault, Université de Montréal
R. Descout, Communication Canada-CCRIT-Laval

Membres: C. Abry, ICP-Grenoble
J. Caelen, ICP-Grenoble
R. De Mori, CRIM-Université McGill-Montréal
N. Ellouze, IRSIT-Tunis
F. Grosjean, Université de Neuchâtel
J.M. Hombert, Université de Lyon
E. Keller, Université du Québec à Montréal
J.S. Liénard, LIMSI-Orsay
P. Martin, Université de Toronto
H. Meloni, Université de Luminy
J.M. Pierrel, CRIN/INRIA-Nancy
J.L. Schwartz, ICP-Grenoble
B. Teston, Université d'Aix-en-Provence
J.P. Tubach, ENST-Paris
M. Wajskop, Université Libre de Bruxelles

Nous remercions les organismes suivants qui ont rendu possible la tenue des XVIII^{èmes} Journées d'études sur la parole

L'Université de Montréal
Faculté des arts et des sciences

Communication Canada
Centre canadien de recherche sur l'informatisation du travail

La Société Française d'Acoustique

Bell Canada

Le Consulat Général de France à Québec
Service culturel, scientifique et de coopération technique

L'École Nationale Supérieure des Télécommunications-Paris

Gouvernement du Québec
Ministère des Affaires Extérieures

INRS-Télécommunications

Les Recherches Bell-Northern

Table des matières

1 PROSODIE

2 PHONÉTIQUE ET PHONOLOGIE

3 PRODUCTION

4 PERCEPTION

5 SYNTHÈSE ET CODAGE

6 OUTILS POUR LE TRAITEMENT DE LA PAROLE

7 RECONNAISSANCE ET DIALOGUE ORAL

8 PROSODIE

9 PRODUCTION ET SYNTHÈSE

10 DIALOGUE ET RECONNAISSANCE

1 PROSODIE

Président: P. MARTIN
Université de Toronto,Canada

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

La désaccentuation des rimes à noyau bref ou long

Laurent Santerre

Université de Montréal

Résumé.
Dans le québécois où le système vocalique comporte le trait obligatoire de durée pour huit (8) sur quinze (15) de ses unités en syllabe entravée, on trouve des systématiques de répartition des durées dans la rime; selon que le noyau est une longue ou une brève par nature et selon que la coda est constituée par une consonne abrégeante, allongeante ou neutre, en position d'accent primaire, on peut prévoir les rimes où l'élément vocalique ou l'élément consonantique a une plus ou moins grande priorité de durée relative, de même que la durée de ces rimes.

Mais qu'en est-il de cette systématique des durées quand la rime s'abrège par désaccentuation? j'ai trouvé que l'opposition phonologique de durée vocalique résiste à tous les traitements accentuels ou syllabiques de la rime, mais que les différences phonétiques de longueur entre les occlusives sourdes et sonores s'affaiblissent, de même que les effets d'abrégement et d'allongement par coarticulation. Les durées relatives voyelle/consonne, qui font nettement dominer la voyelle ou la consonne dans les rimes selon leur composition sous l'accent primaire, ne varient pas de façon linéaire dans la désaccentuation, de sorte que les rimes à dominance vocalique peuvent devenir en dehors de l'accent des rimes neutres ou à dominance consonantique.

En vue de résumer et simplifier ces généralisations, je présente en conclusion un tableau de valeurs calculées en tenant compte des systématiques observées dans le corpus, cela dans le but de faciliter leur mise à l'épreuve auditive au moyen de la synthèse.

Problématique :

Le timbre des voyelles en français se conserve en dehors de l'accent , ce qui n'est pas le cas en anglais; quant à la durée, qu'elle soit de source phonologique comme en français québécois ou le résultat de la coarticulation consonantique, on sait qu'elle est étroitement liée à l'accentuation. On sait aussi que les effets d'abrègement et d'allongement consonantiques sont conditionnés par la structure syllabique, c'est-à-dire qu'une consonne abrégeante ou allongeante exerce son action sur la voyelle qui précède dans la mesure où elle l'entrave et ferme la syllabe; dans ce cas, la consonne constitue la coda de la rime. La répartition des durées relatives entre le noyau et la coda tient à la nature phonologique de longue ou de brève de la voyelle et aux traits phonétiques de la consonne abrégeante, allongeante ou neutre, (Santerre 1987).

C'est sous l'accent que se manifaste le plus clairement la systématique qui règle les durées relatives dans les rimes. Il s'agit de voir si cette systématique se retrouve en dehors de l'accent.

Je rappelle brièvement les conclusions dégagées du corpus sous l'accent de 1987.

	Brèves	Longues	Allongeantes	Abrégeantes
ptk		X		X
bdg	X			
fsʃ		X		
vzʒ	X		X	

Seules les occlusives sourdes sont abrégeantes, et seules les constrictives sonores sont allongeantes. Les sourdes sont longues et les sonores brèves.

Sur les dix-sept (17) voyelles du système québécois, quinze peuvent être entravées.

	brèves	longues	s'allongent beaucoup	s'abrègent peu
i y u	X		X	
ɛ œ a ɔ	X		X	
ə ø ɑ o		X		X
ɛ̃ œ̃ ɑ̃ ɔ̃		X		X

Voici un tableau des durées moyennes des voyelles et des consonnes sous l'accent primaire (frontières 1 et 2 réunies en une moyenne : patte", et "coupe-lui les pattes".)

	V	C		V	C
v + ptk	7.8	15.7	v + fsʃ	10.0	17.4
v̄ + ---	17.3	13.25	v̄ ---	19.7	15.2
v + bdg	10.7	9.5	v+ vzʒ	18.0	9.0
v̄+ ----	19.1	8.6	v̄+ ---	22.4	9.3

(v = voyelles brèves; v̄ = voyelles longues).

Remarques:

1. Les consonnes longues gardent toute leur durée dans les rimes à noyau bref, et elles s'abrègent dans les rimes à noyau long. Les consonnes brèves ne se prêtent pas à ce conditionnement.
2. Les consonnes allongeantes sont nécessairement dans des rimes à noyau long; ces noyaux allongés sont presque aussi longs que les noyaux longs par nature.
3. Les voyelles entravées par bdg ou fsʃ ont sensiblement la même durée, soit 10cs. Par comparaison, ptk sont abrégeants (7.8 contre 10cs) et vzʒ sont allongeants (18 contre 10 cs); ptk abrègent un peu les longues par nature (17.3 contre 19.1 et 19.7); les vzʒ allongent aussi les longues par nature (22.4 contre 19.1 et 19.7).
4. Les brèves s'abrègent peu (7.8 cs contre 10 cs) sous l'effet des occlusives sourdes. Les longues varient peu (de 17.3 à 22.4 cs) sous l'effet de l'abrègement et de l'allongement.
5. Dans ces huit (8) groupes de rimes, il est facile d'identifier celles où la durée vocalique l'emporte en pourcentage sur la durée consonantique, de même que celles où la durée consonantique est prédominante. Il n'y a qu'une rime où les deux durées ne s'opposent pas vraiment, c'est celle d'un noyau bref entravé par une consonne brève non allongeante (v+bdg).

Cette systématique me semble transparente et facile à expliquer sous l'accent; mais qu'en est-il sous l'accent secondaire ou en dehors de l'accent, ou encore dans le cas où la voyelle et la consonne dans la rime d'un morphème sont séparées par une coupe syllabique qui place la coda sous l'accent primaire, comme dans "prends du pâté" [pɑ te], ou dans une
0 1
syllabe inaccentuable, comme dans "de la pâte à tarte" [pɑ ta
2 0
tart]
1

J'ai tenté de mettre la systématique à l'épreuve dans la production d'un seul locuteur mais en focalisant sur les oppositions de /a/ et /ɑ/ et de /ɛ/ et /ə/ entravés par des consonnes des quatre classes, soit t/d et ʃ/ʒ. La segmentation et la mesure des durées ont été faites sur un micro-ordinateur qui affiche l'oscillogramme et le sonagramme synchronisés du signal étalable à volonté. La précision au milliseconde près n'est pas pertinente mais elle n'a pas été systématiquement rejetée. J'ai fais commencer la voyelle après la consonne au moment où naissent les formants vocaliques et je l'ai fait terminer avec la disparition du F2. Le délai d'établissement du voisement (V.O.T.) fait partie de la consonne. Quand il y un bref espace entre la fin de la voyelle et le début d'une constriction qui suit, j'ai attribué ce délai à la voyelle à cause de la rémanence perceptuelle du signal dans la cochlée. Comme le corpus ne comportait pas de rimes terminées par une sonante, la segmentation a été assez facile; pour la délimitation des constrictives finales, j'ai essayé d'être consistant en la faisant terminer au moment où un schwa plus ou moins clair annonce la fin de la constriction.

Il faut s'attendre à retrouver la systématique tirée de l'ensemble du corpus de 1987, puisqu'il s'agit du même locuteur. Mais qu'advient-il de cette systématique dégagée sous l'accentuation, quand les mêmes mots ne sont plus sous le même accent? D'autres systématiques doivent sans doute interférer avec la première; il importe donc de les connaître si l'on veut que l'ordinateur ne prononce pas et ne reconnaisse pas que des mots isolés.

Frontières 1 et 2: la rime en finale absolue (durée moyenne en cs).

		V	C			V	C
patte	/at/	11.5	16.5	fade	/ad/	14.	13.0
pâte	/ɑt/	17.0	13.5	-------		--.--	--.--
chante	/ɑ̃t/	19.5	13.0	amende	/ɑ̃d/	20.2	11.5
tache	/aʃ/	12.2	21.5	nage	/aʒ/	20.0	4.7
tâche	/ɑʃ/	20.5	16.7	âge	/ɑʒ/	22.0	12.2
étanche	/ɑ̃ʃ/	21.5	17.5	mélange	/ɑ̃ʒ/	23.7	12.2
faites	/ɛt/	10.2	18.0	laid	/ɛd/	11.7	12.7
fête	/ət/	19.5	11.5	l'aide	/əd/	21.0	10.7
feinte	/ɛ̃t/	19.7	12.2	scinde	/ɛ̃d/	21.8	8.1
pèche	/ɛʃ/	11.0	20.2	-----		--.--	--.-
pêche	/əʃ/	21.5	18.7	neige	/əʒ/	25.0	17.0
lynche	/ɛ̃ʃ/	21.0	18.2	singe	/ɛ̃ʒ/	21.7	12.5

Observations: On retrouve, ici comme dans le corpus général de 1987, les mêmes prédominances relatives de durée vocalique ou consonantique selon la nature des noyaux et des codas. On pourrait montrer que le rapport des durées relatives associé avec la durée de la rime et la distinction des occlusives et des constrictives faciliteraient la reconnaissance automatique de ces rimes.

Frontière 3: inter-syntagmatique (SN+SV). Le mot patte me plaît

patte	/at/	8.0	15.0	fade	/ad/	10.5	7.5
pâte	/ɑt/	12.5	12.0	------		--.--	--.-
chante	/ɑ̃t/	14.5	10.0	amende	/ɑ̃d/	15.0	8.5
tache	/aʃ/	8.0	10.5	nage	/aʒ/	15.5	8.5
tâche	/ɑʃ/	16.0	10.0	âge	/ɑʒ/	16.5	7.0
étanche	/ɑ̃ʃ/	16.5	12.5	mélange	/ɑ̃ʒ/	16.5	9.0
faites	/ɛt/	7.0	14.0	laide	/ɛd/	10.0	9.0
fête	/ət/	12.5	12.5	l'aide	/əd/	16.5	9.0
feinte	/ɛ̃t/	16.5	12.0	scinde	/ɛ̃d/	15.6	7.9
pèche	/ɛʃ/	9.0	12.5	------		--.--	--.-
pêche	/əʃ/	14.0	10.5	neige	/əʒ/	17.5	9.5
lynche	/ɛ̃ʃ/	15.0	11.5	singe	/ɛ̃ʒ/	16.5	8.5

Observations:

1. Les rimes ici se trouvent sous l'accent secondaire, à la fin du syntagme sujet. Les timbres des voyelles, longues ou brèves, ne changent pas pour l'oreille, mais la durée liée au degré d'accentuation appelle des remarques importantes; l'abrègement des voyelles et des consonnes ne se fait pas dans les mêmes proportions dans toutes les rimes, ce qui change le rapport des durées relatives.
2. Partout les voyelles brèves restent dominées en durée par les consonnes sourdes; les voyelles longues ne dominent plus les consonnes sourdes, ex. dans /ɑt/ et /ət/. Les moyennes réunies des voyelles longues orales et nasales dans le corpus de 1987 ne laissaient pas voir que la durée des orales se réduit davantage que celle des nasales sous l'effet d'une moins grande accentuation; on peut penser que cet effet est dû à l'abrègement plus marqué du /t/ sur les orales que sur les nasales. L'entrave des longues par le /ʃ/ non abrégeant laisse aux voyelles longues leur nette prédominante en durée dans la rime.
3. Dans les rimes à voyelles brèves entravées par la consonne brève non allongeante /d/, j'ai pu observer que la voyelle ou la consonne peut librement dominer en durée. Les noyaux brefs allongés dominent nettement leur coda sonore allongeante..

Je proposerai en conclusion un tableau des durées relatives à mettre à l'essai dans la synthèse de ces différentes rimes.

Frontière 4: intra-syntagmatique; des pâtes maison.

patte	/at/	7.0	13.0	fade	/ad/	8.5	8.5
pâtes	/ɑt/	10.0	11.0	-----		--.---	-.-
chante	/ɑ̃t/	10.5	9.5	amende	/ɑ̃d/	14.0	9.5
tache	/aʃ/	8.5	11.0	nage	/aʒ/	12.5	8.5
tâche	/ɑʃ/	13.0	10.5	âge	/ɑʒ/	13.5	8.0
étanche	/ɑ̃ʃ/	13.5	10.0	mélange	/ɑ̃ʒ/	15.0	8.5
faites	/ɛt/	7.0	10.5	laide	/ɛd/	10.0	6.0
fête	/ət/	11.5	11.5	l'aide	/əd/	15.0	8.5
feinte	/ɛ̃t/	12.5	11.0	scinde	/ɛ̃d/	13.5	8.3
pèche	/ɛʃ/	7.0	12.5	------		--.--	--.--
pêche	/əʃ/	10.5	10.0	neige	/əʒ/	14.0	6.0
lynche	/ɛ̃ʃ/	14.0	10.5	singe	/ɛ̃ʒ/	12.0	6.0

Observations:

1. On remarque que la durée des voyelles longues s'abrège de 2 ou 3 cs par rapport à l'accent en frontière inter-syntagmatique; celle des brèves, déjà assez réduite, ne s'abaisse pas sous 7 cs. Les consonnes sourdes peuvent s'abréger de 1 ou 2 cs, pour s'établir entre 10 et 13 cs; la durée nettement plus grande des constrictives que des occlusives sous l'accent primaire ne se retrouve plus ici.
2. Les voyelles longues ou allongées perdent aussi 2 ou 3 cs par rapport à la frontière inter-syntagmatique. Quant aux consonnes sourdes qui variaient entre 7.5 et 9.5 cs sous l'accent secondaire, elles peuvent descendre jusqu'à 6 cs.

On peut penser que les rimes dans les morphèmes à l'étude ont été peu accentués à cette frontière, ce qui ne veut pas dire qu'on ne pourrait pas faire une prononciation qui donne plus de poids aux morphèmes brefs ou longs à l'étude. Mais mon but a justement été de veiller à ne pas souligner des syllabes pour leur contenu sémantique ou leur position métrique dans les mots.

Le point plus important, c'est que la distinction nette de durée des voyelles brèves ou longues se maintient toujours, comme celle du timbre d'ailleurs, quel que soit le degré de désaccentuation.

Frontière 5: de la pâte à tarte (V/c). [pɑ ta tart] ou "pas ta tarte"?

						0	1
patte	/at/	7.5	8.5	fade	/ad/	12.0	7.5
pâte	/ɑt/	9.0	7.0	------		--.--	--.--
chante	/ɑ̃t/	12.0	7.0	amende	/ɑ̃d/	14.0	5.0
tache	/aʃ/	9.0	11.0	nage	/aʒ/	11.0	8.0
tâche	/ɑʃ/	10.5	10.0	âge	/ɑʒ/	13.5	7.5
étanche	/ɑ̃ʃ/	10.5	10.0	mélange	/ɑ̃ʒ/	15.0	6.5
faites	/ɛt/	7.0	8.0	laide	/ɛd/	8.o	6.2
fête	/ət/	13.0	8.5	l'aide	/əd/	13.0	6.0
feinte	/ɛ̃t/	12.5	8.0	scinde	/ɛ̃d/	11.9	5.0
pèche	/ɛʃ/	6.5	11.5	-------		--.--	--.--
pêche	/əʃ/	11.0	12.0	neige	/əʒ/	14.0	6.5
lynche	/ɛ̃ʃ/	16.5	11.0	singe	/ɛ̃ʒ/	13.0	6.0

Observations:

Cette frontière permet de mesurer les mêmes voyelles et le mêmes consonnes en dérivation syllabique, puisque la voyelle reste accentuable et que la consonne ne peut l'être.

1. Il n'y a pas de différence pertinente de durée entre les voyelles des deux frontières 4 et 5. J'ai prononcé ces phrases une dizaine de fois à divers moments : la variation se déplace mais la moyenne reste à peu près la même; elle est de 10 cs devant les codas sourdes et de 12.5 devant les codas sonores. Ici aussi la distinction de timbre et de durée entre les voyelles longues et les brèves reste intacte. Je précise que ces chiffres ne sont pas des moyennes.
2. Les codas qui tombent dans une syllabe non accentuable se trouvent abrégées de 2 ou 3 cs, si ce sont des occlusives; les constrictives semblent ne pas pouvoir s'abréger davantage, mais la distinction de durée entre les sourdes et les sonores reste toujours pertinente, même en cas de désaccentuation totale.

J'entends souvent dans le langage populaire des réalisations qui ne coupent pas les morphèmes à noyau long par la syllabation et qui font [pɑt a taʀt]; il est alors impossible d'entendre "pas ta tarte". Même une voyelle brève peut garder des traces de la durée qu'elle a sous l'entrave allongeante dans le morphème: "il nage à loisir [il naʒ a ...], différent de "il n'a jamais" [il na ʒɑ].

Frontière 6: un empâté v/C

patte	/at/	8.0	13.0	fade	/ad/	8.5	8.5
pâte	/ɑt/	10.0	14.0	------		--.--	--.--
chante	/ɑ̃t/	12.0	11.5	amende	/ɑ̃d/	14.5	10.0
tache	/aʃ/	6.0	16.5	nage	/aʒ/	13.5	8.0
tâche	/ɑʃ/	12.0	16.0	âge	/ɑʒ/	14.0	10.0
étanche	/ɑ̃ʃ/	13.0	16.5	mélange	/ɑ̃ʒ/	14.5	8.5
faites	/ɛt/	7.5	11.5	laide	/ɛd/	8.5	8.5
fête	/ɜt/	11.0	13.0	l'aide	/ɜd/	13.0	9.0
feinte	/ɛ̃t/	12.0	14.0	scinde	/ɛ̃d/	10.6	10.1
pèche	/ɛʃ/	5.0	16.5	-------		--.--	--.--
pêche	/ɜʃ/	9.5	17.0	neige	/ɜʒ/	15.0	8.5
lynche	/ɛ̃ʃ/	13.0	17.5	singe	/ɛ̃ʒ/	15.0	10.0

Observations:

A l'inverse de la frontière précédente, c'est la voyelle qui n'est pas en position accentuable (pénultième), et la consonne tombe sous l'accent primaire.

1 Les voyelles sont sensiblement de même durée qu'en frontière 5.

2. Les consonnes retrouvent les durées caractéristiques qu'elles ont sous l'accent primaire, mais avec une réduction notable qui vient de la différence de position dans la syllabe accentuée: ici elles ne sont plus codas en finale absolue, mais en position initiale de syllabe.

Frontière 7: l'empâtement [ɑ̃ pɑt mɑ̃]

patte	/at/	7.0	13.0	fade	/ad/	9.0	8.0
pâte	/ɑt/	10.0	12.0	------		--.--	--.--
chante	/ɑ̃t/	11.5	12.5	amende	/ɑ̃d/	14.0	12.5
tache	/aʃ/	8.0	13.0	nage	/aʒ/	12.0	9.0
tâche	/ɑʃ/	13.0	13.5	------		--.-	--.--
étanche	/ɑ̃ʃ/	13.0	16.5	mélange	/ɑ̃ʒ/	14.0	10.0
faites-m'en	/ɛt/	6.5	11.0	aide	/ɛd/	9.0	6.0
fête	/ɜt/	11.5	12.0	-------		--.--	--.--
feinte	/ɛ̃t/	13.0	12.0	scinde	/ɛ̃d/	12.4	11.5
sèche	/ɛʃ/	7.9	16.9	------		--.--	--.-
empêche	/ɜʃ/	12.5	12.0	neige	/ɜʒ/	13.0	11.5
lynche	/ɛ̃ʃ/	15.5	13.5	singe	/ɛ̃ʒ/	15.0	10.0

Observations:

Les morphèmes à l'étude sont en position pénultième normalement inaccentuable.

1. Ici comme ailleurs les distinctions de timbre et de durée vocalique sont toujours nettement respectées. Ces durées sont à peu près semblables à celles qu'on voit en frontière 5 et 6, c'est-à-dire, sans effet accentuel.

2. Quant aux consonnes, la durée des sourdes est légèrement plus grande que celles de la frontière 4 où la rime est sous l'accent secondaire (pâtes maison). Cela donne à penser qu'il s'agit d'une trace d'accentuation des morphèmes longs en pénultième. Ici comme en position 5 (pâte à tarte), le parler populaire fait souvent sentir un allongement, surtout quand le mot se trouve en fin de phrase et que l'accent primaire n'est pas matérialisé par un intonème, comme c'est souvent le cas (Rossi 1979).

A la suite de ces observations dispersées, j'ai pensé utile de ramasser les conclusions sous forme d'un tableau de durées calculées, non plus mesurées, en tenant compte des tendances observées.

Durées calculées des voyelles et des consonnes.

Accent primaire: frontières 1 et 2. (Moyenne en cs).
(v = voy. brèves; V̄ = voy. longues; Ṽ = voy. nasales)

v + t =	12.0	16.0	V̄ + t =	18.5	13.0	Ṽ + t =	20.0	13.0
+ d =	13.0	13.0	+ d =	21.0	10.0	+ d =	21.0	10.0
+ ʃ =	13.0	20.0	+ ʃ =	21.0	17.0	+ ʃ =	21.0	17.0
+ ʒ =	20.0	14.0	+ ʒ =	23.0	13.0	+ ʒ =	23.0	13.0

Accent secondaire inter-synt: "le mot pâte me plaît", (fr. 3).

v + t =	8.0	14.0	V̄ + t =	12.5	12.5	Ṽ + t =	15.0	11.0
+ d =	9.0	9.0	+ d =	15.0	8.0	+ d =	16.0	8.0
+ ʃ =	10.0	12.0	+ ʃ =	15.0	12.0	+ ʃ =	16.0	12.0
+ ʒ =	15.0	8.0	+ ʒ =	17.0	8.0	+ ʒ =	17.0	8.0

Accent secondaire intra-synt.: "pâte maison" et "pâte à tarte", (fr.4 et 5).

		f4/f.5			f.4/f.5			f.4/f.5
v + t =	7	11/8	v + t =	11	11/8	Ṽ + t =	11.	11/8
+ d =	8	8/6	+ d =	13	8/6	+ d =	13	8/6
+ ʃ =	8	11/11	+ ʃ =	13	10/10	+ ʃ =	13	10/10
+ ʒ =	12	7/6	+ ʒ =	14	7/6	+ ʒ =	14	7/6

En syllabe libre et en dehors de l'accent: "empâté /ɑ̃ pɑ te/ (fr. 6)

v + t =	7	13	V̄ + t =	11	13	Ṽ + t =	12	13
+ d =	7	9	+ d =	11	9	+ d =	12	9
+ ʃ =	7	16	+ ʃ =	11	16	+ ʃ =	12	16
+ ʒ =	13	9	+ ʒ =	13	9	+ ʒ =	14	9

A la pénultième: "empâtement" / ɑ̃ pɑt mɑ̃/ (fr. 7).

v + t =	7	12	V̄ + t =	11	12	Ṽ + t =	12	12
+ d =	8	8	+ d =	--	--	+ d =	12	18
+ ʃ =	8	15	+ ʃ =	11	15	+ ʃ =	12	15
+ ʒ =	13	10	+ ʒ =	13	10	+ ʒ =	14	10

Commentaires en conclusion.

Le tableau des durées vocaliques et consonantiques calculées est le résultat d'une simplification ou d'une régularisation des mesures observées; il gomme de légères variations pour mieux faire ressortir les principaux traits de la systématique que j'ai cru observer. Ainsi il se prête mieux à une formalisation en vue de la synthèse par ordinateur.

Le tableau montre clairement que l'opposition brève/longue des voyelles du québécois se retrouvent dans tous les contextes accentuels ou syllabiques examinés ici, à une exception près qui tient à l'allongement des brèves par la constrictive sonore, ce qui est un fait de toutes les langues à divers degrés.

On peut aussi conserver l'opposition de durée des consonnes sourdes et des consonnes sonores; elle est plus ou moins grande ou respectée selon que les consonnes se trouvent sous l'accent en position forte de début de syllabe, ou en fin de syllabe non accentuée ou non accentuable.

J'avais remarqué (Santerre 1987) que les occlusives sourdes, abrégeantes, occupaient une durée relative nettement prioritaire dans les rimes à noyau bref, et qu'elles cédaient de la place au noyau long , comme pour ne pas trop allonger les rimes longues. Ce phénomène se trouve sous l'accent seulement et il est lié au degré d'accentuation; on ne le retrouve plus dans les rimes plus courtes.

Les durées telles que calculées ici me semblent respecter ce que Dell (1984) appelle une prononciation neutre. Il va sans dire qu'il faudra les mettre à l'épreuve dans la synthèse pour juger de leur convenance ou de leur naturel à l'oreille; de plus, elles devraient être remodelées par le passage de quelques règles d'accentuation de la parole spontanée, comme par ex. le déplacement de l'accent sur la première syllabe d'un groupe (arc accentuel, Fonagy 1979): de l'épatement, ou maladie grave

2 0 1 2 0 1

Le corpus ne contenait pas de contiguïtés accentuelles.

Cet essai s'en tient aux rimes à attaque et à coda simples. Il est bien évident qu'on ne pourra pas faire l'économie de la recherche sur les attaques et les codas complexes. On peut seulement espérer avec le temps mettre en lunière les généralisations les plus utiles et ne pas les gauchir en voulant les simplifier.

On pourrait formaliser ce tableau en prenant comme base de durée les rimes à noyau bref et à noyau long en dehors de l'accent, et en multipliant par un facteur d'allongement séparé pour la voyelle et pour la consonne en tenant compte du degré d'accentuation. Mais il ne me semble pas possible d'échapper à la nécessité de faire des règles séparées pour plusieurs espèces de rimes (2 noyaux: long ou bref; 3 sortes d'entrave: abrégeante, allongeante, neutre; deux ou trois degrés d'accentuation; 2 positions des constituants des morphèmes longs par rapport à la coupe syllabique et à l'accent.

Références

Dell, f. 1984. "L'accentuation dans les phrases en français", D. Hirsh et J.-R. Vergnaud, Forme sonore du langage. Hermann, Paris.

Fonagy, F. 1979. "L'accent français: accent probabilitaire". Fonagy, F et P.R. Léon, L'accent en français contemporain. Studia Phonetica, no 15. Didier.

Rossi, M. 1979. "Le français, langue sans accent?" Fonagy, I. et P.R. Léon, L'accent en français contemporain. Studia Phonetica, no 15, Didier.

Santerre, L. 1987. "Systématique des durées segmentales dans les rimes syllabiques à voyelles longues et brèves par nature". Proceedings of XIth International Congress of Phonetics Sciences. Tallinn, U.S.S.R, Vol 5 p. 126-129.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

APPORT D'INFORMATION LEXICALE ET MARQUES PROSODIQUES

Pascal ROMEAS

Institut de Phonétique d'Aix-en-Provence/ CNRS URA 261, FRANCE.

Résumé : Un corpus de dialogue homme-machine a donné lieu 1- à un étiquetage prosodique, 2- à une étude psycho-sémantique portant sur l'organisation de l'information lexicale. Les jugements de cette étude (2) concernent l'apport d'information relatif des mots du texte transcrit. L'étiquetage (1) renvoie à l'hypothèse d'un double mode d'organisation tonale : local et global. Nous vérifions l'hypothèse d'une stratégie énonciative, consistant à réaliser, au moyen du mode global d'organisation tonale, une alternance marqué/ non-marqué à l'échelon lexical

Abstract : Both prosodic labelling (1) and psychosemantic judgements (2) have been carried out from a man-machine dialog corpus. Judgements (2) involved the relative information content of the words in the written text. Labels (1) distinguish between two main types of tonal organization : local and global. We bring to light the existence of a discourse strategy which consists in realizing a marked/ unmarked alternance at the lexical level, using the global organization of prosody.

Dans le but de décrire les relations qui existent entre l'organisation de l'information d'origine lexicale et la distribution de configurations tonales observées en français sur des corpus de dialogue homme-machine, nous proposons ici de confronter les résultats d'une étude psycho-sémantique avec un étiquetage prosodique exhaustif.

Le matériau commun est un corpus de dialogue homme-machine (corpus "météo", du GRECO-"Communication parlée"). L'étude psycho-sémantique porte sur la transcription du corpus, l'analyse prosodique sur les tracés Fo.

1. étiquetage prosodique et analyse fonctionnelle.

L'étiquetage prosodique consiste en une stricte application du corps d'indices proposés dans Roméas (1988) pour la distinction de deux types configurationnels. Ces deux catégories ont été dénommées "concave" et "non-concave" en raison d'une différence d'aspects des tracés, que l'expert peut aisément coder en un complexe d'indices acoustiques hiérarchisés. Cependant, il s'agit là d'une commodité qui ne doit pas masquer une différence de comportement linguistique entre les deux configurations tonales

Nous souscrivons à l'hypothèse d'une coexistence de deux modes d'organisation tonale dans la parole. Ces deux modes recouvrent l'opposition global/local observée pour le traitement perceptif des patrons tonals par House (1987, a, b). Ils consistent, pour le français, d'une part, en une variation tonale locale, intra-syllabique, démarcative par un marquage ponctuel des frontières finales de constituants syntaxiques, et, d'autre part, en une variation tonale globale, potentiellement supra-syllabique, démarcative par le marquage mélodique d'une séquence de syllabes. Cette variation globale assume, en cela, une fonction intégratrice du niveau syllabique vers le niveau lexico-syntaxique (Roméas, 1990).

Nous avons montré que les configurations concaves (ou intra-syllabiques) sont des marqueurs de frontières qui permettent de détecter une finale de mot dans plus de 92.9% des cas, (moyenne sur 5 locuteurs), et que les configurations non-concaves portent, 9 fois sur 10, soit sur une seule unité lexicale, soit sur deux unités lexicales entretenant une relation de dépendance syntaxique directe. Les configurations non-concaves sont suprasyllabiques dans 2/3 des cas, le tiers restant se réalisant fréquemment sur des monosyllabes. Du point de vue de l'étendue exacte de ces configurations sur la chaîne syllabique, on observe une large coextensivité avec les unités lexicales (Roméas, 1990).

Nous rappelons succintement que, du point de vue des indices, la configuration intrasyllabique se caractérise par un important glissando positif de Fo de pente généralement croissante, par la présence d'un noyau vocalique unique et allongé, et, éventuellement, par la présence d'une pause, silencieuse ou voisée, subséquente au glissando.

La configuration suprasyllabique porte sur une séquence de syllabes

- -regroupées autour d'un pic mélodique ne répondant pas aux critères du premier type configurationnel,
- -et dont les valeurs Fo sont significativement supérieures à la valeur Fo, au même temps, de la droite de régression appliquée à l'ensemble des points bas de la courbe (base-line).

Si la séquence se limite à 1 syllabe (1/3 des cas), la configuration ne répond pas aux critères d'allongement et de concavité, et reste donc identifiable.

2. Le test psycho-sémantique.

2.1. Présentation

A partir du corpus transcrit, 6 sujets ont effectué une tâche de classification des mots, sur le critère de leur apport informatif relatif. Les consignes étaient :

1- d'éliminer le maximum de mots de chaque énoncé, jusqu'à ce que la demande d'information perde une trop grande partie de son sens et risque ainsi de ne plus être comprise par la machine (ces mots devaient être encadrés);

2- de sélectionner, parmi les mots non-encadrés, ceux qui complètent utilement le sens, sans être pour autant strictement indispensables (ces mots devaient être soulignés).

A l'issue du test, chaque sujet a donc défini 3 classes de mots : les mots strictement indispensables à la compréhension de l'énoncé (1), les mots porteurs d'information mais non-indispensables (2), et les mots sans importance informative (3).

Le test portait sur les 15 premiers énoncés des 5 premiers locuteurs du corpus, soit 890 mots.

2.2. Résultats.

Le tableau 1 montre la distribution relative des 3 classes de jugement, pour chun des 6 sujets. La classe (3) est la plus importante. La majorité des mots fait donc l'objet d'un jugement "apport informatif nul ou négligeable". La distribution parmi les deux autres classes fait apparaître une prédominance du jugement "apport informatif indispensable".

Tableau 1 : répartition, en 3 classes, des 890 mots du corpus en fonction de leur apport informatif relatif. Le jugement "apport nul ou négligeable" est consigné dans la classe 3. Les mots sélectionnés par les 6 sujets se répartissent entre les jugements "apport indispensable" (classe 1) et "apport complémentaire" (classe 2).

sujets	S1	S2	S3	S4	S5	S6
sélections	324	354	386	380	354	297
% sur 890	36.4	39.8	43.4	42.7	39.8	33.4
% classe 1	29.1	26.1	33.2	28.6	27.5	24.4
% classe 2	7.3	13.7	10.2	14.1	12.3	9
% classe 3	63.6	60.2	56.6	57.3	60.2	66.6

La faible part réservée au jugement intermédiaire (11% du total en moyenne) indique la perception d'un contraste dans l'information provenant du matériau lexical. Les unités mises en contraste sont les termes lexicaux porteurs d'une information importante et ceux qui n'apportent aucune information. Sous réserve de contraintes émanant d'autres niveaux d'organisation linguistique (syntaxe, dialogue), nous pouvons, du point de vue de la production, former l'hypothèse d'une stratégie énonciative consistant à réaliser, au moyen de traits prosodiques, une alternance marqué/non-marqué à l'échelon lexical.

3. La mise en relation des deux niveaux d'organisation.

Nous considérons désormais uniquement les mots de la classe (1). La présence d'une configuration tonale suprasyllabique (ci-après: /supra/) y est un phénomène majoritaire, et d'une grande stabilité inter-sujets (tableau 2).

Tableau 2 : Cas où l'on trouve une configuration /supra/ sur les items de classe (1), exprimés en pourcentage.

sujets	S1	S2	S3	S4	S5	S6
(%)	64.5	69.2	65.9	62.3	64.5	63.6

On ne peut cependant pas conclure que les unités lexicales apportant une information indispensable sont marquées de façon systématique par les configurations /supra/. En revanche, on peut dire que le recours à un procédé tonal déterminé est attesté de façon prépondérante, mais qu'il entre probablement en conflit avec d'autres facteurs d'organisation des traits prosodiques.

Le pourcentage recueilli ne constitue en soi qu'une indication qu'il serait sans doute coûteux d'intégrer à une source de connaissance, étant donné la probabilité d'échec de l'hypothèse (1/3 des cas). En revanche, si l'on parvient à fournir un petit nombre d'explications linguistiques pour une détermination aisée des 35% restants, la conjonction des deux informations se révèlera d'un certain intérêt. Nous savons que la présence de /supra/ sur les mots de classe (1) est loin d'être aléatoire, mais qu'elle se limite à 65%. Nous allons donc nous intéresser aux spécificités linguistiques des cas où /supra/ n'apparaît pas.

D'une part, il convient d'observer les modalités de la concurrence entre les deux types configurationnels, là où celle-ci apparaît. D'autre part, les sujets ayant sélectionné, à de très faibles fluctuations près, le même sous-ensemble de mots, la faible variabilité inter-sujets tend à prouver qu'il y a là l'effet systématique de contraintes liées à la structure même des énoncés.

Considérons d'abord le problème de la concurrence entre les deux configurations. Deux cas sont à étudier : celui où l'on trouve une configuration intrasyllabique (ci-après: /intra/) à la finale du mot, et pas de /supra/ sur la partie antérieure (codage (-s+i)); et celui où le mot présente /supra/ sur les premières syllabes, puis /intra/ sur la finale (codage (+s+i)).

Pour le cas (-s+i), nos données indiquent que 39.8% des mots de classe (1) ne présentant pas /supra/ ont, en revanche, une configuration /intra/ en syllabe finale. Nous posons donc la question suivante : est-ce l'affectation prioritaire de /intra/ en syllabe finale qui vient bloquer l'occurrence de /supra/ sur la partie initiale? En réalité, il semble qu'il faille intégrer ici une condition phonotactique de proximité. En termes de nombre de syllabes, il s'avère que ce pourcentage est la somme des trois suivants:

13.1% concernent des mots monosyllabiques (où la coexistence de /supra/ et de /intra/ est donc impossible).

18.9% concernent des mots de 2 à 3 syllabes.

7.8% concernent des mots de plus de 3 syllabes (c'est-à-dire moins de 1/5 des 39.8%).

Etant donné que la configuration concave occupe à elle seule un noyau vocalique, on peut penser que lorsque /supra/ n'apparaît pas, c'est simplement par manque d'une séquence syllabique de réalisation potentielle. Ceci expliquerait le cas des mots de 1 à 3 syllabes.

Le cas (+s+i) représente 13.5% des mots de la classe (1). Même si la coexistence de ces deux configurations sur un même mot ne représente guère plus de 1 mot de classe (1) sur 7, ce phénomène vient confirmer qu'un choix exclusif n'est pas requis a priori, au niveau de l'unité lexicale, entre les deux marques tonales. Il semble, du reste, que de ces conditions émane un profil assez typique : près de 4/5 des mots où coexistent /supra/ et /intra/ ont leur syllabe initiale marquée par le début de /supra/ (première syllabe d'une séquence non-basse). Dans 58.2% des cas (+s+i), l'enchaînement des deux configurations se fait sans retour au niveau de la ligne de base (ligne reliant les minima de la courbe Fo, à l'exception des variations micromelodiques). Dans 29.1% des cas, on compte une seule syllabe basse intermédiaire. Ce profil-type est illustré par les figures 1 et 2. Il concerne 158 sélections de classe (1) sur 1505, ce qui n'est pas négligeable, compte tenu du nombre de conditions requises.

Nos resultats antérieurs (Roméas, 1990) sur la coextensivité des marques tonales et des unités lexicales sont confirmés. Considérons la réunion des deux groupes suivants : 1) les mots démarqués à l'initiale par /supra/ et à la finale par /intra/ ; 2) les mots démarqués aux deux bornes par /supra/. Ce nouvel ensemble représente 45.9% des mots de la classe (1) présentant /supra/.. Sans pouvoir prétendre, à partir de ces résultats, à la généralisation d'un accès lexical à base prosodique, cette donnée vient toutefois compléter de façon appréciable les connaissances établies sur le marquage prosodique, qui concernent principalement la détection de frontières finales de syntagme (Di Cristo *et al* 1982, Vaissière 1988, Carbonell & Bonin 1988). Les phénomènes tonals localisés à l' intiale ou à l'intérieur d'une unité lexicale ont été observés, et décrits (Rossi 1985, Vaissière 1974, Hirst & Di Cristo 1984, Pasdeloup 1988). Nous en faisons ici une étude quantifiée.

Notre approche présente une spécificité : contrairement aux phénomènes tonals localisés en fin de syntagmes, les proéminences non-finales ne sont pas analysées comme des réalisations locales, mais globales. Leur domaine n'est donc pas la syllabe. Il est suprasyllabique.

Ceci équivaut à prendre une certaine distance à l'égard de l'association habituellement faite entre proéminence et syllabe. En ce sens, notre approche n'est pas accentuelle, puisque nous ne posons pas la syllabe comme unité de contraste. Peu enclin à une approche qui préjugerait du statut de la syllabe dans l'organisation prosodique, nous considérons que la prise en compte d'un mode d'organisation local et d'un mode d'organisation global est une voie de recherche qui offre d'intéressantes perspectives. En conséquence, nous proposons de saisir la spécificité de structuration thématique que présente le dialogue homme-machine afin d'envisager l'organisation prosodique dans sa relation au lexique. C'est dans ce cadre que doivent être interprétées les congruences évoquées ci-dessus.

Ayant considéré les modalités de coexistence des deux types de configurations tonales sur les unités lexicales, nous examinerons les particularités linguistiques des items de classe (1) ne présentant ni /supra/ ni /intra/.

Sur les 317 sélections concernées (soit 21% de la classe (1)), la particularité linguistique la plus répandue (75.7%) réside dans le fait d'être localisé en finale d'énoncé. Le dernier rang dans l'énoncé explique donc les 3/4 des cas d'absence de configurations sur les mots de grand apport informatif.

Si l'on retient, parmi les 77 sélections restantes, celles sur lesquelles s'accordent au moins 4 sujets sur 6, nous parvenons à 62 sélections portant sur 13 mots, dont la plupart présentent la particularité linguistique de former, avec une autre unité lexicale, un groupe syntaxique au sein duquel on analyse une relation déterminant/déterminé. Le groupe "acide sulfurique", par exemple, où "acide" est encadré sans toutefois présenter de configurations tonales, fait l'objet de 20 sélections, réparties sur 4 occurrences dans le texte. On peut citer également :

"vent prévu", "temps prévu", "humidité atmosphérique".

4. Conclusion.

Après avoir remarqué une tendance à la bipolarisation du jugement sur l'apport informatif des mots du corpus transcrit, nous avons observé que la réalisation d'une configuration tonale suprasyllabique était un phénomène majoritaire, mais non systématique (65%), sur les unités lexicales jugées indispensables. Il se confirme que le domaine privilégié du patron tonal suprasyllabique est l'unité lexicale, ou éventuellement le groupe syntaxique à 2 termes dont 1 nominal.

Dans ce cadre, la réalisation d'un tel patron, sur des mots d'apport informatif important, peut se trouver bloquée pour plusieurs types de raisons :

1- une contrainte phonotactique, portant sur le nombre de syllabe, en cas de présence d'une configuration intrasyllabique en finale (ce qui, du reste, suppose une affectation prioritaire de /intra/, marqueur de frontières syntaxiques, sur /supra/, marqueur lexical contrastif),

2- une contrainte de rang dans l'énoncé : pour des raisons bien connues de chute finale de Fo, la dernière unité lexicale peut difficilement être porteuse d'une proéminence mélodique;

3- une contrainte de cohésion sémantico-syntaxique : certains groupes syntaxiques, formant des unités de sens, semblent réfractaires à une mise en contraste tonal d'un seul de leurs termes.

En l'absence de ces 3 contraintes, il apparaît que l'apport d'information d'origine lexicale est très largement marqué par la prosodie, au moyen d'une configuration tonale suprasyllabique.

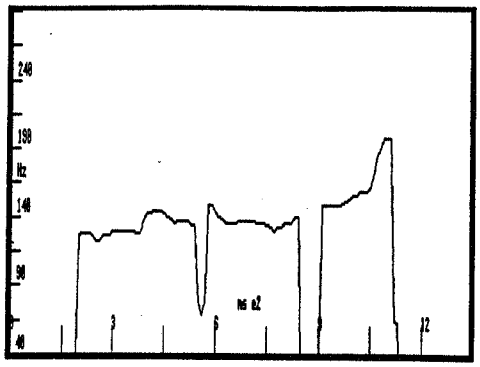


fig 1

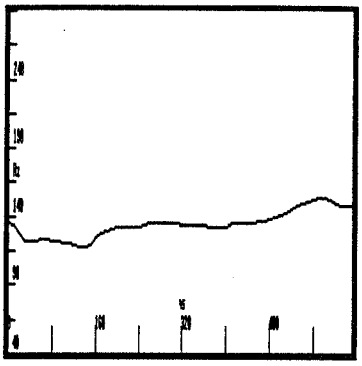


fig 2

Figures 1 et 2.
Les séquences dont la courbe Fo est représentée sont respectivement. la température et à Remiremont. Dans les deux cas, la syllabe finale porte une configuration concave (/intra/). En faisant abstraction des effets micromélodiques des segments consonantiques, nous observons une configuration /supra/ sur tempé et sur Re. Notre conception du cadre de référence tonal nous amène à considerer comme basses toutes les autres syllabes.
N.B.: le voisement détecté sur /t/ n'est pas erroné.

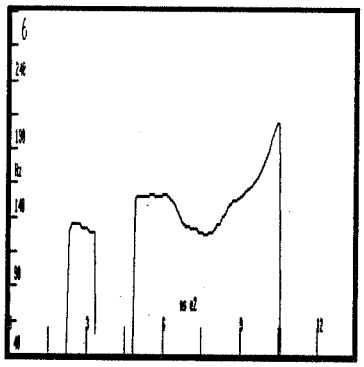


fig 3

Figure 3.
Courbe Fo de la séquence le treize juin. La séquence est précedée et suivie par des pauses. La syllabe finale porte une configuration concave (/intra/). Le monosyllabe treize porte une configuration non-concave (/supra/). Le est une syllabe basse.

REFERENCES

Carbonell, N., Bonin, J.J., (1988),Utilisation d'informations prosodiques en reconnaissance de la parole continue, 17emes Journées d'Etude sur la Parole, Nancy, 20-22 Sept. 1988, pp163-167.

Di Cristo, A., Haton, J.P., Rossi, M., Vaissière, J., (1982). Prosodie et reconnaissance automatique de la parole, GALF (séminaire)

Hirst, D., Di Cristo, A., (1984), French intonation : a parametric approach, Die neueren Sprachen, 83, Frankfurt.

House, D., (1987,a), Speech perception, intonation, and memory. Reports from Uppsala University, Dep. of Linguistics, 17, pp72-77.

House, D., (1987,b), Perception of tonal patterns in speech: implication for models of speech perception. Proceedings of the 11th ICPhS, Tallinn, 1-7 Août 1987, pp76-79.

Pasdeloup, V., (1988), Essai d'analyse du système accentuel du français: distribution de l'accent secondaire, 17èmes Journées d'Etude sur la Parole, Nancy, 20-22 Sept. 1988, pp65-70.

Roméas, P., (1988). Statut de la prosodie dans les recherches relatives au dialogue homme-machine, Travaux de l'Institut de Phonétique d'Aix-en-Provence, 12, pp153-183.

Roméas, P., (1990), Prosodie et lexique : tendances majeures observées en dialogue homme-machine, 1er Congrès Français d'Acoustique, Lyon 10-13 Avril 1990 (à paraître).

Rossi, M., (1985), L'intonation et l'organisation de l'énoncé, Phonetica, 42, 2-3, pp135-153.

Vaissière, J., (1974), On french prosody, Res. Lab. Elect., Q. Prog. Report, M.I.T., 115, pp212-223.

Vaissière, J., (1988), The use of prosodic parameters in automatic speech recognition. Recent advances in speech understanding and dialog systems, Nieman, Lang & Sagerer (eds.), NATO ASI Series, 46, pp71-100.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

FREQUENCE FONDAMENTALE ET DUREE POUR LA DETECTION DE FRONTIERES SYNTAGMATIQUES EN PAROLE CONTINUE.

Jean-Jacques BONIN Jean-Marie PIERREL

CRIN-INRIA Lorraine
BP 239 - 54506 VANDOEUVRE-LES-NANCY CEDEX
FRANCE

ABSTRACT

The present communication concerns the use of prosodic parameters in automatic continuous speech recognition and understanding. It is a matter of studying prosodic knowledges and informations for their uses in a natural language task oriented spoken dialogue system (multi-levels knowledge sources system developping in CRIN). Results are discussed concerning the detection of syntagm boundaries from the use of fundamental frequency and rhythm variations. There are presented in a statistical study form which combine F0, vocalic duration and pause cues corresponding to three differents speech production conditions : reading, memorized sentences (short-term memory), quasi-spontaneous information dialogues.

KEYWORDS

automatic continuous speech recognition / understanding - natual language - spontaneous speech - prosody - prosodic parameters - fundamental frequency - pauses - duration - vocalic nucleus - syntagm / lexical boundaries detection

1. INTRODUCTION

Cette étude est un premier pas vers l'utilisation d'une composante prosodique dans un système de compréhension de la parole continue (SCPC). En effet, il est maintenant admis que la prise en compte d'informations linguistiques contenues dans les paramètres prosodiques puisse accroître sensiblement les performances qualitatives de tels systèmes [15]. La prosodie a vraissemblablement son rôle à jouer dans la segmentation et la détection de l'organisation syntaxique et sémantique de l'énoncé. L'importance d'un module prosodique pour les systèmes de compréhension est aujourd'hui bien admise. On trouvera dans Vaissière 88 [19], une bonne bibliographie commentée des derniers travaux sur l'utilisation de la prosodie dans les systèmes de reconnaissance. Dans l'étude de la parole continue spontanée en vue d'applications futures, on observera toute l'importance du rôle linguistique de la prosodie.

Le travail relaté ici a pour but d'étudier informations et connaissances prosodiques afin de prévoir leurs intégrations dans un système de compréhension de dialogues oraux en langue naturelle [4]. Différents caractères doivent entrer dans la spécification d'un module prosodique :

- quasi totale indépendance avec les autres modules d'un SCPC. Avec le signal à analyser, seules les marques de segmentation obtenues par le décodage acoustico-phonétique seront fournies au module prosodique,

- fournir des informations directement utilisables par les autres modules et considérées en tant qu'hypothèses émises par celui-ci. Il s'agit principalement d'hypothèses sur les frontières de mots (frontières lexicales/syntagmatiques) et éventuellement sur le type d'énoncé.

La mise en oeuvre d'une telle composante améliorerait sensiblement les résultats de ces systèmes, voire même fortement suivant la qualité des indices prosodiques ainsi calculés.

L'idée n'a rien de "révolutionnaire", pourquoi ne pas informer les composantes directement en relation avec la composante prosodique que, moyennant un certain score, celle-ci a su placer une frontière lexicale, syntagmatique ou encore déterminer le type d'énoncé (affirmatif, interrogatif ...) de la phrase à reconnaître. Si de telles informations étaient fournies à ces composantes, cela réduirait d'autant plus le nombre d'hypothèses émises par celles-ci qu'il y a d'indices prosodiques trouvés. L'approche de type paramétrique et le système réalisé ont conduit dans un premier temps à limiter les objectifs fixés par la composante prosodique : extraire du signal des informations prosodiques et les utiliser pour limiter les hypothèses lexicales et syntaxiques. Cette approche se fonde sur des corpus d'enregistrement de parole spontanée ou de lecture, conçus comme échantillons de la performance du locuteur, mais l'interprétation des faits s'inspire néanmoins des théories linguistiques.

Le travail relaté ici sert de base à l'élaboration de la composante prosodique au sein du SCPC développé au CRIN. Il montre, à travers les résultats d'un premier objectif restreint et sur un échantillon de parole correspondant aux applications potentielles envisagées (voir § 2, les corpus utilisés et principalement le corpus METEO), la possibilité réelle d'intégration d'une composante prosodique dans un système de reconnaissance et compréhension de la parole continue (SRCPC).

Ce premier objectif était le suivant :

- les paramètres accoustiques retenus pour la recherche d'indices prosodiques ont été limités aux variations temporelles de F0 portant sur les noyaux vocaliques (l'intonation constitue le paramètre prosodique actuellement le mieux connu des phonéticiens), ainsi qu'à la mesure de durée (noyaux vocaliques et silences). L'intensité quant à elle n'a pas été retenue, ce paramètre moins étudié par les phonéticiens semble beaucoup moins important en français,

- afin d'étudier au mieux ces paramètres à partir de données rigoureuses, les corpus tests ont été segmentés manuellement.

Actuellement une étude est en cours sur l'utilisation exclusive des résultats fournis par le module de décodage acoustico-phonétique du SCPC (composante APHON) et principalement ceux obtenus lors du prétraitement par NOVOCA [7] pour la détection des noyaux vocaliques. Celle-ci, dans l'état actuel du module de décodage, correspondrait à environ 90 % de noyaux trouvés (tests effectués sur le corpus de "La Bise et le Soleil" par D. Fohr). L'utilisation de NOVOCA modifierait sensiblement les résultats discutés dans cette étude,

- à partir des données fournies (signal et étiquetage manuel) et des données calculées (F0, durée), effectuer la détection automatique des frontières syntagmatiques et lexicales.

2. CORPUS UTILISES

Les résultats de cette étude ont été obtenus sur trois corpus de langage naturel dont un orienté vers le dialogue homme-machine, correspondant chacun à un type de parole différent. On retrouve dans ces corpus les principales familles d'applications potentielles des SCPC :

- mémorisation puis énonciation de phrases lues au préalable (mémoire à court terme),
- lecture de textes,
- dialogues finalisés de type demandes de renseignements.

Ces trois corpus ont une fréquence d'échantillonnage de 16KHz, leur segmentation et étiquetage phonétique ont été effectués manuellement par des experts phonéticiens à partir du logiciel Snorri developpé au CRIN par Y. Laprie [8]. Chaque expert a eu à sa charge la segmentation et l'étiquetage complet d'un des corpus.

Corpus COMBESCURE : composé de 50 phrases différentes prononcées par 5 locuteurs non professionnels (4 hommes et 1 femme). Le texte d'une phrase du corpus de Combescure [5] est présenté au locuteur qui, après l'avoir lu et mémorisé, le dit aussi naturellement que possible. L'acquisition a été effectuée dans l'atmosphère "ronronnante" des ventilations des machines d'une salle de travail.

Corpus LABISE : il s'agit de la lecture d'un texte relativement simple "La Bise et le Soleil". Le corpus fait partie de la base de données (BDSONS) mise en oeuvre par le GRECO Communication parlée. Il a été enregistré dans d'excellentes conditions de qualité "studio". Pour l'instant, segmentation et étiquetage manuels n'ont été réalisés que pour 9 locuteurs hommes et femmes.

Corpus METEO : il s'agit d'un corpus de parole continue quasi-spontanée en situation de dialogue oral finalisé : consultation d'un centre de renseignements météorologiques [3]. Dix locuteurs masculins non sélectionnés, à partir de mots-clés affichés sur l'écran d'un micro-ordinateur, interrogent un centre automatique de renseignements simulé par un compère; les mots-clés précisent la nature des informations à obtenir mais la structure du dialogue n'est pas imposé.

Exemples : continuation du froid (jusqu'à quand),
on peut avoir les phrases suivantes :
- Heu, donnez-moi onze endroits où le froid va encore continuer.
- Le froid durera-t-il encore longtemps ?
- Jusqu'à quand continuera le froid ? ...

L'enregistrement a été effectué dans l'atmosphère relativement calme d'une salle légèrement insonorisée. Il est à noter que tous les locuteurs présentent un accent local des régions nancéiennes. Il faut aussi signaler l'importance de ce corpus et la source de données qu'il constitue pour l'étude de la parole continue spontanée (la segmentation et l'étiquetage complet du corpus portent sur 290 phrases qui comptent 5802 noyaux vocaliques).

3. MISE EN PLACE D'UNE STRUCTURE DE DONNEES

3.1. Traitement automatique des résultats : étiquetage syntaxique

Pour l'étude complète et détaillée des résultats sur la détection automatique des marqueurs prosodiques (cf § 5), outre les données de base disponibles (signal, segmentation et étiquetage phonétique), une segmentation en classes syntagmatiques est disponible sur l'ensemble des trois corpus. Chaque marqueur trouvé est répertorié suivant la place qu'il occupe à l'intérieur des groupes syntagmatiques. Le traitement automatique des marqueurs prosodiques détectés permet de multiples classifications facilitant le dépouillement des résultats.

Les classes syntagmatiques correspondant aux énoncés analysés sont les suivantes :

SN:	syntagme nominal	SAI:	syntagme adverbial inachevé
SV:	syntagme verbal	CNT:	connecteur
SA:	syntagme adverbial	SAJ:	syntagme adjectival
SNI:	syntagme nominal inachevé	SAJI:	syntagme adjectival inachevé
SVI:	syntagme verbal inachevé	SI:	syntagme verbal à l'infinitif

De plus, une décomposition supplémentaire a été effectuée pour les mots lexicaux (ML) non situés en fin de groupe. Il faut également ajouter le repérage des pauses remplies marquant l'hésitation du locuteur (H pour "euh").

Exemples pris dans le corpus METEO :

- | Le juge | veut prolonger | l'interrogatoire |.
  SN SV SN

- | Quel est | euh | le degré | d'humidité | sur Lunéville |.
  SV H ML SN SN

Ces décompositions ne représentent en aucun cas un quelconque niveau de l'arbre syntaxique des énoncés. Une analyse des résultats respectant la hiérarchie des constituants sera l'objet d'une autre étude.

3.2. Classes de marqueurs prosodiques

Les paramètres prosodiques intervenant dans la détection des frontières syntagmatiques et lexicales sont la fréquence fondamentale, la durée des noyaux vocaliques et la durée des pauses. On verra dans la discussion des résultats l'importance du dernier paramètre en parole continue spontanée, source principale d'informations et facilement interprétable par la composante prosodique. La combinaison de ces paramètres associés à chaque noyau vocalique étudié a conduit à une classification des indices prosodiques potentiels détectés par le système. Ainsi chaque noyau vocalique se voit affecté ou non d'un indice F0 et/ou d'un indice de durée variable et/ou d'un indice de pause signifiant que celui-ci est suivi par une pause dans l'énoncé. On trouvera [figure 1] les différentes combinaisons possibles de ces trois indices.

Avant toute analyse des résultats, les classes 1 et 2 sont d'ores et déjà exclues de la catégorie des classes susceptibles d'être retenues comme contenant les marqueurs prosodiques (frontières syntagmatiques/lexicales). La classe 1 représente les noyaux vocaliques ne contenant aucune information prosodique, quant à la classe 2, l'indice de durée est trop peu significatif. L'indice de durée vocalique est pris en référence à

la durée vocalique moyenne (valeur médiane des durées des noyaux vocaliques de l'énoncé), notion globale, relative à l'énoncé entier.

Classe du noyau vocalique	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Noyau suivi d'une pause	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
Indice F0	-	-	-	-	+	+	+	+	-	-	-	-	+	+	+	+
Durée vocalique de 1.5 à 2 DVM	-	+	+	+	-	+	+	+	-	+	+	+	-	+	+	+
de 2 à 3 DVM	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+	+
≥ 3 DVM	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	+

- : abscence d'indice
+ : présence d'indice

figure 1 : Classes des indices prosodiques étudiés.

4. OUTILS LOGICIELS D'ANALYSE

4.1. Calcul de la fréquence fondamentale

Le principe de l'algorithme de détection de pitch est celui du calcul de la fonction d'autocorrélation de Sondhi [16] repris dans Rabiner [11] et adapté pour les besoins. Cette méthode est relativement résistante au bruit de fond présent sur le signal et le calcul en est assez simple. Le signal temporel est d'abord filtré (filtre passe-bas FIR) puis sous-échantillonné à 8 KHz. Pour chaque signal correspondant à un énoncé analysé, on détermine un seuil de bruit de fond calculé sur 50 ms de silence précédent le début de parole. Après avoir effectué un "center clipping", on applique la fonction d'autocorrélation. A partir des deux pics maxima détectés, une décision voisement / non voisement est prise. Pour les parties voisées, une correction à posteriori est nécessaire. L'algorithme de correction de Bristow [2] effectue le rattrapage d'erreurs sur les valeurs du pitch ainsi calculées. L'efficacité de l'algorithme de calcul du fondamental est obtenue par accélérateur vectoriel sur Masscomp. La [figure 2] fournit un exemple pris dans le corpus METEO.

4.2. Détection des pics linguistiquement significatifs de la fréquence fondamentale [1]

L'analyse porte uniquement sur les valeurs de F0 correspondant aux noyaux vocaliques. Après lissage de la courbe des variations de F0, la valeur du maximum fréquentiel est sélectionnée pour chaque noyau. Ce principe a été comparé à d'autres (valeur de F0 au deux tiers du noyau, valeur moyenne de F0 sur le dernier tiers du noyau) et donne les meilleurs résultats.

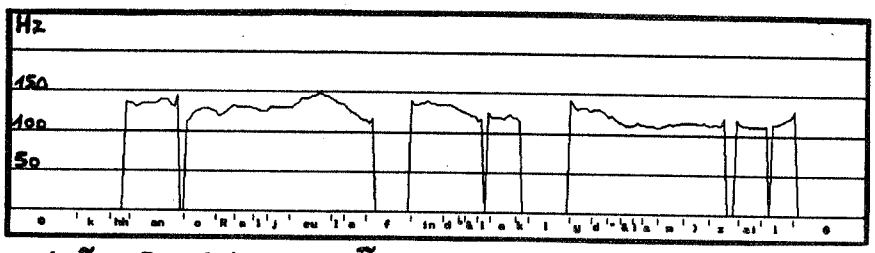
L'algorithme sélectionne ensuite les pics de la courbe les mieux marqués à l'aide de seuils relatifs calculés par fonctions homographiques dont les paramètres dépendent des mesures effectuées sur le signal analysé. Pour chaque phrase étudiée et quelque soit le corpus utilisé, ces seuils seront donc ajustés automatiquement aux mesures prises sur le signal. Les fonctions homographiques sont naturellement indépendantes du type de corpus analysé.

Un exemple de marquage, effectué sur l'énoncé de la [figure 2] par cet algorithme, est donné en [figure 3]. Ces divers algorithmes sont intégrés au système d'étude interactif de la parole Snorri [8].

5. DISCUSSION ET RESULTATS

Comme il a été précisé dans l'introduction, les résultats portent uniquement sur la détection de frontières syntagmatiques et lexicales à partir de l'analyse des variations de la fréquence fondamentale et de la durée. L'algorithme de détection des pics significatifs de F0 est susceptible de trouver certaines fins de mots (dernier noyau vocalique) appartenant à une des deux catégories : "fin de syntagme" (frontière syntagmatique), "fin d'unité lexicale" uniquement (frontière lexicale).

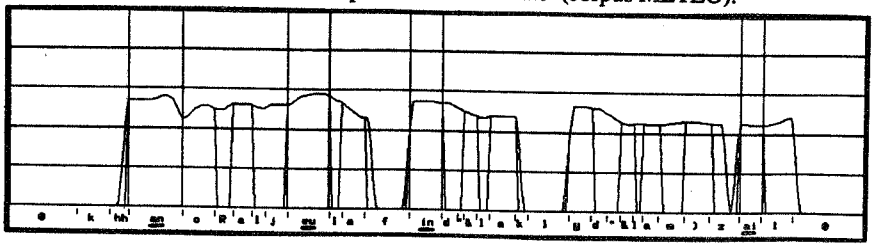
Cet algorithme a été testé sur les trois corpus décrits précédemment. Les résultats sont présentés dans les tableaux [1, 2, 3, 4], conjointement avec la structure de données mise en place. Les frontières lexicales ou syntagmatiques non détectées automatiquement sont celles qui n'ont pas de marques prosodiques remarquables (selon nos critères, classes 1 et 2). les classes 3 à 16 représentent les noyaux vocaliques retenus comme marqueur prosodique de frontières syntagmatiques ou lexicales.



k ɑ̃ ɔ R a l j ø l a f ɛ̃ d ə l a k R y d ə l a m ɔ z ɛ l

Quand aura lieu la fin de la crue de la Mosell(e) ?

figure 2 : Calcul de la fréquence fondamentale (corpus METEO).



Quand aura lieu la fin de la crue de la Mosell(e) ?

figure 3 : positionnement des marques de frontières syntagmatiques et lexicales.

Corpus	METEO	LABISE	COMBESCURE
VV détectées prosodiquement (détection automatique, classes 3 à 16)	1189	319	131
VV détectées correspondant à FS ou FL	1076	301	127
% de fiabilité dans la détection des frontières	90 %	94 %	97 %
VV détectées ne correspondant pas à FS ou FL ("erreurs")	113	18	4
"erreurs" correspondant à des frontières de mots outils	42	6	2
VV détectées correspondant à FL	215	40	9
VV de type FL	882	158	36
VV détectées correspondant à FS	861	261	118
VV de typr FS	1282	479	159
VV par corpus	5802	1615	546

Tableau 1 : Résultats de la détection automatique de marqueurs prosodiques pour les trois corpus et sur l'ensemble des noyaux vocaliques (VV).
VV détectées prosodiquement : noyaux vocaliques marquées prosodiquement par la détection automatique
FL : frontière lexicale de l'étiquetage syntaxique
FS : frontière syntagmatique de l'étiquetage syntaxique

Une première remarque est à observer quant à l'utilisation des pauses dans le langage parlé. Les tableaux de résultats sur les trois corpus d'étude montrent l'importance des insertions de pauses lors de l'élocution et principalement dans le corpus METEO. Sur ce corpus de dialogue simulé, la pause constitue la source d'informations principale directement utilisable (ou presque) pour le marquage de frontières syntagmatiques ou même lexicales. Les différents types de pauses rencontrés sur ce corpus (pauses silencieuses ou de respiration, pauses remplies) se localisent en des points de jonction grammaticale où on retrouve alors facilement les frontières des constituants syntaxiques. J. Vaissière [18] signale d'ailleurs que ce phénomène n'est pas propre au français mais a été observé pour d'autres langues. Les pauses d'hésitation (ou reprise) marqueront quant à elles les frontières lexicales. Une des particularités prosodiques des pauses remplies ("euh") qui peut être très intéressante, est une remarquable stabilité de la courbe de F0 à chaque émision prolongée. Comme il est relaté dans [12], la valeur F0 moyenne observée sur ces pauses remplies fournit certainement une bonne indication de la F0 usuelle du locuteur ; valeur pouvant servir de référence et facilement repérable dans un énoncé. Pour ce même corpus (METEO), la durée de ces phénomènes varie de 200 ms à au plus 850 ms, alors que la durée des voyelles finales les plus allongées ne dépasse pas 400 ms. Lors de l'élocution, chaque pause importante marque une rupture prosodique (cassure dans la forme prosodique, rupture de linéarité) dûe soit à une hésitation dans le discours, soit à une ou plusieurs pauses réfléchies conduisant sur un apport de précision (lieu, date...).

Aucune remarque particulière ne concerne les autres corpus. Dans le corpus LABISE, de part sa nature, le nombre des pauses est réduit aux seules frontières entre groupes de souffle; tandis que pour COMBESCURE on ne les trouve qu'en fin de phrase.

Dans les tableaux de résultats, on peut noter l'importance de l'indice F0 seul (classe 5 par rapport aux autres classes) pour la détection des frontières lexicales. C'est aussi justement dans cette classe que le nombre des marqueurs ne correspondant ni à des frontières syntagmatiques, ni lexicales, est le plus élevé. Ces "erreurs" seront étudiées dans le prochain paragraphe. Il est à noter également que les différentes combinaisons des indices suivants : allongement en fin de syntagme, augmentation de F0, pause, participent toutes à la détermination des limites des constituants syntaxiques.

Les frontières syntagmatiques et lexicales sont détectées avec une bonne fiabilité de l'ordre de 90% pour METEO (94% pour LABISE, 97% pour COMBESCURE). A ce niveau des résultats, cela représente environ 1 noyau vocalique sur 5 1/2 marquant une frontière lexicale ou syntagmatique (respectivement 1 sur 5 et 1 sur 4). Sont comptées comme "erreurs" les noyaux des classes 3 à 16 qui ne correspondent ni à une fin de syntagme ni à une fin de mot lexical de l'étiquetage syntaxique des corpus. Parmi ces "erreurs", il faut noter un grand nombre de "mots outils" (articles, prépositions, auxilliaires, mots de liaisons ...) qui n'ont pas été étiquetés comme mots lexicaux (ML) : 42 noyaux vocaliques marquent ainsi la fin de "mots outils" pour METEO (respectivement 6 et 2). En tenant compte de ces noyaux, le pourcentage de fiabilité dans la détection des frontières est de l'ordre de **94%** (respectivement **96%** et **98%**).

Les autres "erreurs" proviennent essentiellement de la détection d'accents :
- accents affectant les mots longs (3 à 4 syllabes) correspondant aux accents lexicaux,
- accents d'emphase résultant à plusieurs reprises de l'accent régional (exemple corpus METEO : Nancy, 16 fois).

Exemples : (les "erreurs" sont en carractère gras)

(LABISE)

...	quand	ils ont vu	un voyageur	qui	s'avançait	...
	CNT	SV	SN	Cnt	SV	

...	le voyageur	réchauffé	a oté	son manteau	.
	SN	SAJ	SV	SN	

classes de marqueurs prosodiques des noyaux vocaliques	3 4	5 6 7 8	9 10 11 12	13 14 15 16	Total
VV détectées prosodiquement	55 29	323 99 42 7	260 75 54 33	66 55 80 11	1189
VV détectées correspondant à FS	32 20	131 63 29 5	244 72 49 28	55 49 74 10	861
VV détectées correspondant à FL	6 0	129 22 12 2	10 3 5 2	11 6 6 1	215
VV détectées ne correspondant pas à FS ou FL ("erreurs")	17 9	63 14 1 0	6 0 0 3	0 0 0 0	113
"erreurs" correspondant à des frontières de mots outils	8 9	13 2 1 0	6 0 0 3	0 0 0 0	42

Tableau 2 : Résultats corpus METEO par classe de marqueurs prosodiques

classes de marqueurs prosodiques des noyaux vocaliques	3 4	5 6 7 8	9 10 11 12	13 14 15 16	Total
VV détectées prosodiquement	6 0	106 41 11 0	42 37 22 2	9 23 20 0	319
VV détectées correspondant à FS	5 0	60 34 10 0	40 37 22 2	9 22 20 0	261
VV détectées correspondant à FL	0 0	32 6 1 0	1 0 0 0	0 0 0 0	40
VV détectées ne correspondant pas à FS ou FL ("erreurs")	1 0	14 1 0 0	1 0 0 0	0 1 0 0	18
"erreurs" correspondant à des frontières de mots outils	0 0	3 1 0 0	1 0 0 0	0 1 0 0	6

Tableau 3 : Résultats corpus LABISE par classe de marqueurs prosodiques

classes de marqueurs prosodiques des noyaux vocaliques	3 4	5 6 7 8	9 10 11 12	13 14 15 16	Total
VV détectées prosodiquement	5 0	33 29 10 0	24 11 13 3	1 1 1 0	131
VV détectées correspondant à FS	4 0	24 27 9 0	24 11 13 3	1 1 1 0	118
VV détectées correspondant à FL	0 0	6 2 1 0	0 0 0 0	0 0 0 0	9
VV détectées ne correspondant pas à FS ou FL ("erreurs")	1 0	3 0 0 0	0 0 0 0	0 0 0 0	4
"erreurs" correspondant à des frontières de mots outils	1 0	1 0 0 0	0 0 0 0	0 0 0 0	2

Tableau 4 : Résultats corpus COMBESCURE par classe de marqueurs prosodiques

(METEO)

J'aimerais	connaître	l'ensoleillement	ou	les précipitations
ML				
SV		SN	Cnt	SN

... | actuellement | au centre | de l'agglomération | messine |.
| | ML | ML | |
| SA | | SN | |

... | dans la Moselle |.
| SN |

Y avait-il	des brumes	matinales	ce matin	dans les vallées
	ML			ML
SV	SN		SN	SN

vosgiennes |.

Euh	quelle est	actuellement	la pression	euh
			ML	H
H	SV	SA		SN

atmosphérique | à Nancy |.
| SN |

Il est à noter enfin la part importante des critères de F0 et de durée vocalique (corpus METEO) à la détermination des limites des constituants syntaxiques ; le critère de pause pouvant ou non intervenir dans la détection de telles frontières (voir le nombre des noyaux vocaliques correspondant à ces différentes classes dans les tableaux de résultats). On retrouve ainsi à travers ces corpus les réalisations possibles des différents critères caractérisant les fins de syntagmes décrits par Vaissière [17], Rossi [13], Meloni et Guizol [9], Rossi et Di Cristo [14], Perennou et Caelen [10].

6. CONCLUSION

L'utilisation de tels paramètres prosodiques comme aide aux modules lexical, syntaxique-sémantique est donc envisageable ponctuellement sans recours à la médiation de données prosodiques structurelles. Les résultats obtenus ont montré qu'il était possible de prendre en considération certains paramètres prosodiques et d'établir un système automatique relativement fiable de détection des frontières syntagmatiques et lexicales. Quelque soient les critères d'utilisation de ces paramètres prosodiques, lorsque l'on aborde le problème de la détection automatique, on est confronté à un autre problème qui est celui de la normalisation et de la recherche de seuils généralement calculés sur deux dimensions temporelle et fréquentielle. Les seuils utilisés dans cette étude sont en totale dépendance avec les paramètres calculés sur le signal. Leurs choix ont permis de maintenir un excellent score dans le marquage des frontières syntagmatiques et lexicales tout en conservant un bon rapport de détection sur ces noyaux de l'ordre de 1 sur 5. Le pourcentage d'erreurs (mots outils exclus) est en moyenne inférieur à 4 %. Ces chiffres ont été obtenus sur la totalité des classes 3 à 16 ; on peut maintenant s'interroger sur le problème du choix des classes dont les éléments serviront de marqueurs prosodiques fiables utilisés par le système de détection, ou bien un problème équivalent qui est celui d'affectation de coefficients de certitude aux éléments des différentes classes. Dans les corpus COMBESCURE et LABISE, les seules erreurs obtenues par le système (mots outils exclus) se sont produites en classe 5 (2 pour COMBESCURE et 11 pour LABISE). Comme pour le corpus METEO, c'est principalement cette classe qui introduit le plus grand nombre d'erreurs ; il faudra donc en tenir compte lors de l'attribution de scores.

En ce qui concerne l'utilisation exclusive des paramètres F0, durée pris sur l'ensemble des noyaux vocaliques, ainsi que les pauses, on peut se demander si l'utilisation d'indices relatifs à F0 (comportement micro-mélodique) tel que l'effet de fréquence co-intrinsèque à l'attaque de la voyelle dans les syllabes CV, ne pourrait pas être pris en compte comme information supplémentaire dans un système d'analyse de la macro-prosodie (système actuel). Par exemple, on observe un dépassement de la cible tonale sur les premières période de la voyelle à la suite d'une obstruante non-voisée [6]. Cette information n'étant utilisable que dans la mesure où le système de décodage acoustico-phonétique est assez fiable sur ce type de phonème.

Références

[1] Bonin J.-J., "Détection d'indices prosodiques linguistiquement significatifs", Mémoire de DEA informatique, Université de Nancy I, 1987.

[2] Bristow G.J., Fallside F., "An autocorrelation Pitch detector with error correction", IEEE, pp 184-187, 1982.

[3] Carbonell N., Haton J.-P., Pierrel J.-M., "Elaboration expérimentale d'indices prosodiques pour la reconnaissance, application à l'analyse syntaxico-sémantique dans le système Myrtille II", Acte du séminaire "Prosodie et reconnaissance automatique de la parole", pp 59-91, Aix-en-Provence, 1982.

[4] Carbonell N., Bonin J.-J., "Détection de frontières syntagmatiques en parole continue : utilisation de la fréquence fondamentale", dans XVIIème JEP, pp 163-167 Nancy, 1988.

[5] Combescure P., "Vingt listes de dix phrases phonétiquement équilibrées", Revue d'Acoustique 14, 1981.

[6] Di Cristo A., "De la microprosodie à l'Intonosyntaxe", Thèse de Doctorat, Université de Provence, 1978.

[7] Fohr D., "APHODEX : un système expert en décodage acoutisco-phonétique de la parole continue", Thèse de Doctorat d'Université, Université de Nancy I, 1986.

[8] Laprie Y., "Snorri : un système interactif d'étude de la parole", XVIIème JEP, pp 71-76, Nancy, 1988.

[9] Meloni H., Guizol J., "Utilisation de paramètres prosodiques dans un système de reconnaissance automatique de la parole continue", Acte du séminaire "Prosodie et reconnaissance automatique de la parole", pp 93-120, Aix-en-Provence, 1982.

[10] Perennou G., Caelen G., "Utilisation de la prosodie pour la reconnaissance de la parole dictée", Acte du séminaire "Prosodie et reconnaissance automatique de la parole", pp 25-57, Aix-en-Provence, 1982.

[11] Rabiner L.R., Schafer R.W., "Digital Processing of Speech Signals" by Bell Laboratories, Prentice-Hall, 1978.

[12] Roméas P., "Statut de la prosodie dans les recherches relatives au dialogue homme-machine", 12ème TIPA, 1988.

[13] Rossi M., Di Cristo A., Hirst D., Martin P., Nishinuma Y., "L'intonation, de l'accoustique à la sémantique", Klincksieck, PARIS, pp 184-233, 1981.

[14] Rossi M., Di Cristo A., "Enquête des indices de segmentation prosodique de l'énoncé", Acte du séminaire "Prosodie et reconnaissance automatique de la parole", pp 141-164, Aix-en-Provence,1982.

[15] Rossi M., "Prosodie et technologies vocales", Journées Nationales du GRECO-PRC, Communication homme-machine, EC2 Editeur, pp 63-80, 1988.

[16] Sondhi M. M., "New methods of pitch extraction", IEEE Trans., AU-16, pp 262-266, 1968.

[17] Vaissière J., "Utilisation de paramètres supra-segmentaux comme aide à la segmentation en phonèmes", Acte du séminaire "Prosodie et reconnaissance automatique de la parole", pp 123-139, Aix-en-Provence, 1982.

[18] Vaissière J., "Language-independent prosodic features ", dans Cutler A., Ladd D.R. editors, "Prosody : Models and measurements", Springer-Verlag, Berlin Heidelberg New York Tokyo, 1983.

[19] Vaissière J., "The use of prosodic parameters in automatic speech recognition", dans Neiman, Lang, Sagerer (eds.), "Recent advances in speech understanding and dialog systems", NATO ASI Series, Springer-Verlag, 1988.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Les marqueurs acoustiques de l'énoncé en discours québécois spontané

Conrad Ouellon

Université Laval

Les recherches sur la prosodie du français québécois ont traditionnellement porté sur des corpus lus. Il demeure toutefois important de connaître le comportement des locuteurs en discours oral spontané. A partir d'extraits d'enregistrements découpés en énoncés par un test de perception (60 juges), nous tentons de définir les marqueurs acoustiques caractéristiques des fins d'énoncés. Le rôle particulier de la pause en discours spontané sera mis en évidence. Certains aspects de la distribution de l'énergie propres à favoriser la perception d'un énoncé ressortiront également de l'analyse de l'idiolecte retenu.

Au Québec, les recherches sur le phonétisme du français parlé datent d'à peine vingt-cinq ans. Pendant une première période, jusqu'à 1975 approximativement, les travaux des phonéticiens ont permis l'établissement d'un bon inventaire des réalisations phonétiques des Québécois et de leurs caractéristiques articulatoires. Les corpus étaient alors lus et les informateurs étaient, pour l'époque, très scolarisés. La plupart des chercheurs travaillaient selon les méthodes alors en usage à l'Université de Strasbourg (avec Simon P. et Straka G.) et tiraient leurs données des techniques cinéradiologiques (Gendron 1966; Charbonneau 1971; Santerre 1971, 1974; Rochette 1973; Ouellon 1978; etc). A partir de 1970, s'ajoutent à ces travaux les contributions des phonologues (Morin 1979, 1982; Dumas 1972, 1977, 1981), des dialectologues (Juneau 1972; Juneau et Poirier 1973), puis des sociolinguistes (Sankoff et Cedergren 1971; Deshaies 1974; Thibault 1979; Paradis 1985; etc). Ces derniers ouvrent à leur tour de nouvelles perspectives dans les études sur le français québécois. On tiendra maintenant davantage compte de la variation linguistique, des caractéristiques socioculturelles des locuteurs.

Au cours de ces trois décennies, on ne recense que peu d'études sur la prosodie du québécois. On peut en citer trois (Marcel Boudreault, Gilles Lavoie, Normand Beauchemin) qui toutes portent sur du texte lu. Les travaux du groupe de Pierre Léon ont également abordé la question.

Depuis quelques années, on constate un intérêt de plus en plus marqué pour des études sur les aspects suprasegmentaux de la parole. De nombreux laboratoires de recherche français s'intéressent à ces questions, en particulier à Lannion (CNET), à Grenoble (Institut de la communication parlée), à Aix-en-Provence (Institut de phonétique). Au Québec, on peut signaler entre autres les travaux menés à l'Université de Montréal, à l'UQAM, à l'INRS, au CCRIT. Une simple description des sons de la parole humaine, de leurs caractéristiques articulatoires ou acoustiques ne peut suffire à rendre compte de la complexité du discours oral. Les chercheurs s'entendent sur la nécessité de scruter d'autres dimensions et d'étudier plus à fond le phénomène d'intonation, le rôle de l'accent et sa nature, l'influence des pauses, des hésitations...

Le projet de recherche sur la Prosodie du français québécois (PROSO) veut faire progresser les connaissances sur la prosodie du québécois à partir de l'étude d'un corpus en discours oral spontané. Il s'inscrit aussi dans une approche qui n'a que peu considéré les phénomènes suprasegmentaux (sauf Cedergren H. et Simoneau L.), celle de la sociolinguistique variationiste.

Dans cette communication, je tenterai d'apporter quelque éclairage sur un point particulier de la question, celle de la définition de l'énoncé en discours spontané suivi. Plus précisément, je dégagerai certains marqueurs acoustiques qui permettent de délimiter l'unité "énoncé". A l'écrit, l'unité de discours se définit relativement bien. Elle est normalement marquée par des ponctuations fortes. Elle se distingue aussi par une structure syntaxique complète et une certaine unité sémantique. Ces caractéristiques ne valent cependant plus en discours oral, spontané de surcroît. L'unité syntaxique peut être rompue, l'unité sémantique peut très bien ne pas exister. L'oral spontané n'est pas exempt de reprises, d'hésitations, de pauses imprévues, il renferme aussi des éléments redondants et non fonctionnels tels euh, t'sé, j'sais pas, ben, ... D'un point de vue acoustique, on peut de façon temporaire dire que l'énoncé est un segment de discours qui débute à la frontière d'un changement marqué dans la courbe de Fo ou dans la distribution de l'énergie. L'intention interrogative, énonciative, exclamative influencera les paramètres acoustiques, en particulier la courbe de Fo qui sera ascendante ou descendante selon le cas.

Cette définition d'énoncé subira forcément quelques modifications dans les pages qui suivent. C'est à partir de l'étude d'énoncés perçus à l'intérieur d'un discours oral spontané que nous en suggérerons une nouvelle description acoustique. Notre approche ne suppose donc pas une définition préalable de l'énoncé; c'est à partir du jugement d'un groupe de témoins que nous le délimiterons. Une fois analysés quelques-uns des marqueurs acoustiques de début et de fin d'énoncé en français québécois, nous examinerons la possibilité de les hiérarchiser.

A) MÉTHODOLOGIE UTILISÉE

La méthodologie que nous avons retenue s'inspire de celle que décrivent Lehiste I., puis Kreiman J.; nous présenterons ultérieurement les résultats d'une analyse comparative du même corpus en version filtrée (signal résiduel).

-1) Le corpus

Le corpus est constitué de dix extraits d'une durée de dix secondes chacun repiqués d'une entrevue linguistique faite à Chicoutimi en 1982. Cette entrevue, de même que les soixante autres de l'enquête de Claude Paradis, se déroulait chez le sujet interviewé et de tierces personnes pouvaient intervenir. L'informateur est un professionnel âgé de 32 ans, natif de la région. Les dix extraits sont indépendants les uns des autres et peuvent contenir plus d'un énoncé. On n'y relève toutefois aucune intervention d'un tiers, ni de l'enquêteur.

-2) Le test de perception

Les extraits retenus ont été numérisés et traités à l'aide de Micro Speech Lab. MSL Audio, un utilitaire du système MSL, a servi à l'élaboration du test de perception. Nous avons enregistré les dix extraits sur bande magnétique de la façon suivante: chacun a été répété trois fois avec insertion d'une pause de 5 secondes, cette pause ayant été allongée à dix secondes entre les extraits. L'enregistrement final durait donc environ dix minutes.

Nous avons soumis le test à deux groupes d'auditeurs en leur demandant d'indiquer par un trait sur le texte qui leur était fourni ce qui était perçu comme le début d'un énoncé. Le texte fourni (voir annexe 1) ne laissait voir ni ponctuation, ni majuscule, ni autre signe susceptible d'influencer la décision. Le groupe A se composait de 60 étudiants n'ayant reçu aucune formation universitaire en phonétique. Le second groupe d'auditeurs, groupe B, était formé de 14 étudiants du programme de linguistique inscrits à un cours de phonétique acoustique.

-3) Le traitement acoustique du corpus

Les tracés oscillographiques fournis par MSL ont tous été délimités et chacun des éléments phonétiques a été repéré sur les listes de valeurs numériques de Fo et d'énergie, ces paramètres étant mesurés par tranches de 25 ms.

B) RÉSULTATS DE L'ANALYSE

-1) Le test de perception

Comme nous l'avons mentionné plus haut, l'énoncé que nous voulons analyser est l'énoncé **perçu**. A partir du test de perception, nous ajouterons une précision à notre description de l'énoncé: dans le cadre de cette recherche, ce que nous entendrons par énoncé de discours oral spontané, **c'est ce qui est perceptuellement senti comme une unité par au moins 10% d'auditeurs**. Cette unité ne peut aisément se décrire du point de vue de la perception; c'est une partie d'un extrait de discours qui commence là où les juges ont eu "l'impression que débutait une phrase". Le corpus (Annexe) s'est ainsi trouvé découpé en 50 énoncés; notre étude ne retiendra pas le dernier énoncé de chacun des dix extraits, puisqu'ils ne sont pas complets (la durée de chaque extrait a été limitée à dix secondes exactement). Les taux de reconnaissance d'énoncé que nous avons obtenus font voir à l'évidence que certains énoncés sont mieux perçus, d'autres moins bien perçus. Nous avons donc décidé d'examiner de quelle façon se comportaient les facteurs acoustiques en fonction des taux de reconnaissance. Dans la suite de l'exposé, nous comparerons donc deux classes d'énoncés, ceux perçus par 50% (n=30) et plus d'auditeurs et ceux retenus par moins de 50% des juges. Cette façon de procéder nous permettra de dégager les facteurs qui font qu'un énoncé est mieux ou moins bien perçu comme une unité dans le discours oral spontané.

-2) L'analyse des données numériques

Après avoir délimité les éléments phonétiques sur les tracés oscillographiques et sur les listes numériques, nous avons analysé et mis en rapport les paramètres acoustiques suivants:
- présence et durée des pauses
- courbe de Fo
- courbe et valeurs d'énergie (En)

Notre attention, en regard de nos objectifs, s'est davantage portée sur l'analyse du signal en début et en fin d'énoncé. Nous avons quand même considéré le mouvement général des courbes Fo et En pour en tirer quelques observations. Considérant aussi que notre définition d'énoncé se fonde sur la perception, nous tenterons de dégager des facteurs acoustiques qui favorisent une meilleure perception de cet énoncé.

a) La pause:

Nous appelons pause ce qui correspond à un silence véritable. Les hésitations, les périodes d'occlusion de consonnes n'en font donc pas partie. Sur les listes numériques, les pauses se caractérisent par des valeurs d'Én inférieures à 100, ce qui équivaut à une perte de 20 dB. Dans la plupart des cas où elles ont été observées, elles durent plus de 150ms.

Pause > 150ms

Dans l'ensemble de notre corpus, nous relevons 24 pauses dont la durée varie de 150ms à 1900ms. Dix-sept suivent un énoncé perçu à 50%+ (50% et plus d'accord chez les témoins). Trois autres énoncés perçus par 48% d'auditeurs précèdent une pause.

On relève cependant quatre pauses en position interne des énoncés du corpus. Les auditeurs ne les ont pas jugées comme des limites d'énoncés.

Par ailleurs, l'effet d'une pause intervenant après un énoncé dont la courbe de Fo est décroissante est univoque. Les quatorze cas relevés ont été perçus à 50%+. Toutefois, si la pause suit un énoncé dont le Fo est croissant, cet énoncé peut être senti à 50%+ (1D, 5C, 10C) ou à 50%- (5A, 7B, 7C); il faut signaler que le taux de perception de ces derniers exemples avec Fo croissant est de 48%. On peut donc conclure que peu importe la direction de la courbe de Fo, la présence d'une pause subséquente suggère fortement à l'auditeur une fin d'énoncé.

L'importance de la fonction de marquage de la pause est déjà connue.

> " The acoustic analysis of speech shows that speakers insert a large number of pauses when talking. **Respiratory pauses** represent only a part of all pauses; there are also **hesitation pauses**, but the majority of pauses (including the respiratory pauses) are located at **grammatical junctures**. The use of pauses as major boundary markers between and within sentences seem to be similar across thoses languages for which there is available data." (Vaissière J. 1988: 82)

Pause <150ms

Si l'énoncé n'est pas suivi d'une pause, ou si cette dernière dure moins de 150ms, la perception de fin d'énoncé chez les auditeurs du groupe témoin ne paraît pas aussi assurée. Il convient toutefois d'apporter quelques précisions.

Si l'absence de pause se combine à un autre facteur négatif comme une augmentation des valeurs de Fo en finale, l'énoncé est perçu à 50%; c'est le jugement porté sur onze des treize exemples qui répondent aux deux conditions. Seuls 4C et 8A font exception et sont perçus à 50%+; les témoins sont quand même loin de s'entendre, avec des taux de perception de 55% et 67% respectivement.

Par contre, la combinaison d'une absence de pause et d'une décroissance de la courbe de Fo, c'est-à-dire d'un facteur négatif et d'un facteur positif, semble produire des résultats ambigus à la perception: trois énoncés sont en effet retenus à 50%-, trois autres à 50%+. L'analyse plus poussée permet cependant de croire que, dans ces conditions, l'auditeur aura tendance à percevoir un énoncé à 50%+ si n'interviennent pas d'autres facteurs. L'exemple 3C, avec un taux de reconnaissance de 23%, se caractérise par une structure syntaxique tronquée: dans ce cas, il est possible qu'ait pu jouer le sentiment linguistique des témoins; de plus, l'énoncé se termine sur une longue voyelle d'hésitation de 1050ms. L'exemple 10A (38%) est marqué par deux facteurs négatifs que nous traiterons plus loin, Fof > Foi et une courbe de valeurs d'Énergie croissante. Le dernier exemple, 6D, avec un taux de perception de 48%, aurait tout aussi bien se retrouver dans le groupe des énoncés à 50%+; dans ce cas-ci, l'absence d'une véritable pause (>150ms) semble avoir contrebalancé l'influence positive d'une courbe de Fo décroissante.

b) La courbe de Fo

Les caractéristiques intonatives de la phrase énonciative sont relativement bien connues. L'énonciation est normalement marquée par une décroissance de Fo en finale (Fouché 1952; Boudreault 1966; Rossi et DiCristo 1982; Vaissière 1988). Est-ce qu'on dénote le même comportement de la courbe de Fo en discours oral spontané?

Fo décroissant

Vingt énoncés se terminent sur une diminution des valeurs de Fo. Ils sont normalement perçus à 50%+ comme des énoncés. Nous ne notons que trois contre-exemples (3C, 6D, 10A) où, malgré un Fo décroissant,

l'énoncé est perçu à 50%- (23%, 47% et 38% respectivement). Nous avons traité ces exemples plus haut.

Fo croissant
La courbe de Fo est ascendante à la finale de dix-neuf énoncés. Conformément à ce qu'on peut attendre, 74% d'entre eux (n=14) sont perçus à 50%-. Parmi ceux-ci toutefois, trois sont suivis d'une pause (5A, 7B, 7C) et le pourcentage de reconnaissance d'énoncé est alors de 48% pour chacun des cas, ce qui illustre une fois de plus l'importance du marqueur pause.
Des cinq énoncés jugés à 50%+, quatre précèdent une pause (1D, 4C, 5C, 10C), ce qui confirme encore le rôle prédominant du facteur pause dans la perception d'énoncé. L'exemple 8A possède les caractéristiques de l'énoncé perçu à 50%-, mais 67% des auditeurs l'ont retenu; le seul facteur parmi ceux considérés qui puisse expliquer ce jugement de perception est le fait que le Fo de la syllabe finale (Fof) est plus élevé que le Fo de la syllabe initiale de l'énoncé suivant (Foi).

c) Le rapport entre Fof et Foi
Compte tenu des caractéristiques de notre corpus qui regroupe au moins trois énoncés par extrait, il nous a paru intéressant de considérer les rapports qui pouvaient exister entre la valeur Fo de la syllabe finale d'un énoncé (Fof) et celle de la syllabe initiale de l'énoncé suivant (Foi), les mesures de Fo étant prises sur la voyelle.

La tendance à la baisse de la courbe de Fo pendant la réalisation d'une phrase est déjà connue (Pierrehumbert 1979; Vaissière 1988). Bien que le phénomène de déclinaison ne signifie pas qu'il y ait chute continue de la courbe de Fo, il laisse entendre que la valeur de Fo en finale est plus basse qu'en initiale d'énoncé. On a aussi remarqué que dans une longue phrase, la ligne de déclinaison se brisait aux frontières prosodiques des sous-phrases (correspondant aux frontières syntaxiques); elle se caractérisait alors par une légère remontée ("reset") avant de continuer à décroître (Collier 1989:38).
Nous avons posé comme hypothèse que si, de part et d'autre d'une frontière prosodique, on pouvait noter une remontée de Fo, de la même façon le Fof d'un énoncé perçu par une majorité d'auditeurs devrait être plus bas que le Foi de l'énoncé subséquent.

Fof < Foi
On note vingt-quatre énoncés pour lesquels Fof < Foi
- Si la courbe de Fo est décroissante, ce qui se produit en dix-huit (75%) occasions, l'énoncé est perçu à 50%+ à deux exceptions près (3C, 6D) que nous avons déjà commentés en 2)b); l'ajout du facteur positif Fof < Foi ne réussit pas à annuler l'effet des facteurs négatifs évoqués.
- Si la courbe de Fo remonte en finale (6 cas), le jugement des témoins sera partagé. Les énoncés seront sentis à 50%+ si s'additionne l'influence d'une pause subséquente, même brève (1D et 4C). Par contre, l'absence de pause tend à provoquer un faible taux de reconnaissance de l'énoncé (< 30%) en 2D, 2F et 4A. Ce dernier exemple représente cependant un énoncé tronqué et il est possible qu'ait également joué le sentiment linguistique des auditeurs dans leur décision. Le seul contre-exemple est l'énoncé 7B, perçu par 48% des juges, malgré la présence d'une longue pause.

Fof > Foi
Quinze finales d'énoncés affichent un Fo plus élevé que celui de la syllabe initiale de l'énoncé suivant.
- Si cette situation survient en contexte de Fo croissant, l'énoncé sera normalement perçu à 50%- (10 cas sur 13). Les énoncés 5C, 10C et 8A font exception; dans les deux premiers cas, il a fallu déterminante l'influence de la pause d'une durée respective de 650ms et 975ms. En 8A, tous les marqueurs déjà étudiés sont négatifs et nous ne sommes pas en mesure d'expliquer le jugement porté; d'autres paramètres non examinés permettraient peut-être de résoudre ce cas. Toutefois, le pourcentage de perception d'énoncé n'est pas très haut pour ces trois exemples, à 73%, 72% et 67%.
- Si la courbe de Fo décroît en finale, avec Fof > Foi, il semble nécessaire qu'apparaisse une pause après l'énoncé pour qu'il soit retenu à 50%+. C'est ce qui se produit en 1B où la pause subséquente dure 1150ms. En 10A, en absence de pause, la chute de Fo n'a pas suffi à démarquer l'énoncé, perçu par 38% des témoins seulement; précisons qu'intervient ici une remontée de la courbe d'énergie, facteur négatif sur lequel nous reviendrons.

Le facteur que nous venons d'analyser (rapport entre Fof et Foi) paraît jouer un rôle de marqueur d'énoncé si on l'étudie en conjonction avec la direction de la courbe de Fo en finale.

d) La courbe d'Énergie (Én)
L'examen des valeurs d'énergie ne révèle à prime abord que fort peu de renseignements sur le rôle qu'elle pourrait jouer dans la démarcation des énoncés. L'énergie décroît en finale d'énoncé dans trente-et-un exemples sur trente-neuf (79%), ce qui n'est guère surprenant; dix-huit d'entre eux sont perçus à 50%+. On pourrait peut-être s'attendre à ce qu'une montée d'énergie ne marque que des énoncés retenus par 50%-des auditeurs, mais dans trois cas sur huit, ils sont sentis à 50%+.

Le regroupement des données fait à partir des taux de perception révèle cependant des tendances intéressantes.

Tableau 1: Perte d'énergie des énoncés perçus à 50%+ et 50%-

	Én-	Én+	DÉn-	DÉn+
Énoncé à 50%+	-9(18)	-8.5(3)	2.5(18)	1(3)
Énoncé à 50%-	-7(12)	-5.5(5)	3 (12)	2 (5)

Én : Perte moyenne d'énergie (dB) sur la voyelle en syllabe finale
- + : Sens décroissant ou croissant de la courbe Én en fin d'énoncé
DÉn : Écart (dB) entre la perte d'Én sur l'avant-dernière et celle sur la dernière syllabe de l'énoncé

Tableau 2: Perte d'énergie (DÉn) fonction du taux de perception d'énoncé

Perception (%)	Én- (dB)	Én+ (dB)	Én (dB)	DÉn- (dB)	DÉn+ (dB)
75 - 100	-9 (12)	-8 (1)	-9 (13)	-2.5(12)	+2 (1)
50 - 75	-8 (6)	-8.5 (2)	-8 (8)	-2 (6)	+1 (2)
25 - 50	-7 (8)	-7.5 (3)	-7 (11)	-3 (8)	+2 (3)
10 - 25	-7 (4)	-3 (2)	-5.5 (6)	-2.5 (4)	+2.5(2)

n.b.: L'énoncé 9C n'a pas été considéré à cause de difficultés de segmentation.

Si l'on considère les énoncés selon qu'ils ont perçus par plus de 50% (50%+) ou moins de 50% (50%-) des juges (tableau 1), on constate que la perte d'énergie en finale est plus forte lorsque l'énoncé est perçu à 50%+, peu importe le sens de la courbe d'Én. Un niveau d'Én relativement haut laisse entendre que l'énoncé n'est pas complet, qu'on peut donc attendre une poursuite du discours. Cette hypothèse se trouve partiellement confirmée par l'examen du Tableau 2, plus nuancé: plus la perte d'Én est élevée sur la syllabe finale, peu importe la direction de la courbe d'Én, plus l'énoncé tend à être perçu à 50%+. Inversement, une faible perte d'Én calculée sur le noyau de la dernière syllabe entraîne la perception d'un énoncé par moins de 50% des témoins.

La comparaison de la mesure de perte d'énergie (DÉn) entre l'avant-

dernière et la dernière syllabe ne révèle rien de particulier si on l'analyse en regard du taux de perception d'énoncé (tableau 2). Le caractère restreint de l'échantillon en est peut-être responsable. Le regroupement des énoncés en deux classes seulement (50%+ et 50%) laisse en effet entrevoir (tableau 1) que l'écart entre les pertes d'Én semble plus faible dans le cas des énoncés à 50%+, en contexte de courbe d'Én croissante ou décroissante. Des tests sur un plus vaste corpus permettraient de vérifier la valeur de ces tendances.

Les pertes d'Énergie calculées en fin d'énoncé (DÉn) sont-elles perceptibles? Sorin C. (1981) a démontré qu'une augmentation d'intensité de 1 à 2 dB par rapport au signal naturel peut être détectée, si elle survient sur une durée d'au moins 200ms. Ces valeurs correspondent à ce qu'a mesuré Rossi (1971) comme seuil différentiel d'intensité pour de voyelles longues prononcées isolément.

C) LES MARQUEURS ACOUSTIQUES DE L'ÉNONCÉ

Malgré le nombre relativement restreint des exemples que nous avons traités, nous pensons être en mesure de donner quelques caractéristiques acoustiques qui permettent de délimiter les énoncés du discours oral spontané.

L'énoncé perçu à 50%+ aurait les marques suivantes (facteurs positifs):

- présence d'une pause subséquente,
- courbe de Fo décroissante,
- Fo de la syllabe finale (Fof) plus bas que Fo de la syllabe initiale (Foi) de l'énoncé subséquent,
- perte d'Én relativement forte sur la dernière syllabe.

L'énoncé perçu par moins de 50% d'auditeurs se caractériserait par les marques suivantes (facteurs négatifs):

- absence de pause subséquente,
- courbe de Fo croissante,
- Fo de la syllabe finale (Fof) plus haut que Fo de la syllabe initiale (Foi) de l'énoncé suivant,
- perte d'Én relativement faible sur la dernière syllabe.

D) LA HIÉRARCHISATION DES MARQUES

Malgré l'inévitable présence de cas particuliers, peut-on penser à une hiérarchisation des marques acoustiques de l'énoncé dans l'idiolecte analysé? Est-il possible de préciser la contribution de chacun des facteurs étudiés?

> As stated before, the relative contribution of phonetically-conditionned variations and of each of the many functions of prosody to the determination of the observed quantitative values of the three physical variables (Fo, duration and energy) is not easily determined. One specific function (such as stress marking or juncture marking) is generally not defined by a single prosodic parameter but by a combination of all prosodic cues: duration, intensity and Fo (and eventually by the insertion of a pause). Furthermore, the exact contribution of each parameter varies as a function of the context. (Vaissières J. 1988: 76)

Nous avons pu observer au cours de cette recherche que l'effet d'une longue pause comme marqueur d'énoncé pouvait être annulé par la présence d'une courbe de Fo montante. Inversement, une courbe de Fo décroissante, traditionnellement reconnue comme indice de fin d'énoncé, peut ne plus jouer ce rôle si il y a absence d'une pause subséquente. Il ne semble guère possible de déterminer le rôle exact de chaque facteur. Quelles conclusions tirer de ces faits?

> There are three consequences: first, a combination of at least three parameters (pitch, duration and intensity) is desirable to achieve more reliable decisions, even for isolated words; second, contextual rules are necessary to achieve reliable decisions in continuous speech; third, a prosodic event (such as a vowel lenghtening or an Fo rise) will often receive more than one interpretation. (Vaissières J. 1988: 77)

Ces considérations, faites dans le cadre de la reconnaissance de la parole, valent probablement tout aussi bien pour un test de perception. Dans notre recherche cependant, certains paramètres semblent plus importants que d'autres (nous rappelons que nous n'avons pas traité des durées, ni de la distribution des valeurs de Fo et d'énergie sur la totalité de l'énoncé):

1) La présence d'une véritable pause, d'un silence, semble un facteur déterminant dans la perception de fin d'énoncé. Sa présence peut contrebalancer l'effet d'une courbe de Fo croissante.
2) Dans le cas de phrases énonciatives (typiques de notre corpus), l'intonation descendante en finale caractérise le plus souvent l'énoncé perçu par 50%+ des témoins.
3) La comparaison de la valeur de Fo de la syllabe finale de l'énoncé (Fof) et de celle de la syllabe initiale de l'énoncé suivant (Foi) peut aussi servir à marquer la finale d'énoncé quand Fof < Foi. Cette marque paraît aussi importante que la direction de la courbe de Fo. Ces deux facteurs peuvent jouer un rôle déterminant s'ils agissent dans le même sens, en absence de pause:
- si [Fo- + (Fof < Foi)]: l'énoncé sera retenu par 50%+ des témoins,
- si [Fo+ + (Fof > Foi)]: l'énoncé sera saisi par 50%- des auditeurs.
Si les facteurs s'opposent, la fréquence ne semble plus jouer un aussi grand rôle dans le marquage de l'énoncé et c'est alors la présence ou l'absence d'une pause qui fera qu'un énoncé sera retenu par une majorité ou non d'auditeurs.
4) Considéré pour un énoncé en particulier, le rôle de la distribution d'énergie dans l'identification de la fin d'énoncé n'est pas paru important. Les données regroupées du tableau 1 permettent toutefois de croire qu'une forte baisse d'Én combinée à un faible écart des mesures de perte d'Én des noyaux vocaliques des deux dernières syllabes de l'énoncé favoriseraient la perception d'un énoncé à 50%+.

Plusieurs questions demeurent évidemment sans réponse, faute de données quantitativement suffisantes pour faire l'objet d'un traitement statistique. Il faudra, dans une autre étape de la recherche, accroître la taille de notre corpus, mesurer le degré de corrélation des marqueurs, tester la fiabilité de nos résultats par la synthèse notamment. Il faudra aussi vérifier si la seule ligne prosodique peut suffire à la reconnaissance de l'énoncé, abstraction faite des aspects sémantiques et syntaxiques. Il faudra enfin étudier plus à fond le rôle de la distribution d'énergie dans la prosodie.

BIBLIOGRAPHIE

Boudreault, Marcel (1968), *Rythme et mélodie de la phrase parlée en France et au Québec*, Québec, Presses de l'Université Laval.

Cedergren H.J., et Louise Simoneau (1985), «La chute des voyelles hautes en français de Montréal. 'As-tu entendu la belle syncope?'», in M. Lemieux et H.J. Cedergren (1985), *Les tendances dynamiques du français parlé à Montréal*, tome 1, 57-144.

Charbonneau, René (1971), *Étude sur les voyelles nasales du francais canadien*, Québec-Paris: Presses de l'Université Laval-Librairie C. Klincksieck.

Collier, R, (1989), «Intonation Analysis: the Perceptionof Speech Melody in relation to Acoustics and Production», in *EUROSPEECH 89*, European Conference on Speech Communication and Technology, 2 vol. 1989.

Deshaies-Lafontaine, Denise (1974), *A socio-phonetic study of a Quebec French community: Trois-Rivières*, Thèse inédite de l'University of London.

Dumas, Denis (1972), *Le français populaire de Montréal. Description phonologique*, Thèse de maîtrise, Université de Montréal.

Dumas, Denis (1977), *Phonologie des réductions vocaliques en français québécois*, Thèse de doctorat inédite, Université de Montréal, 183p..

Dumas, Denis (1981), «Structure de la diphtongaison québécoise», *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 26:1, pp.1-61.

Fouché, P. (1952), *Phonétique historique du français*, Klincksieck, Paris.

Gendron, Jean-Denis (1966), *Tendances phonétiques du français parlé au Canada*, Paris-Québec: Klincksieck-Presses de l'Université Laval.

Juneau, Marcel (1972), *Contribution à l'histoire de la prononciation française au Québec*, Québec: Presses de l'Université Laval.

Juneau, Marcel, et Claude Poirier (1973), *Le livre de comptes d'un meunier québécois (fin XVIIe - début XVIIIe siècle). Édition avec étude linguistique*, Québec: Presses de l'Université Laval.

Kreiman, J (1982), "Perception of sentence and paragraph boundaries in natural conversation" *Journal of Phonetics*, 10.

Léon, Pierre et al. (1970), *Analyse des faits prosodiques/Prosodic feature analysis*, coll. Studia Phonetica, Montréal: Didier

Léon, Pierre R. et M. Rossi (1979), *Problèmes de prosodie 1*, coll. Studia Phonetica 17, Ottawa:Didier.

Léon, Pierre R. et M. Rossi (1980), *Problèmes de prosodie 2*, coll. Studia Phonetica 18, Ottawa:Didier.

Morin, Yves-Charles (1979), «La morphophonologie des pronoms clitiques en français populaire», *Cahiers de linguistique*, 9:1-36.

Morin, Yves-Charles (1982), «De quelques [l] non étymologiques dans le français du Québec», *Revue québécoise de linguistique*, 11-2: 71-93.

Ouellon, Conrad (1978), *L'enchaînement des voyelles suivies de consonnes occlusives en français*, thèse de doctorat inédite de l'université Laval.

Paradis, Claude (1985), *An acoustic study of variation and change in the vowel system of Chicoutimi-Jonquière (Québec)*, Thèse inédite de l'University of Pennsylvania, 326p.

Pierrehumbert, Janet (1979), «The Perception of Fundamental Frequency declination» in *J.A.S.A.*, no 66, 363-369.

Rochette, Claude (1973), *Les groupes de consonnes en français. Étude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie*, Paris-Québec: Klincksieck-Presses de l'Université Laval, 2 vol.

Rossi, Mario (1980), «Le cadre accentuel et le mot en italien et en français», in *Mélanges Faure. Problèmes de prosodie*, vol.1, 9-22.

Rossi, M. et A. Di Cristo (1982), «En quête des indices de segmentation prosodique de l'énoncé», in *Actes du séminaire Prosodie et reconnaissance automatique de la parole*, Institut de phonétique d'Aix-en-Provence, 141-164.

Sankoff, G. et H. Cedergren (1971), «Some results of a sociolinguistic study of Montréal French», in *Linguistic Diversity in Canadian Society*, Regna Darnell (ed.), Edmonton Linguistic Research, Inc. pp.61-87.

Santerre, Laurent (1974), «Deux E et deux A phonologiques en français québécois. Étude phonologique, articulatoire et acoustique», in *Le français de la région de Montréal. Aspects phonétique et phonologique*, en collaboration, Montréal, Presses de l'Université du Québec.

Santerre, Laurent (1971), *Les voyelles orales dans le français parlé à Montréal*, Thèse inédite de l'Université des sciences humaines de Strasbourg.

Sorin, C. (1981), «Functions, Roles and Treatment of Intensity in Speech», dans *Journal of Phonetics*, no 9, 359-374.

Thibault, Pierrette (1979), *Le français parlé. Études sociolinguistiques*, Sociolinguistic Series 5, Carbondale (USA)-Edmonton (Canada): Linguitic Research, Inc.

Vaissière, Jacqueline (1988), «The Use of Prosodic Parameters in Automatic Speech Recognition», dans *Recent Advances in Speech Understanding and Dialog Systems*, NATO ASE Series, vol.F46, éd. H. Niemann et al, Springer Verlag, Berlin.

CORPUS

(Apparaît entre barres obliques une lettre majuscule qui identifie l'énoncé perçu subséquent; entre parenthèses est indiqué le pourcentage de reconnaissance de début d'énoncé.)

Phrase 1: A/ ouais je fais (25%) /B/ ben ces temps-ci je fais deux jours (98%) /C/ en hiver je fais deux jours (60%) /D/ ça dépend de la clientèle (92%) /E/ au printemps pis à l'automne je fais quatre jours (30%) /F/ puis à entre les deux je fais trois

Phrase 2: A/ ouais (80%) /B/ disons ici au Québec on peut dire ça (85%) /C/ ah je fais de la phlébologie (30%) /D/ on peut pas dire je suis phlébologue (20%) /E/ parce que en principe c'est une spécialité ça (85%) /F/ mais c'est une spécialité européenne (27%) /G/ ca existe pas euh

Phrase 3: A/ ceux qui ont juste de l'esthétique y viennent moins longtemps parce qu'y en ont moins (98%) /B/ pis les autres ben qui en ont plus ben y viennent plus longtemps (40%) /C/ ça fait qu'au point de vue disons a (23%) /D/ je sais pas séance par séance euh j'en ai pas tant

Phrase 4: A/ quatre cinq autos mettons là de (20%) /B/ pas quatre cinq autos mais quatre cinq par voiture là (77%) /C/ tsé tout euh un petit village du qui se connaissant tsé (55%) /D/ une disait à telle autre on y va (10%) /E/ pis elle amenait sa soeur

Phrase 5: A/ ça c'était un c'était un groupe de Montréal (48%) /B/ qui avait ouvert une clinique ici dans la région (85%) /C/ ça fait qu'elle s'est trouvé quelqu'un ce qui était le vietnamien (73%) /D/ mais ça prenait une deuxième personne (22%) /E/ c'est moi qui suis

Phrase 6: A/ elle avait pas de cours (35%) /B/ elle a été formée chez un dentiste euh (95%) /C/ remarque qu'elle ne pratique pas chez les patients euh (95%) /D/ elle fait seulement qu'assister à la chaise (47%) /E/ alors c'est

Phrase 7: A/ ah oui (53%) /B/ de toute façon c'est une profession excessivement exigeante (48%) /C/ excessivement stressante (48%) /D/ que les gens les gens s'imaginent pas que c'est comme ça (96%) /E/ ben tu vois les gens

Phrase 8: A/ euh alors on travaille avec des petits instruments (67%) /B/ on travaille près de la langue euh (80%) /C/ on travaille près du plancher de la bouche près du palais près des joues euh (60%) /D/ on travaille avec des instruments rotatifs qui tournent euh à

Phrase 9: A/ ça fait que j'ai un moment donné j'ai j'ai j'ai dû n'en laisser tomber (88%) /B/ alors j'ai laissé tomber euh mes cours de guitare la liturgie (65%) /C/ non j'ai conservé la liturgie c'est-à-dire (28%) /D/ j'ai laissé tomber mes

Phrase 10: A/ y en a qui m'ont jugé comme opportunisme (38%) /B/ mais ça c'est toujours ça hein (32%) /C/ c'est-à-dire que tu donnes de ton temps (72%) /D/ parce que y a ben des après-midi de bureau que j'ai pas pu travailler à cause de ça (85%) /E/ tu donnes ton ton

2 PHONÉTIQUE ET PHONOLOGIE

Président: L.J. BOÉ
ICP-Grenoble,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

SYSTEMES VOCALIQUES : TYPOLOGIE ET TENDANCES UNIVERSELLES

VALLEE N. BOE L.J. & SCHWARTZ J.L.

INSTITUT DE LA COMMUNICATION PARLEE
INPG / ENSERG - UNIVERSITE STENDHAL - URA CNRS 368
Domaine Universitaire BP 25X 38040 Grenoble Cedex

RESUME.
Le but de cette étude est d'établir une typologie des systèmes vocaliques de différentes langues à partir d'une vaste base de données (UPSID), d'en dégager des régularités et des implications. Cette classification sera comparée et discutée à la lumière des travaux de Jakobson (1941), de Liljencrants & Lindblom (1972), de Crothers (1978) et de Lindblom (1975... 1988).

1. INTRODUCTION.

De nombreuses études ont été consacrées à la description phonétique et phonologique de différentes langues du monde. Il est possible aujourd'hui de disposer d'inventaires relativement importants pour proposer des typologies afin de mettre à jour les grandes tendances de développement des systèmes vocaliques.

Nous avons établi notre typologie sur un échantillon représentatif de 317 langues, réuni dans le cadre du projet UPSID : UCLA Phonological Segment Inventory Data Base ; Maddieson, 1986). Cette base de données a déjà servi à l'établissement d'un certain nombre d'hypothèses et de tendances phonologiques universelles (Crothers, 1978), que nous allons discuter (cf. §5 universaux de Crothers en annexe).

La tendance à la spécification en systèmes type constitue le problème majeur que l'on rencontre dans les typologies car n'y sont traitées que les voyelles orales et nasales. Cette approche exclut tout type de combinaison entre une structure de base et des phénomènes plus complexes tels que la longueur, l'aspiration, l'assourdissement, etc..

La littérature offre donc des typologies basées sur la régularisation d'inventaires plus ou moins importants, dans lesquels certaines données initiales sont écartées. Il en résulte une perte certaine et non négligeable d'information. La prise en compte de toutes les données pourrait mettre en évidence des informations nouvelles.

2. TYPOLOGIE - REGULARISATION DES SYSTEMES VOCALIQUES.

Nous avons pris en compte l'intégralité de la description des systèmes vocaliques fournie par Maddieson : nous n'avons pas utilisé d'équivalence typologique pour assimiler une voyelle x à une voyelle y ; nous n'avons pas non plus subdivisé les systèmes en sous-systèmes (Crothers, 1978). Toutes les voyelles sont donc prises en compte dans les arrangements.

Les systèmes ont été classés en types, suivant le nombre de voyelles qu'ils contiennent : entre 3 et 24 voyelles (il n'a pas été attesté de systèmes à 21, 22, ou 23 voyelles). Nous avons représenté chaque système sur une "grille" vocalique telle qu'elle a été proposée par Lindblom (1986). Les voyelles autres que orales et brèves sont localisées avec leurs correspondantes de même timbre vocalique. Il peut donc se trouver plusieurs voyelles en un même point de la grille vocalique (marquées ou non du diacritique de longueur, de nasalité, de pharyngalisation, d'assourdissement, de rétroflexion, de laryngalisation, etc.). L'identification des types a été réalisé par un codage basé sur le nombre de voyelles et la fréquence d'apparition du système dans les différentes langues (par exemple le système 3.1 /i a u/, 3.2 /e a o/, 4.2 /i ɐ a ɔ/ ; cf. figure 1).

3. TENDANCES UNIVERSELLES.

3.1. Occurrences.

Notre classification permet de mettre en évidence des tendances universelles de fréquence d'apparition.

Pour reprendre la terminologie de Lindblom (1988), toutes les langues, sans exception, possèdent des phonèmes vocaliques qui sont des segments de base (ex. /a/, /i/, /ɨ/ /a̟/... cf. figure 2). La plupart des systèmes à nombre élevé de voyelles ne développent pas de nouveaux timbres mais ajoutent plutôt un ou plusieurs traits à des segments de base (ex. /ã/, /ũː/, /ɹ̃ː/, /ã˞ː/).

Cette constatation ne confirme pas les résultats obtenus avec la théorie de la dispersion maximale, développée pour rendre compte de la répartition des voyelles dans l'espace vocalique. Elle peut s'expliquer par le fait que, dans les systèmes à nombre réduit de voyelles, les articulations de base sont suffisantes pour distinguer l'ensemble des voyelles, alors que les systèmes plus larges requièrent - pour une distinction susceptible de rester symétrique dans l'espace vocalique - plusieurs dimensions en combinant les formes.

Les 317 systèmes se répartissent en 220 arrangements avec une très nette dominance des systèmes à 5 voyelles (23% des langues)ce qui confirme la 8ème tendance universelle établie par Crothers (1978). Les systèmes de 3 à 10 voyelles représentent 80% de l'échantillon. Les systèmes larges sont donc minoritaires (cf. figure 3). On relève peu de types dominants à l'intérieur de ces systèmes : plus le nombre de voyelles est important, plus les chances de trouver plusieurs arrangements semblables s'amenuisent. Les systèmes les plus fréquents semblent avoir une meilleure dispersion dans l'espace des voyelles. Ils sont composés de voyelles stables, indépendamment de la taille du système.

Pour les voyelles cardinales extrêmes /i̟ a u̟/, la hiérarchie établie par Jakobson (1941) et confirmée par Greenberg (1966) : /a/>/i/>/u/ doit être ici réordonnée : /u/ implique /a/, /a/ implique /i/ (99% d'occurrences pour /i/, 97.8% pour /a/ et 94% pour /u/) ce qui n'aurait pas été le cas si on avait regroupé les occurrences de toutes les voyelles basses : /a̟, a, ɑ, ɒ/ (cf. figures 4a et 4b).

Bien que 44% des langues aient développé au moins une voyelle intérieure, cette catégorie est nettement minoritaire par rapport aux voyelles périphériques. Par ailleurs un déséquilibre apparaît entre les voyelles intérieures arrondies et non arrondies : les antérieures arrondies /ʉ, ʏ, ø, œ/ sont, d'une part, plus fréquentes que celles de la série centrale non arrondie /ɨ, ɪ, ə, ɘ, 3, ɐ/. D'autre part, les postérieures non arrondies /ɯ, ɯ̽, ɤ, ɤ̞, ʌ/ sont plus fréquentes que celles de la série centrale arrondie /ʊ, ʊ̽, ɵ, ɵ̞/. La comparaison entre les deux séries centrales montre que, dans tous les cas, les centrales non arrondies sont les plus fréquentes.

Environ 9% des langues ont développé une ou plusieurs antérieures arrondies donnant, par ordre de fréquence décroissante : /ʉ/>/ø/>/œ/>/ʏ/. Seules 5 combinaisons de ces voyelles - sur les 11 possibles - sont attestées dans 22 langues : la série la plus complète n'est pas la plus rare. Les séries d'antérieures arrondies à deux phonèmes apparaissent dans les systèmes ayant au moins 7 voyelles ; celles ayant 3 et 4 phonèmes apparaissent dans les systèmes à 16 et 19 voyelles.

Les voyelles postérieures non arrondies apparaissent dans 34 langues suivant l'ordre /ɯ/>/ɯ̽/>/ɤ/>/ɤ̞/>/ʌ/. A la différence des voyelles antérieures arrondies, les éléments de cette série sons parfois rencontrée sans la voyelle postérieure périphérique de même aperture. Les séries attestées contiennent au plus trois voyelles.

Nous pouvons conclure que :

- les voyelles antérieures arrondies apparaissent davantage par série (brève et/ou longue) ;
- les voyelles postérieures non arrondies figurent généralement seules dans leur catégorie (cf. figure 5 et figure 6).

Les voyelles nasales ne se rencontrent que dans 71 langues et elles n'apparaissent que dans les systèmes de plus de 7 voyelles, dans l'ordre décroissant : /ɑ̃/>/ũ/>/ĩ/, avec respectivement 20.2% d'occurrence, 19.6% et 18%, les autres voyelles nasales ayant une occurrence très faible. La diversité des voyelles nasales est loin d'être aussi importante en nombre que celle des voyelles orales (cf. figures 7a et 7b).

3.2. Implications.

Ce classement nous permet de mettre en évidence une hiérarchie dans le développement des systèmes vocaliques. Elle consiste en fait à déduire d'un système naturel à 3 voyelles, des systèmes à 4, 5, puis 6 voyelles, etc.. Autrement dit, il s'est agit de rechercher un système à n voyelles parmi des systèmes à n+1 voyelles. Notre étude nous a permis de construire un arbre des "tendances implicationnelles" pour les systèmes de 3 à 10 voyelles en prenant en compte les notions développées dans la théorie de la dispersion. Par exemple,l'insertion du phonème /æ/ n'existe dans la base de systèmes étudiés, qu'avec un passage de /a/ vers /ɑ/. Mis à part cet ajout, le système reste identique. Nous avons limité les modifications à un ou deux phonèmes au plus, car la prise en compte d'une restructuration plus vaste aurait été beaucoup plus complexe. Tout au long de la progression, les systèmes sont identifiés par leur code typologique. Une lecture ascendante du type A => B signifie que si une langue possède A, elle peut aussi posséder B qui apparaît nécessairement dans un système plus réduit que le système dans lequel se trouve A.

En partant de deux "triplets de base" /i a u/ et /e a o/ (systèmes à 3 voyelles les plus fréquents), apparaissent deux progressions qui fusionnent à partir des systèmes à 9 voyelles et qui ne peut se poursuivre au delà de 10 voyelles - entrent alors en jeu d'autres dimensions telles que la nasalité ou la longueur, ce qui entraînne moins de distinctions de timbres vocaliques correspondant au nombre total des voyelles des systèmes considérés (cf. figure 8).

Notre hiérarchie n'est donc pas généralisable à l'ensemble des systèmes vocaliques : d'une part elle se limite aux systèmes ne comportant que des timbres distinctifs de voyelles orales, et d'autre part elle est construite sur le principe de l'existence d'un système à n voyelles dans un système à n+1 voyelles. Il s'agit donc plutôt d'un arbre de tendances observées à l'intérieur de l'ensemble des systèmes vocaliques naturels.

Les voyelles de base sont donc privilégiées dans la hiérarchie. Ce sont celles qui apparaissent dans les plus petits systèmes et que l'on peut définir comme voyelles stables indépendamment de la taille du système. Parmi celles-ci, on peut noter /i, e, ɛ, a, ɔ, o, u/ (cf. figure 9). Les voyelles périphériques sont donc bien les premières voyelles à être développées par les systèmes. On les rencontre toutes dans des systèmes à 3 ou 4 voyelles.

Nous avons appliqué notre arbre des tendances implicationnelles à la parenté génétique afin d'observer si les deux progressions restent plus ou moins dans une même famille de langues voire dans un même groupe de langues, ou s'en éloignent nettement. Mais de manière générale, en passant d'un système à n voyelles à un système à n+1 voyelles, on ne retrouve pas systématiquement les mêmes familles de langues et encore moins les mêmes groupes de langues. Notre hiérarchisation se veut donc représentative de l'ensemble des langues du monde.

La comparaison avec celle de Crothers (1978a) traitant des systèmes naturels, celle de Liljencrants et Linblom (1972), celles de Lindblom (depuis 1975) et celle de Crothers (1978b) fondées sur des systèmes théoriques, montre que la plus voisine de notre arbre des tendances implicationnelles, reconstituant de manière générale la croissance des systèmes vocaliques, est celle de Crothers (1978a) qui, rappelons-le, a été établie sur une base de données, sous-ensemble d'UPSID (209 langues au lieu de 317).

3.3. Régularités.

L'étude des implications dans le développement des systèmes vocaliques, ainsi que quelques unes des régularités que nous avons pu mettre en évidence, montrent que la plupart des universaux proposés par Crothers (1978) ont besoin d'être nuancés, car les résultats de nos observations en font plus des tendances que des phénomènes universels. Travaillant à partir de données phonétiques, celui-ci a rangé bon nombre des réalisations phonétiques parfois très variables sous un même symbole vocalique. Ces données sont moins exposées aux exceptions et donc plus souples et plus malléables pour la démonstration de l'existence de phénomènes proprement universels, au risque de simplifications.

La recherche de régularités entre les deux dimensions du triangle vocalique, c'est-à-dire entre la distinction de hauteur et la distinction d'antériorité / postériorité, montre que toutes les langues distinguent au moins deux voyelles sur chacune des deux dimensions : dans tout système, il existe au moins 2 voyelles qui s'opposent par le degré d'aperture, et 2 qui s'opposent sur la distinction avant / arrière. La tendance générale dans les systèmes vocaliques est de 3 ou 4 degrés de distinction par l'aperture (haut, moins haut, moyen haut, moyen, moyen bas, moins bas, bas), et de 3 distinctions sur l'axe antéro-postérieur, et ceci quelle que soit la taille du système. L'absence d'une voyelle centrale basse explique le plus souvent l'existence de systèmes avec seulement 2 distinctions avant / arrière (un peu plus de 5% des systèmes étudiés). La suprématie des contrastes haut / bas sur les contrastes antérieur / postérieur est plutôt à considérer ici comme une tendance universelle des systèmes vocaliques naturels. Seulement 6% des systèmes qui échappent à cette tendance.

Même sans avoir fait d'équivalences typologiques, on peut confirmer que les voyelles périphériques sont toujours dominantes dans les systèmes. Le nombre de degrés d'aperture dans les voyelles antérieures tend à être supérieur (67.5% des systèmes) ou égal (24%) au nombre de distinctions de hauteur dans les voyelles postérieures.

Les voyelles centrales, arrondies /ʉ, ʉ̞, ɵ, ɵ̞/ ou non arrondies /ɨ, ɨ̞, ə, ə̞, ɜ/, s'organisent plus régulièrement sur l'axe haut / bas plutôt que sur l'opposition arrondi / non arrondi. Leur nombre n'excède pas celui des voyelles antérieures ou postérieures, sauf pour deux langues. A la différence de Crothers, et même si le phonème /ɨ/ est très fréquent dans les séries centrales orales (50% des 28 langues attestant une série centrale), on ne peut pas dire que /ɨ/ et /ʉ/ soient présents systématiquement dans les systèmes comportant une série centrale. On n'a jamais l'une ou l'autre de ces deux voyelles centrales hautes sans la présence d'une voyelle périphérique haute. Nous n'avons pas relevé d'antérieures arrondies dans les systèmes de moins de 5 voyelles dans le cas du /ø/ et dans les systèmes de moins de 7 voyelles pour les autres. Elles sont inexistantes dans les petits systèmes par rapport aux types non arrondis /ɯ, ɯ̞, ɤ, ɤ̞/. La quasi totalité des systèmes qui ont développé la voyelle /ʏ/ ou /ʏ/ possèdent la voyelle postérieure arrondie de même aperture : /u/ dans le cas du /ʏ/ et /o/ dans le cas de /ʏ/. D'autre part, /ʏ/ et /ʏ/ existent toujours en présence de /ɪ/ ou /ʅ/. La tendance générale des systèmes qui ont développé des voyelles postérieures non arrondies est de posséder l'antérieure non arrondie de même aperture. Les deux voyelles, haute et moins haute de cette catégorie (/ɯ/ et /ɯ̞/), peuvent exister sans la voyelle postérieure arrondie de même aperture. Du point de vue de la théorie de la dispersion, les positions spaciales stables de /ʏ/ et /ɯ/ d'une part, /ʏ/ et /ɯ̞/ d'autre part, en font des pôles d'attraction pour les voyelles dans le cadre de cette théorie.

Nous avons résumé, sous forme d'arbres implicationnels, la constitution des séries antérieures arrondies et postérieures non arrondies. Les flèches pleines indiquent une implication au sens fort du terme : la suite est observée dans tous les cas, alors que les flèches en pointillé indiquent une suite possible mais non systématique (cf. figure 10).

Les deux dimensions nasalité et longueur - caractéristiques des systèmes larges (et que l'on peut d'ailleurs rencontrer combinées dans certains systèmes avec des voyelles nasales longues) - sont, après les voyelles orales brèves, les plus développées par les systèmes. Davantage de langues développent la dimension nasale (22.4%) plutôt que la dimension de longueur (16.7%).

Le nombre de voyelles nasales est toujours inférieur au nombre de voyelles orales d'un système, ce qui confirme le 14ème universel de Crothers. Elles sont souvent développées à partir de timbres vocaliques présents dans les voyelles orales, mais elles ne sont pas pour autant systématiques. Contrairement à l'assertion de Crothers selon laquelle dans un système, les voyelles nasales qui font défaut par rapport aux voyelles orales, correspondent de voyelles d'aperture moyenne qui ne seraient pas nasalisées, il ressort de notre étude que, pour plusieurs langues, c'est une voyelle plus fermée qui fait défaut.

Les voyelles longues peuvent être en nombre supérieur aux brèves (79% des langues concernées). La majorité des systèmes comportant des voyelles longues développent une opposition de durée pour un même timbre vocalique. Par ailleurs, le fait que les voyelles brèves soient plus ouvertes que les longues n'est pas un phénomène prédictible et universel, mais relève plutôt d'une tendance. A ce propos, le 13ème universel de Crothers n'est pas toujours vérifié : s'il est vrai que les langues développent la dimension de longueur pour des timbres vocaliques intermédiaires à ceux des voyelles brèves, d'autres langues possèdent aussi un contraste de durée pour un même timbre vocalique.

Les autres dimensions n'ont pas été représentées de manière assez significative pour permettre une systématisation.

4. CONCLUSION.

Le nouvel angle d'attaque que nous proposons dans ce travail typologique ne remet pas en cause les typologies antérieures, mais plutôt permet une nouvelle lecture peut-être plus détaillée et nourrit la discussion sur les modèles théoriques de prédiction des voyelles dans l'espace acoustique, voir par exemple SCHWARTZ & AL, 1989.

5. ANNEXE.

Voyelles périphériques : /i, ɪ, e, ɘ, ɛ, æ, a₊, a, ɑ, ɒ, ɔ, ɤ, o, ɷ, u/.

Voyelles intérieures : /ʏ, ʏ, ø, œ, ɨ, ɨ̞, ɵ, ɵ̞, ɜ, ɐ, ʉ, ʉ̞, ɵ, ɵ̞, ɯ, ɯ̞, ɤ, ɤ̞, ʌ/.

	Universaux de CROTHERS (1978)	Comparaisons	
1	Tous les langages possèdent les 3 voyelles de base /i a u/		Discutable
2	Les langages à au moins 4 voyelles possèdent soit /ɨ/ soit /ɛ/		Discutable
3	Les langages à au moins 5 voyelles possèdent /ɛ/ Généralement /ɔ/		Discutable
4	Les langages à au moins 6 voyelles possèdent /ɔ/et soit /ɨ/ soit /ɵ/, en général /ɨ/		Discutable
5	Les langages à au moins 7 voyelles possèdent soit /e o/, soit /ɨ ɵ/		Discutable
6	Les langages à au moins 8 voyelles possèdent /ɵ/		Discutable
7	Les langages à au moins 9 voyelles possèdent en général /o/	Confirmé	
8	Le système de base est un système à 5 voyelles	Confirmé	
9	Le nombre de degrés de hauteur dans un système ≥ à celui des degrés de postériorité		Discutable
10	Les langages à au moins 2 voyelles intérieures en possèdent toujours une qui est haute		Discutable
11	Le nombre de voyelles dans une série intérieure ≤ à celui des séries antérieures ou postérieures	Confirmé	
12	Le nombre de degrés de hauteur dans les voyelles antérieures ≥ à celui des voyelles postérieures		Discutable
13	Dans les syst. à voyelles courtes, les hautes et basses sont plus centrales que dans les syst. à voyelles longues		Discutable
14	Le nombre de voyelles nasales ≤ à celui des voyelles orales	Confirmé	
15	et s'il est strictement inférieur, c'est la nasalisation d'une voyelle moyenne qui fait défaut, relativement par rapport à l'opposition haut / bas		Discutable

Figure 1. **Exemples de systèmes fréquents tirés de notre régularisation en type des systèmes vocaliques.**
Le nombre d'occurrences dans les langues figure à la gauche immédiate du système. Le code d'identification se compose du nombre total de voyelles n, et de la fréquence d'apparition du système (1 = le plus fréquent pour le nombre de voyelles donné, 2 = un peu moins fréquent... x = le moins fréquent et correspond au nombre de types pour un n considéré).

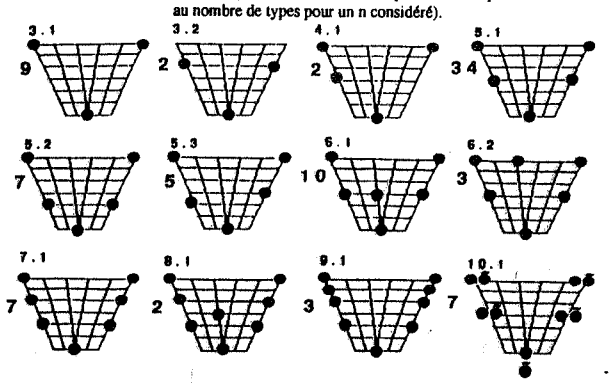


Figure 2. Symboles vocaliques et leur description .

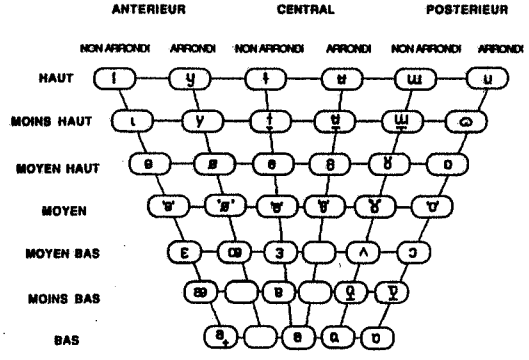


Figure 3. Répartition des systèmes selon le nombre de voyelles.

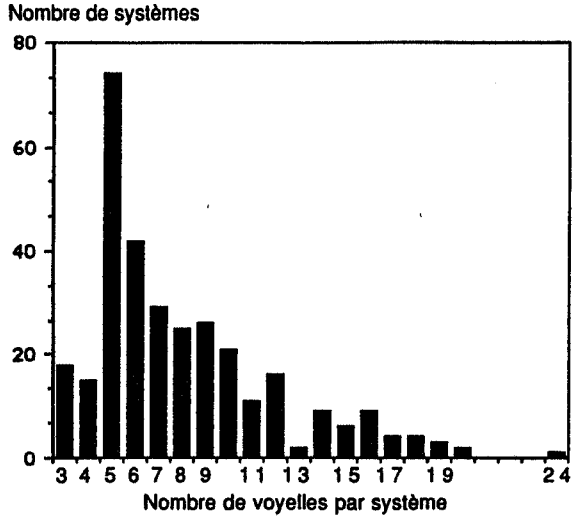


Figure 4.a. Occurrences des voyelles orales, en pourcentage sur le total des 317 systèmes (les valeurs inférieures à 1% n'ont pas été retenues).

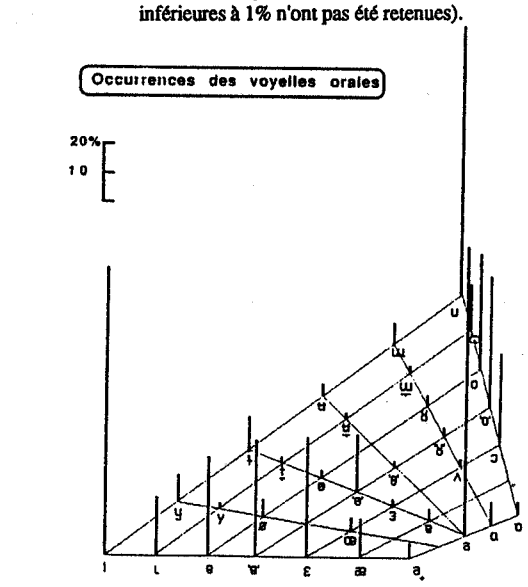


Figure 4.b.

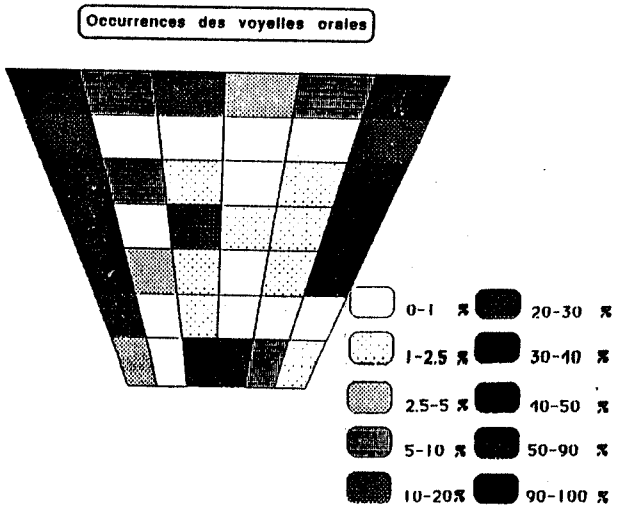


Figure 5. Histogramme des fréquences des séries de voyelles antérieures arrondies.

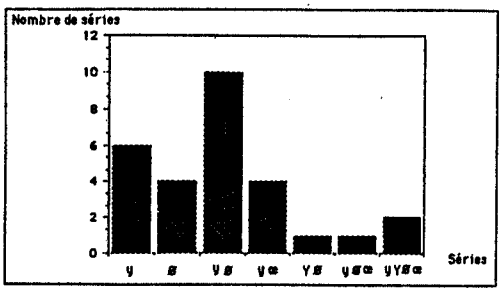


Figure 6. Histogramme des fréquences des séries de voyelles postérieures non arrondies.

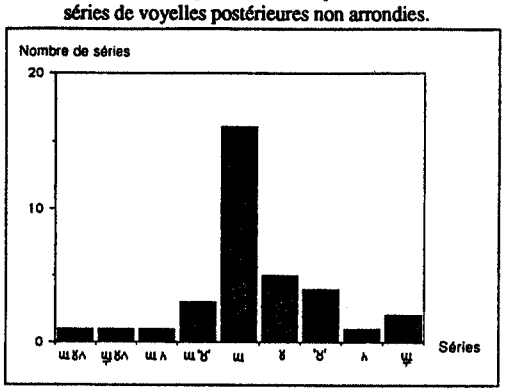


Figure 7.a. Occurrences des voyelles nasales en pourcentage (les valeurs inférieures à 1% ne sont pas reportées).

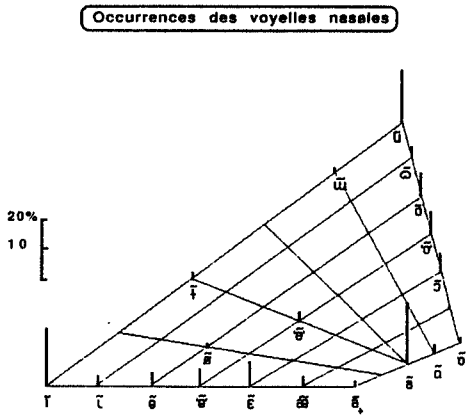


Figure 7.b.

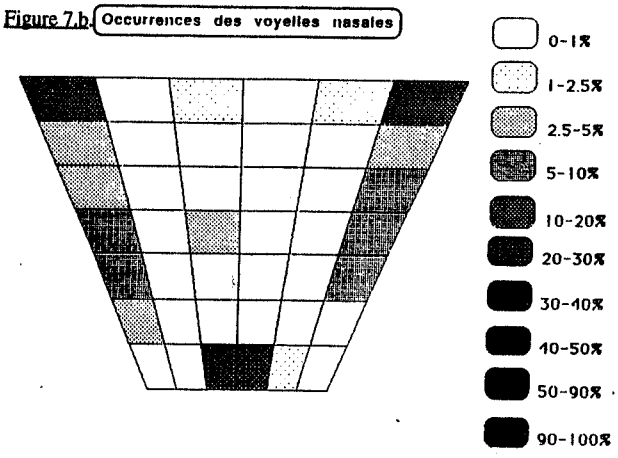


Figure 8. Nombre de timbres vocaliques N, par rapport au nombre de voyelles n contenues dans chaque système.

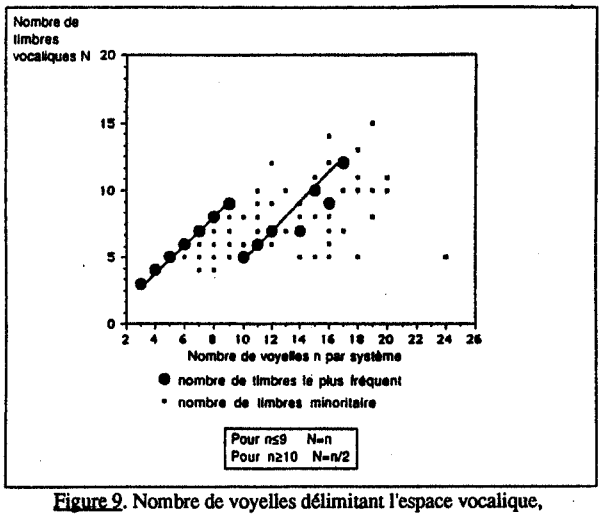


Figure 9. Nombre de voyelles délimitant l'espace vocalique, par rapport au nombre total n des voyelles du système. On observe le même décrochement que celui de la figure 9.

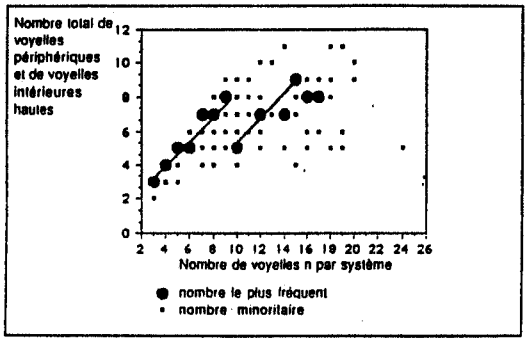


Figure 10. a et b.

Figure 10. a. Composition des séries antérieures arrondies.

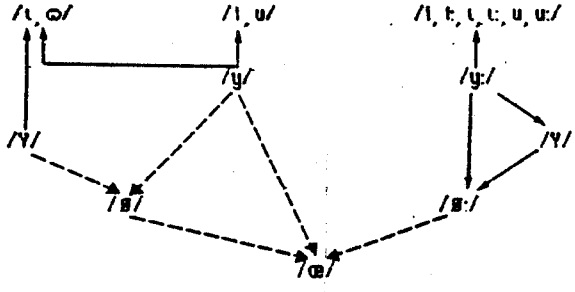
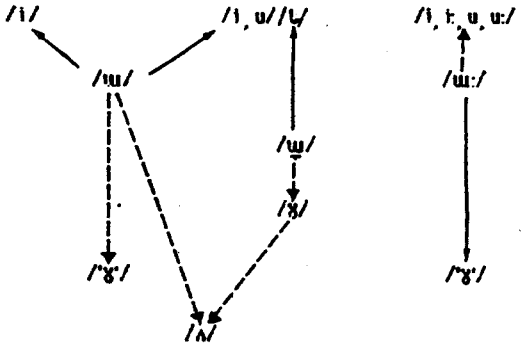


Figure 10. b. Composition des séries postérieures non arrondies.



REFERENCES.

CROTHERS J. (1978). Typology and Universals of Vowel Systems in Phonology. Stanford University Press, Stanford, California, Vol 2, 95-152.

FERRARI DISNER S. (1983). Vowel Quality : the Relation between Universals and Language Specific Factors. 1st ed., University of California, Los Angeles.

HAGEGE C. (1982). La structure des langues. 2e édition, "Que sais-je ? ", P.U.F., Paris, 13-25.

HAGEGE C. (1986). L'Homme de parole. Contribution linguistique aux Sciences Humaines. 2e édition, Fayard, Paris, 51-74.

LASS R. (1984). Vowel System Universals and Typology : Prologue to Theory. Phonology Years Book, Vol 1, Cambridge, 75-111.

LASS R. (1984). Phonology. An Introduction to Basic Concepts. Cambridge University Press, Cambridge, 75-147.

LILJENCRANTS J. LINDBLOM B. (1972). Numerical Simulation of Vowel Quality Systems : the Role of Perceptual Contrast. Language 48, 839-862.

LINDBLOM B. (1986). Phonetic Universals in Vowel Systems. Experimental Phonology J.J. OHALA, Academic Press, New York, 13-44.

LINDBLOM B. (1988). The Elusive Phoneme. Phonetic Experimental Research at the Institute of Linguistics University of Stockholm, Stockholm, Vol 8, 1-19.

MADDIESON I. (1986). Patterns of Sounds. 2nd ed., Cambridge University Press, Cambridge.

MALHERBE M. (1983). Les langages de l'humanité. Une encyclopédie des 3000 langues parlées dans le monde. Seghers, Paris.

PETITOT COCORDA J. (1985). Les catastrophes de la parole. Collection Recherches Interdisciplinaires, Maloine, Paris, 284-293.

SCHWARTZ J.L. (1987). A propos des notions de forme et de stabilité dans la perception des voyelles. Bulletin du Laboratoire de la Communication Parlée, Vol 1A, 159-190.

SCHWARTZ J.L. BOE L.J. PERRIER P. GUERIN B. ESCUDIER P. (1989). Perceptual Contrast and Stability in Vowel Systems : A 3-D Simulation Study. Eurospeech, vol 1, 63-66.

VALLEE N. (1989). Typologie des systèmes vocaliques. T.E.R. de Sciences du Langage, Université Stendhal, Grenoble.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

PLACE DE L'ACCENT SOUS-JACENT, ORGANISATION SYLLABIQUE, ET DISTRIBUTION DES TIMBRES VOCALIQUES EN FRANÇAIS DE MARSEILLE

D.AUTESSERRE & J.Ph.WATBLED

**Laboratoire Parole et Langage URA 261 CNRS
& UFR LAG-LEA, Université de Provence**

Résumé

L'objet de la présente étude est la distribution des timbres vocaliques d'aperture intermédiaire en syllabes non accentuées dans la variété de français parlée à Marseille. Le cadre théorique adopté est un modèle de phonologie non-linéaire et multi-dimensionnelle, dans lequel il s'agit de rendre compte des relations de dépendance qu'entretiennent les différentes unités segmentales et suprasegmentales.

Les éléments entrant en jeu sont essentiellement:
— l'accent, plus ou moins éloigné de la voyelle constituant l'objet de notre étude
— la nature de la syllabe
Plus généralement, la structure suprasegmentale joue un rôle crucial.

I. La situation en syllabes accentuées

Avant d'examiner les données en syllabes non accentuées, il est logique de rappeler ce qui se produit sous l'accent. Dans ce contexte, les timbres mi-fermés et mi-ouverts sont de toute évidence en distribution complémentaire:
(1) *thé* [te] *sept* [sɛt]; *pot* [po] *port* [pɔʁ]; *peu* [pø] *peur* [pœʁ]
On trouve les variantes fermées en syllabe ouverte, et les variantes ouvertes en syllabe fermée; ce principe ne souffre aucune exception en français de Marseille, lorsque l'accent frappe la syllabe finale.

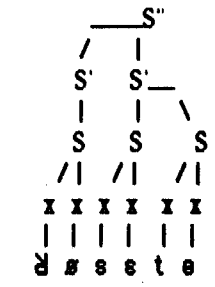
On sait que dans la variété de français qui nous concerne ici l'accent peut frapper l'avant-dernière syllabe: c'est le cas quand le noyau de la syllabe finale est un schwa. On constate alors que ce sont les variantes ouvertes des voyelles d'aperture moyenne qui apparaissent sous l'accent:
(2) *cette* ['sɛtə]; *pore* ['pɔʁə]; *beurre* ['bœʁə]
Sous l'accent, on se retrouve donc avec les variantes fermées en syllabe finale ouverte, et avec les variantes ouvertes dans les deux situations suivantes:
(i) lorsque la voyelle est suivie d'une ou plusieurs consonnes finales
(ii) lorsqu'elle est suivie d'une ou plusieurs consonnes + schwa final
Dans le deuxième cas la voyelle qui nous concerne se trouve pourtant en syllabe ouverte, contrairement à ce qui se passe dans le premier cas. Il est clair qu'une description qui ne s'accompagnerait pas d'une tentative d'explication ne serait pas adéquate, et c'est pourquoi nous proposons le principe explicatif suivant: la voyelle d'aperture moyenne est réalisée ouverte quand elle 'gouverne' une unité à sa droite, soit dans la syllabe, soit dans le pied. Les exemples cités en (2) sont des mots dissyllabiques, mais qui sont pourtant constitués d'un seul pied, dans la mesure où le schwa final est non accentué. Nous postulons les représentations suivantes pour *sept* et *cette*:

(3) *sept* *cette*

Dans ces représentations, chaque x constitue un point phonémique; le symbole S représente la syllabe, et S' une unité supérieure à la syllabe, généralement appelée 'pied' par les partisans de la phonologie non-linéaire. A chaque "étage", l'unité dominante est reliée au symbole supérieur par une barre verticale, et les unités dépendantes le sont par une barre oblique. Ainsi on dira que dans *cette* la première syllabe gouverne la seconde, ou encore que la seconde dépend de la première.

Dans le premier exemple ci-dessus (*sept*) la voyelle gouverne la consonne finale; dans le deuxième cas (*cette*) la voyelle accentuée gouverne le schwa final. Une unité en gouverne une autre, dans le même constituant, si elle est supérieure dans la hiérarchie prosodique, ou si elle est supérieure sur l'échelle de sonorité (un noyau syllabique gouvernera donc les 'marges'). Ce principe de gouvernement / dépendance rend compte de façon unifiée du timbre ouvert de la voyelle accentuée dans les deux mots, que la syllabe accentuée soit fermée ou bien ouverte. Quant à la règle de placement de l'accent, elle est très simple à formuler: l'accent frappe le dernier pied du mot; dans le cas d'un pied dissyllabique, c'est automatiquement la syllabe forte du pied qui sera accentuée:

(4) *recette*



En (4), le niveau S" représente le groupe accentuel, le niveau S' le pied, et le niveau S la syllabe. En descendant verticalement à partir de S", on trouve sur le parcours l'unité dominante à chaque étage, jusqu'à la voyelle de la syllabe accentuée (ici: [ɛ]). On peut, dans ce cadre théorique, formuler de façon simple les règles de placement de l'accent, ainsi que les règles de distribution des timbres vocaliques.

II. Syllabes non accentuées

La question qui se pose en syllabe non accentuée est la suivante: la distribution des timbres obéit-elle aux mêmes principes qu'en syllabe accentuée? Si les timbres fermés et ouverts des voyelles d'aperture moyenne sont toujours en distribution complémentaire, sont-ils régis par les mêmes règles? On s'attend *a priori* à une réponse positive. C'est effectivement le cas lorsque le mot comporte un pied dissyllabique non-final:

(5) *événement, céleri, prestement, omelette*

Dans ces exemples les pieds dissyllabiques sont: - *véne* - - *cèle* - - *preste* - - *ome* - et la voyelle forte du pied est ouverte: [ɛ]. Quand la voyelle d'aperture moyenne se trouve dans une syllabe ouverte, et que cette syllabe est le constituant unique d'un pied monosyllabique, la réalisation de la voyelle est fermée (comme sous l'accent):

(6) *soleil* [sɔlɛj] *feutré* [føtʁe]

On note que pour la norme du français standard, la première syllabe de *soleil* devrait comporter un [ɔ] ouvert et non un [o] fermé comme en français méridional. On note également en français méridional des alternances telles que *neutre, neutraliser*: [nœtrə] [nøtralize] (cf. aussi *beurre* avec 'eu' ouvert et *beurrier* avec 'eu' fermé). De la même façon le français de Provence ne connaît pas d'oppositions comme *ossement, haussement* (avec 'o' ouvert dans *ossement* et 'o' fermé dans *haussement* en français standard): dans les deux cas on a un 'o' ouvert dans la variété de français régional qui nous concerne. Il est également intéressant de comparer le comportement d'un mot tel que *heureusement* en français standard et en français de Provence. En français standard, la norme fournit la réalisation [œʁøzmɑ̃], tandis qu'en Provence la prononciation est régulièrement [øʁœzemɑ̃n]. Il semble bien que des facteurs morphologiques soient ici à l'oeuvre en français non-méridional: la première voyelle est ouverte par analogie avec *heur, bonheur* ; ces facteurs morphologiques ne jouent jamais en français de Provence; en outre la deuxième voyelle est fermée parce que le français standard admet une opposition entre 'eu' fermé et 'eu' ouvert en syllabe fermée (cf. *jeûne, jeune*), opposition que refuse le français méridional (dans cette variété de langue, on a toujours la variante mi-ouverte en syllabe fermée ou quand la syllabe suivante comporte un schwa suivi d'une frontière morphologique): il est curieux de constater par conséquent l'inversion des timbres dans les deux variantes, mais cette inversion est tout à fait régulière dans la mesure où les règles propres à chaque parler sont totalement respectées.

En revanche un problème se pose avec des mots du type suivant:

(7) *esprit, escroc, Eustache, ostrogoth, omnibus, optique, ophtalmologiste, obnubilé, ostréiculture* etc...

Dans ces exemples, la voyelle qui nous intéresse (à l'initiale) est suivie d'un groupe consonantique, et elle se trouve dans un pied monosyllabique (puisque la syllabe suivante ne comporte pas de schwa). Contrairement à ce qui se passe en syllabe finale de mot, on ne peut ici prédire de façon certaine la place de la coupe syllabique. En outre on constate une intéressante variation de timbre, qui va de [e] à [ɛ], de [o] à [ɔ], et de [ø] à [œ], alors que l'on s'attendrait à un timbre ouvert dans tous les cas, étant donné que les locuteurs syllabent ces mots de la façon suivante, dans une prononciation lente (en accentuant chaque syllabe):

(8) *es - prit es - croc Eus - tache os - tro - goth*

Dans cette prononciation syllabée, la voyelle est toujours réalisée ouverte, puisque — précisément — chaque syllabe est accentuée: le locuteur applique en effet les règles citées plus haut pour les syllabes accentuées (chaque syllabe fonctionne comme un mot si elle est détachée des autres). Si la variante syllabée ne nous apprend donc rien sur le timbre des voyelles en syllabes non accentuées, elle nous révèle pourtant la syllabation sous-jacente. Il nous incombe donc d'expliquer pourquoi, en discours suivi, le timbre est très souvent intermédiaire (plus fermé que [ɛ ɔ œ] sans toujours aller jusqu'à [e o ø]), surtout si la syllabe reste fermée.

III. Accent sous-jacent, désaccentuation en discours, et timbres vocaliques

Comparons les deux exemples suivants:

(9)(i) *un ostrogoth* (ii) *un os trop beau*

On constate que la voyelle *o* se ferme parfois, mais seulement en (9)(i), jamais en (9)(ii), et ceci même quand le mot *os* est non accentué en discours. Ce fait nous paraît crucial. En effet, cette différence est liée non pas à l'accentuation en discours, mais à l'accent lexical sous-

jacent. En discours, les deux suites peuvent avoir la même structure suprasegmentale, avec un seul accent sur la syllabe finale. La fermeture de la voyelle est donc conditionnée par la structure suprasegmentale sous-jacente, avant l'enchaînement des mots en discours. La structure initiale des deux suites est:

(10)(i) *'un ostro'goth* (ii) *'un 'os 'trop 'gros*

La syllabe initiale de *ostrogoth* n'étant pas accentuée à ce stade, la voyelle peut se fermer, tandis que celle de *os*, ne peut se fermer, puisque la syllabe unique du mot est accentuée. C'est seulement ensuite, après les règles d'ajustement de la hauteur vocalique, que les règles de désaccentuation s'appliquent:

(11)(i) *un ostro'goth* (ii) *un os trop 'gros*

Cette différence est doublement révélatrice: c'est l'accent lexical sous-jacent qui conditionne l'ajustement; il existe donc un accent lexical sous-jacent en français. En effet, si l'unité lexicale n'était pas accentuée, et s'il existait seulement un accent de groupe dans le discours, on ne pourrait rendre compte de la différence entre les deux suites ci-dessus. Ces principes s'appliquent évidemment à tous les mots dont la structure est analogue à celle de *ostrogoth*: *ostréiculture, omnibus, opticien, auxiliaire, Australie* etc...

Plus généralement, les règles de réorganisation en discours n'affectent pas le timbre des voyelles d'aperture moyenne. Si ce timbre est conditionnée par l'accent sous-jacent et non par l'accent en discours, de même il est conditionné par la syllabation sous-jacente et non par la syllabation en discours: dans *un os à moëlle*, la voyelle de *os* reste ouverte, même si l'enchaînement produit la réorganisation syllabique suivante:

(12) [ɔs][ə] ----› [ɔ][sə]

En ce qui concerne la nature exacte du conditionnement, deux facteurs sont ici en conflit: d'après la syllabation sous-jacente de *os - tro - goth*, la voyelle initiale devrait être ouverte; cependant, son éloignement de l'accent lexical sous-jacent cause une fermeture. Nous sommes donc amenés à penser que le facteur accentuel prévaut dans le cas de la fermeture vocalique, et qu'il l'emporte sur le facteur syllabique.

Il faut ajouter que ce qui vient d'être dit ne s'applique pas aux cas suivants:

(13) *perlé pendant boldoflorine meurtri* etc...

Dans ce type de mots, la voyelle d'aperture moyenne reste ouverte: [ɛ ɔ œ]. Cette fois, c'est le facteur syllabique qui prévaut, et pour la raison suivante: la voyelle qui nous concerne reste ouverte chaque fois qu'elle se trouve dans une syllabe fermée par une liquide, en d'autres termes par une consonne 'vocalique'.

IV. Généralisation

On a vu ci-dessus que deux facteurs sont en conflit. Il est nécessaire, du point de vue théorique, de revenir sur cette question. On sait qu'une voyelle gouverne la consonne qui la suit dans la même syllabe; on sait aussi qu'une syllabe non accentuée est gouvernée par la syllabe accentuée. Les principes suivants s'appliquent en français:

(14) Si une voyelle gouverne une unité à sa droite, elle tend à s'ouvrir
(15) Si une voyelle est gouvernée, elle tend à se fermer

Dans un mot tel que *ostrogoth* les deux principes sont à l'oeuvre: le *o* initial gouverne la consonne qui le suit, mais il est lui-même gouverné par la voyelle accentuée. Les deux principes sont donc en conflit, et (15) l'emporte sur (14) la plupart du temps, bien que la fermeture n'aille pas jusqu'à l'allophone [o].

V. Conclusion

Nous avons traité le problème des timbres des voyelles d'aperture intermédiaire en syllabes non accentuées dans le cadre d'une phonologie multi-linéaire, en tenant compte des relations de gouvernement / dépendance, et en cherchant les principes sous-jacents. Ce modèle, même s'il demeure génératif au sens large du terme, ne saurait être mécaniste: plus que jamais, une conception dynamique de la phonologie s'impose.

On note que le conflit entre les principes sous-jacents ne se produit que dans le contexte de la syllabe non accentuée, par définition, puisque l'accent ne peut causer la fermeture de la voyelle. Il est important de préciser qu'il s'agit de l'accent sous-jacent, dont nous espérons avoir apporté la preuve de l'existence en français.

REFERENCES BIBLIOGRAPHIQUES

Clements, G. & S.J. Keyser (1983):
CV Phonology : A Generative Theory of the Syllable
Linguistic Inquiry Monograph 1, MIT Press
Cambridge, Massachusetts

Durand, J. ed. (1986):
Dependency and Non-Linear Phonology
Croom Helm, Londres & Sydney

Giegerich, H.J. (1985):
Metrical Phonology and Phonological Structure : German and English
Cambridge University Press

Hogg, R. & C.B. McCully (1987):
Metrical Phonology : A Coursebook
Cambridge University Press

Hulst van der, H. & N. Smith eds (1982):
The Structure of Phonological Representations Vol. I & II
Foris Publications, Dordrecht

Nespor, M. & I. Vogel (1986):
Prosodic Phonology
Foris Publications, Dordrecht

Verluyten, P.S., ed. (1988):
La Phonologie du Schwa Français
Collection Linguisticae Investigationes
Suppl. vol.16
John Benjamins, Amsterdam

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Pour une prise en considération de la variation sociolectale dans la parole de synthèse

Jean Dolbec et Claude Paradis

Université du Québec à Chicoutimi et Université Laval

Le développement de systèmes de synthèse plus performants exige nécessairement la prise en compte de la variation linguistique sous toutes ses formes. Si les systèmes existants intègrent déjà depuis longtemps la variation conditionnée par l'environnement phonétique, la variation liée aux caractéristiques socio-géographiques des locuteurs (origine, âge, sexe, éducation, classe sociale, etc.) est par contre encore négligée même si l'on sait par les travaux de la sociolinguistique variationniste qu'elle est hautement systématique et qu'elle conditionne directement les attitudes des locuteurs, et ce faisant l'efficacité de la communication. Le système PHONO est un premier essai de modélisation de la variation sociolectale en français du Québec permettant d'obtenir, à partir d'une représentation phonologique abstraite d'un mot, la réalisation phonétique la plus probable compte tenu d'un profil de locuteur choisi.

1. INTRODUCTION

L'objectif de la recherche en synthèse ou en reconnaissance de la parole est ultimement la mise au point de systèmes dont la performance s'approcherait maximalement de celle du sujet humain en termes d'intelligibilité, de fiabilité, de flexibilité ou de naturel. Outre la nécessité d'intégrer des informations relevant d'autres dimensions du langage (syntaxe, sémantique, pragmatique), un des problèmes majeurs rencontrés sur cette voie est celui de la prise en compte de la variation sous toutes ses formes (Fant 1989).

On sait en effet depuis longtemps que ce que le sens commun identifie comme un *son* et que les phonologues ont décrit systématiquement sous le nom de *phonème* est en fait susceptible de réalisations articulatoires très différentes, cette différence se reflétant dans la composition physique des unités sonores (O'Shaughnessy 1987:56). Les facteurs qui conditionnent cette variation sont de divers ordres, mais peuvent généralement être reliés à trois pôles principaux (Lee 1989): le fonctionnement du système articulatoire, les contraintes du système linguistique et les caractéristiques du locuteur. Si la plupart des systèmes existants intègrent déjà à des degrés divers les deux premiers types, la variation liée aux caractéristiques du locuteur a été généralement négligée. Nous nous proposons ici d'examiner avec un regard de linguistes le traitement réservé aux divers types de variation, avant de centrer notre attention sur la variation sociolectale dont on évaluera l'intérêt de la prise en compte par un système de synthèse avant de considérer les moyens pour y arriver. Le système PHONO est un premier essai pour modéliser une partie de la variation sociolectale en franco-québécois en vue de certaines applications didactiques. Nous croyons toutefois que cette approche pourrait ultérieurement être étendue à d'autres applications pratiques en synthèse et en reconnaissance (cf. le système GEPH -GEnérateur PHonologique- (Pérennou et al. 1989) pour le français de l'Hexagone)

2. LA VARIATION DANS LA PAROLE DE SYNTHÈSE

Le problème de l'intégration de la variation se pose différemment, et a conséquemment reçu des solutions différentes, selon la nature des applications (reconnaissance vs synthèse, mots isolés vs parole continue, etc.) et selon la source de cette variation (physiologique, linguistique, idiolectale). On se limitera ici à donner un aperçu sommaire de la situation dans une perspective de synthèse.

2.1 La variation d'origine articulatoire

La variation qui découle des contraintes physiologiques sur le fonctionnement du système phonatoire est probablement la plus importante et la mieux connue. Alors que les phonèmes comme réalités abstraites du système linguistique sont des unités discrètes, leur réalisation physique dans la parole est un phénomène continu, ce qui implique inévitablement des effets de coarticulation entre sons contigus, effets qui dans certains cas dépassent même la contiguïté immédiate (O'Shaughnessy et al. 1988). Toute tentative de fonctionner par simple concaténation directe des différents sons conçus comme unités fixes non seulement est perçue comme absolument non naturelle, mais encore affecte

de façon importante l'intelligibilité dans la mesure où les mécanismes de perception sont conditionnés à l'existence de tels effets de coarticulation. Les différents systèmes se doivent d'intégrer cette variation, soit en travaillant avec des unités discrètes de taille intermédiaire, plus larges que les sons isolés (demi-syllabes, diphones, etc.), ce qui a pour effet de réduire -mais non pas d'éliminer- la nécessité de procédures de lissage et d'interpolation assurant la continuité spectrale des transitions, soit en assumant directement cette continuité entre unités contiguës et en essayant de la modéliser par un jeu complexe de règles qui déterminent les paramètres acoustiques à intervalles très courts. La qualité des résultats obtenus dans un cas comme dans l'autre (Logan et al. 1989) permet de conclure que le problème de la variation d'origine purement articulatoire a déjà trouvé des solutions, sinon définitives, du moins encourageantes.

2.2 La variation conditionnée par le système linguistique

Une deuxième forme de variation dont doit tenir compte un système de synthèse met en cause les interactions qui existent entre l'output sonore, c'est-à-dire la réalisation phonétique de l'énoncé, et d'autres composantes du système de la langue en question (graphie, morphologie, syntaxe, sémantique, etc.). Parmi ces facteurs conditionnants, on peut signaler:

- la non-univocité de la relation graphème-phonème et la nécessité pour l'algorithme de conversion de faire appel à des indications morphologiques ou syntaxiques -pas toujours disponibles- pour lever certaines ambiguïtés;
- la structure syllabique et les frontières de mots qui, en conjonction avec la force relative des articulations en cause, ont une influence sur la portée et l'orientation des phénomènes de coarticulation;
- la classe grammaticale du mot et sa structure morpho-phonologique en tant qu'ils conditionnent la possibilité de porter l'accent, de même que la place de celui-ci dans les langues à accent mobile;
- la structure syntaxique de l'énoncé, et dans une moindre mesure son contenu sémantique de même que sa portée pragmatique, qui déterminent la réalisation des éléments prosodiques (intonation, accentuation, rythme et débit, etc.).

Il s'agit ici encore d'un type de variation que la parole de synthèse a dû nécessairement prendre en compte sous peine d'atteintes sérieuses au naturel et à l'intelligibilité; il suffit de rappeler que les facteurs prosodiques assument une fonction démarcative essentielle, sans compter les cas où ils jouent un rôle distinctif (valeur pragmatique de l'intonation, place de l'accent en anglais, etc.). Deux types de problèmes se posent pour l'intégration de cette variation d'origine linguistique. Le premier concerne la possibilité d'obtenir l'information pertinente sur les facteurs de variation. Les progrès à cet égard sont tributaires du développement d'outils performants dans les disciplines linguistiques concernées (dictionnaires, analyseurs syntaxiques, etc.); entre temps, beaucoup de systèmes s'en remettent à l'interprétation d'autres indices (par exemple les marques de ponctuation), ce qui dans la majorité des cas constitue un compromis acceptable, une analyse plus poussée étant requise dans d'autres. Le second problème a trait aux moyens de réaliser acoustiquement la variation requise. Celle-ci mettant en cause pour une large part des modifications de fréquence du fondamental, de durée ou d'intensité susceptibles de s'étendre sur plus d'une unité, il devient nécessaire de faire appel à des systèmes de synthèse assez sophistiqués qui seuls possèdent la flexibilité requise pour supporter la variation continue de ces paramètres.

2.3 La variation qui origine du locuteur

Un troisième type de variation dans la parole naturelle est en relation avec la singularité du locuteur, voire la singularité de chaque événement de parole. Entrent ici en ligne de compte des facteurs comme 1) les propriétés physiologiques du conduit vocal, elles-mêmes conditionnées par le sexe, l'âge ou la taille, 2) l'état physique ou psychologique du locuteur, 3) ses habitudes articulatoires, 4) son origine géographique et son profil social, 5) les conditions de la prise de parole et le type de discours, et d'autres encore.

Contrairement à la variation d'origine articulatoire et à la variation d'origine linguistique, la variation liée au locuteur, malgré quelques appels en ce sens (Bladon et al. 1987; Carlson et Nord 1989), a été pratiquement ignorée dans la parole de synthèse, à de rares exceptions près (cf. les différentes voix du système Klattalk- DECtalk (Conroy 1986) et l'accent partiellement québécois du système LOQUAX (Bernardi 1985)). Les raisons de ce peu d'intérêt ont généralement à voir avec le caractère prétendument asystématique de la variation de même que son incidence moins grande sur l'intelligibilité, et même le naturel, de la voix synthétique (Robert, Choinière et Descout 1989). La pertinence de ces raisons en ce qui a trait à un sous-ensemble de la variation liée au locuteur, celle qui découle de ses caractéristiques socio-géographiques, aussi appelée variation sociolectale, sera examinée dans la section qui suit.

3. LA VARIATION SOCIOLECTALE DANS LA PAROLE DE SYNTHÈSE

Il n'est pas nécessaire d'être linguiste pour observer que, sur le plan de la production, les membres d'une communauté n'ont pas tous les mêmes comportements linguistiques. Les constats relatifs à cette variation remontent d'ailleurs pratiquement à la nuit des temps (cf. Le Livre des Juges 12: 4-6). Or, on a longtemps considéré que cette variation linguistique attestée dans une communauté donnée était asystématique. C'est un des grands mérites des travaux menés dans la perspective variationniste à partir des années soixante d'avoir

démontré que l'hétérogénéité linguistique rencontrée dans toute communauté n'était pas dysfonctionnelle, mais, bien au contraire, pouvait être structurée et utilisée à des fins d'expression. On sait maintenant que les locuteurs acquièrent non seulement un ensemble de règles phonologiques et grammaticales, mais aussi un certain nombre de règles d'usage qui leur permettent d'utiliser la bonne forme dans le bon contexte. Les sociolinguistes se sont donc employés à démontrer et à cerner la systématicité de la variation et ont, pour ce faire, développé des techniques quantitatives de cueillette et d'analyse des données qui permettent d'une part d'observer cette variation et d'autre part de préciser les rapports qu'elle entretient avec la structure socio-discursive (Sankoff 1974). C'est ainsi qu'on a accumulé au cours des dernières années, en Europe comme en Amérique, une masse de données quantitatives sur la variation linguistique, notamment sur la variation phonétique. Pour le français québécois, les données et les analyses disponibles sont particulièrement nombreuses (cf. Lamarche et Daoust 1988).

Il est un autre aspect du phénomène de la variation qui présente également un caractère hautement systématique et qui aurait avantage à être pris en considération par les spécialistes de la synthèse de la parole. Il ressort en effet des études faites par Lambert et son équipe (Lambert et al. 1960) sur le français du Québec et par Labov (1976) sur l'anglais américain que les membres d'une communauté linguistique ont tendance à évaluer de façon assez homogène les productions linguistiques de tel ou tel type de locuteur. Mais Labov va plus loin en montrant que l'évaluation d'un locuteur par des sujets de la même communauté linguistique, même si elle est relativement uniforme, diffère en fonction du point de vue considéré. Ainsi, le locuteur de classe moyenne est mieux évalué que le locuteur de la classe ouvrière sur une échelle d'aptitude professionnelle, et ce par tous les sujets peu importe leur groupe social d'appartenance. Sur une échelle «d'amitié» ou d'«amabilité», le locuteur de classe moyenne est encore mieux jugé que le locuteur de classe ouvrière par les sujets des trois groupes sociaux les plus élevés dans la structure sociale alors que le locuteur de la classe ouvrière est mieux évalué par les sujets des deux classes ouvrières les moins élevées. Par contre, sur une échelle de «coriacité», le locuteur de la classe moyenne perd aux dépens du locuteur de la classe ouvrière, et ce pour tous les groupes sociaux d'évaluateurs. En résumé, un locuteur possédant les traits phonétiques de la classe moyenne est jugé par tous les membres d'une communauté comme étant plus capable sur le plan professionnel; toutefois, seulement trois des cinq groupes sociaux de sujets-évaluateurs l'estiment plus "aimable" que le locuteur de la classe ouvrière alors qu'aucun groupe ne considère le locuteur «bourgeois» comme étant plus apte que le locuteur «ouvrier» à tirer son épingle du jeu dans une situation périlleuse. À partir de ces résultats, Labov postule l'existence de deux normes sociales, dont l'une, à savoir celle qui attribue certaines valeurs positives aux variétés vernaculaires est occultée, c'est-à-dire masquée par la formalité de la situation d'entrevue ou de test, mais néanmoins tout aussi effective (Labov 1976:339).

À notre connaissance, les spécialistes de la synthèse de la parole n'ont que rarement pris directement en considération les résultats des études variationnistes si on excepte le travail de Dave W. Bernardi sur la *Synthèse par ordinateur du français montréalais* (1985). Dans ce mémoire, Bernardi présente une partie des aménagements nécessaires à LOQUAX, le synthétiseur de parole mis au point par D. O'Shaughnessy de l'INRS--Télécommunications, pour que celui-ci s'exprime comme un bon Québécois de Montréal. Une étude récente (Robert, Choinière et Descout 1989) visant à évaluer en contexte québécois trois systèmes de synthèse --dont LOQUAX-- fait par ailleurs ressortir que les différentes variétés (québécoise vs parisienne, populaire vs instruite, soutenue vs familière) sont assez nettement perçues par les locuteurs, mais auraient, selon les auteurs, peu d'influence sur l'évaluation globale du système; cette dernière conclusion demanderait sans doute à être réévaluée à la lumière d'une étude qui contrôlerait mieux des facteurs comme l'origine sociale des récepteurs, la situation de communication ou la fonction du message et tiendrait compte de l'existence d'une norme occulte évoquée plus haut.

En fait, si on se fie aux résultats des études en linguistique variationniste ou en théorie de la communication, comme aux usages en publicité ou en marketing, il appert que, tôt ou tard, les synthétiseurs ne pourront plus tous parler d'une seule et même voix. D'une part, les voix synthétiques devront s'adapter aux réalités linguistiques des divers pays de la francophonie. D'autre part, pour chacune des communautés linguistiques, certaines applications (dans la vente ou la promotion de produits, par exemple) exigeront un type de voix bien adaptée aux divers paramètres de l'interaction en cause. C'est ainsi qu'un synthétiseur dont la tâche consisterait à donner de l'information à des élèves de niveau secondaire (lycée) sur les maladies transmises sexuellement ou sur les drogues aurait avantage, du moins dans le contexte québécois, à s'exprimer dans une forme relativement près de celle des usagers, c'est-à-dire dans une variété familière. De même, on peut croire qu'un synthétiseur chargé de donner des renseignements sur les matchs d'une équipe de hockey locale (calendrier des rencontres, disponibilité des billets, commentaires sportifs, messages aux amateurs, etc.) ne pourrait certainement pas utiliser des variantes phonétiques trop formelles.

Jusqu'à maintenant, les chercheurs dans le domaine de la synthèse ont négligé cet aspect de la communication linguistique, soit parce qu'ils étaient trop absorbés par des problèmes qu'on peut certainement concevoir comme plus fondamentaux, soit parce qu'ils étaient insuffisamment conscients du caractère systématique de la variation sociolectale et de son rôle. Celle-ci n'a cependant pas de raison d'être indéfiniment négligée (Carlson et Nord 1989) maintenant que la qualité des systèmes de synthèse s'est améliorée sensiblement.

D'autant que de nombreuses études ont montré qu'étant donné sa nature probabiliste, ce type de variation pouvait très bien être formalisé, ce qui facilite son intégration dans des modèles de synthèse ou de reconnaissance.

4. LE SYSTÈME PHONO

Sous sa forme actuelle, PHONO est un instrument didactique fonctionnant sur PC qui permet de modéliser et d'illustrer certains phénomènes de variation en français du Québec à des fins de support à l'enseignement dans des cours de phonétique du franco-québécois ou de sociolinguistique. Le système opère de la façon suivante: l'utilisateur fournit la transcription normalisée ou phonologique d'un mot en utilisant les symboles de l'*Alphabet Phonétique International* et définit un profil de locuteur en fonction de certains traits socio-géographiques (origine, sexe, âge, classe sociale, éducation, etc.); le système dérive alors une transcription phonétique précise de la prononciation caractéristique du profil de locuteur choisi, avec une identification des phénomèmes de variation ayant joué. Dans certains cas, le système peut aussi permettre une oralisation de la prononciation obtenue et une comparaison avec la forme standard.

Outre les modules qui assurent l'interface avec l'usager et la gestion des caractères phonétiques, PHONO comprend essentiellement un module de caractérisation du locuteur, un module de syllabation, un ensemble de modules correspondant aux différents phénomènes de variation et finalement un module permettant l'oralisation.

4.1 Module de caractérisation du locuteur

Ce module permet de spécifier les caractéristiques d'un type de locuteur potentiel en fonction des différents facteurs de variation généralement reconnus dans les études sociolinguistiques soit le sexe, l'âge, l'origine géographique, le niveau d'éducation, l'activité professionnelle. Parmi ces facteurs, l'origine géographique, l'âge et la classe sociale sont ceux qui sont responsables des différences les plus évidentes et les mieux documentées alors que les variations liées par exemple au sexe du locuteur, bien que réelles, sont le plus souvent trop subtiles pour être d'un intérêt primordial dans un cadre comme celui-ci. D'autres facteurs de variation liés par exemple au caractère plus ou moins formel de la situation de la communication ont une pertinence restreinte dans une approche par mots isolés et ont été simplement laissés de côté dans la réalisation de PHONO.

Plutôt que de spécifier un profil de locuteur particulier, l'usager peut aussi choisir d'utiliser PHONO en mode CATÉGORIQUE, auquel cas le programme cherche à appliquer le plus grand nombre de règles possible, ce qui peut conduire à un résultat quelque peu artificiel, mais présente un certain intérêt si l'on veut illustrer le plus grand nombre de phénomènes ou si l'on cherche à caractériser globalement le français du Québec par rapport à d'autres variétés de français.

4.2 Module de syllabation automatique.

Le mot soumis à l'analyse doit d'abord être divisé en syllabes étant donné qu'un nombre important de faits de variation sont conditionnés par la structure syllabique; la conception d'un algorithme de syllabation se trouve facilitée par le fait que PHONO opère pour le moment sur des mots isolés; l'algorithme mis au point est très compact et n'utilise que cinq règles associées à une routine d'extraction et de réintroduction des semi-consonnes et des liquides.

4.3 Modules de modélisation de la variation.

Le noyau central de PHONO est un ensemble de modules visant à rendre compte des principaux phénomènes de variation pouvant affecter chaque phonème ou classe de phonèmes en français du Québec; le traitement de chaque phénomène se fait sous forme de règles dont les conditions d'application sont déterminées conjointement par des facteurs linguistiques (nature du phonème en cause, environnement phonétique, structure syllabique, etc.) et des facteurs extralinguistiques (i.e. le profil socio-géographique du locuteur tel que défini par le module de caractérisation du locuteur). S'il s'agit d'une règle variable, c'est-à-dire d'une règle dont l'application n'est pas contraignante dans toutes les occasions pour tous les locuteurs d'un profil donné, celle-ci s'accompagne également d'un coefficient de probabilité qui est pris en compte pour la prévision de la prononciation résultante. Ce coefficient d'application a été, dans la mesure du possible, établi à partir des résultats des études sociolinguistiques disponibles; à défaut de données quantitatives précises sur certains phénomènes, la probabilité d'application a été déterminée plus ou moins arbitrairement en tenant compte des évaluations qualitatives fournies par d'autres types d'études moins explicites.

PHONO fonctionne en vérifiant séquentiellement et récursivement pour chacun des phonèmes de la transcription phonologique qui constitue l'input quelles sont les règles de variation qui sont susceptibles de s'appliquer compte tenu de l'environnement linguistique, des caractéristiques socio-géographiques du locuteur et du coefficient de probabilité attaché à chaque règle. L'output est une nouvelle chaîne de symboles phonétiques reflétant les changements provoqués par l'application des règles de variation et correspondant à la réalisation phonétique particulière probable pour le profil de locuteur choisi. Dans une perspective didactique, PHONO signale et identifie aussi chacun des phénomènes de variation intervenus.

4.4 Module d'oralisation

PHONO peut aussi, sous certaines conditions, fournir à l'usager un output vocal lui permettant d'entendre successivement et de comparer la forme standard

et la variante particulière correspondant à la prononciation d'un certain type de locuteur. Le mode "avec oralisation" impose cependant certaines contraintes sur l'input qui ne peut plus être libre mais doit être choisi par l'usager dans une liste de mots présentée par le système; chaque mot renvoie à un certain nombre de fichiers qui correspondent chacun à un enregistrement numérisé (fréquence d'échantillonnage de 10 Khz et précision de 10 bits) de la prononciation d'une des variantes. La liste des mots et des variantes disponibles pour oralisation, quoique finie, est aisément extensible ou modifiable en ajoutant simplement de nouveaux fichiers numérisés en fonction des besoins didactiques.

La reconversion analogique des fichiers est actuellement faite à l'aide du circuit Micro Speech Lab développé par le *Centre for Speech Technology Research*, Univ. of Victoria, qui est déjà utilisé à des fins d'enseignement dans les laboratoires des chercheurs. Une version de PHONO qui tirerait avantage des possibilités d'un circuit de traitement de signal (plus spécifiquement le TMS320C25 Development System de *Loughborough Sound Images Ltd* permettant l'utilisation d'une échelle de quantification jusqu'à 16 bits) est en cours de développement en collaboration avec Speech Technology Research Ltd, Victoria, B.C. Cela devrait permettre dans l'immédiat une amélioration sensible la qualité de l'output sonore à partir des fichiers numérisés en plus d'ouvrir la voie, à moyen terme, à des applications plus ambitieuses faisant appel à une véritable parole de synthèse.

5. CONCLUSION

Dans son état actuel, PHONO présente des lacunes évidentes: absence d'un algorithme de conversion texte-phonème, fonctionnement au niveau de mots isolés, output vocal limité à partir de mots numérisés. Si ces limitations ne constituent pas un handicap important pour le type d'utilisations didactiques considérées jusqu'à maintenant, il en va autrement pour le développement d'autres applications pratiques prenant en compte la variation sociolectale dans la parole de synthèse. La voie la plus intéressante à court terme serait sans doute l'intégration de PHONO dans un système existant qui fournirait les composantes manquantes, le module de variation fonctionnant alors comme un filtre sur l'input du composant de synthèse.

RÉFÉRENCES

Bernardi, Dave W. (1985), *La synthèse par ordinateur du français de Montréal*, mémoire de maîtrise en génie électrique, Université McGill.

Bladon, A., Carlson, R., Granstroem, B., Hunnicutt, S. et Karlsson, I. (1987), «A text-to-speech system for British English, and issues of dialect and style», in *European Conference on Speech Technology*, vol. 1, Edinburgh, pp. 55-58.

Carlson, R. et Nord, L. (1989), «Positional variants of Swedish sonorants in an analysis synthesis scheme», in Tubach et Mariani 1989, vol. 1, pp. 458-461.

Conroy, D., Vitale, T. et Klatt, D.H. (1986), *DECtalk DTC03 Text-to-Speech System Owner's Manual*, Nashua (New Hampshire), Digital Equipment Corporation.

Labov, William (1976), *Sociolinguistique*, Paris: Les Éditions de Minuit.

Lamarche, R.M. et D. Daoust (1988), *Bibliographie de travaux québécois*. Volume 2: *Linguistique générale, linguistique computationnelle, terminologie, traduction*, Office de la langue française au Québec.

Lambert W.E, Hodgson, R.C., Gardner, R.C., Fillenbaum, S. (1960), «Evaluational reactions to spoken languages», *Journal of Abnormal and Social Psychology*, 60, pp. 44-51.

Lee, Kai-Fu (1989), «Hidden Markov Models: Past, Present, and Future», in Tubach et Mariani 1989, vol. 1, pp. 148-155.

Logan, J.S, Greene, B.G. et Pisoni, D.B (1989), «Segmental intelligibility of synthetic speech produced by rules», in *J. Acoust. Soc. Am.* 86(2), pp. 566-581.

O'Shaughnessy, D. (1987), *Speech Communication*, Reading, Mass., Addison-Wesley.

O'Shaughnessy, D., Barbeau, L., Bernardi, D. et Archambault, D. (1988), «Diphone Speech Synthesis», in *Speech Communication* 7, pp. 55-65.

Pérennou, G., de Calmès, M., Ferrane, I. et Tihoni, J. (1989), «Automated Phonotypical Transcription through the GEPH Phonology Expert-System», in Tubach et Mariani 1989, vol. 2., pp. 364-367.

Robert, J.M., Choinière, A. et Descout, R. (1989), «Subjective evaluation of the Naturalness and Acceptability of Three Text-to-Speech Systems in French», in Tubach et Mariani 1989, vol. 2, pp. 640-643

Sankoff, Gillian (1974), «A Quantitative Paradigm for the Study of Communicative Competence», in R. Bauman et J. Sherzer (eds), *Explorations in the Ethnography of Speaking*, Cambridge University Press, pp. 18-49.

Tubach, J.P. et Mariani J.J. (eds) (1989), *Eurospeech 89, European Conference on Speech Communication and Technology*, Paris, Septembre 1989, Edinburgh, CEP Consultants Ltd, 2 vol.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Détection d'erreurs phonémiques - I: effets des types d'erreur, de l'âge et de la nature de la cible chez des sujets normaux; II: Performance des sujets aphasiques.

P. Lemieux, G. Ladouceur, L. Sabourin, P. Villiard, S. Valdois, J. Gagnon, J.-L. Nespoulous et Y. Joanette

Laboratoire Théophile-Alajouanine, Centre de recherche Côte-des-Neiges, Montréal

RÉSUMÉ : Cette étude vise l'évaluation de la capacité de détection des erreurs phonémiques (EPm) de sujets normaux et aphasiques. Pour ce faire, 10 sujets jeunes (21-26 ans), 10 sujets âgés (60-69 ans) et 4 sujets aphasiques (âgés entre 56 et 68 ans) ont été soumis à une tâche de détection d'EPm portant sur des mots et des non-mots. Les six types d'EPm sont produites par (1) *addition*, (2) *omission*, ou par substitution. Dans la catégorie des substitutions, il y a celles (3) par *assimilation*, (4) à *DI>2*, (5) à DI=1 par *voisement* et (6) à DI=1 par un *autre* des traits distinctifs retenus (antérieur, coronal, continu). Aucune différence marquée n'est retrouvée entre sujets jeunes et âgés. Pour les mots et les non-mots, les résultats montrent que les substitutions sont mieux détectées à DI>2 qu'à DI=1, tant chez les normaux que chez les aphasiques. Les substitutions sont mieux détectées que les omissions chez les aphasiques tant pour les mots que pour les non-mots alors que les sujets normaux détectent mieux les omissions que les substitutions uniquement dans le cas des mots. Chez les sujets aphasiques, les résultats montrent plus de bonnes détections dans le cas des additions que dans le cas des substitutions pour les mots et inversement pour les non-mots, alors que chez les normaux, aucune différence significative n'est observée. La nature de la cible interagit avec la population et le type d'EPm. Certains résultats s'expliquent par des phénomènes de restauration («phoneme restauration») et par des phénomènes d'attente cognitive («top-down»).

Les processus qui sous-tendent la perception phonémique ne sont pas encore très bien identifiés, malgré le fait qu'ils aient un rôle important à jouer dans la production du langage. La perception des erreurs phonémiques (EPm) chez le sujet normal est nécessaire à leurs corrections et est d'autant plus importante chez le sujet aphasique (cérébrolésé) dont le discours peut comporter un plus ou moins grand nombre de paraphasies phonémiques, i.e. de productions déviantes dans lesquelles des phonèmes sont ajoutés, omis, déplacés ou substitués. La détection des erreurs se fait à partir d'indices acoustiques mais, alternativement et à un niveau plus abstrait, à partir de paramètres phonémiques. Peu d'études ont tenté une comparaison des capacités de détection de divers types d'erreurs basées sur des paramètres phonémiques. De plus, ces études ont surtout été effectuées auprès d'une population de sujets jeunes non cérébrolésés.

De façon générale, la présente étude s'intéresse à la détection d'EPm afin de mieux comprendre le rôle et l'importance des processus de contrôle de la production phonémique. Plus précisément, l'objectif de ce travail est d'évaluer, à l'aide d'une approche psycholinguistique, dans quelle mesure la capacité de détection d'EPm chez des sujets normaux et aphasiques varie en fonction du type d'erreur, de la nature de la cible (mot ou non-mot). Elle vise également à évaluer dans quelle mesure l'âge des sujets normaux influence la capacité de détection. Pour ce faire, cette étude se propose d'analyser la performance des sujets normaux et aphasiques dans une tâche de détection d'EPm présentées à partir de bandes magnétiques.

ADRESSE DE CORRESPONDANCE: Patricia Lemieux, Centre de recherche du Centre hospitalier Côte-des-Neiges, 4565 Chemin de la Reine-Marie, Montréal, Québec, Canada, H3W 1W5
Recherche subventionnée par le F.C.A.R. (8-EQ-3418)

CADRE THÉORIQUE

La capacité de détection d'EPm semble varier selon plusieurs facteurs, tels que le nombre de traits phonologiques, le type de trait phonologique et le type d'EPm présenté.

Un premier facteur qui influence la détection d'EPm est le nombre de traits phonologiques qui différencient les phonèmes-cibles et les phonèmes effectivement produits. Cet écart (calculé en nombre de traits phonologiques) se nomme *distance interphonémique (DI)*. Par exemple, les phonèmes [p] et [b] ont une DI de 1 parce qu'un seul trait les différencie (voisé); les phonèmes [p] et [d] ont une DI égale à 2 parce que deux traits les différencient (voisé et antérieur). Dans le cadre de cette étude, les différentes EPm s'appuient sur une représentation théorique des phonèmes en termes de traits distinctifs. Quatre traits distinctifs ont été pris en compte: antérieur, coronal, continu, voisé. La DI est évaluée à travers le nombre de traits qui diffèrent entre la cible et le stimulus présenté.

Chez des sujets normaux, Cole (1973) rapporte que les EPm de type DI=1 sont plus difficilement détectées que celles de type DI=2 ou DI=4 dans une expérience où les sujets doivent détecter les erreurs dans un texte présenté oralement. Il explique ce résultat par le fait que les sujets ne portent pas attention à toute l'information contenue dans le signal sonore d'un mot. Il conçoit ainsi la perception d'un mot comme l'appariement d'un ensemble de traits identifiés dans le signal acoustique et d'un ensemble de traits spécifiques, emmagasinés en mémoire à long terme. La reconnaissance se fait quand un nombre minimal de ces traits est présent. Si on assume qu'une certaine variation par rapport à une cible donnée est tolérée dans ce processus de reconnaissance, on peut imaginer qu'un stimulus qui ne diffère de la cible que par un seul trait peut alors se retrouver dans les limites d'acceptation de la cible, et être ainsi identifié comme la cible. Également en rapport avec le mécanisme de perception d'un mot, Marslen-Wilson (1987) parle de processus de sélection selon lequel la représentation du mot qui correspond le mieux à ce qui a été entendu est sélectionnée. Ainsi, le système de détection est "tolérant" vis-à-vis de déviations mineures dans l'«input» sensoriel (Marslen-Wilson, 1987).

McInish et Tikofsky (1969) obtiennent des résultats similaires à ceux de Cole (1973). Dans une expérience de discrimination où des sujets normaux doivent juger si les membres d'une paire de stimuli sont identiques ou non, ils constatent que la détection d'EPm est beaucoup plus facile lorsque la différence est de deux traits (DI=2) plutôt que d'un seul trait (DI=1).

Milberg et al. (1988) observent, cette fois chez des sujets aphasiques, plus d'erreurs de décision lexicale dans le cas de substitutions à DI=1 que dans le cas de substitutions à DI=2. Donc les résultats obtenus chez les sujets normaux et aphasiques, permettent d'avancer l'hypothèse selon laquelle les EPm de type DI=1 soient moins bien détectées que celles de type DI>2.

Par ailleurs, la nature du trait phonologique distinguant les phonèmes entre eux représente un autre déterminant qui influence la perception. Ainsi, Miller et Nicely (1955), étudient la confusion perceptive qui peut exister entre différents phonèmes dans une tâche d'identification. Ils comparent les phonèmes prononcés avec les phonèmes perçus par leurs sujets et observent que le lieu d'articulation est le trait le plus difficilement discriminable en situation de bruit et de filtrage alors que les traits de voisement et de nasalité sont les plus facile à percevoir.

Cette observation a également été notée chez certains aphasiques. En effet, Blumstein et al. (1977), utilisant des paires de stimuli dont une des deux consonnes dans chaque paire est altérée, rapportent que certains aphasiques[1] éprouvent plus de difficulté à distinguer deux stimuli lorsque le trait de lieu d'articulation est en cause que lorsqu'il s'agit du trait de voisement. Dans le cadre de cette étude, il est donc attendu que la substitution du voisement soit plus facile à détecter que les autres erreurs de substitution à DI=1 tant chez les sujets normaux que chez les sujets aphasiques.

Enfin, l'effet du type d'EPm (substitution, omission, addition) sur la capacité de détection peut être évalué. Jusqu'à présent aucune étude n'a analysé l'importance du type d'EPm présenté. En effet, la littérature ne traite que de détection d'EPm par «substitution», sans traiter des types «omission» et «addition». Néanmoins, les études s'intéressant à la production d'EPm chez les sujets aphasiques peuvent renseigner indirectement sur l'effet du type d'erreur en détection. En effet, Basso et al. (1977) ont suggéré l'existence d'un parallélisme entre les désordres phonémiques observés en production et les déficiences dans l'identification phonémique relevés en perception.

Des études (Nespoulous et al., 1987; Blumstein, 1973) s'intéressant à la production d'EPm chez les sujets aphasiques ont montré que dans l'aphasie de Broca, de conduction ou de Wernicke, les substitutions s'observent plus fréquemment que les autres types de transformations. Acceptant l'existence d'un certain parallélisme entre expression et perception, on pourrait donc s'attendre à une plus grande difficulté à détecter les EPm impliquant une substitution qu'une addition ou une omission.

Finalement, deux autres variables doivent être prise en compte: la nature des stimuli (mot ou non-mot) et l'âge des sujets.

Le recours aux mots et non-mots est nécessaire pour évaluer l'importance d'une représentation lexico-sémantique pré-existante lors de la détection d'EPm. Un non-mot est un segment inventé qui ne fait pas partie de la langue du sujet, mais qui en respecte les contraintes phonologiques. Les processus qui permettent de détecter une EPm dans un mot ou dans un non-mot pourraient en effet être différents. D'après Blumstein et al. (1977), il est plus difficile de conserver la représentation d'un non-mot en mémoire à court terme (où l'information se dégrade rapidement) que de maintenir activée la représentation d'un mot déjà présente dans le lexique mental. Ainsi, la détection d'une EPm dans un non-mot demanderait, en plus des processus de détection, le maintien en mémoire à court-terme de l'information phonologique relative au non-mot. On peut donc s'attendre à ce que la détection d'EPm comporte un plus grand nombre de bonnes détections lorsqu'elles apparaissent dans les mots que lorsqu'elles apparaissent dans les non-mots.

La performance de sujets âgés doit également être considérée pour voir si les processus de détection d'erreurs sont sensibles au vieillissement. Chez les personnes âgées, on note que la sensibilité auditive absolue décroît avec l'âge et que les habiletés de discrimination pour la parole diminuent aussi, dans des situations normales d'audition (Marshall, 1981). Il a été avancé que la perte auditive ne peut à elle seule rendre compte de la diminution des capacités de discrimination pour la parole chez la personne âgée (Lutherman et al., 1966; voir Marshall, 1981). On s'attend donc à ce que le pourcentage de bonnes détections d'EPm soit plus grand chez les sujets jeunes que chez les sujets âgés[2].

En résumé, les hypothèses retenues dans la présente étude sont les suivantes: (1) les EPm à DI>2 seront plus faciles à détecter que celles à DI=1; (2) les EPm par voisement (à DI=1) seront plus faciles à détecter que les autres EPm à DI=1; (3) les EPm par addition et omission seront plus faciles à détecter que celles par substitution; (4) les mots donneront lieu à davantage de bonnes détections que les non-mots; et finalement (5) les sujets âgés feront moins de bonnes détections que les sujets jeunes.

MÉTHODOLOGIE

Sujets

Tous les sujets sont droitiers, francophones et ont un seuil d'audition de 30 dB HL ou moins pour les fréquences de la parole (de 500 à 2000 Hz). Trois types de sujets ont été retenus pour cette expérimentation. Un groupe de 10 sujets «jeunes» (groupe J; âge moyen de 23,8 ans, écart-type de 1,3 ans), un groupe de 10 sujets «âgés» (groupe A; âge moyen de 63,3 ans, écart-type de 3,1 ans) ainsi que quatres sujets aphasiques (âgé entre 56 et 68 ans) dont les performances seront comparées au groupe de sujets «âgés». La description des quatres sujets aphasiques est la suivante:

Sujet 1 (S1). Ce sujet a été examiné 4 1/2 ans après son accident cérébro-vasculaire (ACV). Il s'agit d'un homme de 56 ans qui occupait un poste de directeur-adjoint. Il présente une aphasie de Wernicke de type réduit (type I; voir Lecours et Lhermitte, 1979) secondaire à une lésion cérébrale située dans la région temporo-pariétale gauche.

Sujet 2 (S2). Ce sujet a été vu 1 an après son ACV pariétal gauche. Il s'agit d'un homme de 58 ans qui était propriétaire d'une compagnie. Il présente une aphasie de conduction avec apraxie bucco-faciale.

Sujet 3 (S3). Ce sujet a été évalué 3 1/2 ans après son ACV. Il s'agit d'un homme de 65 ans qui occupait un poste de directeur d'école. Il présente une aphasie de Broca secondaire à une atteinte frontale gauche avec possibilité d'un infarctus postérieur ancien.

Sujet 4 (S4). Enfin, le dernier sujet a été vu 5 ans après son ACV. C'est un homme de 68 ans, couvreur de profession. Ce sujet présente une aphasie de Wernicke avec prédominance des troubles du langage écrit (type III; voir Lecours et Lhermitte, 1979), secondaire à une lésion temporo-pariétale gauche; de plus, il est possible que ce sujet ait souffert d'un deuxième ACV gauche, 6 mois après le premier.

Matériel

Douze cibles différentes ont été utilisées dans cette tâche: 6 mots et 6 non-mots trisyllabiques. La structure syllabique des cibles est toujours la même: CCVCVCVC. Les mots et les non-mots sont appariés et renferment les mêmes segments consonantiques. Chaque non-mot est formé en changeant la voyelle de la syllabe initiale et la voyelle de la syllabe finale du mot auquel il est apparié (Tableau 1).

Tableau 1. Liste des cibles.

mots	non-mots
1. privilɛʒ	1. pryviløʒ
2. brikɔlaʒ	2. brykɔliʒ
3. klavikyl	3. klovikal
4. prɔfesœr	4. prafesir
5. krɔkɔdil	5. krakɔdel
6. friʒidɛr	6. fryʒidɔr

Pour chaque cible, six productions erronées ont été construites, renfermant une des six EPm retenues. Les erreurs sont produites par (1) *addition*, (2) *omission*, ou encore par substitution. Dans la catégorie des substitutions, il y a celles (3) par *assimilation*, (4) à *DI>2*, (5) à DI=1 par *voisement* et (6) à DI=1 par un *autre* des traits distinctifs retenus (antérieur, coronal, continu; ce type d'erreur sera appelé dorénavant ADI=1). L'erreur par addition correspond à l'ajout d'une liquide (/l/ ou /r/) dans la deuxième ou la troisième syllabe du mot, alors que l'erreur par omission correspond dans tous les cas à l'élision de la liquide dans la première syllabe. L'assimilation est obtenue en remplaçant une consonne par une autre consonne déjà représentée dans le stimulus. L'assimilation peut ainsi être antérograde (ex.: /klavikyl/ → /klakikyl/) ou rétrograde (ex.: /brikɔlaʒ/ → /bribɔlaʒ/).

Tel que décrit précédemment, la DI est évaluée à travers le nombre de traits (parmi les quatre traits distinctifs retenus) différenciant la cible du stimulus. Ainsi, une substitution à DI>2 correspond au remplacement de la consonne de la deuxième syllabe par une consonne qui diffère par au moins deux traits de la cible. Finalement, une substitution à DI=1 se

manifeste par une modification du voisement ou d'un seul autre des traits distinctifs retenus (Tableau 2).

Tableau 2. Exemples des types d'erreurs phonémiques pour les mots et les non-mots.

	mots	non-mots
CIBLE	brikɔlaʒ	brykɔliʒ
DI=1 (vois.)	brigɔlaʒ	brygɔliʒ
assimilation	bribɔlaʒ	brybɔliʒ
ADI=1	bripɔlaʒ	brypɔliʒ
DI>2	brizɔlaʒ	bryzɔliʒ
addition	brikrɔlaʒ	brykrɔliʒ
omission	blkɔlaʒ	bykɔliʒ

Pour chacune des 12 cibles, 48 stimuli ont été présentés binauralement: la moitié de ces stimuli correspond à la cible alors que l'autre moitié renferme une EPm (cf. supra). Ainsi, 4 occurrences de chacune des 6 EPm et 24 occurrences de la cible sont présentées au sujet, donnant un total de 48 stimuli, pour chacun des mots ou non-mots cible.

Les stimuli correspondant aux mots et aux non-mots sont contenus sur deux bandes magnétiques différentes. Chaque bande commence par une période de familiarisation de 9 essais. L'ordre de passation des deux bandes a été interverti pour la moitié des sujets. Pour chaque cible, l'enregistrement commence par deux présentations de la cible. Un signal sonore («bip») est ensuite présenté, suivi des 48 items correspondant à la cible. Ces 48 stimuli sont présentés dans un ordre quasi-aléatoire, afin qu'il n'y ait jamais plus de trois stimuli successifs qui soient conformes ou non conformes à la cible. L'intervalle (2 ou 3 secondes) entre chacun des stimuli est gardé constant tout au long de la passation; un intervalle inter-stimuli plus court (de moins de 1,5 secondes) aurait pu favoriser un traitement auditif du stimulus entendu, plutôt qu'un traitement phonémique (Werker & Logan, 1985).

Sur une boîte-réponse, deux interrupteurs à haute sensibilité permettent au sujet de signaler si le stimulus entendu est conforme ou non à la cible. Cette boîte est reliée à un micro-ordinateur Macintosh Plus et à un magnétophone Revox B77 servant à présenter les stimuli. Le casque d'écoute du sujet est branché à ce magnétophone. Le micro-ordinateur enregistre, pour chaque stimulus, la réponse ainsi que le temps de détection de l'erreur. Le temps de détection est mesuré entre le moment où est produit le phonème erroné et le moment où le sujet appuie sur l'interrupteur. Cette mesure a été réalisée à l'aide d'un système MacAdios, et incorporée au programme d'analyse des données.

Déroulement

Le sujet est assis devant la boîte-réponse. Il est informé de la nature (mot ou non-mot) des cibles composant la bande qu'il entendra, et une explication concernant les non-mots lui est fournie. La consigne est donnée oralement au sujet. Il doit juger si chaque stimulus entendu est conforme (CORRECT) ou non conforme (INCORRECT) à la cible donnée à deux reprises avant le début des 48 stimuli. Pour répondre, le sujet doit appuyer, le plus rapidement possible, sur l'interrupteur approprié (CORRECT ou INCORRECT). Tous les sujets utilisent leur main gauche puisque certains sujets aphasiques présentent une hémiplégie droite. Avant et après chaque réponse, le sujet ramène son doigt à une position de départ, équidistante des deux boutons. La tâche dure environ 60 minutes, et peut être effectuée en une seule session ou en deux sessions de 30 minutes.

RÉSULTATS

Le temps de détection et le nombre de bonne détection (BD) ont été mesurés pour chacun des sujets. Toutefois, seuls les pourcentages de BD seront rapportés dans la présente étude. Le nombre de BD pour chaque sujet est transformé en pourcentage de BD pour chacun des types d'erreurs, ce dernier correspondant au rapport: nombre de BD sur nombre total (N=24) de stimuli erronés. Dans le cas des substitutions à ADI=1, les pourcentages de BD sont toutefois calculés sur un nombre total de 16 occurrences plutôt que 24.

Chez les sujets normaux, les rapports «nombre de BD/nombre total de stimuli erronés» pour chacun des sujets ont été soumis à une transformation angulaire (Winer, 1971) en raison de leur distribution ne suivant pas une courbe normale. Ces données ont été soumis à trois analyses de variance à mesures répétées GROUPE X (NATURE X TYPE) dans lesquelles le «groupe» comprend les groupes de sujets jeunes et âgés, la «nature de la cible» correspond aux mots et aux non-mots, alors que la variable «type d'erreur» varie en fonction de l'analyse effectuée. Dans la première analyse, la variable «type d'erreur» se limite aux EPm DI>2 et DI=1 (incluant voisement et ADI=1). Dans la seconde analyse, la variable «type d'erreur» comprend chacun des six types d'erreurs. Enfin, dans la dernière analyse, la variable «type d'erreur» est composé des EPm «omission», «addition» et «substitution»; ce dernier incluant le voisement, ADI=1, DI>2 et assimilation.

Quant aux performances des sujets aphasiques, des analyses statistiques ne peuvent être menées. En effet, il n'est pas possible de regrouper les sujets aphasiques en sous-groupes homogènes (Schwartz, 1984). Ainsi, il est fortement recommandé, à la suite notamment de Caramazza (1986), de considérer chaque sujet comme un cas particulier. Par conséquent, un plan expérimental du type «étude de cas multiples» a été retenu. Un tel plan expérimental ne suppose aucunement l'homogénéité d'un groupe de sujets, mais repose plutôt sur la compatibilité de la performance de chacun des sujets avec les hypothèses ayancées. Il s'agira dans des recherches ultérieures de définir, le plus précisément possible, les capacités de détection de chacun des sujets aphasiques et de noter d'éventuelles dissociations.

Le tableau 3 présente les résultats en pourcentages de bonnes détections obtenus pour les groupes de sujets jeunes, âgés et pour chacun des sujets aphasiques en fonction de la nature de la cible et des différentes EPm.

Tableau 3: Pourcentage de bonnes détections pour les groupes de sujets jeunes, âgés et pour chacun des sujets aphasiques en fonction de la nature de la cible et du type d'erreur

	J	A	S1	S2	S3	S4
MOTS						
omission	94.2	93.8	54.2	37.5	87.5	0
addition	97.9	99.6	100	91.7	100	70.8
substitution	97.5	97.3	93.2	73.9	94.4	65.9
assimilation	100	99.2	95.8	75	95.8	58.3
DI>2	99.6	100	100	100	100	100
DI=1	94.7	94.5	87.5	57.5	92.5	50
voisement	95.4	94.6	95.8	70.8	87.5	58.3
ADI=1	93.8	94.4	75	37.5	100	37.5
NON-MOTS						
omission	99.6	98.7	83.3	54.2	87.5	41.7
addition	97.5	96.7	100	75	75	58.3
substitution	98.9	96.2	89.8	81.8	88.6	72.7
assimilation	99.2	96.7	100	75	83.3	83.3
DI>2	100	100	100	100	100	91.7
DI=1	98	93.7	77.5	75	85	55
voisement	98.7	97.1	95.8	87.5	87.5	54.2
ADI=1	96.9	88.7	50	56.3	81.2	56.3

1- Substitutions de type DI=1 et DI>2

Selon l'hypothèse (1), le pourcentage de BD sera plus grand pour les substitutions à DI>2 que pour les substitutions à DI=1, incluant celles par voisement et celles par ADI=1.

Les résultats pour les sujets normaux (analyse 1) montrent que les EPm de type DI=1 sont plus difficiles à détecter que celles de type DI>2 (95,2% contre 99,9%) et ceci, quel que soit le groupe d'âge des sujets ou la nature de la cible.

Les pourcentages de BD des sujets aphasiques révèlent également une difficulté plus marquée pour les substitutions à DI=1. Il est à noter que seul le sujet 4 fait deux mauvaises détections sur les substitutions à DI>2 dans le cas des non-mots, alors que les autres sujets aphasiques ont une performance parfaite.

De façon générale, il semble donc que les données recueillies chez tous les sujets appuient clairement l'hypothèse (1) et corroborent les résultats de McInish et Tikofsky (1969), Cole (1973) de même que ceux de Milberg et al. (1988). Il ressort que deux items s'apparentant fortement quant à leur constitution phonémique, i.e. se distinguant par un petit nombre de traits distinctifs, sont plus souvent confondus.

2- Substitution à DI=1: voisement et autres

D'après l'hypothèse (2), le pourcentage de BD sera plus grand pour les erreurs de substitution à DI=1 ayant trait au voisement que pour les autres (ADI=1).

Pour ce qui est des mots, les résultats chez les sujets normaux (analyse 2) ne révèlent pas de différence significative entre les deux types d'EPm à DI=1 (voisement et autres). La détection des substitutions à DI=1 de voisement est semblable à celle des ADI=1 (95% contre 94,1%), peu importe le groupe d'âge. Toutefois, et conformément à l'hypothèse, les EPm impliquant le voisement sont significativement ($p \leq 0,05$) mieux détectées que les ADI=1 (97,9% contre 92,8%), pour ce qui est des non-mots.

Les performances des sujets aphasiques sont diversifiées. En effet, les sujets 1 et 2, conformément à l'hypothèse, présentent plus de BD sur les EPm ayant trait au voisement que sur les EPm de type ADI=1 dans le cas des mots (S1: 95,8% contre 75% et S2: 70,8% contre 37,5%) et dans le cas des non-mots (S1: 95,8% contre 50% et S2: 87,5 contre 56,3%). Quant au sujet 3, un aphasique de Broca, il éprouve plus de facilité avec les erreurs concernant le voisement qu'avec les ADI=1 dans le cas des non-mots alors que des résultats contraires sont observés dans le cas des mots (non-mots: 87,5% contre 81,2% et mots: 87,5% contre 100%). Pour le sujet 4, c'est dans le cas des mots que les EPm par voisement sont plus facilement détectées que les ADI=1 (58,3% contre 37,5%) alors qu'on note que légèrement moins de BD pour les erreurs concernant le voisement pour les non-mots (54,2% contre 56,3%).

En conclusion, nous constatons que, pour la détection des EPm par substitution à DI=1 pour les non-mots, les EPm par voisement sont plus facilement détectées que les EPm ADI=1 tant chez les sujets normaux qu'aphasiques. Cependant, pour ce qui est des mots, cette observation est notée seulement chez certains sujets aphasiques (sauf S3). Ainsi, les résultats de Miller et Nicely (1955) obtenus chez des sujets normaux lors de présentation de mots en condition de bruit et de filtrage ne sont pas corroborés alors que ceux de Blumstein et al. (1977) obtenus chez des sujets aphasiques le sont.

3- Omission, addition et substitution

Selon l'hypothèse (3), les EPm par addition et omission seront plus facile à détecter que celles par substitution.

La comparaison des moyennes de bonnes détections impliquant les EPm par addition, omission et par différents types de substitution (i.e., assimilation, substitution à DI>2 et à DI=1), est faite à l'aide de la méthode de Tukey-A. De l'analyse portant sur les mots, il ressort, et ceci peu importe le groupe de sujets, que (a) les EPm par substitution sont significativement ($p \leq 0,05$) mieux détectées que les EPM par omission (97,4% contre 93,9%); (b) les EPm par addition sont mieux détectées que les EPm par omission (98,7% contre 93,9%); alors que (c) la détection des EPm par substitution et addition ne diffèrent pas entre elles.

Par ailleurs, l'analyse portant sur les non-mots n'indique aucune différence significative quant à la détection de ces trois types d'EPm tant chez les sujets jeunes qu'âgés.

Chez les sujets aphasiques, les résultats montrent que lorsque la cible présentée est un mot (a) les EPm par substitution sont mieux détectées que les EPm par omission (S1: 93,2% contre 54,2%; S2: 73,9% contre 37,5%; S3: 94,4% contre 87,5%; et S4: 65,9% contre 0%); (b) tous les sujets détectent mieux les EPm par addition que par omission (S1: 100% contre 54,2%; S2: 91,7% contre 37,5%; S3: 100% contre 87,5%; et S4: 70,8% contre 0%); (c) tous les sujets détectent mieux les EPm par addition que les EPm par substitution (S1: 100% contre 93,2%; S2: 91,7% contre 73,9%; S3: 100% contre 94,4%; et S4: 70,8% contre 65,9%).

Chez ces mêmes sujets, pour les non-mots cette fois, on remarque (a) les EPm par substitution sont mieux détectées que les EPm par omission (S1: 89,8% contre 83,3%; S2: 81,8% contre 54,2%; S3: 88,6% contre 87,5%; et S4: 72,7% contre 41,7%) (b) les EPm par addition sont mieux détectées que les EPm par omission (S1: 100% contre 83,3%; S2: 75% contre 54,2%; et S4: 58,3% contre 41,7%), sauf pour le sujet 3 (S3: 75% contre 87,5%); (c) les EPm par substitution sont mieux détectées que les EPm par addition (S2: 81,8% contre 75%; S3 88,6% contre 75%; et S4: 72,7% contre 58,3%), sauf pour le sujet 1 (S1: 89,8% contre 100%).

En résumé, contrairement à l'hypothèse, tous les sujets détectent mieux les EPm par substitution que les EPm par omission et ce, indépendamment de la nature de la cible. Pour ce qui est de la comparaison entre EPm par substitution et par addition, on note aucune différence significative chez les sujets normaux, tandis que chez les sujets aphasiques, les performances de détection de ces EPm varient en fonction de la nature de la cible. Enfin, les EPm par addition sont plus facilement détectées que les EPm d'omission, à l'exception des sujets normaux où aucune différence n'est notée pour les non-mots.

4- Nature de la cible

Les résultats des analyses 2 et 3 indiquent que les types d'EPm influencent différemment la performance de détection, selon que la cible est un mot ou un non-mot. Ces analyses révèlent une différence significative entre les mots et les non-mots uniquement pour les EPm impliquant le voisement ($F(1,18)=6,4$; $p=0,021$) et l'omission ($F(1,18)=13,52$; $p=0,002$). Toutefois, cette différence va dans le sens contraire de l'hypothèse: pour ces deux types d'erreurs, le nombre de bonnes détections est plus grand pour les non-mots que pour les mots.

La comparaison inter-sujet aphasiques met également en évidence des différences dans la détection des EPm lorsque le cible est un mot ou un non-mot. En effet, la détection des EPm par omission et par ADI=1 est mieux réussie lorsqu'il s'agit de non-mots que de mots, alors que les EPm par addition sont facilitées par le statut lexical de la cible. Pour ce qui est des autres types d'EPm, il n'y a aucune cohésion inter-sujet aphasique.

En résumé, contrairement à l'hypothèse (4), les EPm portant sur les non-mots sont mieux détectées que celles portant sur les mots (a) pour les EPm par omission et par voisement chez les sujets normaux et (b) pour les EPm par omission et par ADI=1 chez les sujets aphasiques. Seules les EPm par addition, et ce uniquement chez les sujets aphasiques, semble appuyer l'hypothèse (4) selon laquelle les EPm portant sur des mots sont mieux détectées que celles portant sur des non-mots.

5- Age et pourcentage de bonnes détections

L'hypothèse (5) avancée sur le groupe d'âge prédit que les sujets âgés feront moins de BD que les sujets jeunes.

Contrairement à l'hypothèse, les analyses 2 et 3 menées sur les proportions de BD ne démontrent aucune différence significative entre les sujets jeunes et âgés. La performance des sujets jeunes et âgés est comparable quant au nombre de BD réalisées. Notons toutefois, lorsqu'on compare les EPm DI>2 et DI=1 (analyse 1), les sujets jeunes ne se comportent pas de la même façon que les sujets âgés concernant les mots et les non-mots. Chez les sujets jeunes, le nombre de BD pour les non-mots est supérieur à celui des mots alors que chez les sujets âgés, la détection de ces deux types d'EPm ne varie pas en fonction du statut lexical.

Discussion

La présente étude a mis en évidence une détection plus difficile pour les EPm par substitution à DI=1 que pour celles à DI>2 en ce qui concerne les mots. Les résultats de Cole (1973), de McInish & Tikofsky (1969) de même que ceux de Milberg et al. (1988) sont réitérés.

D'après le modèle TRACE de Mc Clelland et Elman (1986), les traits distinctifs excitent davantage certains phonèmes plutôt que d'autres. Avec l'accroissement de l'activation, un phonème domine normalement les autres. Alors que ce processus est en marche, la rétroaction provenant du niveau phonémique tend à imposer son patron canonique d'activation au niveau des traits distinctifs. Le modèle TRACE amène à poser l'hypothèse selon laquelle les erreurs de détection des substitutions seraient dues à un trop fort effet de la rétroaction provenant du niveau phonémique par rapport à l'excitation provenant du niveau des traits distinctifs. Le fait que les EPm par substitution à DI=1 soient moins bien détectées que les EPm par substitution à DI>2 appuie cette hypothèse. En effet, un phonème substituant à DI=1 active davantage le phonème-cible qu'un phonème substituant à DI>2. Et par conséquent, la rétroaction provenant du niveau phonémique a plus de chance d'agir lorsque la substitution est à DI=1 que lorsqu'elle est à DI>2.

Une meilleure détection des substitutions à DI=1 pour le voisement comparativement aux ADI=1 a été montrée chez les sujets normaux et aphasiques en ce qui concerne les non-mots. Dans le cas des mots, il n'existe de différence entre les substitutions à DI=1 que chez les sujets aphasiques. La distinction observée par Miller et Nicely (1955) et Blumstein (1977) entre voisement et ADI=1 n'apparaît peut-être pas en raison des traits distinctifs utilisés dans notre catégorie ADI=1. En effet, ces auteurs ont opposé voisement et lieu d'articulation, alors que la présente étude a opposé voisement et ADI=1. Ainsi, il semble que ADI=1, regroupant les traits de lieu d'articulation «antérieur» et «coronal» et le trait de mode d'articulation «continu», ne soit pas équivalent au seul trait de lieu d'articulation.

Quant aux comparaisons effectuées entre les trois classes d'EPm soit l'omission, l'addition et la substitution, elles n'ont pas confirmé notre hypothèse. On remarque en effet que pour les mots, les omissions sont les EPm les moins bien détectées, tant chez les sujets normaux qu'aphasiques. Toutefois lorsque la cible est un non-mot, la plus grande difficulté dont fait l'objet l'omission n'est observée que chez les aphasiques.

Il semble qu'un phénomène de restauration intervienne dans le cas des erreurs d'omission. La restauration agit de telle sorte que le sujet croit percevoir un phonème bien qu'il soit manquant. Warren et Obusek (1971) se sont tout spécialement intéressés au phénomène de la restauration phonémique et soulignent qu'elle est un mécanisme essentiel, utilisé fréquemment pour la compréhension en situation de bruit. La restauration pourrait expliquer la difficulté à détecter les EPm par omission pour les mots chez les sujets normaux et aphasiques. Par ailleurs, il semble que les sujets aphasiques utilisent également la restauration dans le cas des non-mots, quoique de façon moins importante que dans le cas des mots. De ce fait, le domaine d'application de la restauration ne serait pas strictement réservé aux stimuli ayant une représentation lexicale, du moins chez les sujets aphasiques. Il faut cependant souligner ici que, parmi les 48 stimuli présentés au cours de la tâche, 24 sont identiques et conformes à la cible présentée deux fois en début de série. Ceci permet une consolidation de la représentation phonologique en mémoire à court terme sans toutefois constituer l'équivalent d'une représentation lexicale.

Il faut souligner que les EPm par omission des stimuli présentés sont d'un type particulier. En effet, la consonne omise correspond toujours à la liquide du groupe consonantique initial (ex. : /brikɔlaʒ/ → /bikɔlaʒ/). Le nombre de BD serait peut-être plus élevé pour ce type d'EPm si l'omission portait sur l'obstruente du groupe consonantique plutôt que sur la liquide (ex. : /brikɔlaʒ/ → /rikɔlaʒ/).

Ce qui était attendu concernant la plus grande facilité à détecter les EPm par addition que les EPm par substitution, n'est pas rencontré. En effet, les sujets normaux ont une performance similaire pour ces deux types d'EPm, aussi bien dans le cas des mots que dans le cas des non-mots. L'examen des résultats des sujets aphasiques révèle une disparité entre la détection des EPm par addition et par substitution lorsqu'il s'agit de mots et lorsqu'il s'agit de non-mots. Alors que pour les mots l'hypothèse semble appuyée, pour les non-mots, c'est la situation inverse qui est observée. L'erreur d'addition consistant de toute évidence à remplir une case vide par un nombre x de traits distinctifs définissant un phonème, est mieux détectée qu'une erreur de substitution plus fine puisqu'elle ne diffère par rapport à la cible que par quelques traits distinctifs.

Il est intéressant de noter que les EPm par omission sont détectées plus difficilement que les EPm par addition, et ceci aussi bien chez les sujets normaux qu'aphasiques. L'hypothèse d'une restauration phonémique peut à nouveau rendre compte de la difficulté marquée des EPm par omission.

La nature de la cible (mot/non-mot) a une grande influence sur le nombre de BD d'EPm comme le montrent les interactions de cette variable avec le type d'EPm et avec le groupe de sujets. Étant donné l'existence d'une représentation lexicale pour les mots, il était attendu que la performance pour ceux-ci soit supérieure à celle des non-mots. La seule donnée appuyant cette hypothèse provient des résultats observés chez les aphasiques lorsque l'EPm est de type addition. Plusieurs types d'EPm, tels l'omission et le voisement chez le sujet normal, et l'omission et l'ADI=1 chez les sujets aphasiques, vont à l'encontre de cette prédiction.

La représentation lexicale aide à la détection seulement lorsque le stimulus avec EPm est très différent de la cible, comme dans le cas de l'addition, car lorsque le stimulus avec EPm est semblable à la cible, il y a un phénomène d'attente cognitive, appelé traitement «top-down» (Rumelhart, 1977). Celui-ci entraînent à son tour un effet de compensation qui viendrait interférer dans la détection d'erreur. Bien que l'omission est également une EPm très différente de la cible, elle est néanmoins sujette au phénomène de compensation appelé ici «restauration de phonème».

L'hypothèse selon laquelle les sujets âgés éprouvent plus de difficulté à détecter les EPm que les sujets jeunes n'est pas confirmée. De façon générale, la performance des sujets jeunes et âgés ne diffèrent pas entre elles. Ainsi, les sujets âgés ne présentent pas de diminution des habiletés de discrimination pour la parole telle qu'observée par Lutherman et al. (1966; Marshall, 1981) auprès d'un même population. Toutefois, la seule observation permettant de distinguer les sujets âgés et jeunes est mise en évidence dans l'analyse portant sur les types d'EPm, DI=1 et DI>2. En effet, il est possible de rapporter que les sujets jeunes se comportent différemment lorsque la cible est un mot ou un non-mot.

On peut tenter d'expliquer l'absence de différence entre les deux groupes de sujets par le fait que les sujets âgés ont des temps de détection plus longs pour les EPm par omission et par substitution à DI=1 pour les mots. Or, ces types d'erreurs sont les plus difficiles. On peut penser que les sujets âgés en présence d'une EPm difficile ont tendance à délaisser la rapidité d'exécution en faveur de l'exactitude de la réponse.

Conclusion

En conclusion, l'étude de la détection de différents types d'EPm révèle effectivement des degrés de «détectabilité» différents. Notamment, les EPm de type *substitution à DI>2* sont très bien détectées, tant chez les sujets normaux qu'aphasiques, alors que les EPm de type *omission*, de même que les substitutions à DI=1 *(voisement et autres)*, sont de façon générale moins bien détectées.

Lorsque la substitution implique plusieurs traits (DI>2), l'erreur est mieux détectée, alors que lorsque le stimulus et la cible sont semblables (substitution à DI=1), la détection de l'EPm est plus difficile, tant chez les sujets normaux qu'aphasiques. L'efficacité du mécanisme permettant la mise en correspondance d'un stimulus erroné avec une cible déterminée est limitée par l'activation d'autres mécanismes tels que l'activation du patron canonique d'activation du phonème-cible provenant de la rétroaction du niveau phonémique et la restauration. Ainsi, la restauration semble opérationnelle dans une tâche hors contexte comme la détection d'EPm alors qu'elle devrait être inhibée. Tous ces mécanismes ne semblent pas s'appliquer de la même façon dans le cas des mots et des non-mots, puisque les sujets ont plus de facilité avec les EPm par addi-

tion lorsqu'il s'agit des mots mais plus de facilité avec les EPm par omission lorsqu'il s'agit des non-mots.

La comparaison des performances de détection des EPm en fonction des groupes d'âge ne révèle pas de différence. Il est donc possible de croire que le facteur âge ne semble pas avoir un rôle prépondérant dans la perception des EPm.

Par ailleurs, l'existence d'une représentation lexico-phonologique et lexico-sémantique dans le cas des mots n'étant pas nécessairement associée à une plus grande facilité de détection des EPm, il faudra donc mener d'autres études pour tenter de mieux saisir le rôle de la représentation lexicale. Aussi, au-delà de l'étude des mécanismes de détection des EPm dans une tâche de détection, il faudra également s'intéresser aux auto-détections d'erreurs effectuées tant par des sujets aphasiques que par des sujets normaux en situation de discours spontané.

Il reste que ce travail montre l'intérêt que peut avoir l'étude des capacités de détection d'erreurs phonémiques. L'identification exacte de ces capacités est importante non seulement pour notre compréhension des processus psycholinguistiques chez le sujet normal ou aphasique, mais également pour l'identification de stratégies de compensation/rééducation susceptibles d'optimiser les aspects les moins touchés du comportement verbal chez ces malades.

NOTES

1. Il s'agit ici de sujets aphasiques présentant une lésion rétro-rolandique.
2. En ce qui concerne les temps de détection, une différence de performance, du moins pour cette tâche, dépend d'au moins deux choses: soit du temps nécessaire à la réponse motrice (le mouvement lui-même), soit du temps de traitement nécessaire à la détection de l'erreur phonémique. Il importe toutefois de préciser que dans cette étude, la distinction entre le temps de traitement de l'information auditive perçue et le temps de réponse motrice n'a pas été considérée.

BIBLIOGRAPHIE

BASSO, A., CASATI, G. et VIGNOLO, L. A. (1977). Phonemic identification defect in aphasia. *Cortex*, 13, 85-95.

BLUMSTEIN, S. (1973). *A phonological investigation of aphasic speech.* The Hague : Mouton.

BLUMSTEIN, S. E., BAKER, E. et GOODGLASS, H. (1977). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia*, 15, 19-30.

COLE, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, 1, 153-156.

CONNINE, C. M. et CLIFTON, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 291-299.

FELDMAN, R. M. et REGER, S. N. (1967). Relations among hearing, reaction time, and age. *Journal of Speech and Hearing Research*, 10, 479-495.

FROMKIN, V. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47, 27-52.

LACKNER, J. R. et TULLER, B. H. (1979). Role of efference monitoring in the detection of self-produced speech errors. In: W.E. Cooper et E.C.T. Walker (Éds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Lawrence Erlbaum Associates, 281-294.

LECOURS, A. R., ET LHERMITTE, F. (1979) *L'aphasie.* Montréal: Les Presses de l'Université de Montréal.

MARSHALL, L. (1981). Auditory processing in aging listeners. *Journal of Speech and Hearing Disorders*, 8, 226-240.

MARSLEN-WILSON, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25, 71-102.

MCINISH, J. et TIKOFSKY, R. S. (1969). Distinctive features and response latency: A pilot study. *Perception and Psychophysics*, 6(5), 267-268.

MILBERG, W., BLUMSTEIN, S. et DWORETZKY, B. (1988). Phonological processing and lexical access in aphasia. *Brain and Language*, 34, 279-293.

MILLER, G. A. et NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of Acoustical Society of America*, 27(2), 338-352.

NESPOULOUS, J.-L., JOANETTE, Y., SKA, B., CAPLAN, D. et LECOURS, A. R. (1987). Production deficits in Broca's and conduction aphasia: Repetition versus reading. In: E. Keller et M. Gopnik (Éds.)*Motor and sensory processes of language*, E. KELLER et M. GOPNIK, Hillsdale, NJ: Lawrence Erlbaum Associates, 53-81.

RUMELHART D.E. (1977). *Introduction to human information processing.* New York: Wiley.

WARREN, R. M. et C. J. OBUSEK (1971). Speech perception and phonemic restoration. *Perception & Psychophysics*, 9, 358-362.

WERKER, J. F. et J. S. LOGAN (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics*, 37(1), 35-44.

WINER, B. J. (1971). Statistical principles in experimental design. (2e éd.). NY: McGraw-Hill.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

PERCEPTION DE LA DUREE SYLLABIQUE DANS UNE PHRASE EN ANGLAIS

Yukihiro, NISHINUMA

CNRS:URA 261, Laboratoire "Parole et Langage"
Institut de Phonétique, Université de Provence,

Résumé

Afin d'étudier la différentiation perceptuelle des durées syllabiques dans le cadre de la phrase, nous avons effectué deux expériences. Une phrase naturelle en anglais comportant 8 syllabes a été re-synthétisée en modifiant la durée de chaque syllabe, de -60% à +60% par pas de 15%. Les stimuli ainsi préparés ont été présentés à 20 sujets pour juger le naturel de la phrase (Expérience I). Les mêmes stimuli ont été utilisés pour repérer les modifications temporelles (Expérience II). Les résultats mettent en évidence que le jugement sur la qualité rythmique de la phrase (tâche psycho-linguistique) et la différentiation temporelle (tâche psycho-acoustique) se fait sur la même base fonctionnelle.

1. INTRODUCTION

L'analyse de l'organisation temporelle de la phrase en français, s'appuyant sur la manipulation de la durée syllabique nous a permis de mettre en évidence un certain nombre de faits perceptuellement importants (Duez & Nishinuma 1986). Entre autres, nous avons remarqué au niveau de la production, (1) une particularité sur la durée de la syllabe initiale, (2) l'alternance des durées syllabiques et (3) une certaine stabilité de la durée des syllabes inaccentuées. La vérification perceptuelle de ces effets nous a conduit à entreprendre cette étude ayant trait à la discrimination de la durée syllabique dans le cadre de phrases courtes en plusieurs langues. Nous présentons ci-après les premiers résultats expérimentaux concernant l'anglais.

2. EXPERIENCE I

2.1. Protocole expérimental

Afin d'obtenir les stimuli expérimentaux, nous avons procédé à la synthèse au moyen d'un logiciel (Espesser 1988), en manipulant uniquement la durée syllabique de la phrase originale lue par une femme: "Some nice young girls drink bad root beer". Cette phrase comportent huit syllabes, un groupe nominal de quatre syllabes et un groupe verbal de quatre syllabes. Compte tenu de la difficulté dans la manipulation temporelle des mots fonctionnels courts, souvent inaccentués et extrêmement brefs, nous avons choisi une phrase constituée uniquement de mots pleins.

A partir des durées syllabiques initiales montrées dans la **Figure 1**, nous avons modifié une seule syllabe dans chaque stimulus-phrase. Le taux de variation va de -60% à +60% par

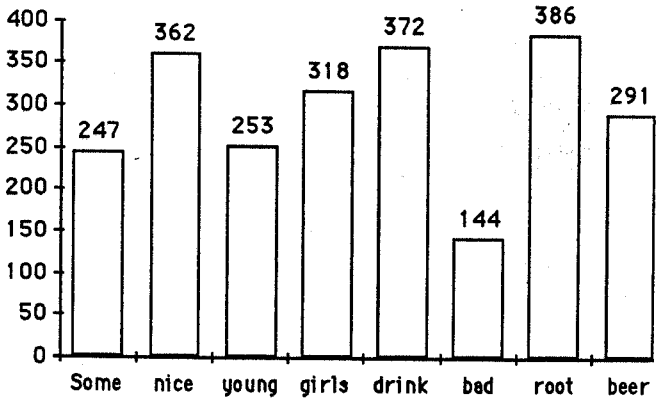


Figure 1
Durées syllabiques initiales de la phrase expérimentale

palier de 15%. Une syllabe se présente donc en neuf versions différentes (avec 4 abrègements, 4 allongements et 1 inchangée). En tout 72 stimuli-phrases (8 syllabes * 9 versions) ont été engendrés et enregistrés sur une bande magnétique, dans un ordre aléatoire, avec un intervalle inter-stimuli de 2 secondes et un intervalle de 5 secondes tous les 18 stimuli. Une série de dix-huit stimuli destinée à l'entraînement, ainsi que les stimuli expérimentaux à proprement parler, ont été présentés une fois à vingt sujets dans une chambre anéchoïque. Leur tâche consistait à juger la qualité rythmique de la phrase entendue; ils avaient trois réponses possibles pré-inscrites sur le questionnaire: *naturelle*, *acceptable* ou *inacceptable*.

2.2. Analyse des résultats

Les résultats bruts sont soumis à des analyses statistiques en excluant bien entendu les séries d'entraînement et les stimuli factices. Nous avons converti en valeurs numériques les trois réponses *naturelle*, *acceptable*, et *inacceptable* en leur affectant les coefficients: 2, 1 et 0 respectivement (ainsi on neutralise les réponses inacceptable). L'analyse de variance sur les scores *naturelle*, *acceptable* montre que la différence entre les sujets ($F(19,71)=29,26$) et la différence entre les stimuli sont significatives ($F(8, 152)=18,80$, $p<0,001$). Par contre, la différence entre les syllabes et l'interaction *sujets* * *syllabes* étant non significative, les sujets semblent se comporter

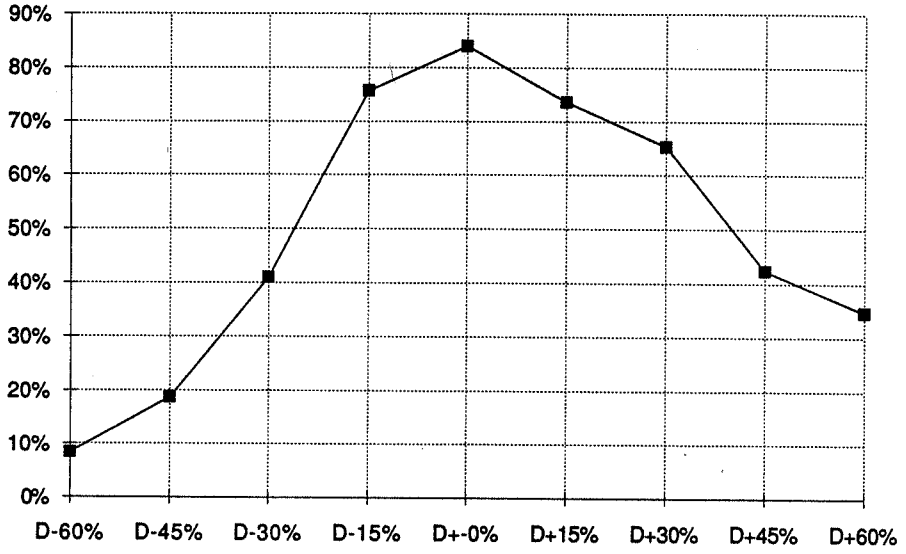


Figure 2
Réponses *naturelle* et *acceptable*

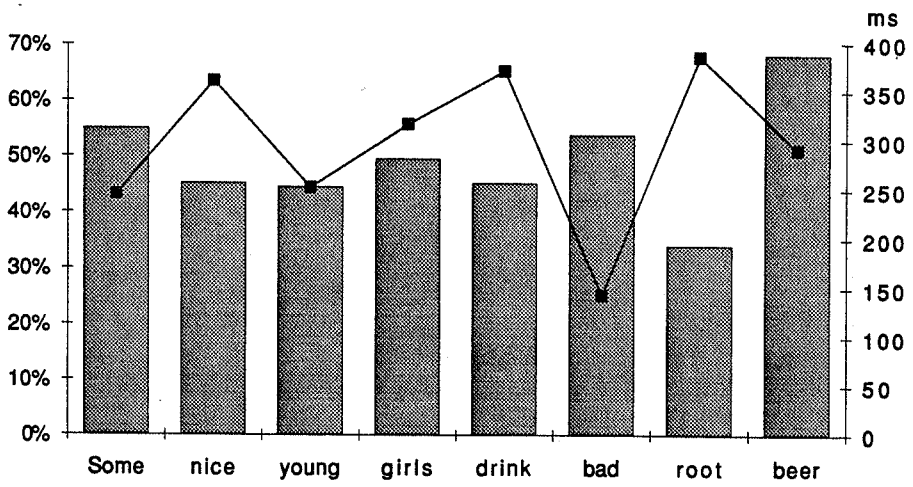


Figure 3
Réponses *naturelle* et *acceptable* par syllabe pour 20 sujets. Les points reliés entre eux représent la durée originale de la syllabe.

avec une relative homogénéité vis-à-vis des syllabes.
La majorité des réponses positives (phrases jugées *naturelle, acceptable*) se concentrent sur les stimuli les moins affectés, c'est-à-dire ceux du milieu comme on le voit sur la **Figure 2** où un point de la courbe représente la réponse de vingt sujets qui ont participé à l'expérience.

Bien que cette différence ne soit pas statistiquement significative, nous avons observé que les réponses par syllabe sont également variables suivant la position, notamment la syllabe initiale et les trois dernières syllabes (cf. **Figure 3**) En ce qui concerne la syllabe initiale, nous avons observé la même tendance en français, à savoir qu'elle supporte un raccourcissement et un allongement prononcés (Nishinuma & Duez, 1989). Plusieurs hypothèses sont possibles pour expliquer ce phénomène. Cette syllabe peut être relativement longue pour des raisons phono-stylistiques: hésitation, accent d'insistance ou accent de contraste, etc, bien que dans le cas normal, elle reste brève. Ainsi les modifications temporelles sur cette syllabe auraient-elles été jugées favorablement par les sujets. Une autre hypothèse est que la notion de débit n'étant établie qu'à partir de la seconde syllabe, la syllabe initiale laisserait un grand degré de liberté perceptive.

On remarquera également que les sujets supportent peu de modifications dans les deux sens (raccourcissement et allongement) sur la syllabe "root", accentuée et ayant une durée très allongée dans la phrase originale. Ce qui est contraire au français pour lequel nous avons constaté une rigidité de la durée sur les syllabes inaccentuées. L'adjectif étant éventuellement *accentuable* aussi bien en français qu'en anglais, ici "bad" accepte une certaine élasticité dans les modifications temporelles

La syllabe finale "beer" a recueilli les meilleurs scores; l'allongement a été jugé favorable, car dans la position qu'il occuoe dans la phrase une durée longue est normale. Les réponses positives sur cette même syllabe pour des durées réduites, tiendraient vraisemblablement au fait que les sujets attendaient une durée longue et corrigeaient d'eux-même la réalité perçue.

Dans cette expérience, on ignore ce que représente réellement la réponse *naturelle*, car elle peut tout aussi bien désigner un stimulus allongé qu'un stimulus raccourci.

3. EXPERIENCE II

3.1. Protocole expérimental

Afin de répondre à cette question, nous avons effectué une deuxième expérience avec les mêmes stimuli. Cette fois-ci, nous avons demandé aux sujets si une modification de durée syllabique était perceptible ou non. Ils ont été invités à inscrire sur les feuilles de réponses, sous la syllabe correspondante (Précisons que les sujets ignoraient quelle syllabe était la syllabe cible), le signe + dans le cas d'un allongement perçu, le signe - pour un raccourcissement, et = si aucune modification n'était entendue. Les stimuli ont été présentés à 20 sujets, qui ont effectué l'expérience I.

3.2. Analyse des résultats

Nous avons converti les réponses: "-", "=", et "+", en valeurs numériques -1, 0, et +1 respectivement. L'analyse de variance sur ces données codées nous indique d'une part, la différence significative entre les sujets ($F(19,71)=3,02$, $p<0,001$) et entre les stimuli ($F(8, 152)=106,69$, $p<0,001$), En revanche les scores entre les syllabes et l'interaction *sujets* * *syllabes* se sont avérés non significatifs.

La **Figure 4** illustre les fonctions psycho métriques représentant les trois catégories, de gauche à droite: abrègements perçus, modifications non-perçues et allongements perçus. A l'intersection des courbes au niveau de 50%, les valeurs correspondantes sur les échelles des stimuli sont -DD = 24,3% et +DD = 27,2%. Le seuil à partir duquel on peut percevoir une modification de la durée syllabique est donc de 25,7% en moyenne dans cette expérience.

Si nous regardons de près la courbe de réponses "=" (**Figure 5**), la syllabe "root" a une meilleure précision perceptive de DD = 18%. Ces résultats peuvent être expliqués par le fait de la durée extrêmement longue de cette syllabe (386 ms), durée supérieure au temps d'intégration de la durée en général.

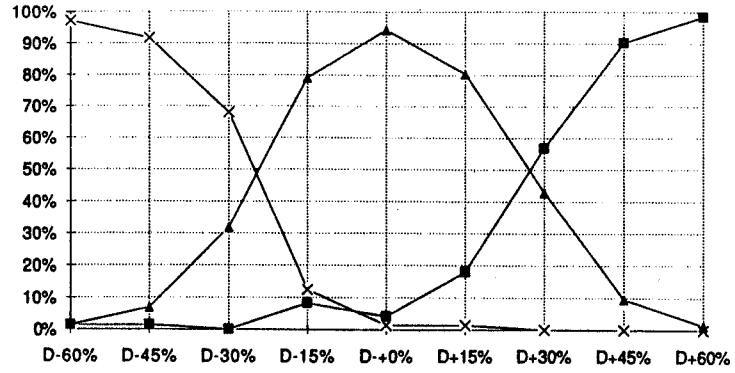


Figure 4
Réponses -, = et + pour chaque stimulus

La même explication ne tiendrait malheureusement pas pour la syllabe "beer" qui a une durée moyenne, mais montre les scores les plus médiocres. A l'examen de l'intensité globale de la phrase, nous avons remarqué que toutes les syllabes, sauf la dernière, ont une intensité homogène de 68 dB en moyenne; celle de la syllabe finale a une intensité décroissant de 62 à 52 dB. Une différence allant de -6 à -16 dB ne peut pas être sans incidence sur la perception de la durée, car le niveau sonore influence la discrimination temporelle (Rochester, 1971). Par conséquent, les sujets semblent avoir eu des difficultés pour apprécier avec précision les modifications temporelles de cette syllabe.

Il nous a semblé que les sujets réagissaient mieux pour les stimuli abrégés que pour les stimuli allongés. Toutefois cette asymétrie observable sur toutes les syllabes ne s'est pas révélée significative d'après l'analyse de variance sur les scores des stimuli variant de 15% à 30% ($F(1,62) = 0,172$).

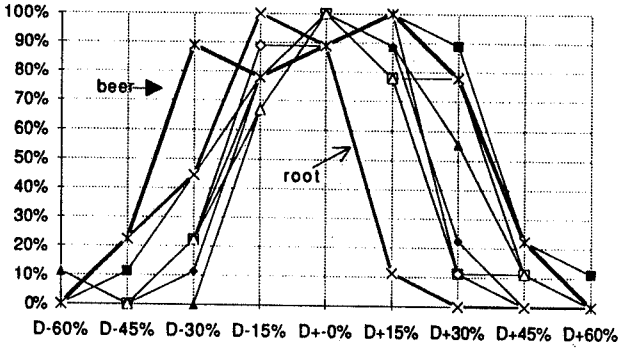


Figure 5
Réponses "= " pour chaque syllabe

La constatation faite par Klatt et Cooper(1975) que le seuil est meilleur en position non-finale est ainsi vérifiée. D'autre part, en l'absence de la différence significative entre les syllabes (surtout celles du milieu) nous partageons la conclusion de Eilers et al (1984) qui ont trouvé la variabilité du seuil peu significative avec les stimuli pluri-syllabiques.

Nous nous sommes aperçu que les réponses erronées ne se distribuaient pas aléatoirement; en effet les erreurs consistent en une disconcordance dans la localisation du stimulus modifié et son repérage. Parmi les erreurs, 72% des cas montrent un décalage d'une syllabe à droite (vers la fin de la phrase) par rapport au stimulus-cible. Par ailleurs les discordances de plus de deux syllabes n' atteignent même pas un dixième des erreurs. Ainsi dans nos expériences, l'effet de combinaison de certaines syllabes n'a pas eu lieu contrairement à l'observation de Huggins (1971).

4. DISCUSSION

Le seuil différentiel de durée a été étudié par certains spécialistes avec les stimuli de la parole. Etant donné la différence de procédure expérimentale dans les travaux, la comparaison des résultats n'a qu'un sens relatif, toutefois à titre indicatif, nous pouvons retenir que le seuil va de quelque pour-cent à plus de 30% pour une durée comparable à la taille d'une syllabe. Pour une durée de 200 ms, le seuil le plus fin se trouve bien inférieur à 10% (Fujisaki et al, 1975; Bovet et Rossi, 1977) et le seuil le moins fin est de plus de 30% (Eilers et al, 1984). La majorité des travaux indiquent un seuil situé entre 10% et 20% (Huggins, 1971; Rossi, 1972;

Nooteboom, 1973; Klatt & Cooper, 1975, Bochner et al, 1988).

Notre estimation du seuil se trouve assez large et s'approche des résultats de Eilers et al. L'explication à la relative imprécision de nos résultats proviendrait de plusieurs facteurs. Le stimulus-cible étant noyé dans un élément de support long de 2,4 secondes environ, l'effet de masque doit être important comme il a été signalé par Klatt et Cooper (1975). D'autre part, la tâche infligée aux sujets était plus ou moins complexe pour un intervalle de 2 secondes: repérer l'endroit (Quelle syllabe?) et le sens de la modification (allongé ou raccourci) en même temps. Par ailleurs, le fait d'avoir réduit à son minimum le nombre des présentations des stimuli linguistiques afin d'éviter l'effet néfaste de la répétition (Warren, 1968), serait en partie responsable des résultats.

Dans l'expérience I, la limite inférieure des réponses *naturelle* se situe à 66,66% à cause des coefficients appliqués aux réponses traitées. Les valeurs calculées à ce niveau de la courbe psychométrique se trouvent entre 19% et 28% de modification temporelle (moyenne = 23.5%). Or le seuil estimé dans l'expérience II est d'environ 25.7%. Avec les phrases en français, nous avons observé la même tendance et des résultats similaires entre les deux expériences (Cf. **Tableau 1**, extrait de Nishinuma & Duez, 1989). Ce qui nous laisse imaginer que la tâche psycho-linguistique et la tâche psycho-acoustique semblent avoir été exécutées sur la même base mécanique différentielle de durée. Si l'acuité est plus grande dans l'Expérience I, elle devrait être de nature apprise et renforcée par l'usage intense de la langue.

	Exp I	Exp II
Phrase anglaise (Exp. actuelles)	23.5%	25.7%
Phrase française I	30.0%	25.3%
Phrase française II	22.8%	24.4%

Tableau 1
Estimation du seuil de durée syllabique

5. CONCLUSIONS

Nos deux expériences psycho-linguistique et psycho-acoustique ont permis de dégager les faits suivants:

1) Une variation temporelle de +/-25% par rapport à la durée d'origine pourrait être le seuil séparant le *naturel* du *non-naturel* de la phrase. Ce seuil correspond au seuil de durée différentiel défini par une tâche psycho-acoustique de l'Expérience II.

2) Dans la première expérience, la durée critique pour la syllabe accentuée s'est trouvée significativement réduite; la syllabe accentuée ne supporte pas de grandes variations temporelles en anglais, contrairement à ce qui se passe en français.

BIBLIOGRAPHIE

Bochner, J. H., Snell, K. B. & MacKenzie, D. J. "Duration discrimination of speech and tonal complex stimuli by normal hearing and hearing-impared listeners", JASA, 84(2), pp. 493-500, 1988

Bovet, P, & Rossi, M. "Etude comparée de la sensibilité différentielle à la durée avec un son pur et avec une voyelle", *Du temps biologique au temps psychologique*, PUF, Paris, pp.289-306, 1977

Duez, D. & Nishinuma, Y. "Influence de la vitesse d'articulation sur la durée des syllabes et des groupes consonantiques en français", Actes des 15e Journées d'Etudes sur la Parole, pp.97-99, 1986

Eilers, R. E., Bull, D. H., Oller, D. K. & Lewis, D. C. "The discrimination of vowel duration by infants", JASA, 75(4), pp.1213-1218, 1984

Fujisaki, H., Nakamura, K. & Imoto, T. "Auditory perception of duration of speech and non speech stimuli", in *Auditory Analysis and Perception of Speech* (eds. G. Fant & M. A. A. Tatham), Academic Press, London, pp.197-219, 1975

Huggins, A. W. F. "Just noticeable differences for segment duration in natural speech", JASA, 51(4), pp. 1270-1278, 1971

Huggins, A. W. F. "On the perception of temporal phenomena in speech", JASA, 51(4), pp. 1279-1290, 1971

Klatt, D. H. & Cooper, W. E. "Perception of segment duration in sentence contexts", in *Structure and Process in Speech Perception* (eds. A. Cohen & S. G. Nooteboom), Springer, New York, pp.69-89, 1975

Nishinuma, Y. & Duez, D. "Perceptual optimization of syllable duration in short French sentences", *Proc. Eurospeech 89* (Eds. J. P. Tubach & J. J. Mariani), 2, pp. 694-697, 1989

Nooteboom, S. G. "The perceptual reality of some prosodic durations", *Journal of Phonetics*, 1(1), pp.25-46, 1973

Rochester, S. "Detection and duration discrimination of noise increments", JASA, 49(2), pp. 1783-1794, 1971

Rossi, M. "Le seuil différentiel de durée", *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, (éd. A. Valdman), Mouton, La Haye, pp.435-450, 1972

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

REALISATIONS TONALES ET CONTRAINTES SEGMENTALES EN FANG

Jean-Marie HOMBERT

(LAPHOLIA, Université LYON 2 et LACITO-CNRS)

RÉSUMÉ

Les études phonétiques concernant l'évolution des systèmes tonals se sont surtout intéressées à l'influence des consonnes - généralement occlusives - sourdes et sonores sur la Fo des voyelles adjacentes [1, 2]. Nous présentons ici un autre type de conditionnement : l'influence du support segmental sur les réalisations tonales.

1. LE SYSTEME TONAL DU FANG

Le fang est une langue bantu parlée au Sud Cameroun et au Nord Gabon. Les variations dialectales à l'intérieur du domaine fang sont mal connues ; les données sur lesquelles repose l'analyse présentée ici proviennent du parler de Bitam (Nord Gabon)(1).

Comme le montrent les exemples des tableaux 1 et 2, la très grosse majorité des substantifs de cette langue sont mono- ou bisyllabiques ; dans ce dernier cas la première syllabe est un préfixe qui peut avoir la structure V, CV ou N(2) (nasale syllabique). Les exemples des tableaux 1 et 2 n'épuisent pas complètement l'inventaire des réalisations tonales de cette langue ; en effet on trouve aussi des mots tels que "la corde" ŋ̀kɔ̄ɔ́ [_ ⟋] ou "la noix de cola" àbɛ̀ɛ́ [_ ⟋] qui contiennent une mélodie montante. Il faut noter que ces tons montants ne se rencontrent que sur voyelles longues.

lune	ŋgʷàn	[⟍]
caïman	ŋgān	[—]
enfant	mʷân	[⟍]
rat (var.)	mvə́n	[‾]

Tableau 1 - Contrastes tonals sur monosyllabes en fang (les tons sont indiqués par les conventions habituelles : ` = ton bas, ¯ = ton moyen, ´ = ton haut, ^ = ton descendant)

banane plantain	èkʷàn	[— ⟍]
oiseau	ōnʷān	[— —]
honte	ɔ̀sʷân	[_ ⟍]
écureuil (var.)	ɔ̀sə́n	[_ ‾]

Tableau 2 - Contrastes tonals sur bisyllabes en fang

2. ORIGINE DE CES REALISATION TONALES

Il est clairement établi que la langue-mère (proto-langue) des langues bantu parlées actuellement était une langue à deux tons : haut et bas. Comment est-on passé d'une langue à deux tons ponctuels à une langue telle que le fang où les réalisations tonales sont à la fois plus variées et plus complexes ?

En proto-bantu, la plupart des substantifs avaient une structure $-C_1V_1C_2V_2$ précédée par un préfixe. Or, un certain nombre de langues du Nord-Ouest du domaine bantu (Cameroun, Gabon) n'ont pas préservé la voyelle finale V_2 ; toutefois le ton porté par V_2 n'a pas systématiquement été perdu lors de la chute de cette voyelle.

Puisque le proto-bantu avait deux tons (haut et bas), nous avions quatre possibilités de schèmes tonals sur les structures bisyllabiques $-C_1V_1C_2V_2$: bas-bas (BB), bas-haut (BH), haut-bas (HB) et haut-haut (HH). Le tableau 3, où nous avons rassemblé les exemples des tableaux 1 et 2 et leurs reconstructions en proto-bantu, indique clairement que *BB > B et *HH > H(3), ce qui n'est guère surprenant. Les deux autres correspondances sont p!us intéressantes : *BH > M et *HB > Descendant.

	Fang	Proto-bantu
lune	ŋgʷàn	*gɔ̀ndè
caïman	ŋgān	*gàndʊ́
enfant	mʷân	*yánà
rat (var.)	mvə́n	*bɛ́ndɛ́
banane plantain	èkʷàn	*kɔ̀ndè
oiseau	ōnʷān	*nɔ̀nɪ́
honte	ɔ̀sʷân	*cɔ́nɪ̀
écureuil (var.)	ɔ̀sə́n	*cɪ́ndɪ́

Tableau 3 - Correspondances tonales Fang - Proto-bantu

Pour comprendre l'origine des réalisations tonales montantes, il suffit de se réputer aux reconstructions proto-bantu correspondantes pour constater que ces formes proviennent toutes de formes avec d en position C_2 et un schème tonal *BH (exemples :'corde' *gɔ̀dɪ́ > -kɔ̄ɔ́, 'noix de cola' *bɪ̀dʊ́ > -bɛ̀ɛ́).

Un examen plus complet des correspondances tonales fang - proto-bantu fait apparaître des exceptions aux règles présentées ci-dessus ; ainsi 'froid' *pɛ́pɔ̀ > -və́p et non pas və̂p et 'hameçon' *dóbò > lɔ́p et non pas lɔ̂p. Toutes ces "exceptions" ont une structure CVC où la consonne finale est sourde.

Si nous reclassons l'ensemble de nos correspondances tonales fang - proto-bantu en prenant en compte la structure syllabique (CVC (où la consonne finale est sourde), CVN (N = consonne nasale) et CVV) comme nous l'avons fait dans le corpus présenté en annexe, nous aboutissons au tableau récapitulatif suivant :

	*B B [_ _]	*B H [_ ¯]
CVC	1. [—]	4. [—]
CVN	2. [⁀]	5. [—]
CVV	3. [⁀]	6. [⁀]
	*H B [¯ _]	*H H [¯ ¯]
CVC	7. [¯]	10. [¯]
CVN	8. [⁀]	11. [¯]
CVV	9. [⁀]	12. [¯]

Tableau 4 - Tableau récapitulatif des correspondances tonales proto-bantu - fang en fonction de la structure syllabique fang.

3. CONSIDERATIONS PHONETIQUES

Trois considérations phonétiques se dégagent du tableau 4.

a. Lors de l'élision de la voyelle finale V_2, le ton de cette voyelle a été reporté sur la voyelle précédente et a donné naissance à un ton modulé lorsque :
   - le ton porté par V_1 était différent de celui porté par V_2
   - le matériel segmental portant le ton résultant était de durée suffisamment longue.

b. Les consonnes nasales finales permettent la réalisation d'un ton modulé descendant (provenant d'une séquence *HB) (Tableau 4,case 8), mais elles ne permettent pas la réalisation d'une séquence BH (case 5), suggérant ainsi une plus grande difficulté (et par conséquent nécessitant une plus grande durée) à produire une modulation montante qu'une modulation descendante. Cette constatation est en accord avec des résultats expérimentaux sur les vitesses de variation de Fo (modulations ascendante et descendante) [4, 5, 6].

c. Lorsque le ton modulé n'a pas assez "de temps" - faute de support segmental voisé suffisamment long - (cases 4 et 5 pour les *BH et case 7 pour *HB), il se réalise comme un ton ponctuel dont la valeur de Fo est proche de la valeur de la Fo de $*V_1$.

4. CONCLUSION

Ce travail rend compte des correspondances (et par conséquent de l'évolution) tonales entre le proto-bantu et le fang. Ces correspondances font apparaître que des contraintes d'origine segmentale (durée du matériel segmental voisé) peuvent influer sur l'évolution des réalisations tonales. Elles mettent également en évidence une plus grande difficulté à réaliser (et peut-être à percevoir) une modulation montante de Fo par opposition à une modulation descendante.

BIBLIOGRAPHIE

[1] Hombert J.M. (1978) "Consonant types, vowel quality and tone", in V. Fromkin (ed.) *Tone : a Linguistic Survey*, Academic Press, pp. 77-111.

[2] Hombert J.M., J.J. Ohala et W. Ewan (1978) "Phonetic explanations for the development of tones", *Language* 55, (1), pp. 37-58.

[3] Hombert J.M., P. Medjo et R. Nguema (1989) "Les Fang sont-ils bantu ?", *Pholia* 4, pp. 133-147.

[4] Sundberg J. (1973) "Data on maximum speed of pitch changes", *STL-QPSR* 4, pp. 39-47.

[5] Ohala J.J. et W. Ewan (1973) "Speed of pitch changes", *Journal of the Acoustical Society of America* 53, 345.

[6] Hombert J.M. (1977) "Difficulty of producing different Fo in speech", *UCLA Working Papers in Phonetics* 36, pp. 12-20.

---

1 Je remercie Monsieur Pither MEDJO pour sa collaboration dans ce travail ; pour une présentation plus complète de la phonologie de cette langue, voir [3].

2 On remarquera que dans notre corpus certaines nasales en position initiale sont syllabiques et par conséquent portent un ton alors que d'autres (voir tableau 1) ne le sont pas ; cette différence peut être expliquée par des considérations morphologiques (appartenance à des classes nominales différentes).

3 Nous utilisons ici les conventions habituelles de la linguistique comparative : * représente une forme reconstruite et > indique la transformation de cette forme ancienne.

ANNEXE

GLOSE	FANG (Bitam)	PB
1. peau	(ɛ̀)-kɔ̀p	gùbɔ̀
araignée	(à)-bɔ̀p	bùbɪ̀
molaire	(ɛ̀)-kɔ̀'	gɛ̀gɔ̀
vent	(m̀)-fɔ̀p	pɛ̀ɛ̀pɛ̀
2. abdomen	(à)-bùm	bɥ̀mɔ̀
pancarte	ndɔ̀m	dɪ̀mbɔ̀
saleté	mvɪ̀n	bɪ̀ndù
guérisseur	ŋgā̀ŋ	gàŋgà
3. patte	(à)-kùù	gɥ̀dù
amertume	(à)-yɔ̀ɔ̀	dùdù
sitatunga	mvùù	bɥ̀dɪ̀
force	ŋgùù	gɥ̀dɥ̀
4. veine	(n̄)-sīs	cɪ̀cá
os	(ē)-vēs	pɪ̀cí
aile	(ā)-fāp	pàpá
écureuil (var.)	mvō'	bùgé
5. pigeon (var.)	(ō)-bōŋ	bɪ̀ŋgá
dent	(ā)-sōŋ	cùŋgá
oiseau	(ō)-n'ān	nɔ̀nɪ̄
racine	(ŋ̄)-kā̄ŋ	kàŋgá
6. corde	(ŋ̀)-kɔ̀ɔ́	gɔ̀dɪ́
noix de cola	(à)-bɛ̀é	bɪ̀dú
7. froid	(à)-vɔ́p	pépɔ̀
hameçon	(ǹ)-lɔ́p	dɔ́bɔ̀
ficelle	(ǹ)-dzɪ̄'	dɪ́gɪ̀
sac	(m̀)-fɔ́'	pɥ́kɔ̀
8. mari	(ǹ)-nôm	dúmɪ̀
coeur	(ǹ)-nɔ̂m	tímà
piège	(ɔ̀)-lâm	támbɔ̀
langue	(ɔ̀)-yɔ̂m	dímɪ̀
9. mousse	(à)-vúù	pɥ́dɪ̀
épaule	(ɛ̀)-túù	tɥ́dɪ̀
tortue	kúù	kɥ́dɥ̀
10. pagaie	(ɛ̀)-ŋgáp	gápɪ́
bosse	(ɛ̀)-tút	tɥ́ɥ́tɥ́
tique	kɔ́p	kúpá
animal	tsɪ́t	tɪ́ɪ́tú
11. genou	(à)-bɔ́ŋ	bɔ́ŋgɔ́
écureuil (var.)	(ɔ̀)-sɔ́n	cíndí
palmier	(à)-lɔ́n	téndé
kaolin	fɔ́m	pémbé
12. plaie	fɔ́ɔ́	pútɛ́

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

TONS ET "CREAKY VOICE" EN CHINOIS STANDARD

Michel Grenié, Agnès Belotel-Grenié

Institut de Phonétique, Laboratoire Parole & Langage URA CNRS 261
29 av. Schuman, 13 621 Aix-en-Provence Cedex, France

RESUME
Cette étude porte sur les types de phonation mis en oeuvre en chinois par trois locuteurs masculins, lors de la prononciation de mots isolés contenus dans des phrases porteuses, en fonction des tons qui affectent chacun de ces mots. Les tons dynamiques 3 (niveaux 2-1-4) et 4 (niveaux 5-1) sont souvent associés à un mode particulier de phonation qui correspond à de la "creaky voice". D'après les analyses acoustiques réalisées à l'aide d'un logiciel de traitement numérique du signal, la "creaky voice" est caractérisée en chinois par la modification du cycle glottique au début de la voyelle, par la présence de bruit sur la structure harmonique, par une chute de l'intensité de l'harmonique 2 au niveau du milieu de la voyelle et, dans certains cas, par une disparition complète du voisement. Un certain nombre d'observations conduisent à supposer que la "creaky voice" est utilisée en chinois à des fins linguistiques comme indice des tons 3 et 4 par certains locuteurs.

1. INTRODUCTION

Le chinois standard ou *putonghua* présente quatre tons : le ton 1 qui est statique et se décrit en terme de niveaux par les valeurs 5-5 (ces valeurs correspondent à la représentation des tons sur l'échelle de Chao (1968), représentation schématique qui n'est pas très éloignée de la courbe de fréquence fondamentale), le ton 2 qui est montant (niveaux 3-5), le ton 3 qui est descendant-montant (niveaux 2-1-4), le ton 4 qui est descendant (niveaux 5-1). Le *putonghua* possede un systeme vocalique complexe comprenant à la fois des voyelles simples et des voyelles complexes (diphtongues et triphtongues). Beaucoup de locuteurs chinois n'ont pas toujours une conscience réfléchie des tons qu'ils prononcent pour communiquer. Pour s'en convaincre, il suffit de constater que la question d'un apprenant étranger sur le ton associé à tel ou tel mot se termine généralement par un désaccord entre les interlocuteurs auxquels il s'adresse. Pour les locuteurs chinois, la différence entre deux tons n'est pas secondaire par rapport à la différence entre deux phonèmes. C'est de l'association indissoluble du ton et des phonemes que naît le mot. Voilà pourquoi il nous paraît intéressant de montrer dans le cadre de cette étude qu'il existe au niveau de la phonation des indices segmentaux de la réalisation des tons.

2. PROCEDURE EXPERIMENTALE

Constitution du corpus

Le corpus a été enregistré en chambre sourde par trois locuteurs masculins issus de provinces différentes de l'est de la Chine (Shanxi, Henan, Dongbei). Les trois locuteurs ont un niveau d'études supérieures, tous parlent et lisent couramment le *putonghua* (chinois standard). Lors de l'enregistrement, le corpus est présenté aux locuteurs sous forme de fiches comprenant chacune une phrase écrite en chinois à l'aide d'un traitement de texte. Un premier corpus comprenant 121 mots monosyllabiques a été prononcé par le locuteur 1. Ce corpus a été complété par l'enregistrement de deux autres locuteurs avec une liste plus étendue de 179 monosyllabes. Ces mots sont inclus dans une phrase porteuse analogue à celle utilisée par Howie (1976) ("Le mot ..., c'est Mr Li qui l'a écrit"). Les monosyllabes ont été choisis dans le *Xinhua Zidian* de manière à obtenir le plus possible d'unités segmentales aux quatre tons.

Méthodes d'analyse

Après filtrage, les énoncés ont été numérisés sur MASSCOMP avec une fréquence d'échantillonnage de 10 kHz puis segmentés en utilisant l'éditeur de signal VES (logiciel de Visualisation d'Edition de Signal développé par R. Espesser). Pour chacun des mots correctement lus par les locuteurs, l'oscillogramme, le spectrogramme, les valeurs de F0 et de l'intensité sur tout le mot ainsi que les spectres FFT pris à partir du début et du milieu de la voyelle ont été analysés en vue de l'étude des modes phonatoires associés aux tons.

3. MISE EN EVIDENCE DE LA "CREAKY VOICE"

Chacun des trois locuteurs utilise deux types de phonation qui se distinguent aisément sur les spectrogrammes. La figure 1 donne l'exemple de la syllabe [ja] prononcée aux quatre tons par le locuteur 2. Les réalisations des tons 1 et 2 correspondent à une phonation normale tandis que pour les tons 3 et 4 de la "creaky voice" est nettement visible (nous

preferons conserver le terme "creaky voice car la traduction de ce terme en français sous la forme de "voix friturée", "voix grinçante" ou "voix craquée" peuvent introduire des confusions. Ce type de phonation est carac-térisé par une diminution rapide et irrégulière de la périodicité des impulsions glottiques. Ce phénomène s'accentue au fur et à mesure que l'on se rapproche du centre de la voyelle. La chute de la fréquence fondamentale conduit à l'interruption complète et momentanée de la production de voisement. D'autre part, comme l'illustrent les raies verticales large bande visibles au début des voyelles des réalisations avec tons 3 et 4, le cycle de vibration des cordes vocales est modifié lors de ce type de phonation. Enfin, le rapport d'énergie entre la zone de basse fréquence qui correspond au fondamental et aux premiers harmoniques avec l'énergie de la bande du formant 1 fournit un indice supplémentaire de distinction de ce type de phonation puisque pour les réalisations associées au ton 3 et au ton 4 l'énergie de F1 (formant 1) est très largement supérieure à l'énergie basse fréquence, ce qui n'est pas le cas pour les deux autres tons. Selon Ladefoged (1975) :

> *"When producting creaky voice, the vocal cords are much more tensed, closing rapidly during each glottal cycle. As a result, the vocal tract is excited by a sharper pulse that has more energy in the higher harmonics."*

Lors de la production de la "creaky voice" les cordes vocales sont plus tendues et se ferment plus rapidement pendant chaque cycle glottique. Cela a pour conséquence de produire des ondes glottiques plus abruptes, ce qui explique la présence d'énergie dans les hautes fréquences pour chaque impulsion glottique. D'après Catford (1988), la "creaky voice" est produite avec la glotte complètement fermée sur la quasi-totalité de la longueur des cordes vocales, exceptée une petite portion qui vibre et qui est située sur la partie postérieure des cordes vocales.

La figure 2 illustre les modifications produite par la "creaky voice" dans les structures harmoniques présentes au centre des voyelles [a] des monosyllabes [ja]. Pour les tons 1 et 2, les harmoniques apparaissent très nettement dans la bande 0-2000 Hz. H_1 et H_2 ont une intensité égale à celle de l'harmonique le plus intense dans la zone du formant 1. En revanche, lors des réalisations avec "creaky voice", la structure des harmoniques est nettement moins visible. L'intensité de H_2 est alors largement inférieure, d'une part, à celle de l'harmonique le plus intense qui correspond au formant 1 (environ -10 dB) et, d'autre part, à celle de H_1. Du bruit apparaît également entre les harmoniques. L'analyse de toutes les occurrences de "creaky voice" du corpus confirme ces observations.

4. DISTRIBUTION STATISTIQUE DE LA "CREAKY VOICE"

Le tableau 1 présente la distribution statistique des occurrences de la "creaky voice" selon les locuteurs et les tons. Pour les trois locuteurs la "creaky voice" n'apparaît jamais au ton 1. Au ton 2, elle se produit pour un très petit nombre d'occurences pour un seul locuteur. Par contre, au ton 3 elle est très fréquente puisqu'elle affecte les réalisations des tons3 des trois locuteurs dans les proportions de 58%, 81% et 48%. La présence de la "creaky voice" est également associée, dans une moindre mesure, au 4ème ton, puisque les proportions des voyelles produites en "creaky voice" au 4ème ton sont seulement de 19,6% pour le locuteur 3 et de 21% pour le locuteur 2. Aucune "creaky voice" n'est relevée à ce ton pour le locuteur 1. Ce tableau montre qu'outre la caractéristique du ton, l'identité du locuteur est à prendre en compte pour expliquer l'apparition de la "creaky voice". Cette variabilité entre locuteurs semble toutefois cohérente dans la mesure où elle paraît correspondre à 3 degrés d'utilisation de la "creaky voice". C'est d'abord le ton 3 qui est le plus affecté, puis le ton 4 et enfin le ton 2.

TON	TON 1		TON2		TON 3		TON 4	
	CV	SANS	CV	SANS	CV	SANS	CV	SANS
LOCUTEUR 1	0	28	0	22	12	13	0	31
LOCUTEUR 2	0	56	6	33	39	9	12	44
LOCUTEUR 3	0	57	0	38	32	23	13	53

TABLEAU 1 : Répartition des occurrences avec et sans "creaky voice" selon les tons et les locuteurs.

L'examen des distributions des occurrences de "creaky voice" aux tons 3 et 4 selon les voyelles montre que la voyelle [i] est toujours prononcée avec une phonation normale tandis que la voyelle [a] et les voyelles complexes qui en sont proches entraînent fréquemment l'apparition de "creaky voice" (figures 3 & 4). La taille du corpus analysé ne permet pas de préciser la pertinence statistique de cette tendance.

5. ORGANISATION TEMPORELLE ET "CREAKY VOICE"

L'observation des oscillogrammes et des spectrogrammes montre que l'apparition de la "creaky voice" est liée à une organisation temporelle particulière qui apparaît sur toute la durée de la voyelle et qui peut se découper en quatre phases (figure 5). La voyelle débute toujours par la présence de vibrations, nettement visibles, des cordes vocales. Ces vibrations se manifestent sur les spectrogrammes sous la forme de raies verticales allant de 0 à plus de 3500 Hz. Très brutalement cette périodicité diminue et devient irrégulière. Il s'agit de la seconde phase pendant laquelle le spectre produit par les cordes vocales se modifie et les fréquences aigues visibles à l'initiale s'atténuent. L'arrêt complet du voisement qui

suit cette baisse de périodicité constitue la troisième phase. Le signal est alors beaucoup moins intense et il y a production de bruit. Enfin, des périodicités peu intenses et peu riches en harmoniques réapparaissent parfois en finale absolue de la voyelle au troisième ton.

La durée pendant laquelle le voisement est présent à l'initiale de la voyelle peut être très brève (proche de 50 ms) par comparaison à la durée des voyelles réalisées avec une phonation normale. Dans les cas où les durées des phases 1 et 2 sont courtes, le mouvement des articulateurs semble se poursuivre malgré l'absence de voisement puisque des pôles de bruit sont visibles sur les spectrogrammes dans la continuité des transitions formantiques. La durée totale de la voyelle est alors celle d'une voyelle normale (figure 5).

Du point de vue acoustique, le passage de la voix normale à la "creaky voice" paraît progressif et continu. Plusieurs étapes peuvent être distinguées : 1°/ voix normale ; 2°/ présence de bruit sur la structure harmonique ; 3°/ modification du cycle glottique ; 4°/ chute de l'intensité de l'harmonique 2 avec maintien du voisement ; 5°/ disparition complète du voisement avec présence d'un bruit de souffle.

Toutes les occurrences de "creaky voice" atteignent des valeurs minimales de F0 (voisines de 90 Hz pour nos trois locuteurs). Il semble qu'en dessous de ce seuil, les locuteurs ne puissent plus contrôler précisément les battements des cordes vocales. Cependant, à l'initiale des voyelles produites en "creaky voice" réalisées sur les tons 3 et 4, la F0 dépasse souvent 120 Hz alors que des indices d'une modification de l'onde glottique sont déjà présents sous la forme de raies verticales. Ce n'est donc pas la valeur instantanée de F0 qui peut expliquer l'apparition de la "creaky voice". La "creaky voice" paraît être programmée dès le début de la voyelle à un haut niveau d'encodage. Cela nous conduit à supposer qu'en chinois la "creaky voice" pourrait être, pour certains locuteurs masculins, un indice redondant des tons descendants. Le statut linguistique de ce phénomène reste à confirmer par l'examen de la production d'un plus grand nombre de locuteurs masculins et féminins.

Dans la mesure où l'établissement de voisement dépend très étroitement de la pression sous-glottique et du contrôle des muscles du larynx, il est permis de se demander si l'apparition de la "creaky voice" résulte d'une baisse de la pression sous-glottique ou bien d'une configuration particulière du larynx. La figure 6 montre que la diminution de la périodicité de la F0 et l'apparition d'ondes glottiques irrégulières ne correspondent pas nécessairement à une baisse significative de l'intensité de la voyelle. Le maintien d'un niveau élevé d'intensité tout au long de la voyelle suggère, en dépit de la présence de vibrations irrégulières des cordes vocales, la continuation d'un effort et d'un contrôle articulatoire. Cela est d'ailleurs illustré par la reprise du voisement à la finale de certaines voyelles craquées. La "creaky voice" serait donc principalement dûe à l'état du larynx.

6. CONCLUSION

La "creaky voice" paraît souvent être associée en chinois standard aux réalisations des tons 3 et 4. Elle résulte principalement de la modification du fonctionnement de la source vocale. Elle n'interrompt pas le mouvement des articulateurs dans les résonateurs. Plusieurs observations suggèrent que les variations des contours de F0 et d'intensité ne suffisent pas à expliquer son apparition et qu'elle est programmée à un niveau d'encodage élevé. Une certaine variabilité s'observe entre les locuteurs quant aux mots affectés par ce phénomène et à son importance statistique. Il reste à déterminer si ce type de phonation s'observe de la même manière pour les femmes et s'il se manifeste d'une façon analogue en parole spontanée.

BIBLIOGRAPHIE

CATFORD J.C. (1988) A practical introduction to phonetics. Clarendon press, Oxford.

CHAO Y.R. (1968) A grammar of spoken chinese, Berkeley.

FANG K. (1980) Laryngeal features and tone development, Bulletin of the institute of history and philology, academia sinica, vol 51, p. 1-13.

GARDING E. (1987) Speech act and tonal pattern in Standard Chinese : constancy and variation, Phonetica, 44, p. 13-29.

GARDING E. & al. (1986) Tone 4 and tone 3 discrimination in modern Standard Chinese, Language and speech, vol. 29, p. 281-293.

HOLLIEN H., MICHEL J.F. (1968) Vocal fry as phonological register, J.S.H.R., vol. 11 n°3, p. 600-604.

HOMBERT J.M. (1984) Phonétique expérimentale et diachronie, application à la tonogénèse, Thèse d'état, Université de Provence.

HOWIE J.M. (1976) Acoustical studies of mandarin vowels and tones, Cambridge University Press.

KRATOCHVIL P. (1987) The case of the third tone, in Wang Li memorial volume, Hong Kong.

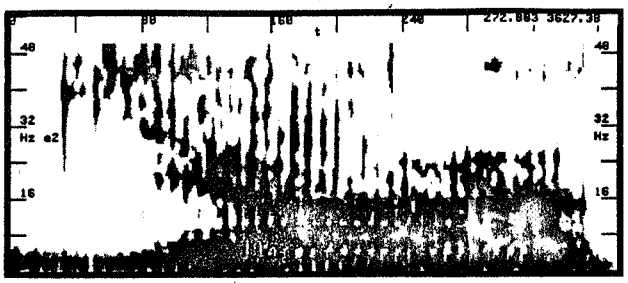
LADEFOGED P. (1975) A course in phonetics, Harcourt Brace Jovanovich, Inc., New York.

LADEFOGED P., MADDIESON I., JACKSON M.(1988) Investigating phonation types in different languages, Vocal folds physiology, voice production, mecanisms and functions ed. Fujimura O., Ravers Press.

SAGART L., HALLE P. (1984) Esquisse de la phonologie d'un dialecte du Jiangxi : Nancheng, C.L.A.O., vol. XIII, n°2, p. 191-215.

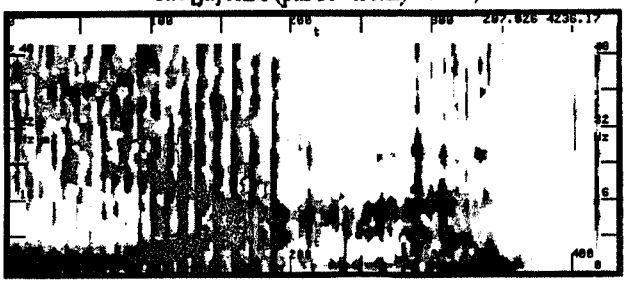
WU Zongji (1980) "Shilun Putonghua de qubie tezheng jiqi xiangdang guanxi" (traits distinctifs et leur corrélation dans la phonologie du Putonghua) Zhonggua Yuwen vol.5, 321-325.

ZEE E. (1978) The interaction of tone and vowel quality,UCLA working papers in phonetics, 41, p. 53-67.



1a : [ja] ton 1 (pas de "creaky voice")

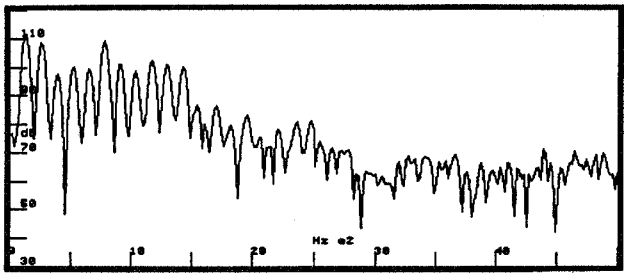
1b : [ja] ton 2 (pas de "creaky voice")



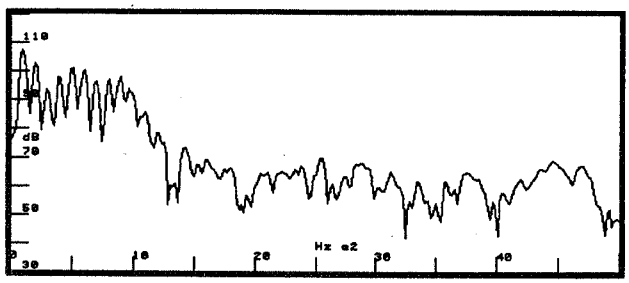
1c : [ja] ton 3 ("creaky voice")

1d : [ja] ton 4 ("creaky voice")

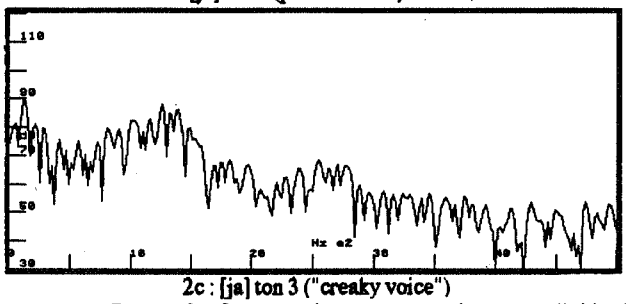
FIGURE 1 . Spectrogrammes numériques obtenus avec un filtre large bande de 350 Hz sur la bande 0-5000 Hz pour les monosyllabes [ja] prononcés au quatre tons par le locuteur 2. La "creaky voice" est visible pour les documents 1c et 1d qui correspondent aux tons 3 et 4.



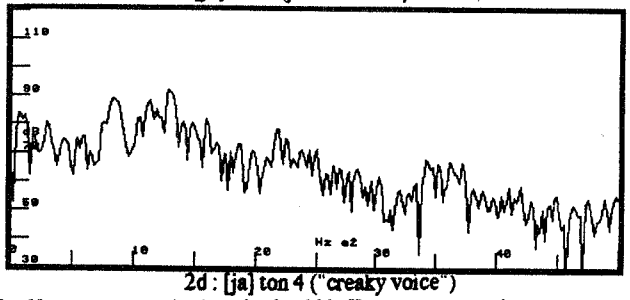
2a : [ja] ton 1 (pas de "creaky voice")



2b : [ja] ton 2 (pas de "creaky voice")



2c : [ja] ton 3 ("creaky voice")



2d : [ja] ton 4 ("creaky voice")

FIGURE 2 . Spectres intantannés obtenus à l'aide de FFT effectuées sur la bande 0-5000 Hz au centre des voyelles pour un filtre d'une largeur de 80 Hz. La "creaky voice" est visible sous la forme d'une baisse de l'énergie H2 par rapport à H1 pour les documents 2c et 2d qui correspondent aux tons 3 et 4.

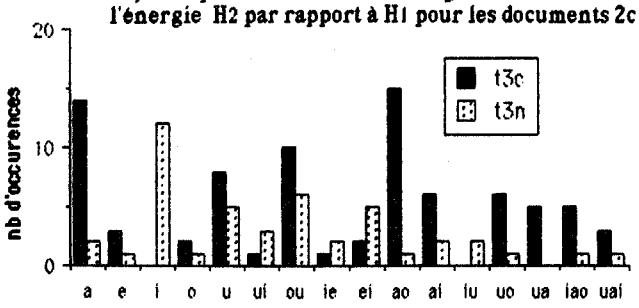


FIGURE 3 : Répartition des occurrences avec "creaky voice" (en noir) et sans "creaky voice" (en blanc) selon les voyelles notées en APC pour les trois locuteurs au ton 3.

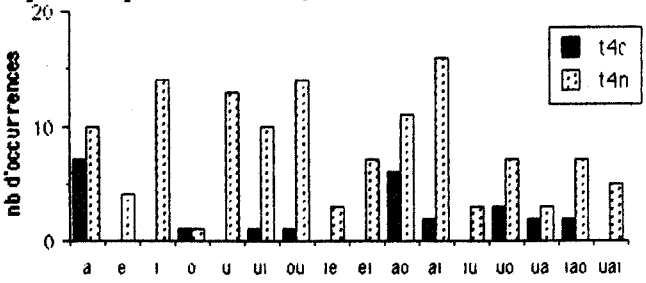
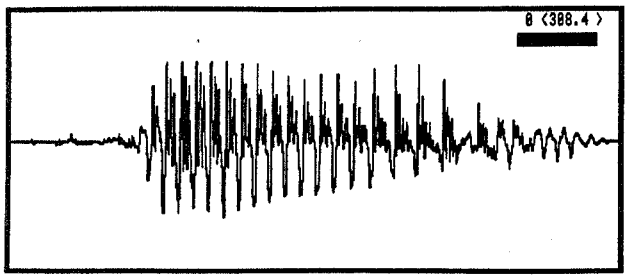
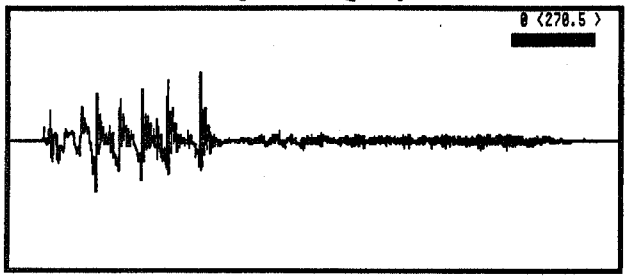


FIGURE 4 : Répartition des occurrences avec "creaky voice" (en noir) et sans "creaky voice" (en blanc) selon les voyelles notées en APC pour les trois locuteurs au ton 4.

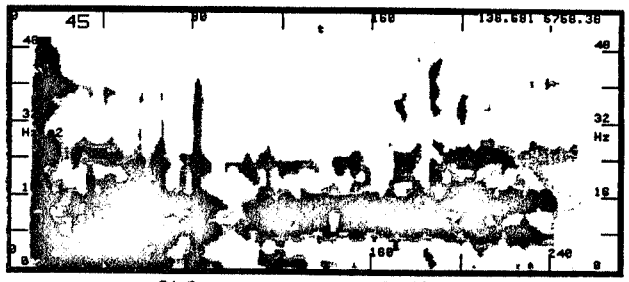


5 a 1 : oscillogramme de [pao 4] locuteur 2

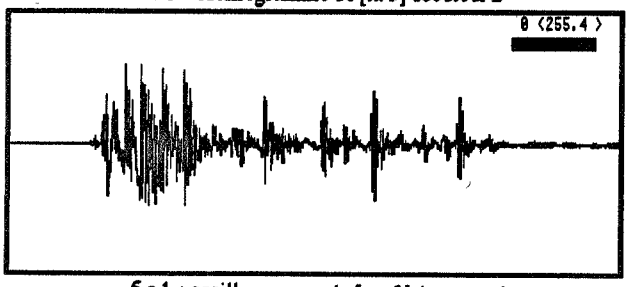
5 a 2 : spectrogramme de [pao 4] locuteur 2



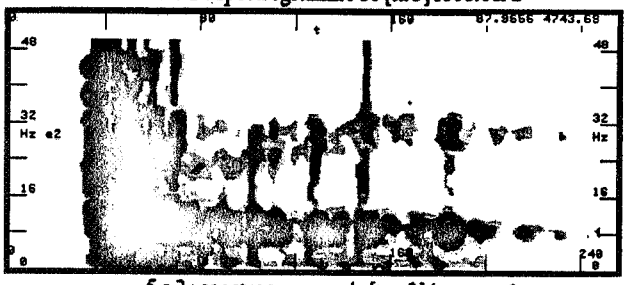
5 b 1 : oscillogramme de [ta 3] locuteur 2



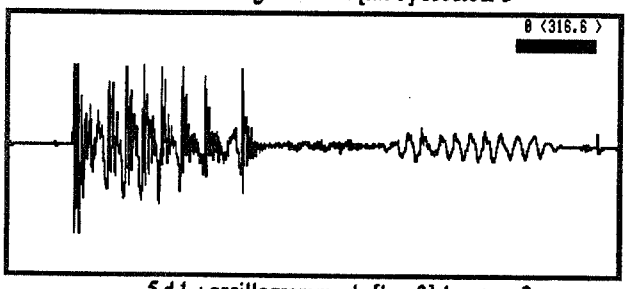
5 b 2 : spectrogramme de [ta 3] locuteur 2



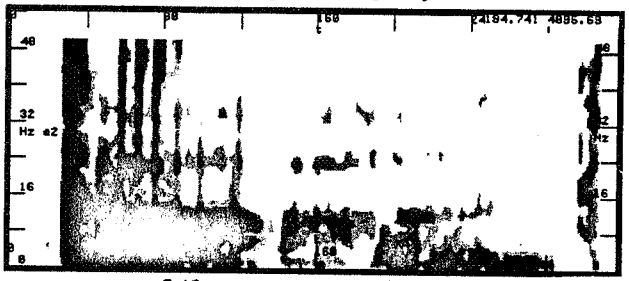
5 c 1 : oscillogramme de [tao 3] locuteur 3



5 c 2 : spectrogramme de [tao 3] locuteur 3

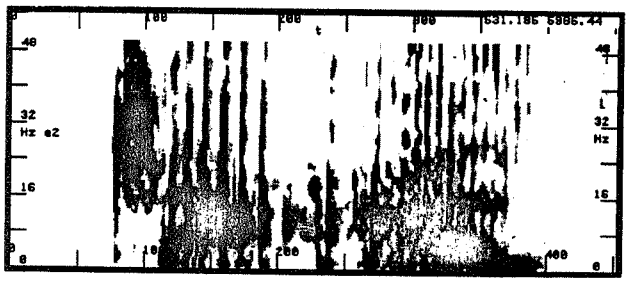


5 d 1 : oscillogramme de [kao 3] locuteur 2

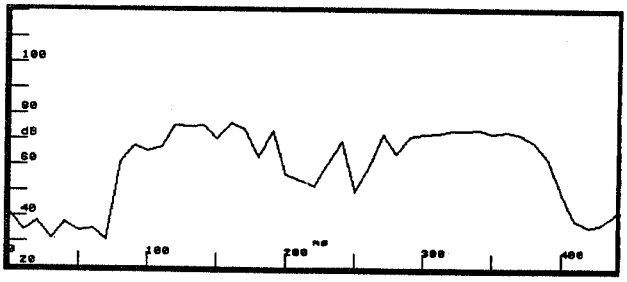


5 d 2 : spectrogramme de [kao 3] locuteur 2

FIGURE 5 . Oscillogrammes et spectrogrammes numériques obtenus avec un filtre large bande de 350 Hz sur la bande 0-5000 Hz pour quatre mots prononcés par les locuteurs Hu et Songjun aux tons 3 et 4. Ces documents illustrent que la production d'une voyelle avec "creaky voice" correspond à l'enchaînement temporel de plusieurs phasesdont la présence et la durée peuvent varier.



6 1 : spectrogramme de [tsa 3] locuteur 2



6 2 : courbe d'intensité de [tsa 3] locuteur 2

FIGURE 6 : Spectrogramme numérique et courbe d'intensité du mot [tsa 3] prononcé par le locuteur 2. L'intensité des impulsions glottales varie peu sur toute la durée de la voyelle.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

M.A.V.L. / V.O.T. ? PROPOSITIONS POUR UN CLASSEMENT PHONETIQUE EN TERMES DE MOMENTS D'APPARITION DES VIBRATIONS LARYNGIENNES DES OCCLUSIVES FRANÇAISES ET QUEBECOISES

Jean-Pierre GOUDAILLIER

Margaret BENTO

Laboratoire de Phonétique
U.F.R. de Linguistique Générale et Appliquée
Université René Descartes, Paris, France

Le V.O.T. ne peut pas à lui seul rendre compte de la grande variété des matérialisations phonétiques des occlusives françaises et québécoises, d'autant plus si l'on prend en considération l'assibilation des dentales.

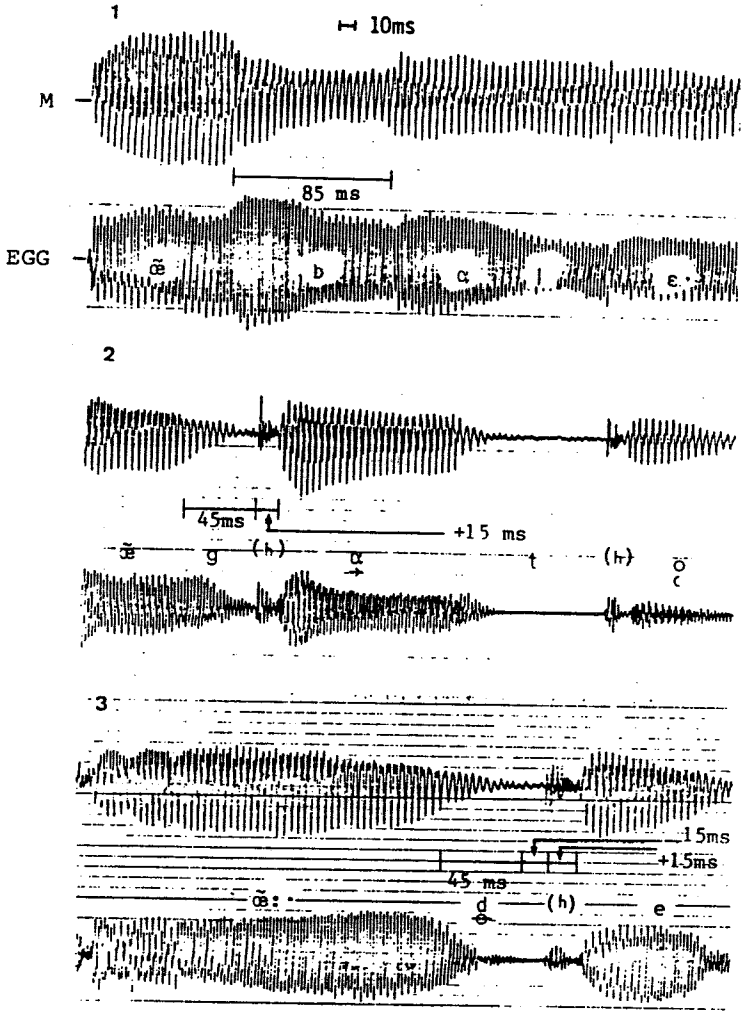
A l'aide du concept de M.A.V.L. (Moments d'Apparition des Vibrations Laryngiennes) peuvent être mis en valeur divers types de réalisations, qui sont au nombre de 4 pour les occlusives phonologiquement 'sonores' et au nombre de 2 pour leurs correspondantes 'sourdes'. Ces types sont respectivement au nombre de 8 et de 3, si l'on tient compte des assibilées.

Depuis un peu plus de 25 ans, la notion de V.O.T. (Voice Onset Time) proposée par Leigh Lisker et Arthur A. Abramson est utilisée tant d'un point de vue phonétique que dans le cadre de la théorie linguistique, plus précisément phonologique (cf., entre autres, l'emploi qui en est fait dans *S.P.E.* de Noam Chomsky et Morris Halle). Au-delà de son utilité même, ce concept trouve cependant ses limites, lorsqu'il s'agit de décrire l'ensemble des réalisations possibles pour les articulations de type occlusif. La distinction entre *VOT+* et *VOT-*, même si elle peut être "ajustée" en *short voicing lead* et *long voicing lead* d'une part et en *short voicing lag* et *long voicing lag* d'autre part, ne peut pas à notre avis englober les divers cas de figure que l'on peut rencontrer, ne serait-ce que pour les occlusives /p/,/t/,/k/ et/b/,/d/,/g/ en français et en québécois. Une description complète de ces consonnes nécessite de tenir compte des cas de devoisement partiel et/ou total pouvant affecter les occlusives phonologiquement 'sonores' tant en français qu'en québécois et de ceux d'assibilation propres aux occlusives dentales 'sourdes' et 'sonores' du québécois, ceci tant d'un point de vue phonologique (Goudaillier, 1986a) que d'un point de vue phonétique. A cet effet, non seulement doit être utilisé le concept de V.O.T. mais aussi celui de M.A.V.L. (Moments d'Apparition des Vibrations Laryngiennes) (Goudaillier, 1986b). Une illustration en est ici-même fournie en ce qui concerne l'analyse des consonnes occlusives d'enfants québécois, pour lesquelles le deuxième de ces concepts s'est révélé être plus performant dans une perspective typologique.

Comment déterminer les M.A.V.L. d'une consonne occlusive ? Une illustration est donnée aux Planches 1 et 2, qui comportent les tracés du phonogramme (ligne M) et de l'éléctroglottogramme (EGG) de 6 séquences prononcées par des enfants québécois. Le [b] de [ɛ̃bɑlɛ·] (Fig.1) est entièrement voisé; ceci veut dire que sa phase d'occlusion et celle de relâchement sont toutes les deux accompganées de vibrations des cordes vocales. La durée de l'ensemble est de *85ms* et aucune interruption des vibrations laryngiennes n'a lieu lors du passage de la consonne [b] à la voyelle [ɑ]. C'est un *type 1* de M.A.V.L.. Le [g(h)] de [ɛ̃g(h)ɑt(h)o](Fig.2) et la consonne dentale [d(h)] de la Fig. 3 : [ɛ̃d(h)e], quant à eux, ne comportent pas de voisement pendant leurs phases de relâchement : l'explosion n'est pas voisée et le bruit de friction suivant cette dernière ne l'est pas non plus. Un V.O.T. positif de *+15ms* est observé dans les deux cas. Pour [g(h)], qui est un *type 2* de M.A.V.L., l'occlusion est entièrement sonore. Elle dure *45ms*. L'occlusion du [d(h)] n'est voisée que partiellement : seuls *45ms* de celle-ci, qui dure *60ms*, sont voisés (*75%*); la fin de cette occlusion est donc sourde pendant *15ms* (Fig. 3). Il s'agit d'un *type 3* de M.A.V.L.. Dans [b(h)ozi](Fig. 4) le [b(h)] a une occlusion sourde. Cette articulation compte par ailleurs un V.O.T. positif de *+5ms*. On est en présence ici d'un *type 4*.

Pour ce qui est des consonnes phonologiquement 'sonores', les quatre types de M.A.V.L. sont récapitulés comme suit : M.A.V.L. 1 (occlusion voisée + relâchement voisé; [b],[d],[g]); M.A.V.L. 2 (occlusion voisée + relâchement non voisé; [b(h)], [d(h)] [g(h)]); M.A.V.L. 3 (occlusion mi-sonore + relâchement non voisé; [b(h)] [d(h)] [g(h)]); M.A.V.L. 4 (occlusion non voisée + relâchement non voisé; [b(h)] [d(h)] [g(h)]).

Pour le [p(h)] de [ɛ̃p(h)anje·] (Fig.5) et le [pʰ] de [pʰo·s] (Fig.6), les durées d'occlusion sont de *80ms* et de *115ms*. Si le V.O.T. positif est inférieur à 40ms, la consonne a un *type 5* de M.A.V.L.. Ceci est le cas du [p(h)] de [p(h)anje·], puisque son V.O.T. ne dure que *+20ms*; si le V.O.T. po-



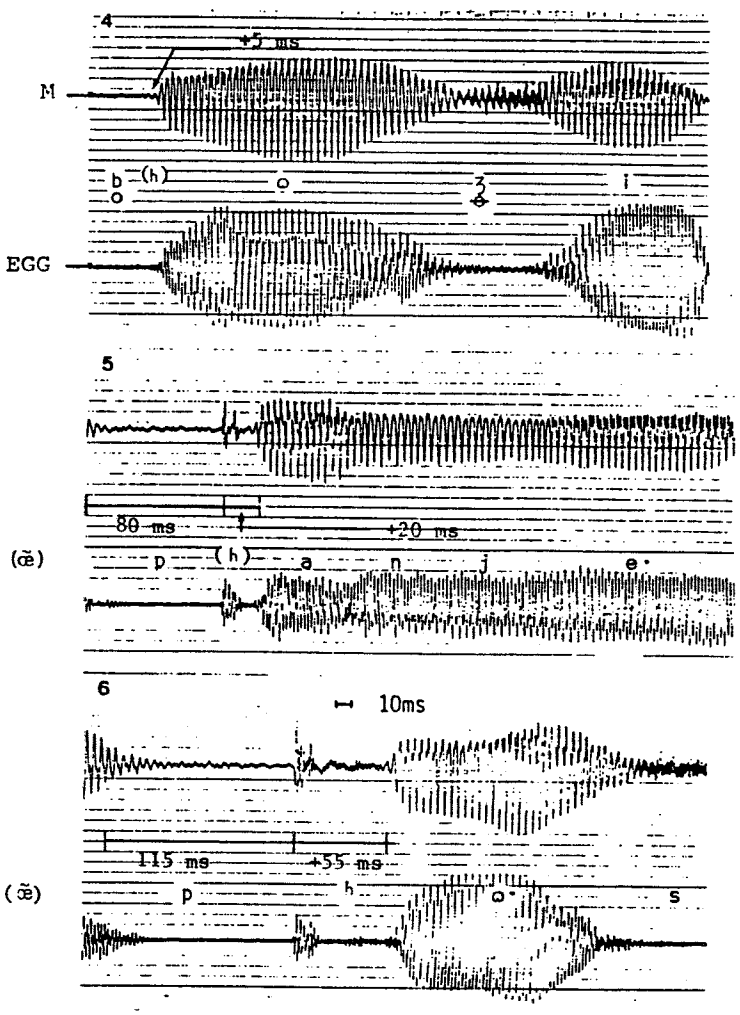


Planche 1 — Les différents types de moments d'apparition des vibrations laryngiennes (début)

Fig. 1 : [œ̃bɑlɛ·] (M.A.V.L. 1)

Fig. 2 : [œ̃g$^{(h)}$ɑt$^{(h)}$ɔ̃] (M.A.V.L. 2)

Fig. 3 : [œ̃:·d$^{(h)}$e] (M.A.V.L. 3)

Planche 2 — Les différents types de moments d'apparition des vibrations laryngiennes (suite)

Fig. 4 : [b$^{(h)}$ozi] (M.A.V.L. 4)

Fig. 5 : [œ̃p$^{(h)}$anje·] (M.A.V.L. 5)

Fig. 6 : [œ̃p^{h}o·s] (M.A.V.L. 6)

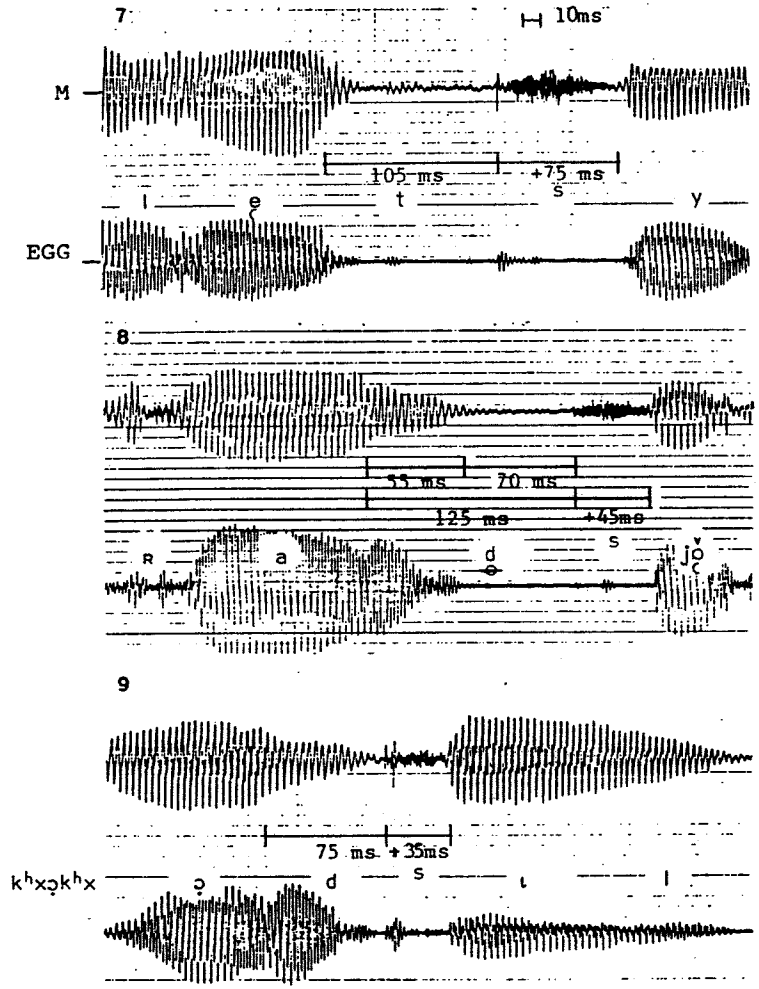


Planche 3 Les différents types de moments d'apparition des vibrations laryngiennes (assibilation en québécois)

Fig. 7 : [lɛt^{s}y] (M.A.V.L. 7)

Fig. 8 : [ʀad^{s}jǒ] (M.A.V.L. 3/7)

Fig. 9 : [kʰxɔ̣kʰxɔ̣d^{s}ɪl] (M.A.V.L. 2/7)

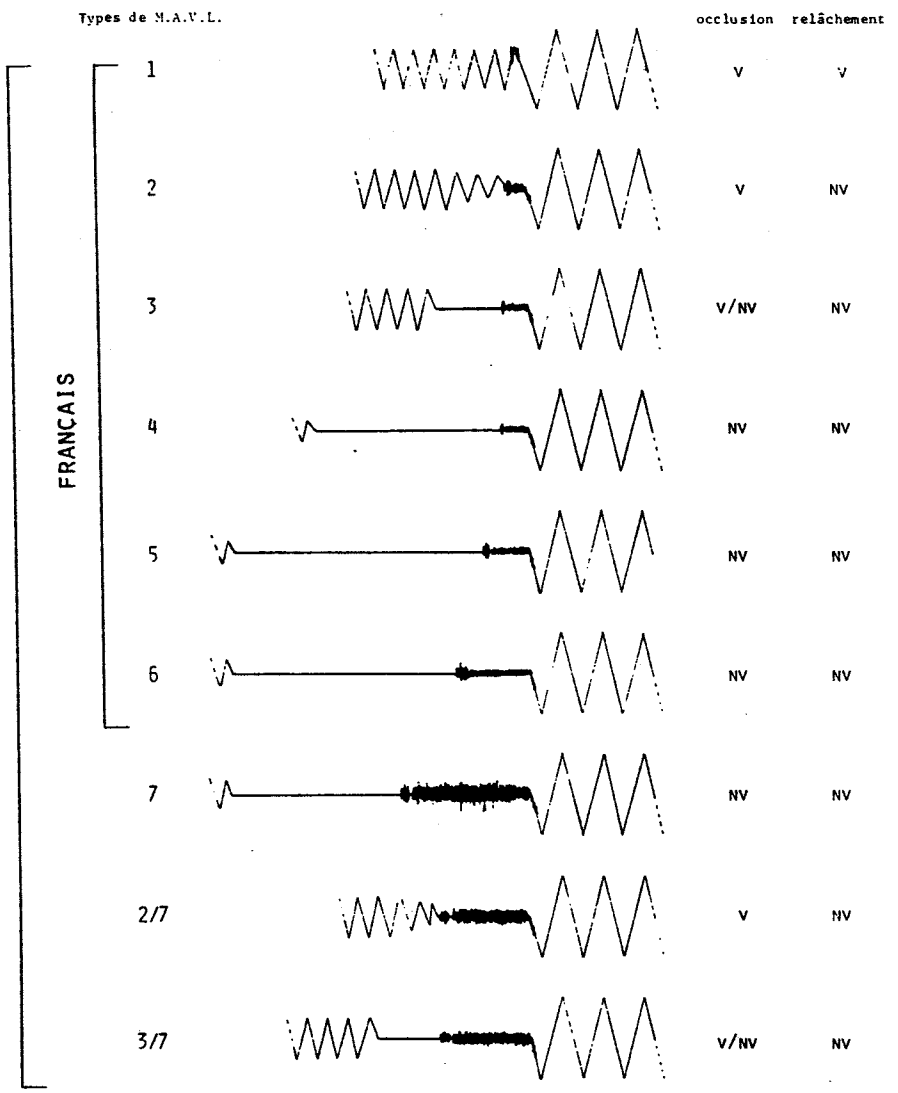


Planche 4 Schématisation des différents types de moments d'apparition des vibrations laryngiennes en français et québécois : M.A.V.L. 1, 2, 3, 4 (occlusives 'sonores'), 5, 6 (occlusives 'sourdes'), 7 (occlusives 'sourdes' assibilées), 2/7 et 3/7 (occlusives 'sonores' assibilées)

sitif est supérieur à 40ms, on est en présence d'un *type 6* : le [pʰ] de [ɛ̃pʰo·s], qui a un V.O.T. de *+55ms*, est donc de *type 6*.

Les deux types de M.A.V.L. attribués aux consonnes phonologiquement 'sourdes' peuvent être récapitulés comme suit : M.A.V.L. 5 (occlusion non voisée + relâchement non voisé avec un V.O.T. inférieur à 40ms;[p],[p(ʰ)],[t],[t(ʰ)],[k],[k(ʰ)]); M.A.V.L. 6 (occlusion non voisée + relâchement non voisé avec V.O.T. supérieur à 40ms;[pʰ],[tʰ],[kʰ]).

Pour des mots tels /lety/,/etyi/,/tigʀ/ ainsi que pour /ɛ̃diɛ̃/,/ʀadio/,/kʀokodil/,/ʃodiɛʀ/,/divɛ̃/, qui peuvent être respectivement prononcés [letˢy], [etˢɥi],[tˢigʀ],[ɛ̃d͡ˢjɛ̃],[ʀad͡ˢjo],[kʀokod͡ˢɩl],[d͡ˢivɛ̃],[ʃod͡ˢɛ̃ɛʀ], etc., le corpus québécois montre toute son originalité par rapport aux faits français; en effet, ainsi qu'il est communément reconnu pour le franco-canadien et le franco-ontarien (cf. Léon, 1979; Marchal, 1980; Thomas, 1985), les dentales /t/ et /d/ comportent une phase articulatoire qui correspond sur le plan acoustique à une assibilation, lorsqu'elles se trouvent devant une voyelle *haute* (de premier degré d'aperture) palatale /i/ ([i],[ɪ] ou [j]) ou /y/ ([y],[Y] ou [ɥ]). Le phénomène a été relevé chez tous les enfants québécois analysés (Goudaillier, 1986b). L'occlusive 'sourde' de /lety/ (Fig.7, Pl. 3) est réalisée [tˢ] avec une phase de friction (après l'explosion) d'une durée de *75ms*; l'occlusion dure dans ce cas 105ms. En termes de M.A.V.L. le *type 7* est attribué à une telle articulation.

Une relation évidente peut être mise au jour entre assibilation et dévoisement des occlusives dentales 'sonores' : ceci constitue un trait particulièrement indexant, propre aux adultes, aux adolescents et aux enfants québécois (Goudaillier, 1987; Bento, 1989). Le [d͡ˢ] de la Figure 8 (Pl. 3), qui a une phase d'assibilation (non voisée) de *45ms*, ne comporte des vibrations que pendant *55ms* au début de son occlusion, i.e., pour seulement *44,0%* (55ms/125ms) de celle-ci; plus de 50% de cette occlusion ne comprend donc aucune vibration des cordes vocales (70ms/125ms). Il s'agit ici d'un M.A.V.L. de type 3/7 (3 indique que la phase d'occlusion est partiellemnt désonorisée et 7 que la phase de relâchement comporte une assibilation). Le type 2/7, cas pour lequel l'assibilation n'occasionne pas de dévoisement de l'occlusion, est moins fréquent (de l'ordre de 20% des cas) que le type 3/7. Pour le type 2/7, 2 indique que la phase d'occlusion est voisée. Le [d͡ˢ] de la Figure 9 a une phase d'affrication sans vibration des cordes vocales de *35ms* de durée et une occlusion entièrement voisée de 75ms. On peut cependant noter que l'amplitude des vibrations décroit très nettement sur le tracé EGG pendant la phase d'occlusion, à tel point qu'aucune vibration des cordes vocales n'est relevée juste avant l'explosion.

Afin de pouvoir rendre compte des faits québécois, il convient donc d'ajouter au classement en termes de M.A.V.L. le type 7 pour les occlusives dentales 'sonores' et de constater, tant d'un point de vue articulatoire qu'acoustique, que l'affrication/assibilation occasionne dans bien des cas un dévoisement important de la phase d'occlusion des consonnes dentales 'sonores'.

Les trois types de M.A.V.L. attribués aux consonnes occlusives dentales 'sourdes' et 'sonores' en cas d'assibilation peuvent être récapitulés comme suit : M.A.V.L. 7 (occlusion non voisée + relâchement non voisé avec phase de friction importante (V.O.T. supérieur à 40ms); [tˢ]); M.A.V.L. 2/7 (occlusion voisée + relâchement non voisé avec phase de friction importante (V.O.T. supérieur à 40ms); [dˢ]); M.A.V.L. 3/7 (occlusion mi-sonore + relâchement non voisé avec phase de friction importante (V.O.T. supérieur à 40ms); [d͡ˢ]).

La notation traditionnelle des articulations occlusives assibilées par les linguistes et les phonéticiens québécois qui utilisent [dˢ] et non [dᶻ] dans le cas des 'sonores' est entièrement justifiée, puisque la friction, qui est présente lors de la phase de relâchement, n'est pas sonore dans la plus grande partie des cas.

Les différents types de M.A.V.L. présentés ci-dessus peuvent être schématisés, ainsi qu'il est indiqué à la Planche 4.

Une récente étude (Bento, 1989), menée dans la même perspective typologique, a permis de mettre au jour 2 autres types de M.A.V.L., à savoir les types 1/7 et 4/7 : M.A.V.L. 1/7 (occlusion voisée + relâchement voisé, y compris phase d'affrication; cf. le [dᶻ] de la Figure 10 (Planche 5)); M.A.V.L. 4/7 (occlusion entièrement dévoisée + relâchement non voisé; cf. le [d͡ˢ] de la Figure 11 (Planche 6)).

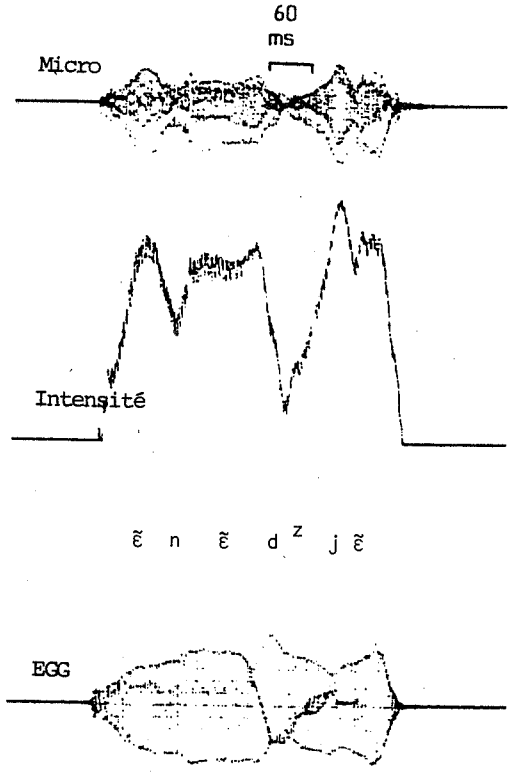


Planche 5. Figure 10.

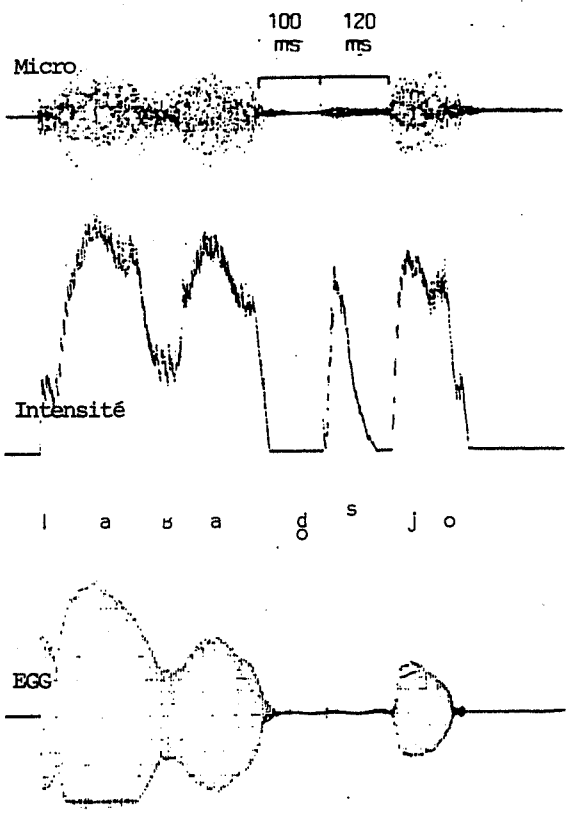


Planche 6. Figure 11.

BIBLIOGRAPHIE

BENTO Margaret (1989). *Assibilation des occlusives dentales. Approche bibliographique et étude de phonétique expérimentale des faits québécois*. D.E.A. de Phonétique. Université René Descartes.

GOUDAILLIER Jean-Pierre (1979). De l'utilisation de la phonétique expérimentale dans le cadre de la théorie phonologique fonctionnelle. *La Linguistique*. 15/1. 91-100.

GOUDAILLIER Jean-Pierre (1981). Exemple de traitement de l'opposition de "sonorité" par des enfants de Cours Préparatoire - Utilisation de la méthode électroglottographique. *12èmes J.E.P.* (Montréal). 377-391.

GOUDAILLIER Jean-Pierre (1981). Voicing contrast by eleven 6-7 year old children of a school in the North of France. *14the Annual Meeting of the Societas Linguistica Europaea* (København) (unpublished).

GOUDAILLIER Jean-Pierre (1982). Etude électroglottographique du voisement. Le cas des occlusives d'enfants scolarisés du Nord de la France. *Phonétique instrumentale et linguistique* (GOUDAILLIER,éd.). 121-124.

GOUDAILLIER Jean-Pierre (1983). Diverses possibilités de matérialisation du trait voisement - Etude électroglottographique des occlusives d'enfants âgés de 7-8 ans d'un Cours Elémentaire 1ère Année du Nord de la France. *11ème Congrès International d'Acoustique (I.C.A.)* (Paris). Vol. 4. 267-270.

GOUDAILLIER Jean-Pierre (1986a). Eléments de phonologie fonctionnelle. *Langues et Linguistique* (Université Laval, Québec). 12. 131-180.

GOUDAILLIER Jean-Pierre (1986b). Voisement et assibilation des occlusives 'sonores' d'enfants québécois (étude électroglottographique). *12ème Congrès International d'Acoustique (I.C.A.)* (Toronto). A3-7.

GOUDAILLIER Jean-Pierre (1990). *Principes théoriques de phonologie fonctionnelle expérimentale (P. F.E.). Théorie, illustrations et application aux occlusives d'enfants francophones français et québécois*. Hamburg. Buske Verlag.

GOUDAILLIER Jean-Pierre (éd.) (1984). *Phonétique instrumentale et linguistique. Actes de la Journée d'Etudes organisée le 15/5/1982 par le Laboratoire de Phonétique de l'Université René Descartes de Paris*. Hamburg. Buske Verlag.

LEON Pierre (1979). Contribution aux études de phonétique au Canada. *Linguistique expérimentale et appliquée au Canada*. 59-132.

MARCHAL Alain (1980). L'affrication de [t] et [d] en français de Montréal. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*. 7. 79-99.

THOMAS Alain (1985). L'assibilation en franco-ontarien. *Information & Communication*. 4. 65-80.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

GROUPES CONSONANTIQUES: PREMIER INVENTAIRE DES REALISATIONS ACOUSTIQUES DES PHASES DE TRANSITION.

Christine MEUNIER

Institut de phonétique
29, av. Robert Schuman 13621 Aix-en-Provence

Abstract

CONSONANT CLUSTERS: PRIMARY STOCKLIST OF ACOUSTIC REALIZATIONS OF TRANSITION PHASES

We undertook a study related to the acoustic organization of consonant clusters in order to draw up a stocklist of the different acoustic realizations according to the types of clusters.

We are especially interested in the boundaries between two consonants which we call transition phases, meaning the segment which indicates the change from a consonant to the following one. We present an acoustic description of these transition phases according to a phonetic classification of consonant clusters already established: from three consonant classes (stops, fricatives, vocalic consonants) we deduce two consonant cluster classes: homogeneous consonant clusters (the two consonants belong to the same consonant class) heterogeneous consonant clusters (the two consonants belong to two different consonant classes)

Introduction

Il existe de nombreuses études articulatoires [1] concernant les groupes consonantiques, mais il n'est pas, à notre connaissance, de travaux portant sur la description acoustique de l'ensemble des groupes de consonnes du français. En conséquence, nous avons entrepris l'étude de l'organisation acoustique des groupes consonantiques afin de dresser un inventaire des réalisations acoustiques selon les types de groupes.

Parallèlement à cette étude, nous travaillons sur la segmentation et l'étiquetage des corpus acoustiques (ACC01 à ACC05) des groupes consonantiques de la Base de Données des Sons du Français (BDSON). Ce type de travail nous a amené à concentrer notre observation sur les frontières entre les consonnes d'un groupe, ce que nous appelons les phases de transition, c'est à dire le segment, plus ou moins étendu, qui marque le passage d'une consonne à la suivante. Nous présenterons ici une description acoustique de ces phases de transition à partir d'un classement phonétique des groupes consonantiques.

1-Les fondements de la methode d'observation [2]

Lorsque l'on observe de façon détaillée le signal de parole, on peut constater que les différences majeures existant entre les unités phonétiques sont dues à des différences de mode d'articulation [3] et [4] les variations dues au lieu étant secondaires. Par conséquent, si l'on veut mettre en évidence les caractéristiques principales des différents types de groupes consonantiques (**GC**), nous sommes amenés à les regrouper à partir d'un classement en consonnes proche d'une division en modes d'articulation

Consonnes occlusives (**OCC**) : p t k
b d g

Consonnes constrictives (**CONST**) : f s ʃ ʁ
v z ʒ

Consonnes vocaliques (**C.VOC**) : j ɥ w l ʁ m n — Toutes ces consonnes ayant une structure spectrale formantique

De ce classement en consonnes, nous déduisons deux types de **GC**

Les groupes consonantiques homogènes (**GC$_{ho}$**) ou les deux consonnes appartiennent à la même classe de consonnes

GC$_{ho}$1 = OCC + OCC
GC$_{ho}$2 = CONST + CONST
GC$_{ho}$3 = C.VOC + C.VOC

Les groupes consonantiques hétérogènes (**GC$_{hé}$**) dans lesquels les deux consonnes appartiennent à deux classes de consonnes différentes:

GC$_{hé}$1 = OCC + CONST
GC$_{hé}$2 = CONST + C.VOC
GC$_{hé}$3 = OCC + C.VOC

Nous espérons ainsi répertorier les tendances très générales d'homogénéité et d'hétérogénéité des **GC**. Il faut noter que ce classement, bien que déduit d'une répartition des consonnes suivant leur similitude acoustique, est purement phonétique ; nous entendons par là que les termes "homogène" et "hétérogène" ne concernent pas la description des réalisations acoustiques des phases de transition mais l'organisation globale du groupe en fonction du type de consonnes qui le composent.

En ce qui concerne la description acoustique, nous nous sommes limités ici à l'observation des phases de transition. Pour décrire ces phases, nous avons retenu deux indices acoustiques: 1) les changements d'amplitude du signal, 2) les discontinuités spectrales; à ces deux indices, nous associons l'étendue de la discontinuité, c'est à dire la

durée, courte (deux périodes maximum) ou longue (plus de deux périodes), du passage d'une consonne à l'autre. Ainsi nous obtenons deux types de discontinuités:

a) la discontinuité majeure (**dmaj**) caractérisée par la combinaison des deux indices ci-dessus pour une durée courte: (1+2) court

b) les discontinuités mineures (**dmin**) caractérisées soit par la présence de l'un ou l'autre indice, pour une durée courte ou longue, soit par la combinaison des deux indices pour une durée longue:

1 long (**dmin1L**) ou 1 court (**dmin1C**)
2 long (**dmin2L**) ou 2 court (**dmin2C**)
(1+2)long (**dminL**)

2-Le corpus

Nous avions antérieurement remarqué l'importance du mode d'articulation pour distinguer les classes de consonnes à l'intérieur d'un **GC**. Afin d'étudier ce phénomène, nous avons conçu un corpus, appelé "corpus mode", dans lequel nous avons confronté les consonnes d'un groupe selon leur mode d'articulation. Il s'agit donc ici de donner une représentation des **GC** conforme à la classification en GC_{ho} et $GC_{hé}$ définie plus haut, et non de concevoir un évantail exhaustif des **GC** suivant les modes et lieux d'articulation. En effet, nous pensons que les modes d'articulation ont une incidence majeure sur la variation des phases de transition à l'intérieur des **GC**; les lieux d'articulation ont une forte incidence sur la globalité du **GC**, mais entraînent peu de variation quant à la phase de transition.

Le corpus mode est composé de 38 mots isolés lus par 10 locuteur. Les **GC** sont regroupés selon notre classement en GC_{ho} et $GC_{hé}$. Sur ces 38 mots nous étudierons ici 12 mots caractéristiques qui sont des exemples représentatifs des 6 types de **GC** que nous avons définis. A partir de chaque exemple nous généraliserons sur les problèmes spécifiques du **GC** correspondant.

Il faut concevoir l'analyse de ce corpus comme un premier état d'une description acoustique de la phase de transition à partir d'un classement en GC_{ho} et $GC_{hé}$

3-Les six classes de GC et leur phase de transition

3-1) $GC_{ho}1$ (OCC+OCC) : [agda] [ebdo]

Le groupe d'occlusives est l'exemple caractéristique du groupe phonétiquement homogène mais acoustiquement discontinu. Cela tient à la structure asymétrique de l'occlusive et à son organisation temporelle gauche-droite: d'abord une tenue (silence ou période simple) puis une explosion. La frontière droite et la frontière gauche ne sont donc pas semblables. La phase de transition est caractérisée par une dmaj (si l'explosion est de forte intensité) ou par une **dmin2C**.

Pour [agda], sur les dix locuteurs, sept ont réalisé une voyelle latente (très courte) entre les deux occlusives. Pour [ebdo], seulement deux locuteurs ont réalisé cette voyelle latente. On peut ainsi penser que la réalisation du [ə] dépend de la place du **GC** dans la syllabe (**GC** implosif ou **GC** explosif). Nous nous sommes également interrogés sur une hypothétique relation entre le débit et la présence de [ə]: la faible étendue de nos données ne nous permet pas d'évaluer statistiquement cette relation; cependant on constatera que l'absence de la voyelle latente est plutôt associée à un débit rapide alors qu'un débit lent semble favoriser sa présence. Toutefois nous n'excluons pas que la réalisation de ce [ə] soit un artefact dû à la lecture de mots isolés.

On considère que, dans ce cas, la frontière entre les deux occlusives se situe à la fin de la voyelle latente et au début du voisement; la voyelle latente serait à concevoir comme une modalité du relâchement de l'occlusive (voir fig. 1).

3-2) $GC_{ho}2$ (CONST+CONST) : [fʁã] [asfalt]

Les phases de transition concernant les groupes de constrictives se manifestent par des discontinuités spectrales généralement très brèves. Elles sont identifiables sur le signal temporel de parole à l'aide d'un zoom assez puissant, cependant il est nécessaire dans ce cas d'utiliser parallèlement un spectrographe (voir fig. 2).

Pour les deux exemples que nous avons choisis on observe une différence très nette concernant l'amplitude de bruit des deux constrictives; [f] est caractérisé par un bruit de faible intensité, alors que pour [s] et [ʁ] le bruit est beaucoup plus intense, et très perturbé en ce qui concerne [ʁ]. Dans le cas de [fʁã], et dans de nombreux cas où l'on a affaire à [ʁ] (on le verra plus loin pour [tʁɛ]), il peut exister un segment transitoire bruité différent de [f] et du bruit perturbé et intense de [ʁ]. Ce segment rend la frontière plus difficile à déterminer. Notons enfin que dans huit cas sur dix, pour les deux exemples, la première constrictive est proportionnellement plus longue que la seconde. Cet élément pourrait signifier qu'il existe une relation entre la durée de la constrictive et sa place dans le **GC**; ceci serait à vérifier dans une étude ultérieure.

Dans un groupe de constrictives non voisées la phase de transition sera marquée par une **dmin2C**, plus rarement par une **dmin2L**.

3-3) $GC_{ho}3$ (C.VOC+C.VOC) : [lyt] [mjø]

On constate le plus souvent une discontinuité longue dans la phase de transition pour les groupes de consonnes vocaliques. Il semblerait que la vocalicité ait une incidence sur la durée de la phase de transition entre les deux consonnes. Nous avons regroupé dans la catégorie **C.VOC** des unités phonétiques qui ont en commun une structure acoustique marquée par la présence de formants. Cependant, il faut noter que le degré de vocalicité n'est pas le même pour toutes les consonnes. Ainsi les **dmin2L** ou **dminL** seront régulièrement présentes dans les groupes [r] + glissantes; alors que les groupes comprenant [l] ou une nasale seront plus facilement marqués par la présence d'une **dmin2C**, voire une **dmaj** (voir fig. 3), à la frontière des deux unités.

Dans seulement deux réalisations sur dix, pour [lyt], on constate une discontinuité minime et difficile à identifier à la frontière. Pour les huit autres cas la discontinuité est importante: on note un décrochage très net de l'amplitude du signal ([l] a une intensité plus faible que les autres consonnes vocaliques), ainsi que le passage rapide d'une périodicité peu complexe [l] à une périodicité très complexe [y]; ici la discontinuité pourra être une **dmaj**. Le passage de la nasale à la glissante est beaucoup plus difficile à identifier: sur les dix réalisations, une seule présente un décrochage de l'amplitude permettant de localiser avec précision la frontière; une autre présente un creux d'amplitude entre les deux consonnes; dans tous les autres cas, il n'y a pas de décrochage mais une augmentation très progressive de l'amplitude vers la glissante (il n'y a pas d'interruption dans l'enveloppe générale de la courbe d'intensité du **GC**); ici nous aurons forcément affaire à un **dminL**.

3-4) **GC$_{hé}$1 (OCC+CONST)** : [tsar] [tʁɛ]

Pour ce **GC** particulièrement, il nous faut différencier les groupes voisés des groupes dévoisés. Dans les groupes voisés, l'analyse acoustique ne pose pas problème: l'occlusive est caractérisée par une tenue voisée et une explosion à laquelle fait suite la périodicité bruitée de la constrictive. Si l'explosion est peu marquée, on peut être amené à s'interroger sur le début de la constrictive, mais l'hypothèse d'une confusion entre explosion voisée et voisement bruité sera généralement écartée par l'information spectrographique. La **dmin2C** est la discontinuité généralement observée.

Dans les groupes non voisés, l'occlusive est réduite à son explosion, la durée de la tenue silencieuse n'étant pas mesurable à l'initiale (on peut éventuellement attribuer une durée arbitraire). Si l'explosion est bien marquée (voir fig.4), la frontière est nette; si l'explosion est de faible intensité, l'identification de l'occlusive sera parfois impossible; notons d'ailleurs que pour [tsaʁ] trois locuteurs sur dix ont réalisé une double explosion. Par ailleurs, un segment transitoire n'appartenant pas à la constrictive vient parfois s'insérer après l'explosion de l'occlusive; ceci est particulièrement fréquent dans le cas d'une occlusive non voisée suivie de [ʁ]: ainsi, pour [tʁɛ] quatre locuteurs sur dix réalisent ce segment transitoire bruité. On intègrera ce segment à l'occlusive comme une modalité de relâchement, conformément à l'attitude que nous avons adopté plus haut (cf. **GChoI**). Les discontinuités présentes dans ce type de contexte seront **dmin2C** ou **dmin2L**.

3-5) **GC$_{hé}$2 (CONST+ C.VOC)** : [ʃlɛm] [ʃjɛ̃]

C'est dans ce type de **GC** que l'on remarquera les discontinuités les plus marquées, particulièrement lorsque la constrictive n'est pas voisée. Dans ce cas (constrictive non voisée), nous avons affaire à deux éventualités très précises: soit la rupture spectrale due à la **C.VOC** est immédiatement accompagné de voisement, alors la discontinuité frontière sera une **dmaj**, soit la **C.VOC** est dévoisée à l'initiale et l'on constate alors que la **dmaj** se démultiplie en deux **dmin**: une **dmin2C**, marquant la discontinuité spectrale frontière entre [ʃ] et [l], et une **dmin1C** au début du voisement de la **C.VOC**. Le deuxième cas est d'ailleurs le plus fréquent, particulièrement lorsque la consonne vocalique est un [l]: neuf locuteurs sur dix ont dévoisé le [l] au contact de la constrictive dans [ʃlɛm]. On notera qu'il est fréquent (huits locuteurs sur dix dans [ʃlɛm]) que le début d'un [l] dévoisé soit manifesté par une explosion (voir fig. 5); ce phénomène est particulier au [l] et nous pensons qu'il s'agit du bruit dû au contact de la langue sur le palais, la position particulière du [l] (latérale) entraînant un mouvement articulatoire important par rapport aux autres consonnes. En ce qui concerne [ʃjɛ̃], le dévoisement de la glissante est moins fréquent (cinq locuteurs sur dix).

Si la constrictive est voisée, la discontinuité est parfois allongée. Il s'agit encore ici de l'incidence du voisement sur la durée de la discontinuité. Dans la plupart des cas, la discontinuité frontière sera une **dmaj**, éventuellement une **dminL**.

3-6) **GC$_{hé}$3 (OCC+C.VOC)** : [blø] [dra]

On identifie sans problème la discontinuité frontière entre une occlusive, qu'elle soit voisée ou non, et une consonne vocalique. Les différences spectrales et les changements d'amplitude écartent toute ambiguïté dans la détermination du début de la consonne vocalique. Même si l'explosion est peu marquée, voire non identifiable, la structure vocalique de la **C.VOC** ne peut en aucun cas être confondue avec l'occlusive. La **dmaj** sera la discontinuité régulièrement observée (voir fig. 6).

Notons toutefois que dans certaines réalisations, peu fréquentes et surtout avec [l], la consonne vocalique peut être dévoisée lorsqu'elle est précédée d'une occlusive non voisée. Nous nous trouvons ainsi dans le cas des **GC$_{hé}$1** où il s'agira de distinguer l'explosion du bruit qui le suit. Le problème délicat sera de faire la distinction entre un possible bruit de relâchement de l'occlusive et le dévoisement de la consonne vocalique (ceci sera possible grâce à une analyse spectrographique).

On peut constater que pour les deux exemples que nous avons choisis, la discontinuité sera plus facilement repérable suivant que l'on a affaire à [r] ou à [l]; en effet, dans ce type de contexte voisé, le [r] est vocalique et battu; sa réalisation la plus courante est la suivante: une première partie vocalique, suivie d'une clusion elle-même suivie d'une deuxième partie vocalique [ʃ]. Ainsi, la partie en contact avec l'occlusive sera une partie vocalique que l'on appelera "partie voyelle". Le contraste avec l'occlusive sera donc plus important si elle est suivie de [r] (périodicité très complexe), que si elle est suivie de [l] (périodicité moins complexe). Sur les dix locuteurs, un seul a réalisé un [ʁ] (constrictif) voisé dans [dra]. On observe également l'insertion d'une voyelle latente entre le [b] et le [l] chez un locuteur sur dix pour [blø]; ce phénomène est associé à un débit lent.

4-Conclusions et perspectives

A travers cette étude, nous nous sommes efforcés de mettre en évidence les tendances acoustiques majeures observées dans la phase de transition des **GC**. On remarque ainsi que les phases de transition varient d'abord en fonction des classes de consonnes (**OCC**, **CONST**, **C.VOC**) qui compose le **GC**. Cependant, il est nécessaire de nuancer ce propos et d'effectuer un sous-classement pour chaque type de consonnes afin d'affiner nos observations: en effet, la présence ou l'absence de voisement pour les occlusives et les constrictives dans les **GC$_{hé}$1** et **GC$_{hé}$2** a une incidence importante sur la discontinuité frontière; d'autre part nous avons constaté que, dans la classe **C.VOC**, le degré de vocalicité des consonnes pouvait jouer un rôle sur la durée de la phase de transition. Enfin, même si cette éventualité ne nous a pas parue pertinente pour l'instant, nous tenons à vérifier l'effet produit par les variations des lieux d'articulation des consonnes du **GC** sur le type de discontinuité de la phase de transition.

Ces remarques nous ont conduit à envisager un corpus permettant une étude détaillée de chacune des six classes de **GC**: pour chaque **GC** on fait varier le lieu d'articulation de chaque consonne, le voisement des occlusives et des constrictives, et le degré de vocalicité des consonnes vocaliques.

L'analyse de ce corpus pourrait nous permettre de prolonger et d'affiner l'étude des phases de transition des **GC**. En fait, elle nous sera particulièrement utile pour l'analyse globale des **GC**, c'est à dire pour l'étude des durees et des phénomènes d'assimilations diverses.

Enfin, ce travail nous a également permis de constater les effets perturbateurs de la lecture de mots isolés: certains locuteurs ont tendance à surarticuler et a insérer des segments vocaliques, les durées des segments sont parfois incohérentes. L'utilisation d'un corpus de

groupes consonantiques insérés dans des phrases porteuses devrait nous permettre d'observer des phénomènes plus réguliers et plus cohérents.

REFERENCES

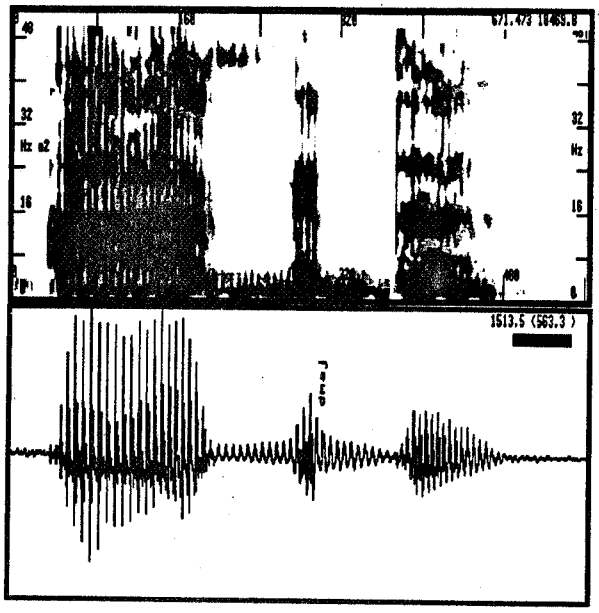
[1] ROCHETTE, C. (1973), Les groupes de consonnes du français, Klincksieck, Québec.

[2] MEUNIER, C. (1989), "Une approche acoustique des groupes consonantiques: étude des phases de transition", Travaux de l'Institut de Phonétique d'Aix, vol. 13, Aix-en-provence (à paraître)

[3] AUTESSERRE, D., ROSSI, M. (1985), "Propositions pour une segmentation et un étiquetage hiérarchisé. Application à la base de données acoustiques du GRECO Communication Parlée", Actes des 14èmes Journées d'Etude sur la Parole, Paris, p.147-152.

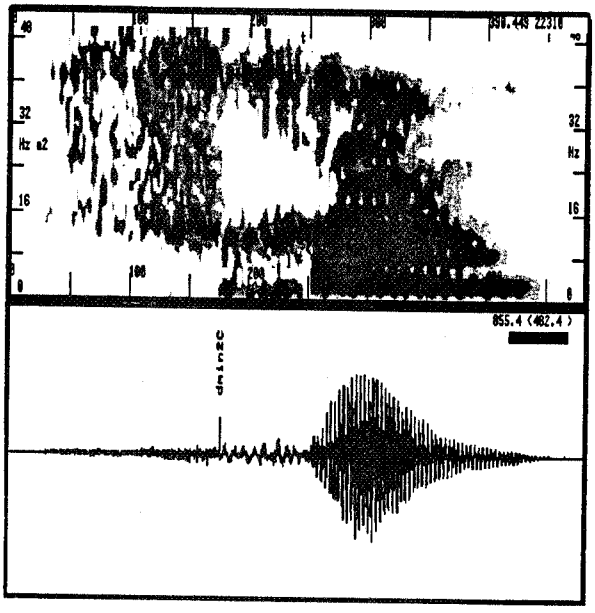
[4] AUTESSERRE, D., ROSSI, M., (1987), "La segmentation et l'étiquetage des groupes consonantiques de la BDsons", Actes des 16èmes Journées d'Etude sur la Parole, Hammamet, p.196-199.

[5] MEUNIER, C., AUTESSERRE, D. (1989), "Variabilité intra et inter individuelle de l'allophone vocalique de (r) en contexte occlusif voisé", Actes du séminaire "variabilité et spécificité du locuteur: études et applications", Marseille (Luminy), éd; H. Méloni, p.36-39.



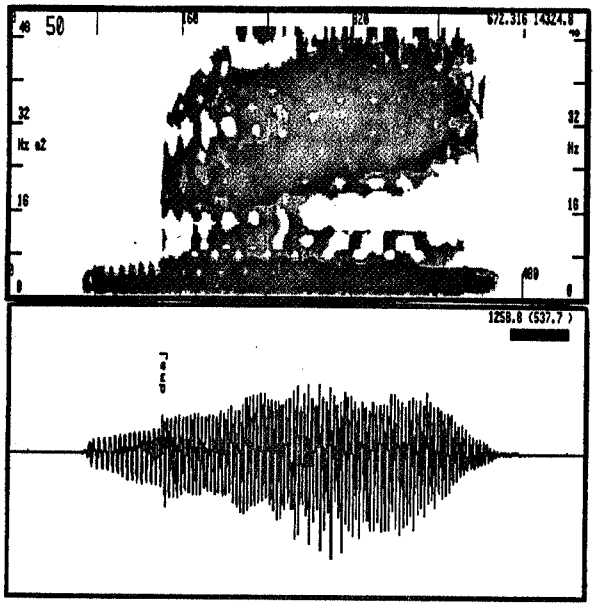
Agde .Loc 2

Figure 1



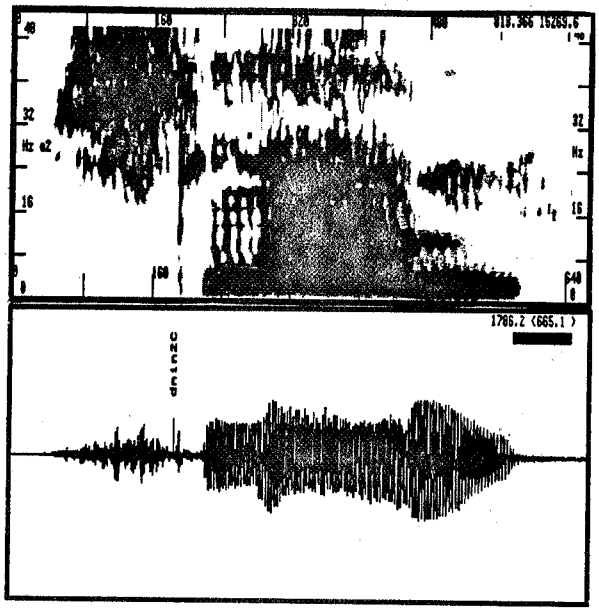
franc.Loc 5

Figure 2



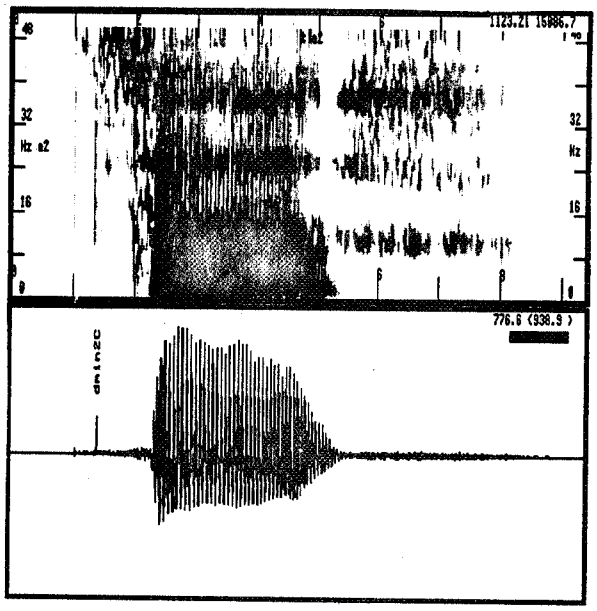
lui. Loc 5

Figure 3



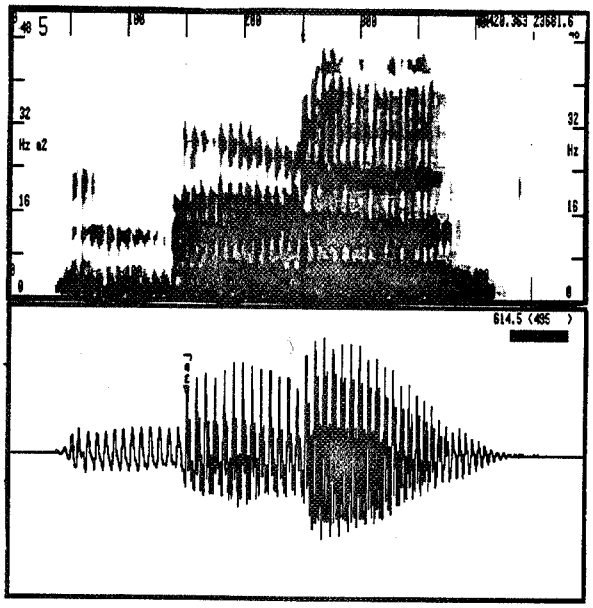
schlem. Loc 6

Figure 5



Tsar. Loc 2

Figure 4



bleu. Loc 2

Figure 6

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

RÉALISATIONS ACOUSTIQUES ET PERCEPTION:

LE CAS DES TIMBRES DU E FRANÇAIS

par Pierre R. Léon
et Jeff Tennant

Université de Toronto

-RÉSUMÉ-

On tente ici d'établir les zones acoustiques correspondant à la perception des deux timbres du E français. On s'en est tenu, selon la méthode classique et quelque peu primitive, au seul examen des deux premiers formants des voyelles. On suggère la possibilité de tenir compte de deux indices (perceptibilité et différenciabilité) permettant de mieux définir acoustiquement les timbres perçus.

-ABSTRACT-

An attempt is made here to establish acoustical zones corresponding to the perception of the two timbres of French E. Following the classic and now somewhat primitive method, we limited ourselves to the analysis of the first two vowel formants. The possibility is suggested of taking into account two indexes (perceptibility and distinguishability), allowing a better definition of perceived timbres.

I PROBLÉMATIQUE

On voudrait tenter d'établir des critères acoustiques permettant la vérification de la perception des timbres vocaliques dans l'étude du "double timbre" du E en français. On a constaté en effet que l'interprétation des timbres vocaliques est souvent sujette à caution lorsqu'il s'agit de timbres voisins tels que [e] et [ɛ] susceptibles de se neutraliser.

Les acousticiens notent les timbres au moyen des formants vocaliques - supposant que F_1 et F_2 sont les principaux, voire les seuls responsables de notre perception de ces timbres. C'est en partant de ce constat acoustique que Matthew Lennig (1978) a tenté de montrer l'évolution des timbres vocaliques dans le système phonique du français parisien, sans toutefois mentionner de corrélations avec la perception.

Pour le français, Pierre Delattre (1965) a établi un tableau des formants vocaliques souvent cité, à partir de paramètres qui lui ont permis d'effectuer une synthèse de type vocoder. Peter Ladefoged (1967) a souligné la difficulté de l'interprétation des valeurs formantiques dans l'analyse comme dans la synthèse. Plusieurs investigations donnent une idée de la difficulté en cette matière, telles celles citées par Pierre Durand (1985): Lafon, Romieu, Debrock & Forrez et la sienne propre (p. 104). Pour le français la dernière étude sur le sujet semble être celle d'André Bothorel et al. (1986). Cette étude offre une vue à la fois plus précise et plus difficile à interpréter en présentant non pas les 2 chiffres habituels pour F_1 et F_2 mais l'ensemble du spectre formantique, indiquant l'évolution des intensités sonores en fonction du temps. La comparaison des spectres pour les quatre sujets émettant les mêmes séquences donne une bonne idée de la variabilité intrapersonnelle.

Malgré des chiffres fluctuants, toutes ces données indiquent une même conclusion, sur quoi tous les phonéticiens s'accordent, pour les timbres des voyelles qui nous intéressent ici: E fermé est une voyelle diffuse et E ouvert une voyelle compacte.

Plus récemment cependant, on a employé des procédures plus sophistiquées, qui tiennent compte d'autres traits acoustiques (voir par exemple Denis Tuffelli et Haïyan Ye (1988). Néanmoins, en raison de nos limitations en ressources et en temps, on a décidé de se cantonner ici à la méthode classique, désormais quelque peu primitive, qui ne tient compte que des deux premiers formants.

II CORPUS ET METHODE

On a utilisé comme corpus un enregistrement des interventions de Bernard Pivot dans trois émissions d'**Apostrophes**. La durée totale du corpus, décrit dans Pierre Léon et Jeff Tennant (1988), est d'environ une heure. Soulignons qu'il s'agit ici d'un corpus spontané. On n'a retenu pour l'analyse que des exemples empruntés à Bernard Pivot, d'origine bourguignonne mais dont le français est très standardisé.

On s'est servi du spectrographe digital couleur RT-1000 de Philippe Martin, pour effectuer l'analyse acoustique du corpus. Cet appareil nous a permis de relever la fréquence (en Hertz) de la zone d'intensité maximale des deux premiers formants (F_1 et F_2) de chaque occurrence de [e] et de [ɛ] en syllabe ouverte, en position finale de mot, en syllabe accentuée.

Avant d'exposer ces résultats et de les discuter, on rappellera les indices formantiques trouvés par Delattre et on proposera deux termes pour désigner les outils d'analyse que nous voudrions utiliser; ce sera: **l'indice de perceptibilité** et **l'indice de différenciabilité**.

III INDICE DE PERCEPTIBILITÉ ET INDICE DE DIFFÉRENCIABILITÉ

On propose d'appeler **indice de perceptibilité** du timbre d'une voyelle donnée, la différence entre les valeurs des deux formants F_1 et F_2. C'est cette différence qui assure en effet la réussite ou l'échec du décodage du timbre. Une petite différence, indice d'un timbre compact, assurera la perception d'un E bien ouvert, alors qu'une grande différence permettra la perception d'un timbre nettement fermé.

On propose en outre d'appeler **indice de différenciabilité** la valeur qui permet de distinguer deux timbres l'un de l'autre. Dans le cas de la comparaison entre [e] et [ɛ], cet indice de différenciabilité E pourra être chiffré en établissant la différence entre les indices de perceptibilité Δe et $\Delta \epsilon$.

Ainsi à partir des chiffres de F_1 et F_2 de Delattre, on établira les indices Δe, $\Delta \epsilon$ et ΔE, comme indiqué dans le tableau 1, ci-dessous.

	F_1	F_2	Indice P	Indice D
E fermé	375	2200	$\Delta e = 1825$	$\Delta E = 575$
E ouvert	550	1800	$\Delta \epsilon = 1250$	

Tableau 1. Valeurs des formants et des indices de perceptibilité (P) et de différenciabilité (D), d'après les chiffres de Delattre.

IV RÉSULTATS DE L'ÉTUDE PERCEPTIVE ET VALEURS ACOUSTIQUES DES TIMBRES DE NOTRE CORPUS

1. Classement acoustique

On a classé dans les tableaux 2,3,4 ci-dessous, toutes les valeurs des 100 E extraits du corpus de Pivot, en ne tenant compte que du point de vue acoustique. Ces valeurs sont présentées dans nos tableaux par ordre de diffusion formantique décroissante. On va donc des timbres théoriquement les plus fermés aux timbres les plus ouverts, selon une échelle de valeurs acoustiques.

2. Vérification auditive

Nous avons écouté attentivement la liste des mots selon leur classement acoustique, tel qu'indiqué ci-dessus. Nous sommes tombés d'accord sur le fait que toutes les voyelles du tableau 1 sont nettement de timbre fermé et que toutes celles du tableau 3 sont nettement de timbre ouvert.

Notre désaccord - ou plutôt notre accord sur l'incertitude des timbres - est reflété par les voyelles du tableau 2.

Nous nous proposons, ultérieurement de faire passer un test d'audition à un auditoire plus large que celui des 2 auteurs de cet article, que nous représentons.

3. Confrontation des timbres perçus et des valeurs acoustiques

MOTS	Formant 1 F1	Formant 2 F2	Indice de perceptibilité (F2 - F1)
guerriers	290	2300	2010
piller	355	2195	1840
arnaqué	265	1985	1720
années	350	2050	1700
initier	275	1950	1675
racontez	310	1970	1660
retrouvez	380	2030	1650
des milliers	350	2000	1650
amassé	345	1985	1640
diriez	300	1935	1635
détournée	330	1965	1635
peignez	350	1975	1625
vanités	360	1970	1610
doué	445	2045	1600
propriété	345	1930	1585
intéressé	310	1885	1575
métier	370	1935	1565
héritiers	315	1880	1565
terminer	405	1965	1560
disiez	330	1890	1560
premier	310	1865	1555
été	355	1905	1550
beauté	375	1925	1550
illustrer	325	1875	1550
gagner	375	1925	1550
mais	380	1915	1535
propriété	315	1840	1525
propriété	290	1810	1520
dirais	320	1835	1515
acheter	345	1860	1515
parlez	360	1875	1515
mais	370	1885	1515
inné	300	1800	1500
rencontrer	425	1925	1500
écrivez	325	1825	1500
démembré	400	1900	1500
emmenait	405	1890	1485
enterrer	405	1885	1480
condamné	425	1900	1475
privée	330	1800	1470
propriété	315	1780	1465
lacérer	360	1825	1465
gratter	375	1840	1465
préférés	340	1800	1460
mais	400	1850	1450
voyait	300	1735	1435
pensez	350	1780	1430
fuyait	325	1750	1425
santé	425	1850	1425
insister	405	1815	1410
honnêteté	370	1780	1410
trouver	405	1815	1410
percé	425	1830	1405
propriété	400	1800	1400
trompés	405	1800	1395
reciter	450	1780	1330
moyenne	357	1895	1538
écart type	44	102	115

Tableau 2. Timbres perçus et valeurs acoustiques de E fermé.

MOTS	Formant 1 F1	Formant 2 F2	Indice de perceptibilité (F2 - F1)
tout à fait	395	1785	1390
naît	375	1750	1375
disais	375	1750	1375
mais	465	1825	1360
albanais	320	1670	1350
sauter	460	1810	1350
mais	375	1725	1350
était	425	1775	1350
moyenne	399	1761	1363
écart type	46	46	15

Tableau 3. Timbres perçus et valeurs acoustiques de E moyen.

MOTS	Formant 1 F1	Formant 2 F2	Indice de perceptibilité (F2 - F1)
disais	310	1655	1345
très	375	1700	1325
vrai	385	1710	1325
dirais	410	1735	1325
engageais	320	1635	1315
disait	425	1740	1315
met	375	1675	1300
albanais	425	1675	1250
français	550	1790	1240
disait	395	1635	1240
est	455	1690	1235
disait	405	1640	1235
albanais	425	1650	1225
c'est	425	1650	1225
albanais	530	1750	1220
mais	470	1685	1215
français	450	1650	1200
dirais	375	1575	1200
connaît	400	1600	1200
dirais	405	1565	1160
emmenait	400	1550	1150
vendait	395	1535	1140
vrai	475	1600	1125
procès	470	1590	1120
essai	450	1535	1085
français	475	1560	1085
vrai	425	1500	1075
exposait	460	1515	1055
succès	425	1480	1055
Albanais	500	1530	1030
disait	530	1540	1010
après	505	1515	1010
dirais	500	1500	1000
dirais	560	1420	860
vrais	650	1500	850
vrai	565	1340	775
moyenne	447	1600	1153
écart type	70	97	139

4. Timbres perçus et valeurs acoustiques de E ouvert.

V DISCUSSION

1. Perception et réalisation acoustique

Dans leur ensemble, tous les spectres formantiques des E **fermés** sont **diffus**, et tous ceux des E **ouverts** sont **compacts**. Les E perçus comme moyens, ont des valeurs acoustiques intermédiaires. Ces résultats confirment bien la théorie classique des acousticiens sur ce plan.

Nous avons trouvé cependant deux exceptions. L'une est le E de "sauter" (6ème du tableau 3), l'autre "re-citer" (tableau 2) dont les formants indiquent respectivement un timbre moyen et un timbre ouvert. Nous n'avons pas trouvé d'explication à cette anomalie, ni en prenant en compte la surface des plages acoustiques ni leurs intensités relatives.

2. Étendue du spectre et indice de perceptibilité

Si nous considérons maintenant **l'indice de perceptibilité**, tel que nous l'avons défini par la valeur de l'écart entre les 2 formants, on s'aperçoit que les résultats de nos analyses sont très différents de ceux décrits, en particulier par Delattre.

Laissons de côté le E moyen et comparons nos chiffres du E fermé et du E ouvert avec ceux de Delattre. Ces valeurs sont assez différentes pour le E **fermé**:

Pivot (moyennes)		Delattre
F_1	357	375
F_2	1895	2200
Δe	1538	1825

Ces valeurs apparaissent encore plus éloignées si l'on considère les extrémités de leur échelle. Le E acoustiquement plus fermé de notre liste est "guerriers": $F_1 = 290$; $F_2 = 2300$, et le E acoustiquement le moins fermé (abstraction faite de l'exception que représente "re-citer"), est (trompés): $F_1 = 405$; $F_2 = 1800$. Ce qui revient à dire que nous percevons presque toujours un E fermé pourvu que le spectre acoustique formantiel soit compris entre e = 2010 et 1395. Le premier de ces chiffres se situe 185Hz au-dessus du Δe de Delattre et le second 430 Hz au-dessous, ce qui est assez considérable et montre bien la fluctuance des valeurs de cet indice.

Quant au E **ouvert**, on note les chiffres suivants:

Pivot (moyennes)		Delattre
F_1	447	550
F_2	1600	1800
Δe	1153	1250

Les chiffres sont plus proches de ceux de Delattre si l'on ne considère que la moyenne. En réalité, l'écart type montre, comme pour le E fermé, une grande dispersion des valeurs. Et l'on va du E ouvert le moins ouvert acoustiquement (**dis**ais, au début de la liste du tableau 4) au E ouvert le plus ouvert acoustiquement (**vrai** en fin de liste du tableau 4), on passe de $F_1 = 310$, $F_2 = 1655$, $\Delta\epsilon = 1345$ à $F_1 = 565$, $F_2 = 1340$ et $\Delta\epsilon = 775$. Tout cela est très loin des chiffres de Delattre. La compacité du E le plus ouvert du Pivot (775) est presque égale à la moitié du même E de Delattre (1250).

Il y a donc, dans la réalité de la parole naturelle, une largeur de spectre formantique beaucoup plus importante que celle utilisée pour la synthèse par Delattre.

3. Indice de différenciabilité

Si l'on considère les moyennes relevées pour les indices de perceptibilité chez Pivot, l'indice de différenciabilité, mesurant la différence entre les timbres, s'établit alors à 1538 - 1153 = 385 Hz alors que chez Delattre il était de 575 Hz.

Cela revient à dire que malgré une dispersion très élevée des valeurs acoustiques des spectres de notre corpus de E (reflétée par les écarts types) la valeur moyenne de l'écart est beaucoup moins élevée, sur ce plan également, que celle relevée chez Delattre. Par contre, si l'on considère les extrémités des échelles formantiques les plus éloignées, les timbres de Pivot peuvent être nettement plus différenciables puisque la différence entre son E le plus fermé et son E le plus ouvert est en termes d'index de différenciabilité: (guerriers) 2010 - (**vrai**) 775 = 1235 Hz ce qui représente (1235 - 575) une différence de 660 Hertz avec les chiffres de Delattre (1965). Notons toutefois que les chiffres de Delattre étaient plus proches des nôtres dans la première version de son article de 1948.

4. Continuum phonique et perception

Il faut noter à la fois la souplesse de l'oreille qui tolère de telles fluctuations dans le continuum phonique et son pouvoir perceptif de diviser ce continuum en catégories discrètes. On remarquera ici que Pivot a moins de 10% de son stock de E que nous avons dû classer dans la catégorie des timbres moyens. Son système vocalique apparaît, sur ce point remarquablement bien structuré avec les deux catégories de timbres décrites par les orthoépistes.

5. Aspects linguistiques

Le classement orthoépique des E de Pivot - que nous avions effectué à priori, avant l'étude acoustique et l'étude perceptive - se trouve confirmé par les réalisations phonétiques à quelques exceptions près (quelques terminaisons en **-ais**, **-ait**, prononcées fermées ainsi que des fluctuations pour la prononciation du connecteur **mais**).

Le yod a une forte influence fermante (visible dans les chiffres des valeurs formantiques).

VI CONCLUSION

En faisant cette recherche, notre but était avant tout de tenter d'établir des indices acoustiques pouvant aider à confirmer la perception auditive. Notre corpus limité à un seul locuteur et à deux auditeurs ne nous permet pas de donner de conclusions définitives sur ce point. Il resterait à prouver que les relations établies ici entre acoustique et perception fonctionnent de la même manière quel que soit l'idiolecte du locuteur et les habitudes perceptives des auditeurs. On peut tout de même imaginer que, si ce type d'outil était affiné, un système de détection et de reconnaissance automatique basé sur les indices acoustiques que nous avons proposés pourrait être utilement employé dans le contrôle de la perception des voyelles "à double timbre".

Une conclusion plus assurée paraît être la mise en évidence d'une grande dispersion des valeurs formantiques pour chacun des timbres naturels examinés, par rapport aux données de la synthèse de la parole. Lorsque l'on constate que des mots comme "propriété" pour le E perçu comme fermé et "**alba**nais", pour le E perçu comme ouvert représentent chacune plusieurs valeurs acoustiques fort éloignées les unes des autres (tableaux 2 et 4), on voit que la parole naturelle est faite de ces fluctuations qui lui donnent le "moelleux" qui manque souvent à la parole artificielle.

Au plan linguistique, il faut rappeler que cette étude porte essentiellement sur les finales vocaliques. Si la parole de Pivot semble aussi clairement définie pour la répartition des E fermés et ouverts selon un modèle orthoépique très classique c'est, d'une part, parce que ces timbres ont été observés en position privilégiée, sous l'accent et, d'autre part, qu'on a affaire à un sujet à l'articulation soignée, habitué à parler en public. Il est bien certain en tout cas que l'étude des timbres inaccentués serait beaucoup plus difficile au plan de la perception et que, dans ce cas, la mesure acoustique des indices de perceptibilité et de différenciabilité pourrait aider de manière beaucoup plus nette à la détermination des timbres vocaliques.

BIBLIOGRAPHIE

BOTHOREL, A., SIMON, P., WIOLAND, F. ET J.P. ZERLING (1986) **Cinéradiographie des voyelles et des consonnes du français**, Strasbourg, Travaux de l'Institut de Phonétique.

DELATTRE, P. (1948) Un triangle acoustique des voyelles orales en français. **French Review**, XXI, May, p. 481.

DELATTRE, P. (1965) **Comparing the Phonetic Features of English, French, German and Spanish**, Philadelphia - New York, Julius Groos.

DURAND, P. (1985) **Variabilité acoustique et invariance en français: consonnes occlusives et voyelles**, Travaux de l'Institut de Phonétique d'Aix-en-Provence, vol. 4, Paris, Editions du CNRS.

LADEFOGED, P. (1967) **Three Aspects of Experimental Phonetics**, London, Oxford University Press.

LENNIG, M. (1978) "Une étude quantitative du changement linguistique dans le système vocalique parisien" in Thibault, Pierrette, dir. **Le français parlé: études sociolinguistiques**, Edmonton, Linguistic Research, 29-39.

LÉON, P. R. & J. TENNANT (1988), "Observations sur la variation phonétique et morphonologique dans **Apostrophes**", in **Information/Communication**, vol. 9, août, 20-47.

LÉON, P. R. & J. TENNANT (à paraître, 1990) "Indices de perceptibilité et de différenciabilité des timbres vocaliques: La variabilité [e] - [ɛ] en français", **Revue québécoise de linguistique**.

3 PRODUCTION

Président: L. SANTERRE
Université de Montréal,Canada

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

ETUDE D'UN MODELE DU SIGNAL DE SOURCE

Jean SCHOENTGEN*

Institut de Phonétique CP 110, Université Libre de Bruxelles, B-1050 Bruxelles

* Fonds National de la Recherche Scientifique, Belgique

RESUME

Nous proposons un modèle du signal de source basé sur une méthode de représentation du signal. Le modèle tient compte du caractère ponctuel de la source voisée (c.-à-d. du fait que ses dimensions caractéristiques sont faibles comparées à une longueur d'onde typique) et de la non-linéarité de son fonctionnement. Les paramètres de contrôle du modèle sont la période et l'amplitude d'un signal d'excitation cosinusoïdal. Nous justifions brièvement le choix du modèle et nous rappelons les relations entre les poids des fonctions de base et des coefficients de Fourier du signal périodique à représenter. Nous étudions ensuite la variation de quelques indices spectraux et temporels avec les paramètres de contrôle du modèle. L'objectif est de montrer comment générer, à partir du signal glottique d'un locuteur, un continuum de signaux qui diffèrent de l'original par le gain et le contenu spectral.

INTRODUCTION

Conventionnellement, le signal de source est modélisé en concaténant un faible nombre de segments de courbes afin d'approcher la forme de l'impulsion glottique (p. ex.: Fant, Liljencrants, 1985). Ce genre de modèle se prête à une synthèse par règles du signal glottique.

Nous avons proposé une alternative (Schoentgen, 1989) qui repose sur le codage de la forme d'une l'impulsion glottique sous l'aspect d'une combinaison linéaire d'un ensemble de fonctions de base. Un exemple très connu d'un tel type de représentation est le développement en série de Fourier d'un signal périodique.

Nous avons pris comme point de départ une représentation du type pôles-zéros d'un signal y(n); x(n) étant un signal d'excitation stéréotypé:

$$y(n) \approx - g_1\, y(n-1) - g_2\, y(n-2) - g_3\, y(n-3) - \ldots\ldots - g_p\, y(n-p) + h_0\, x(n) + h_1\, x(n-1) + h_2\, x(n-2) + \ldots + h_q\, x(n-q). \quad (1)$$

Les hypothèses sous-jacentes à ce genre d'expression sont que le signal y(n) est émis par un système *linéaire* au sein duquel il se propage en un temps *fini*. Ce qui veut dire que les dimensions caractéristiques du système linéaire sont de l'ordre d'une longueur d'onde typique.

Dans le cas de la source, aucune des deux hypothèses relatives à la production du signal n'est vérifiée. Le caractère ponctuel de la source voisée suggère en ce cas de supprimer dans (1) l'ensemble des termes réfléchis et de les remplacer par une somme de puissances de la fonction excitatrice x(n) afin de conférer à l'expression finale le caractère non-linéaire exigé:

$$y(n) \approx c_0 + c_1\, x(n) + c_2\, x^2(n) + c_3\, x^3(n) + . + c_N\, x^N(n). \quad (2)$$

La non-linéarité en x(n) du modèle évoque le fonctionnement non-linéaire du vibrateur laryngé. La non-linéarité de (2) a comme conséquence que la forme du signal produit y(n) dépend de l'amplitude du signal d'excitation x(n). En d'autres termes, le contenu spectral du signal de sortie évolue de façon automatique avec l'amplitude du signal. Le modèle partage cette propriété fondamentale avec l'ensemble des systèmes non-linéaires, le vibrateur laryngé inclus.

Les résultats présentés dans cet exposé constituent un premier pas vers une étude quantitative de la dynamique spectrale de cette famille de modèles. Il s'agit aussi d'établir un lien entre l'expression formelle (2) et des indices susceptibles d'être interprétés physiquement.

Nous avons ainsi modélisé les signaux de source de cinq locuteurs à l'aide d'un modèle non-linéaire et nous avons étudié l'évolution des valeurs de plusieurs indices avec les paramètres de contrôle. Nous montrons, plus précisément, que le rapport entre la pente spectrale et les quotients d'ouverture et d'asymétrie rappelle une loi empirique mise en évidence sur des signaux réels (Tagasuki, 1971; cité par Al-Ansari, 1981).

Nous énonçons ci-dessous les principales relations entre les coefficients de Fourier d'une part et les poids c_i d'autre part. Les relations permettent de déterminer les valeurs des coefficients du modèle pour un signal donné. Nous exposerons ailleurs la preuve mathématique de ces relations, de même que la comparaison des propriétés de ce genre de modèles du signal glottique par rapport au modèle de référence linéaire (1).

CALCUL

Les coefficients c_i peuvent être calculés à condition que le signal y(n) à représenter soit périodique pair et que le signal excitateur x(n) soit cosinusoïdal. On trouve alors la relation

suivante entre les coefficients de Fourier de y(n) d'une part et les coefficients inconnus c_i d'autre part:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \cdots\cdots \\ a_N \end{pmatrix} = 2\ M_e \begin{pmatrix} c_0 \\ c_1/2 \\ c_2/4 \\ c_3/8 \\ \cdots\cdots \\ c_N/2^N \end{pmatrix} \qquad (3)$$

M_e étant une matrice (N+1,N+1) constante égale à:

$$\begin{bmatrix} 1 & 0 & 2 & 0 & 6 & 0 & 20 & \cdots \\ 0 & 1 & 0 & 3 & 0 & 10 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 4 & 0 & 15 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 5 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & 6 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \cdots\cdots & & & & & & & \end{bmatrix}$$

Les coefficients a_j sont les coefficients de Fourier du signal y(n). Les éléments de la matrice sont identiques à ceux d'un demi-triangle de Pascal lorsque la colonne centrale du triangle est identifiée à la première ligne de la matrice.

Il reste à résoudre le cas où y(n) est périodique et impair. Le modèle précédent ne permet que le traitement des signaux pairs parce qu'une somme de puissances entières d'une fonction d'excitation x(n) paire est toujours paire.

En revanche, le modèle suivant tire parti du fait que le *quotient* de deux fonctions impaires est pair:

$$\frac{y(n)}{\sin(2\pi n\ f/fs)} \approx d_0 + d_1\ x(n) + d_2\ x^2(n) + d_3\ x^3(n) + \ldots + d_N\ x^N(n)$$

Les poids d_i inconnus sont calculés dans ce cas à l'aide de la relation suivante:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ \cdots\cdots \\ b_{N+1} \end{pmatrix} = M_o \begin{pmatrix} d_0 \\ d_1/2 \\ d_2/4 \\ d_3/8 \\ \cdots\cdots \\ d_N/2^N \end{pmatrix} \qquad (4)$$

M_o est une matrice constante, les d_i sont les coefficients inconnus et les b_i sont les coefficients de Fourier de y(n). M_o est égal à:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 2 & 0 & 5 & \cdots \\ 0 & 1 & 0 & 2 & 0 & 5 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 3 & 0 & 9 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 4 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & 5 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \cdots\cdots & & & & & & & \end{bmatrix}$$

Les éléments de la matrice sont obtenus par soustraction de deux colonnes adjacentes d'un triangle de Pascal et en retenant la branche positive du triangle anti-symétrique ainsi obtenu.

En toute généralité, le signal de source stationnaire peut être assimilé à un signal périodique mais qui n'est ni pair ni impair. Par conséquent, un modèle non-linéaire du signal glottique comporte deux branches (A et B), l'une modélisant la composante paire du signal et l'autre modélisant la composante impaire (figure 1).

Les relations (3) et (4) établissent un lien entre le spectre du signal à modéliser et les poids des fonctions de base x^i. L'amplitude de la fonction excitatrice cosinusoïdale influence directement le contenu spectral du signal émis par le modèle. Il s'agit d'une conséquence de la propriété suivante: un changement de l'amplitude A de la fonction excitatrice transforme les poids c_i et d_i respectivement en c_iA^i et d_iA^{i+1} (propriété i). Le contenu spectral du signal émis s'appauvrit donc lorsque A diminue; pour A = 0 la sortie du modèle est une constante; pour A = 1 le contenu spectral est celui du signal d'origine, éventuellement limité en sa largeur de bande B par le choix de l'ordre N du modèle (N = 40 et B ≈ 4000 Hz).

Par contre, les amplitudes des harmoniques ne sont pas affectées par la *fréquence* de la cosinusoïde excitatrice; par conséquent, l'enveloppe spectrale évolue avec la fréquence fondamentale.

METHODES ET MATERIAUX

Nous illustrons ci-après le fonctionnement du modèle sur base de signaux glottiques obtenus par filtrage inverse, pour cinq locuteurs masculins qui ont soutenu la voyelle [a] à une force sonore et à une hauteur confortables. Avant la modélisation, les formes d'onde ont été normalisées en amplitude; le minimum du signal est posé égal à zéro et le maximum égal à 2048.

Comme nous l'avons souligné, la dynamique spectrale est une propriété fondamentale qui départage systèmes linéaires et non-linéaires. Dans le cas de notre modèle la pente spectrale est sous contrôle direct de l'amplitude de l'excitation via la propriété (i) citée précédemment.

En toute généralité, les liens entre l'intensité, F0 et la forme d'ondes de débit réelles sont complexes (Baken, 1987). T. Tagasuki a mis en évidence une relation empirique entre pente spectrale d'une part et les quotients d'ouverture et d'asymétrie d'autre part:

$$\alpha = K\frac{QO}{QA}, \qquad (5)$$

avec K égal à -30.

La relation a été obtenue, rappelons-le, à partir de signaux réels.

Nous proposons de vérifier (5) dans le cadre de notre modèle. Pour ce faire nous calculons les signaux de sortie pour l'amplitude A de la cosinusoïde excitatrice admettant des valeurs entre 0.2 et 1 (par pas de $\frac{1}{10}$). Ensuite, nous estimons la pente spectrale, de même que les quotients d'ouverture et d'asymétrie. Les résultats obtenus pour 5 locuteurs sont alors reportés dans un diagramme de dispersion $(\alpha, \frac{QO}{QA})$.

L'indice α est défini comme le rapport exprimé en décibel entre la première et la deuxième harmonique. Le quotient d'ouverture est défini comme la durée pendant laquelle le signal est différent de zéro, divisée par la période. Le quotient d'asymétrie est égal au rapport entre les durées respectives des plages d'ouverture et de fermeture. En pratique, les durées sont calculées en ajustant deux droites, une entre l'instant d'ouverture et le maximum du signal, la deuxième entre le maximum et l'instant de fermeture. Les instants d'ouverture et de fermeture sont assimilés aux points d'intersections respectives des droites avec l'abscisse.

RESULTATS ET DISCUSSION

La figure 2 montre les signaux glottiques modélisés pour chacun des cinq locuteurs (de haut en bas: AS, JS, TT, MP, RV). La fréquence est de 100 Hz dans chaque cas et l'amplitude excitatrice est égale à 1, c.-à-d. que le spectre du signal émis est identique à celui du signal d'origine limité à 40 harmoniques.

Le diagramme de dispersion 3 montre les couples $(\alpha, \frac{QO}{QA})$ mesurés pour les cinq locuteurs.

Comparons les résultats obtenus à l'aide des modèles non-linéaire (2) et linéaire (1). Dans le cas linéaire la figure 3 se limiterait à un nuage de points contentrés à proximité de l'origine. La forme du signal est, en effet, la même quelle que soit l'amplitude du signal - une conséquence immédiate de la propriété de linéarité.

Dans le cas non-linéaire, la pente spectrale évolue avec l'amplitude entre -10 dB et -40 dB à peu près. Une pente plus raide signifie une glotte plus longtemps ouverte et un signal plus symétrique. La droite de régression ajustée à travers le nuage de points (figure 3) a comme équation:

$$\alpha = -31 \frac{QO}{QA} + 3.5. \qquad (6)$$

Cette loi est en accord avec la règle empirique (5). Nous en concluons que l'amplitude A de la cosinusoïde excitatrice permet de contrôler le contenu spectral du signal émis dont la forme évolue en accord avec une loi qui a été déterminée sur base de signaux réels.

CONCLUSION

Nous avons présenté un modèle formel du signal de source qui, contrairement à un modèle pôles-zéros, possède les propriétés générales requises pour rendre compte de la genèse non-linéaire du signal de source. Le rôle du modèle n'est pas d'entrer en compétition - dans le cadre de la synthèse par règles - avec des modèles existants qui ont un faible nombre de paramètres de contrôle et qui fonctionnent (idéalement) dans un cadre phonétique. Pourtant, le modèle a le potentiel pour reproduire une part de la variabilité de la forme glottale. Les résultats présentés ici en constituent une première illustration.

BIBLIOGRAPHIE

Al-Ansari A. (**1981**), "Etude du fonctionnement et simulation en temps réel d'un modèle de la source vocale", Thèse présentée à l'Institut National Polytechnique de Grenoble pour obtenir le grade de docteur-ingénieur

G. Fant, J. Liljencrants, Q. Lin (**1985**), " A four-parameter model of glottal flow", STL-QPSR, 4, pp 1-13

J. Schoentgen (**1989**), "The spectral dynamics of a non-linear model of the glottal waveform", Proceedings European Conference on Speech Communication and Technology, Paris, vol. 2, pp 481-484

Tagasuki T. (**1971**), Analysis by synthesis method, utilizing spectral features of voice source and measurement of glottal waveform parameters", J. of Rad. Res. Lab., 209-220

Baken R. J. (**1987**), "Clinical measurement of speech and voice", Taylor and Francis, London

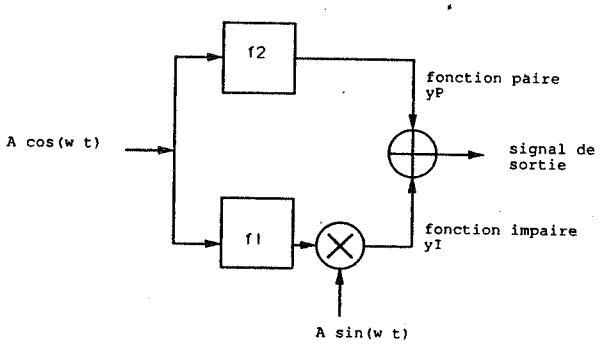


FIGURE 1

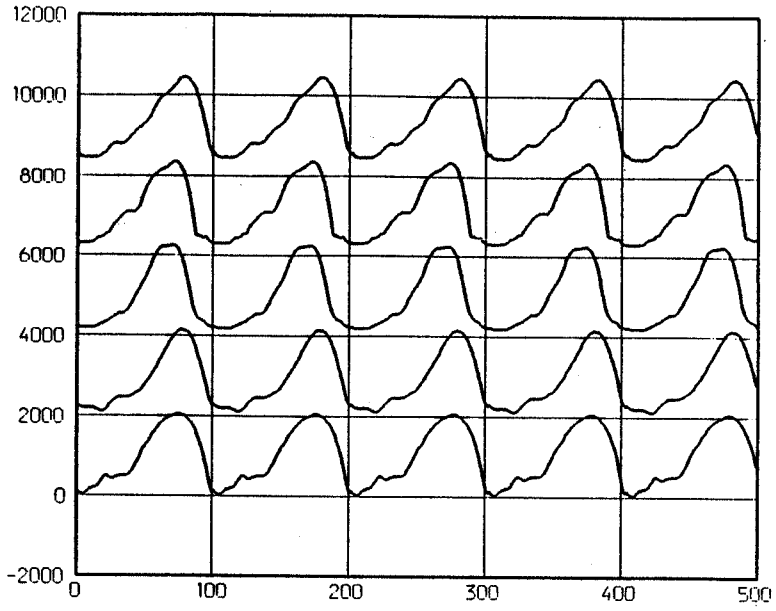
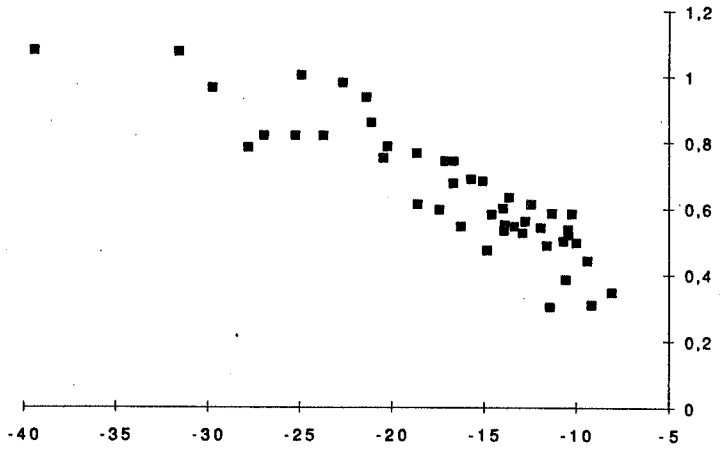


FIGURE 2

Figure 3: QO/QA en fonction de la pente (en dB)



XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

VERS UNE MESURE EN TEMPS REEL DE LA FONCTION DE TRANSFERT DU CONDUIT VOCAL

Amar DJERADI* Pascal PERRIER & Bernard GUERIN

Institut de la Communication Parlée.
INPG/ENSERG & UNIVERSITE STENDHAL
46 Avenue Félix Viallet - 38031 Grenoble Cedex - France
*Détaché de l'Institut d'Electronique de l'USTHB d'Alger - El-Allia - Bab Ezzouar - BP 32 - Alger- Algérie..

ABSTRACT

The measurements of the transfert function of the vocal tract are a very good tool in studying the acoustics of speech production.But the processes usually used for this aim present important limitations: 1) the speaker is told to hold the same articulation position during a long time; 2) the phonation is not possible.We propose here a new process, based on external excitation of the vocal tract through a pseudo random signal, which is not correlated with the speech signal.The process time does not exceed 100ms, and phonation is possible.First attempts to find vocalic and consonantic spectral characterisics are presented.

I. INTRODUCTION

Les mesures de fonctions de transfert du conduit vocal ont permis de faire progresser la connaissance des phénomènes acoustiques mis en jeu au cours de la phonation (cf. en particulier Van den Berg (1955) et Fant (1959) pour les productions vocaliques, et surtout Fujimura & Lindqvist (1971) et Castelli & Badin (1988) pour l'étude de la nasalité). Les techniques les plus récentes reposent sur une excitation extérieure du conduit vocal au niveau du larynx à l'aide d'un petit haut-parleur émettant un signal qui décrit l'ensemble du domaine spectral de la parole. Pour cela on a recours soit à un balayage en fréquence (sweep-tone de Fujimura & Lindqvist, 1971), soit à un bruit blanc dans la gamme de fréquence souhaitée (noise-tone de Castelli & Badin, 1988). Dans tous les cas le temps de la mesure est de l'ordre de quelques secondes, correspondant au temps nécessaire au balayage de la gamme de fréquence ou au calcul de l'estimation spectrale du signal acoustique par moyennage de transformées de Fourier. Or pour des résultats corrects il est nécessaire de maintenir la configuration articulatoire constante pendant toute la durée de la mesure : un léger mouvement du conduit vocal entrainerait en effet une perturbation de la fonction de transfert, ce qui risquerait de fausser la mesure des paramètres recherchés. Par ailleurs, ces techniques de mesure imposent un silence total au cours de l'expérience, et en particulier excluent toute possibilité de production effective d'un signal de parole. Ce dernier point, que l'on peut légitimement négliger dans le cas de mesure sur les voyelles, prend toute son importance dans la caractérisation acoustique des fricatives ; en effet la précision sur l'aire et la position de la constriction est dans ce cas tout à fait fondamentale, et cette précision du geste est en particulier liée à la sensation que procurent sur les articulateurs les turbulences de l'air en ce lieu.

Nous nous proposons dans ce travail de montrer qu'en introduisant une source d'excitation de type pseudo-aléatoire associée à un mode de calcul approprié de la fonction de transfert (technique empruntée à la mesure de l'acoustique des salles, Jullien et al.,1984) on obtient des résultats tout à fait fiables pour un temps de mesure de quelques centaines de millisecondes, avec une production effective du signal de parole.

II. PRINCIPE THEORIQUE DE LA MESURE.

II.1. Démarche théorique générale

Supposons que le conduit vocal se comporte dans les conditions d'expérience comme un filtre acoustique linéaire de réponse impulsionnelle h(t) (échantillons h(n)) et de fonction de transfert H(f) (échantillons H(k)). Excitons celui-ci par une source x(t) (échantillons x(n)) à laquelle se superpose un bruit b(t) (échantillons b(n)) non corrélé avec x(t). Nous obtenons ainsi en sortie un signal y(t) dont les échantillons y(n) sont donnés par la convolution numérique :

$$y(n) = [b(n) + x(n)] * h(n) \qquad (1)$$

soit R_{xy} la fonction d'intercorrelation entre l'excitation x(n) et y(n), on peut l'exprimer sous la forme suivante

$$R_{xy}(n) = R_1(n) + R_2(n) \qquad (2)$$

avec :

$$R_1(n) = \sum_{k=-\infty}^{k=\infty} h(k).\left[\sum_{m=-\infty}^{m=\infty} x(m) \; . \; x(m+n-k) \right] \qquad (3)$$

et

$$R_2(n) = \sum_{k=-\infty}^{k=\infty} h(k) \; . \left[\sum_{m=-\infty}^{m=\infty} x(m).b(m+n-k) \right] \qquad (4)$$

Donc si on pose φ_{xx} et φ_{xb} respectivement autocorrélation de x et intercorrélation de x et b, on peut écrire :

$$R_{xy}(n) = h(n) * \varphi_{xx}(n) + h(n) * \varphi_{xb}(n) \qquad (5)$$

Si on considère que les signaux x et b sont non corrélés, il vient:

$$R_{xy}(n) = h(n) * \varphi_{xx}(n) \qquad (6)$$

Le schéma équivalent de ce traitement est alors donné figure 1.

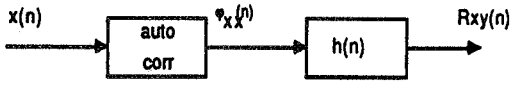


Fig 1. Schéma équivalent

On approchera donc d'autant mieux h(n) par la mesure de Rxy(n) que la fonction d'autocorrélation $\varphi_{xx}(n)$ est proche d'un Dirac. Ceci équivaut à dire que x(n) ait les caractéristiques proches de celles d'un bruit blanc. Une technique numérique (Jullien et al., 1984) consiste à approcher, dans une bande de fréquence donnée, un bruit blanc par une séquence numérique "pseudo-aléatoire",dont la fonction d'autocorrélation est proche, sur une période, de l'impulsion unité (Schroeder, 1979).

II.2. Excitation du conduit vocal par une séquence pseudo-aléatoire périodique.

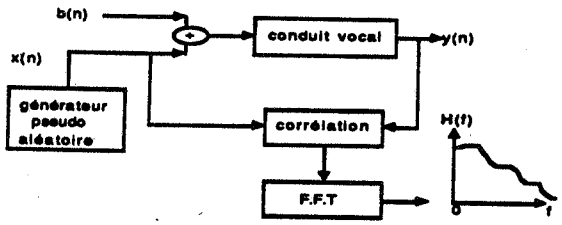


FIG.2. Schéma de principe de la mesure par excitation pseudo-aléatoire

Le signal x(n) étant périodique de période N, sa fonction d'autocorrélation est périodique et elle est donnée par :

$$\varphi_{xx}(n) = \sum_{m=0}^{m=N-1} x(m) \cdot x(m+n) \qquad (7)$$

et

$$\varphi_{xx}(n) = \varphi_{xxN}(n) * P_N(n) \qquad (8)$$

avec

$$\varphi_{xxN}(n) = \varphi_{xx}(n) \text{ pour } n \in [0, N-1] \qquad (9)$$

et

$\varphi_{xxN}(n) = 0$ ailleurs,

$P_N(n)$ étant le train d'impulsion unité de période N.

Par conséquent

$$R_{xy}(n) = \left[\varphi_{xxN}(n) * P_N(n)\right] * h(n) \qquad (10)$$

Supposons maintenant que h(n) est de longueur inférieure ou égale à N. Alors l'équation (13)peut s'écrire :

$$R_{xy}(n) = \varphi_{xxN}(n) * \sum_{k=-\infty}^{k=\infty} h(n-kN) \qquad (11)$$

Par ailleurs on montre (M.R.SCHROEDER, 1979) que si x(n) est pseudo-aléatoire de période N sa fonction d'autocorrélation ne prend que deux valeurs et on a :

$$\varphi_{xxN}(0) = N-1 \qquad (12)$$

$$\varphi_{xxN}(n) = -1 \qquad n \in]0, N-1] \qquad (13)$$

Nous pouvons donc écrire que :

$$R_{xy}(n) = (N-1) \cdot \sum_{k=-\infty}^{k=\infty} h(n-kN) - \sum_{p=1}^{N-1} \left(\sum_{k=-\infty}^{k=\infty} h(n-p-kN) \right) \qquad (14)$$

$R_{xy}(n)$ est donc périodique, et h(n) étant de longueur N, on a pour $0 \le n \le N-1$.

$$R_{xy}(n) = N \cdot h(n) - \sum_{p=0}^{N-1} h(p) \qquad (15)$$

Le second terme de cette différence est, pour un système donné, une constante Cst représentant en fait, au facteur N près, la valeur moyenne de la réponse impulsionnelle.

d'où

$$h(n) = (R_{xy}(n) + Cst) / N \qquad (16)$$

En prenant la T. F. Discrète sur N points de l'équation (16), on trouve

$$R_{xy}(k) / N = H(k) \qquad \text{pour tout } k \neq 0$$

$$R_{xy}(0) / N - Cst / N = H(0) \qquad \text{pour } k=0$$

On voit que mis à part en continu (k=0), $R_{xy}(k)$, TFD de l'intercorrélation circulaire sur N points de x et y, est une image exacte (à un facteur près 1/N) de la TFD sur N points de la réponse impulsionnelle h du filtre acoustique analysé.

III. EVALUATION THEORIQUE DE LA METHODE DE MESURE PAR SIMULATION

III.1. Principe de la simulation

Le schéma de principe est le suivant :

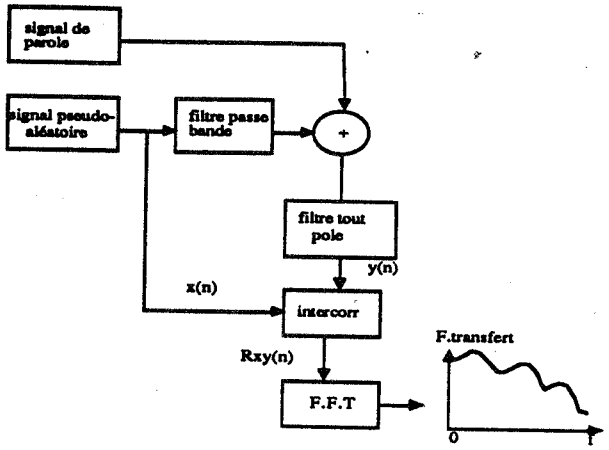


Fig.3. Schéma synoptique

Pour réaliser cette simulation, il faut d'abord disposer du signal pseudo-aléatoire. Ce signal x(n) est construit à partir d'une suite de nombres binaires, constituant un champ de nombre fini nommé 'champ de Galois '(Schroeder 1979). Le conduit vocal est excité par un petit haut parleur dont la réponse spectrale en charge doit être plate dans la bande 50 à 8000Hz. Nous simulons celui-ci à l'aide d'un filtre passe bande. Le conduit vocal est simulé ensuite selon le modèle linéaire de production de la parole sous forme d'une structure série de filtres résonants du second ordre (Gold & Rabiner, 1968).

III.2. Détermination de la longueur de la séquence pseudo-aléatoire

Soit $\Phi_{xx}(k)$ la TFD de $\varphi_{xx}(n)$, c'est-à-dire la densité spectrale de puissance numérique de la séquence pseudo-aléatoire x(n). $\Phi_{xx}(k)$ vaut :

$\Phi_{xx}(0) = 1$ pour $k = L.N$ et $\Phi_{xx}(k) = N+1$ pour $k \neq L.N$

Un des problèmes posé par la conversion numérique-analogique réside dans les modifications des caractéristiques spectrales du signal d'excitation. En effet après conversion par interpolation d'ordre 0 à fréquence Fe (=1/Te) la densité spectrale de puissance $\Phi_{xxa}(f)$ du signal analogique excitant le haut-parleur est : $\Phi_{xxa}(f) = (Te.\mathrm{sinc}(\pi.Te.f))^2 \cdot \Phi_{xx}(f)$

Soit f_{max} la largeur de bande dans laquelle on souhaite que le spectre garde une amplitude constante à -3dB près. Il faut dans ces conditions que : $\mathrm{sinc}(f_{max}/f_e) > 1/2$ soit encore que : $f_e > f_{max}/0,6$.

Ces conditions étant en-decà des conditions élémentaires de Shannon, on peut légitimement considérer que le signal d'excitation après conversion-numérique garde un spectre d'amplitude constante.

La longueur de la séquence déterminera la périodicité du signal d'excitation et donc la résolution spectrale sur la fonction de transfert mesurée. Ainsi dans le cadre de ce papier qui se limite à une étude de quelques voyelles, la largeur spectrale intéressante est égale à 5 kHz, et une résolution spectrale de l'ordre de 10 Hz est suffisante. Nous prendrons ainsi F_e=10kHz, et la durée de la séquence sera de l'ordre de 1023 ($N = 2^m-1$) .

III.3. Cas d'une excitation uniquement pseudo-aléatoire.

La simulation va permettre de retrouver la fonction de transfert du conduit vocal, représenté par un filtre tout pôle.
Le schéma de la mesure est le suivant

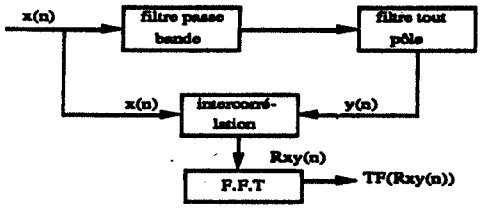


Fig.4. Synoptique de la mesure

On excite le filtre tout pôle que l'on veut caractériser, à travers le filtre passe bande, par le signal pseudo-aléatoire x(n). La figure 5 donne le résultat du calcul de la transformée de Fourier de l'intercorrélation $R_{xy}(n)$.
Il apparaît clairement que l'on retrouve bien la fonction de transfert du filtre tout pôle.

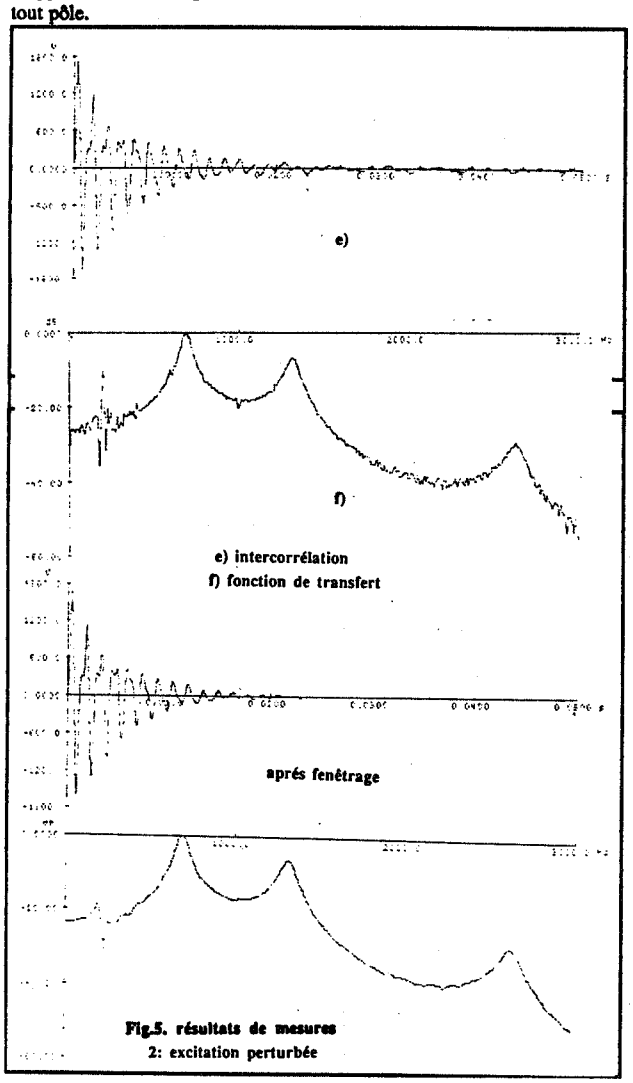


Fig.5. résultats de mesures
2: excitation perturbée

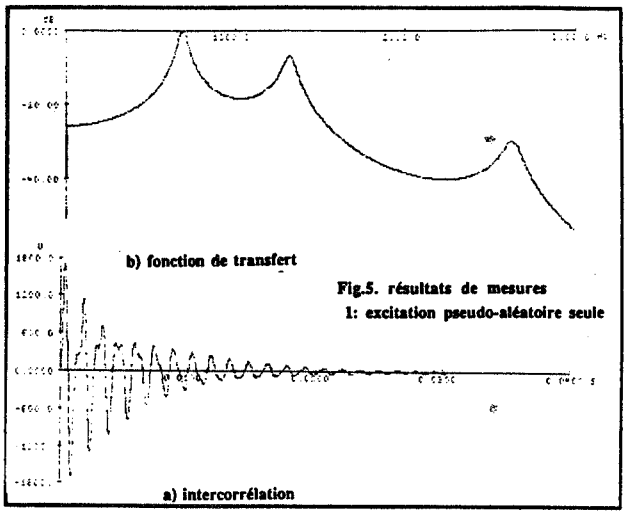


Fig.5. résultats de mesures
1: excitation pseudo-aléatoire seule

III.4. Cas d'une excitation somme d'une séquence pseudo-aléatoire et d'un bruit non corrélés

Supposons maintenant qu'au signal pseudo-aléatoire se superpose un signal parasite non corrélé avec celui-là (signal glottique par exemple). La mesure se fait alors selon le schéma suivant :

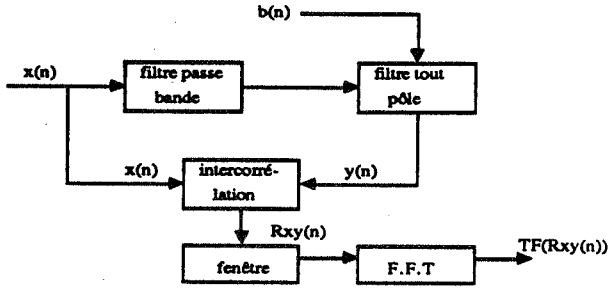


Fig.6.Schéma de mesure dans le cas du signal pseudo-aléatoire perturbé

Le filtre est excité par deux signaux, l'un représentant le signal utile (séquence pseudo-aléatoire) et l'autre simulant le signal glottique (perturbateur). la même technique de calcul que celle du (§.III.3) doit nous fournir théoriquement la fonction de transfert du filtre tout pôle. Selon le schéma équivalent donné figure 1, la sortie du filtre de réponse impulsionnelle h(n) est la superposition de 2 "réponses" du filtre : l'une, correspondant exactement à celle que nous avons trouvée dans le cas précédent et donc à la réponse impulsionnelle de ce filtre, s'atténue au bout d'une certaine durée D ; l'autre, provenant d'une intercorrélation non rigoureusement nulle entre x(n) et le signal parasite, a une amplitude maximale à peu près constante mais faible relativement au maximum d'amplitude de la réponse impulsionnelle, et elle n'apparait vraiment que lorsque la réponse impulsionnelle est suffisamment amortie. Pour garder après TFD une résolution spectrale de 10 Hz, nous devons analyser ce signal sur une durée de 100 ms. Dans ces conditions la perturbation liée à l'intercorrélation non nulle se manifeste par l'apparition d'un pic de résonance à la fréquence du signal perturbateur. Or de manière générale la réponse impulsionnelle du filtre est inférieure à 100 ms. Nous avons donc choisi de pondérer l'intercorrélation entre x(n) et y(n) par une fenêtre de Hanning pour atténuer l'amplitude de ce pic . Mais ceci va aussi modifier les valeurs des bandes passantes. Pour évaluer les conséquences de ce traitement sur les bandes passantes, nous avons observé les effets de 2 fenêtres, l'une de largeur 30 ms et l'autre de 50 ms . Cette étude a été menée pour quatre valeurs de formants, chacun de ces formants pouvant avoir quatre bandes passantes. Elle a montré qu'en moyenne les calculs sur la fenêtre de 50 ms induisent un écart de l'ordre 7 Hz, erreur inférieure donc à la résolution, alors que pour la fenêtre de 30 ms, l'erreur est de l'ordre de 30

Hz (Fig.6), et ceci quelque soit la fréquence centrale du formant.

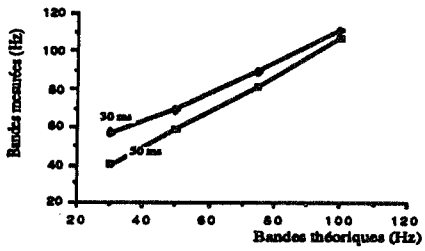


Fig.7. Variation des bandes en fonction de la durée de la fenêtre

Nous voyons donc que l'erreur introduite, dans la mesure des bandes, par l'utilisation d'une fenêtre de Hanning de durée 50 ms est tout à fait acceptable, et que ce traitement permet une amélioration de la précision sur la mesure la fonction de transfert.

III.4. Identification de cinq fonctions de transfert vocaliques.

Par simulation nous avons déterminé les fonctions de transfert et mesuré les paramètres formantiques des voyelles [i, e, a, o, u]. Dans cette manipulation, l'excitation est constituée du signal pseudo-aléatoire et d'un signal de type glottique de période 130 Hz. Le tableau ci-dessous montre bien que les mesures des bandes passantes et des formants concordent bien avec les valeurs théoriques.

voyelle	F1		B1		F2		B2		F3		B3	
	a	b	a	b	a	b	a	b	a	b	a	b
i	246	248	82	85	2285	2285	50	52	3134	3130	245	250
e	423	425	50	52	2139	2139	103	102	2727	2727	261	263
a	685	685	50	53	1317	1316	73	75	2629	2629	101	102
o	406	406	50	55	806	803	50	52	2458	2458	50	53
u	297	300	81	82	790	790	50	51	2322	2330	50	52

Tableau.1.résultats obtenus
a : valeurs théoriques
b : valeurs mesurées par simulation de la méthode.

IV. PREMIERES EVALUATIONS PRATIQUES : mesures sur un locuteur de fonctions de transfert vocaliques et consonantiques.

Cette méthode a été implantée sur IBM PC. La conversion numérique-analogique du signal, ainsi que le calcul de la fonction d'autocorrélation, sont effectués sur une carte conçue dans notre laboratoire (Verdier , 1989) autour d'un processeur de signal TMS 320C25 et placée sur le bus IBM PC.

IV.1 Dispositif expérimental

L'excitation du conduit vocal est assurée de l'extérieur du conduit vocal par un petit haut parleur de 3 cm de diamètre, enrobé d'une boule de pâte à modeler afin d'en diminuer les rayonnements parasites. Le locuteur applique ce haut parleur directement contre la peau de son cou au niveau de la pomme d'Adam. Un microphone directionnel capte à la sortie du conduit vocal le signal modulé par les cavités supraglottiques. Le signal est ensuite numérisé et stocké sur disque dur. Il est ensuite traité en temps différé par notre logiciel implanté sur IBM PC.

Le processus expérimental prévoit la mesure de fonctions de transfert avec ou sans phonation effective. Dans le premier cas le locuteur n'a aucune difficulté à positionner correctement ses articulateurs, puisqu'il dispose d'un feedback auditif naturel. Dans le second cas, il est nécessaire de lui fournir un moyen de contrôle auditif : ceci est assuré par une phase d'excitation par un bruit blanc du conduit vocal, le signal capté par le microphone étant alors renvoyé par un casque vers le locuteur ; après cette phase de positionnement des articulateurs, on commute sur l'excitation pseudo-aléatoire. Les deux signaux d'excitation sont générés par programme et la carte numérique assure la conversion numérique-analogique ainsi que la synchronisation entre excitation et signal recueilli sur le microphone.

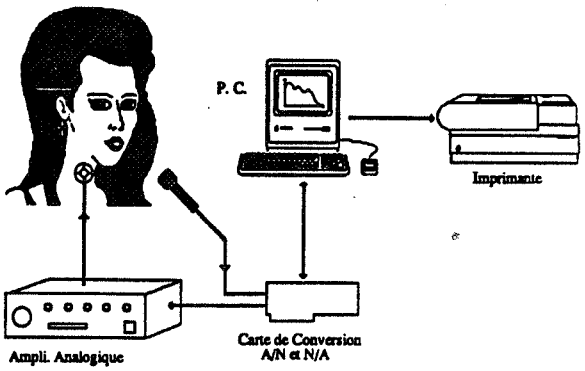


FIGURE 8 : Dispositif expérimental

IV.2 Premières mesures.

L'objectif que nous recherchons par ces premières mesures est double : (1) validation de la technique de mesure par évaluation des caractéristiques formantiques et comparaisons aux valeurs traditionnellement reconnues dans la littérature ; (2) comparaison des résultats obtenus selon qu'il y a ou non phonation effective.

Le locuteur analysé est un locuteur masculin de langue française. Les enregistrements ont été effectués au laboratoire, hors chambre sourde. En l'absence de phonation la mesure est effectuée à glotte ouverte. Les fenêtres d'analyse sont dans tous les cas d'une durée de 30 ms. Précisons dès maintenant que les mesures avec/sans phonation ont toujours été faites après un repositionnement des articulateurs : la configuration articulatoire n'est donc jamais strictement identique. La figure 9 présente les résultats obtenus pour des configurations articulatoires correspondant à [i] et [u] sans phonation (Fig. 9.a) et avec phonation (Fig. 9.b).

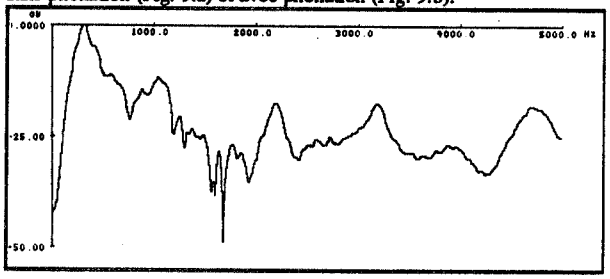


fig.9.a1.[i] sans phonation

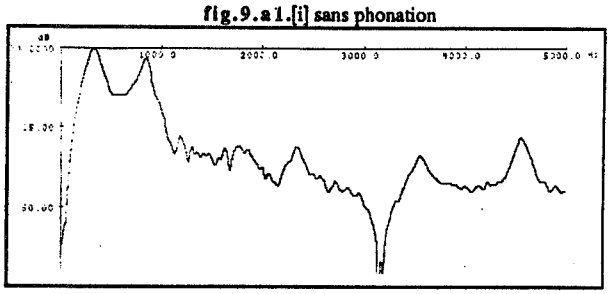


fig.9.a1.[u] sans phonation

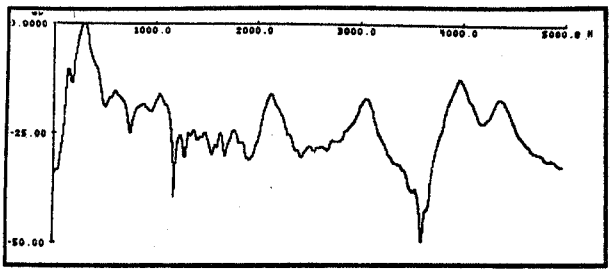


fig.9.b1.[i] avec phonation

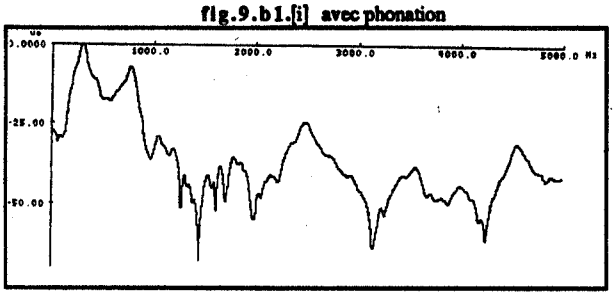


fig.9.b2.[u] avec phonation

La première observation que l'on peut faire est que les valeurs des formants sont dans les deux cas en accord avec les valeurs classiquement connues pour ces voyelles (cf. tableau 1 ci-dessus) :

* pour la voyelle [i] :

F1 ≃ 320 Hz, F2 ≃ 2150 Hz, F3 ≃ 3100 Hz ;

* pour la voyelle [u] :

F1 ≃ 300 Hz, F2 ≃ 800 Hz, F3 ≃ 2400 Hz .

Ceci atteste donc bien de la capacité du système à déterminer correctement les fréquences des trois premiers formants d'une configuration articulatoire vocalique.

La comparaison des fréquences formantiques obtenues avec et sans phonation donne :

* pour la voyelle [i] :

F1 = 315 Hz vs 333 Hz ; F2 = 2111 Hz vs 2185 Hz ;

F3 = 3037 Hz vs 3185 ;

* pour la voyelle [u] :

F1 = 296 Hz vs 296 Hz ; F2 = 778 Hz vs 852 Hz ;

F3 = 2481 Hz vs 2333 Hz;

Compte tenu des différences de configurations articulatoires inhérentes au déroulement de la mesure décrit ci-dessus, les différences observées peuvent légitemement être considérées comme non significatives.

La figure 10 présente les résultats obtenus pour une configuration articulatoire correspondant à la fricative [ʃ] sans phonation (Fig. 10.a) et avec phonation (Fig. 10.b). Il est ici plus difficile d'évaluer dans l'absolu l'exactitude de la mesure, car les connaissances des fonctions de transfert consonantique sont encore limitées. On peut cependant noter la bonne similitude des courbes pour les deux cas étudiés.

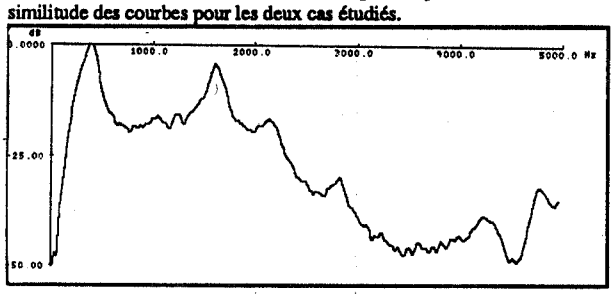


fig.10.a.[ʃ] sans phonation

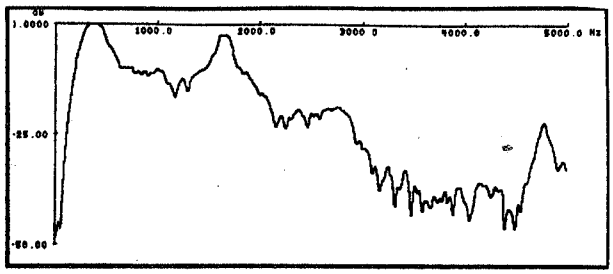


fig.10.b.[ʃ]avec phonation

Au delà de ces premiers résultats encourageants, il faut noter un certain nombre de points qui devront être analysés puis améliorés : (1) les résonances hautes du conduit vocal (au dessus de 2 kHz) émergent peu d'un niveau moyen anormalement élevé (≃ - 40 dB) ; (2) un pic lié à l'excitation glottique périodique (cf. III.4) apparaît systématiquement entre 100 et 300 Hz, et sa présence peut nuire à la détermination de faibles résonances du conduit vocal telles que celles qu'ont relevées Castelli et al. (1989).

LES CONCLUSIONS

Les résultats de simulation, puis de mesure effective sur un locuteur, montrent que la méthode de mesure de fonctions de transfert du conduit vocal par excitation pseudo-aléatoire est fiable.

Mais nos mesures ont permis aussi de mettre en avant certaines limitations à cette technique. En effet l'existence d'un signal parasite, lié à une intercorrélation non rigoureusement nulle entre le signal pseudo-aléatoire et les signaux extérieurs (signal glottique, bruit ambiant...), induit des perturbations sur la fonction de transfert mesurée. L'usage de fenêtres de troncature temporelle permet d'atténuer ces perturbations, sans modifier de manière sensible les caractéristiques formantiques. Dans ces conditions les mesures avec et sans phonation donnent des résultats proches. Mais ceci ne pourra être définitivement confirmé que par une analyse d'un corpus de parole plus large et sur un nombre plus important de locuteurs.

L'analyse avec une résolution spectrale de 10 Hz nécessite une excitation de durée égale à100 ms. Nous pouvons donc faire des estimations de la fonction de transfert du conduit vocal avec des temps de mesure incomparablement plus courts que ceux qui sont généralement nécessaires dans les méthodes classiques.

Le conditionnement analogique des signaux utilisé dans notre processus expérimental doit être sensiblement amélioré du point de vue du rapport signal/bruit dans les hautes fréquences, ce qui permettrait une meilleure évaluation des amplitudes des formants élevés de la fonction de transfert.

REFERENCES :

GOLD B. & RABINER L.R. (1968).Theory and application of digital signal processing, Prentice hall,London.

CASTELLI E. & BADIN P. (1988), Mesure des fonctions de transfert du conduit vocal,17ième. J.E.P. (SFA), 189-196.

CASTELLI E., PERRIER P. & BADIN P. (1989) Caractérisation de la nasalité, bull L.C.P,vol3,187-212.

DE COULON F. (1984), Théorie et traitement du signal, Traité d'Electricité, vol VI, Presses Polytechniques Romandes.

FANT G. (1959), The acoustics of speech, 3th Int. Cong. Acoustics, 180-201.

FUJIMURA O. & LINDQVIST J. (1971), Sweep-tone Measurement of Vocal-Tract Characteristics, J. Acoust. Soc.Am. 19, 511-558.

JULLIEN J.P,GILLOIRE A & SALIOU A. (1984), Mesure de réponse impulsionnelle en acoustique, Note technique NT/LAA/TSS/181 CNET-LANNION.

SCHROEDER M.R. (1979), Number Theorie in Science and Communication,Berlin(1979, 248 à 258.

VAN DEN BERG J.W. (1955), Transmission of the Vocal Cavities, J. Acoust. Soc. Am. 27,161-168

VERDIER P. (1989), Carte d'entrée sortie vocale, mémoire DEA (I.N.P.Grenoble).

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

CARACTERISTIQUES ACOUSTIQUES DE LA DYSARTHRIE DANS LA MALADIE DE FRIEDREICH *

Michèle GENTIL

CHU Pitié-Salpêtrière, URA CNRS 385

The speech of 14 patients with Friedreich disease was studied using acoustic analyses of repeated syllables, sentences and sustained vowels. Marked abnormalities were observed as irregular and inappropriate changes in fundamental frequency and intensity, inconsistency of segment durations (particularly in syllable repetitions), slowness of speech. These speech characteristics reflected impairments of the speech motor control system and are discussed in relation to speech in ataxic dysarthria.

INTRODUCTION

La parole est produite par les mouvements coordonnés des différents organes articulatoires. Des lésions affectant le système nerveux central peuvent perturber le contrôle de production de la parole et être à l'origine de patterns acoustiques anormaux. Ces altérations de la parole dues à des problèmes de contrôle sont regroupées sous le nom de dysarthrie (cf la définition de Darley et al., 1975). Les premières études concernant les dysarthries ont été des analyses perceptuelles qui ont essayé de classer les divers types de dysarthrie (Darley et al., 1969a et b ; Hirose, 1973). L'observation du signal acoustique est une autre approche qui apporte une meilleure connaissance de la dysarthrie. La présente étude concerne la parole de patients atteints de la maladie de Friedreich. C'est une hérédo-dégénérescence des voies spino-cérebelleuses qui évolue progressivement. Peu étudiée jusqu'ici, la parole de ces malades a fait l'objet de descriptions qui donnent les principales caractéristiques suivantes : voix saccadée, explosive avec des changements brutaux de hauteur, troubles articulatoires et respiratoires (Hiller, 1929 ; Dejong, 1979 ; Joanette et al. 1980 ; Murray, 1983 ; Gilman et al., 1985). A partir des observations des auteurs précités, il apparait que la phonation et prosodie sont spécialement affectées, le but de notre étude est donc d'estimer ces caractéristiques au moyen d'une analyse quantitative des paramètres acoustiques de durée, fréquence fondamentale et intensité dans divers exercices de parole.

METHODE

Sujets

14 patients atteints de maladie de Friedreich (10 femmes et 4 hommes) et 3 sujets contrôle normaux (2 femmes et 1 homme) participaient à l'étude. L'âge moyen des patients était 37,1 ans avec une SD de 11,9. La durée moyenne des symptomes au moment de l'examen était 18,4 ans (SD 5,1). La plupart des patients ne se déplaçaient qu'en chaise roulante. Les anomalies de parole étaient jugées perceptuellement de modéré à sévère. Aucune rééducation orthophonique n'avait été jusqu'alors envisagée. Les sujets normaux avaient un âge moyen de 27 ans avec une SD de 3,46.

Enregistrement des données

L'enregistrement de la parole des sujets a été fait en chambre sourde en utilisant un microphone LEM maintenu à une distance constante de 10 cms des lèvres. Le signal acoustique était simultanément enregistré par un magnétophone Revox type A 77 à la vitesse de 19 cms et par un traceur Gould type ES 1000 à la vitesse de 50 mm/s.

Corpus

Les sujets avaient à réaliser quatre différents types d'exercices.

1 - Répéter 7 fois six monosyllabes (consonne-voyelle), /ba/, /pa/, /fa/, /ka/, /pu/, /fu/ à 2 vitesses de parole, conversationnelle et rapide. Ces syllabes étaient destinées à mesurer la durée, fréquence fondamentale et amplitude relative.

2 - Produire les voyelles tenues /i/, /a/, /u/ qui devaient servir également à mesurer fréquence fondamentale et amplitude mais pour une aperture déterminée au cours d'une longue durée.

3 - Produire 3 mots "bas, basse, bassement" à une vitesse conversationnelle, ceci afin de déterminer si les patients atteints de Friedreich ajustent la durée de la syllabe au nombre de syllabes dans le mot.

4 - La tâche finale consistait à lire 8 phrases (précisées ultérieurement) en vue de comparer les durées entre patients et sujets normaux.

Traitement des données

Le traitement des données était effectué sur ordinateur PDP 11/73. On entrait les données en utilisant une fenêtre de 2,8 secondes. L'extraction de la fréquence fondamentale était faite suivant

modèle de prédiction linéaire. Fréquence fondamentale et intensité calculées étaient ensuite visualisées sur écran en même temps que la représentation graphique. Les durées des productions étaient mesurées directement sur l'enregistrement papier fait pendant l'examen.

RESULTATS

Variabilité des durées

La variabilité intraindividuelle des durées segmentales au cours des diverses répétitions d'une même production est un élément d'appréciation du contrôle moteur : une grande variabilité étant significative des défauts de ce contrôle. Les standards déviations ont donc été calculées pour les durées des voyelles /a/ et /u/ dans les syllabes répétées /pa/, /fa/, /ka/, /pu/, /ku/ à 2 vitesses de parole. Ces valeurs figurent dans le tableau 1 qui montre des standards déviations supérieures pour les patients indiquant une plus grande variabilité des durées. La différence des valeurs entre patients et sujets normaux a été estimée significative par une analyse de variance.

Cependant la variabilité des durées peut dépendre de la durée moyenne de production (Lehiste, 1972). Nous avons pour le /a/ de la syllabe /ba/ comparé le contrôle de la durée pour les patients et sujets normaux au moyen d'un graphe (figure 1) qui présente la durée moyenne et standard déviation. Les cercles pleins et vides représentent les résultats respectifs des patients et sujets normaux En ce qui concerne les patients, tous les cercles sont à l'extérieur d'un rectangle renfermant les cercles correspondant aux sujets normaux. La majorité des résultats des patients sont décalés vers le haut et la droite par rapport à ceux des sujets normaux. Si quelques résultats des patients se trouvent à l'intérieur des durées moyennes des sujets normaux, les standards déviations, en tout état de cause, sont nettement supérieures.

Variabilité de la fréquence fondamentale et intensité

Dans cette étude, on note pour tous les patients des changements brutaux de la fréquence fondamentale et de l'intensité au cours des voyelles tenues /a/, /i/, /u/. La figure 2 montre le signal audio, les contours de la fréquence fondamentale et intensité correspondant à la voyelle tenue /u/ pour le sujet normal, C2. La figure 3 montre les mêmes signaux pour le patient P1. Pour C2 (fig. 2), la fréquence fondamentale est représentée par une ligne horizontale qui reflète la constance du Fo autour de 260 Hz. L'intensité est représentée par une ligne qui descend lentement et régulièrement. Au contraire, pour le patient P1, les contours du Fo et de l'intensité (fig. 3) indiquent des changements brutaux de ces paramètres. Cette variabilité est observée chez tous les patients.

Par ailleurs, la variabilité intraindividuelle du Fo et de l'intensité est notée entre les diverses répétitions d'une même production.

Ajustement de la durée au nombre de syllabes dans le mot

La série "bas, basse, bassement" était utilisée pour déterminer si les patients atteints de Friedreich ajustent la durée de la syllabe au nombre de syllabes dans le mot. On sait que la durée de la syllabe de base décroit lorsque le nombre de syllabes augmente de 1 à 3 (Lehiste, 1972). Dans notre étude, les sujets normaux réduisent la syllabe de base comme prévu. De même les patients ajustent la durée de la syllabe de base quand le nombre de syllabes augmentent. Le tableau 2 résume ces résultats. Bien qu'il existe une différence d'ajustement entre patients et sujets normaux, cette différence n'est pas significative.

Durée moyenne des phrases

Le tableau 3 contient les durées moyenne des phrases suivantes : 1 - J'ai passé des vacances à Chicago. 2 - Vous cachez ce cas. 3 - Jean joue à cache-cache. 4 - Papa et maman hâtent le pas. 5 - Cachez-vous ce cas ? 6 - C'est une affaire intéressante. Qu'en pensez-vous ? 7 - Il faut la faire sans aucun regret. 8 - Mets tes beaux habits. Les différences entre patients et sujets normaux sont tout à fait importantes. Les durées moyennes des phrases des patients dépassent celles des phrases des sujets normaux, quelquefois de 2 standards déviations. Par ailleurs, une analyse de variance nous a permis d'estimer significative la différence existant entre durées moyennes des phrases des patients et sujets normaux.

COMMENTAIRE

Cette étude met en évidence les troubles de la parole dans la maladie de Friedreich, à savoir des variations excessives dans les paramètres acoustiques de fréquence fondamentale, intensité et durée, de même qu'une augmentation significative de la durée des segments, syllabes ou phrases. Ces anomalies montrent que la régulation des aspects dynamiques de la phonation est affectée. Cette dysarthrie a des caractéristiques communes avec la dysarthrie ataxique, les lésions cérebelleuses perturbent le contrôle de la fréquence fondamentale, de l'intensité et de la durée (Hirose et al., 1978 ; Kent et al., 1975, 1979 ; Larson et al., 1978). Par ailleurs, dans la dysarthrie ataxique les durées des syllabes sont augmentées, de même que la fréquence et durée des pauses entre les syllabes (Charcot, 1877 ; Kent et al., 1975). Dans notre investigation, un allongement important de la durée des phrases apparait pour les patients en comparaison avec les sujets normaux. La lenteur de la parole est donc un trait commun. Cependant dans la dysarthrie cérebelleuse, il y a à la fois un ralentissement de la parole et des altérations des relations temporelles entre les syllabes des mots. Notamment dans la production de mots qui normalement demande une réduction systématique de la syllabe de base, les ataxiques cérébelleux montrent des réductions inconsidérées, voire même des allongements (Kent et al., 1975). Au contraire, nos données montrent que les Friedreich ajustent la syllabe de base lorsqu'il est nécessaire de le

faire. Ainsi il semble qu'une lésion cérébelleuse proprement dite empêche le fin contrôle exigé par une réduction appropriée de la syllabe de base. Probablement c'est le cervelet lui-même qui est impliqué dans ce contrôle précis.

REFERENCES

CHARCOT, J.M. : Lectures on the disease of the nervous system, vol I (The New Sydenham, London 1877).

DARLEY, F.L. ; ARONSON, A.E. ; BROWN, J.R. : Differential diagnostic patterns of dysarthria. J. Speech Hear. Res. 12 : 246-269 (1969a).

DARLEY, F.L. ; ARONSON, A.E. ; BROWN, J.R. : Clusters of deviant speech dimensions in dysarthrias. J. Speech Hear. Res. 12 : 462-496 (1969b).

DEJONG, R.N. : The neurologic examination. (Harper & Row, Hagerstown, 1979).

GILMAN, S. ; KLUIN, K. : Perceptual analysis of speech disorders in Friedreich disease and olivopontocerebellar atrophy. In J.R. Bloedel, J. Dichgans, W. Precht (Eds), Cerebellar Functions. (Springer Verlag Berlin, Heidelberg, 1985) : 148-163.

HILLER, H. : A study of speech disorders in Friedreich's ataxia. Arch. Neurol. Psychiat. 22 : 75-90 (1929).

HIROSE, H. : Toward a differential diagnosis of the dysarthrias. In I. Kirikae (Ed), Approaches to the disorders of the central nervous system. (Kanehara, Tokyo, 1973) : 214-232.

HIROSE, H. ; KIRITANI, S. ; USHIJIMA, T. ; SAWASHIMA, M. : Analysis of abnormal articulatory dynamics in two dysarthric patients. J. Speech Hear. Dis. 43 : 96-105 (1978).

JOANETTE, Y. ; DUDLEY, J.G. : Dysarthric symptomatology of Friedreich ' s ataxia. Brain Lang. 10 : 39-50 (1980)

KENT, R.D. ; NETSELL, R. : A case study of an ataxic dysarthric : cineradiographic and spectrographic observations. J. Speech Hear. Dis. 40 : 115-134.

KENT, R.D. ; NETSELL, R. : ABBS, J.H. : Acoustic characteristics of dysarthria associated with cerebellar disease. J. Speech Hear. Res. 22 : 627-648 (1979).

LARSON, C.R. ; SUTTON, D. ; LINDEMAN, R.C. : Cerebellar regulation of phonation in rhesus monkey (macaca mulatta). Exper. Brain Res. 33 : 1-18 (1978).

LEHISTE, I. : The timing of utterances and linguistic boundaries. JASA 51 : 2018-2024 (1972).

MURRAY, T. : Friedreich ' s ataxia. In W. Perkins (Ed) Current therapy of communication disorders - dysarthria and apraxia. (Thieme-Stratton, New-York, 1983).

DARLEY, F.L. ; ARONSON, A.E. ; BROWN, J.R. : Motor speech disorders (W.B. Saunders, Philadelphia, 1975).

SUJETS	VITESSE ARTICULATOIRE						VITESSE RAPIDE			
	F0 (/pa/)	F0 (/fa/)	F0 (/ka/)	F0 (/pa/)	F0 (/ka/)	F0 (/pa/)	F0 (/fa/)	F0 (/ka/)	F0 (/pa/)	F0 (/ka/)
P 1	73	60	84	98	86	58	15	28	41	80
P 2	34	76	79	15	37	41	25	48	27	16
P 3	31	66	18	24	24	110	56	52	55	42
P 4	25	46	104	119	72	70	40	22	19	22
P 5	16	37	43	56	41	34	45	30	74	41
P 6	37	30	68	26	77	61	39	11	11	20
P 7	30	44	20	40	52	35	8	30	85	41
P 8	56	36	84	31	37	15	37	43	30	20
P 9	14	67	49	88	36	24	66	30	43	19
P10	54	188	41	169	29					
P11	72	24	66	61	37	32	45	74	40	72
P12	36	73	98	30	23	66	55	50	45	39
P13	126	113	125	89	46	11	41	158		62
P14	33	78	76	73	45	122	121		80	
C 1	15	10	15	19	21	19	10	24	17	16
C 2	0	16	0	15	8	11	8	0	8	0
C 3	8	0	10	11	8	8	10	0	8	0

Tableau 1 - Variabilité intraindividuelle dans la durée des voyelles /a/ et /u/. Les valeurs sont des standards déviations en msec.
P1-P14 : Patients atteints de la maladie de Friedreich
C1-C3 : Sujets normaux

SUJETS	Durée en ms Mot de base (ba) ba(s)	Durée en ms Mot de base (ba) ba(sse)	Proportion	Proportion moyenne	Durée en ms Mot de base (ba) ba(ssement)	Proportion	Proportion moyenne
P 1	567	467	0.82		340	0.60	
P 2	713	600	0.84		380	0.53	
P 3	700	600	0.86		280	0.40	
P 4	420	380	0.90		330	0.79	
P 5	360	340	0.94		300	0.83	
P 6	430	390	0.91		250	0.58	
P 7	493	440	0.89	0.83	246	0.50	0.59
P 8	600	460	0.77		310	0.52	
P 9	680	580	0.85		560	0.82	
P10	750	490	0.65		360	0.48	
P11	580	400	0.69		210	0.36	
P12	390	360	0.92		270	0.69	
P13	730	470	0.64		350	0.48	
P14	860	720	0.84		640	0.74	
C 1	480	320	0.67		220	0.46	
C 2	366	280	0.77	0.77	140	0.38	0.44
C 3	420	360	0.86		200	0.48	

Tableau 2 - Ajustement de la durée au nombre de syllabes dans le mot. Série : bas, basse, bassement. P1-P14 : patients, C1-C3 : sujets normaux

SUJETS	PHRASES							
	1	2	3	4	5	6	7	8
P 1	4200	1940	2460	4400	2000	5680	3520	2240
P 2	2840	2000	2200	2580	1600	3460	2600	1440
P 3	3160	1880	2760	3000	1760	3640	2880	1340
P 4	2220	1020	1440	1760	1920	3440	2200	1180
P 5	2540		2040	2100	1260	3100	2160	1720
P 6	2460	1300		2440	1720	4280	3740	1580
P 7	2620	1440	2360	2460	1300	3400	2200	1400
P 8	3200	1520	2320	2300	1400	3300	2800	1460
P 9	4760	2320	2940	3800	2200	5400	4280	2260
P10	2540	1440	1840	2860	1580	2920	2480	1260
P11	3000	1540	2120	2900	1300	2740	2040	1400
P12	2640	2080	1650		2040	2840		1540
P13	3600					3000	2340	
P14	14500	5500	9400	12100	7860	18800	12200	5660
NORMAUX								
MOY.	1913	1040	1600	1686	973	2660	1600	900
MOY + 2 SD	2313	1110	2128	2406	1199	3924	1966	1014

Tableau 3 - Durée moyenne des phrases pour 14 patients (P1-P14). Moyenne pour le groupe des normaux, de même que moyenne + 2 SD pour comparaison. Toutes les valeurs sont en msec.

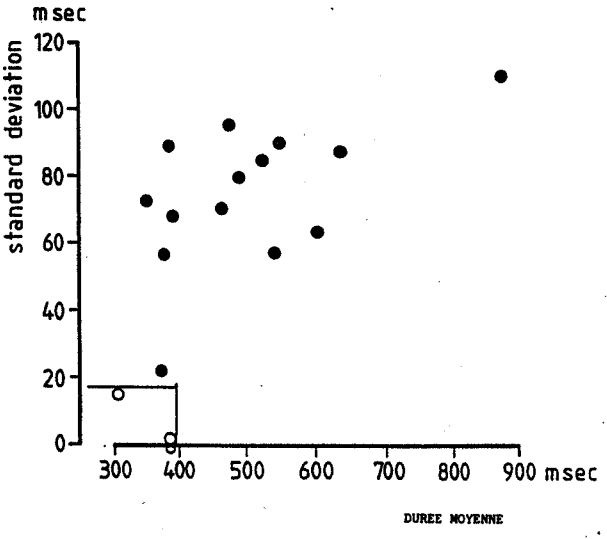


Figure 1 - Graphe représentant la durée moyenne et la standard déviation pour la syllabe /ba/ : sujets normaux (cercles vides), patients (cercles pleins).

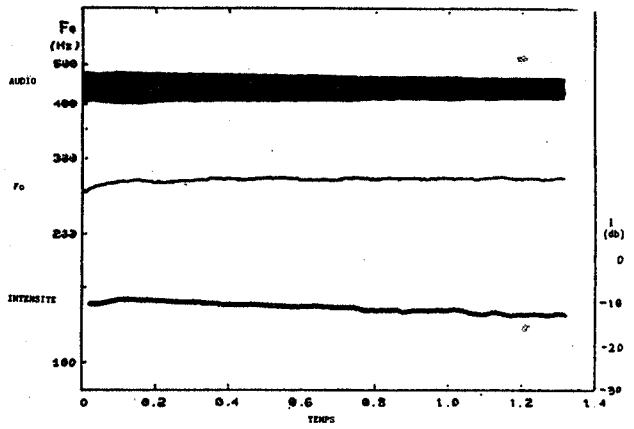


Figure 2 - Signal audio, contours Fo et intensité au cours de la voyelle tenue /u/ pour le sujet normal, C2. Fo est en Hz, intensité relative en db, temps en sec.

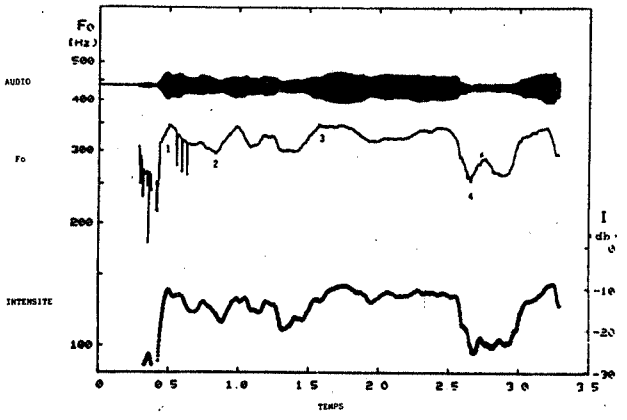


Figure 3 - Signal audio, contours Fo et intensité au cours de la voyelle tenue /u/ pour patient P1. Fo est en Hz, intensité relative en db et temps en sec.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

ANALYSE ET MODELISATION DE TRAJECTOIRES VOCALIQUES.

ETUDE DE TRANSITIONS VOYELLE-VOYELLE.

R. Carré, Télécom Paris, Unité Associée au CNRS, 46 rue Barrault, Paris et Institut de la Communication Parlée, INP-ENSERG, 46 avenue Félix Viallet, Grenoble ;

M. Mrayati, Centre d'Etudes et de Recherches Scientifiques, BP 4470, Damas, Syrie.

Résumé.

On analyse des trajectoires formantiques dans le cas de transitions voyelle-voyelle. Ces trajectoires sont représentées dans les plans F1-F2-F3. La modélisation de ces trajectoires est ensuite étudiée par rapport à un nouveau concept permettant de représenter la fonction d'aire du conduit vocal en régions et modes distinctifs.

Introduction.

Les études effectuées en production de parole sont nombreuses en ce qui concerne les voyelles : études des configurations vocaliques, études acoustiques,... En revanche, les données que l'on peut trouver dans la littérature sont plus éparses dès que l'on s'intéresse aux aspects dynamiques. Or la parole est essentiellement un phénomène dynamique.

L'objet du travail décrit ici vise à analyser et à modéliser des trajectoires formantiques dans le cas des transitions voyelle-voyelle (V-V). Il s'agit, tout d'abord, d'accumuler des données acoustiques (fréquences des trois premiers formants) sur un corpus V-V de parole naturelle. Ensuite, les trajectoires formantiques obtenues par analyse sont formalisées à l'aide d'un nouveau modèle constitué de régions et de modes distinctifs (Distinctive Regions and Modes model : DRM model) (Mrayati, Carré et Guérin, 1988). Les trajectoires dans l'espace F1-F2 ou F1-F2-F3 sont projetées dans l'espace articulatoire à travers un modèle constitué respectivement de quatre (R1, R2, R3, R4) ou huit régions (R1, R2, ..., R8). Cette opération est possible dans la mesure où le modèle peut :

a) reproduire tout son dans l'espace vocalique,

b) reproduire, d'une manière simple, des trajectoires naturelles.

Des premières conclusions sur la formalisation des trajectoires dans l'espace F1-F2-F3 sont discutées.

La représentation des voyelles dans le plan F1-F2 a été vulgarisée par Potter et Steinberg (1950), ensuite par Peterson et Barnay (1952), et puis très souvent utilisée grâce à sa simplicité et à son pouvoir explicatif. En effet, les deux premiers formants sont reconnus, en première approximation, comme étant nécessaires et suffisants pour représenter les voyelles soutenues. Ils représentent les deux premières fréquences de résonance du conduit vocal et ils permettent un synthèse intelligible des voyelles. Par ailleurs, une information sommaire peut être donnée en ce qui concerne la correspondance articulatoire : on qualifie les voyelles représentées dans le plan F1-F2 par voyelles d'avant ou d'arrière, voyelles ouvertes ou fermées. En revanche, la description de trajectoires formantiques entre deux voyelles V1 et V2 dans l'espace F1-F2-F3 a été assez peu étudiée (Holbrook and Fairbanks, 1962 ; Lehiste et Peterson, 1961 ; Gay, 1968 ; Pols, 1977 ; Peeters, 1987 ; ten Bosh, 1987 ; Repp, 1989 ; Gottfried, 1989). Les travaux portent principalement sur des diphtongues et l'interprétation des trajectoires n'est pas approfondie. Il est vrai aussi qu'une représentation correcte de ces trajectoires demande une analyse précise des formants.

Dans cet article, il s'agit d'étudier :

a) si les trajectoires se focalisent à l'intérieur de "conduits" correspondant à des réalisations V1-V2 spécifiques,

b) un modèle quantifié de la fonction d'aire qui formalise d'une façon simple la reproduction de ces trajectoires.

c) l'effet des variations de l'aire de certaines régions du conduit vocal sur les trajectoires,

d) les stratégies sous-jacentes des mouvements de la fonction d'aire pour réaliser ces trajectoires,

e) le rôle du dynamique pour différencier des voyelles ayant les mêmes caractéristiques acoustiques statiques (F-pattern).

I. ANALYSE DES DONNEES DE PAROLE NATURELLE

Les données acoustiques sur les trajectoires vocaliques ont déjà été étudiées pour des objectifs précis. Carré (1971) a utilisé des trajectoires vocaliques pour des applications en vérification de locuteurs. Menon et al. (1971) a donné une schématisation de trajectoires naturelles dans le plan F2-F3 pour des réalisations CV, VC. Mermelstein (1973), quant à lui, a présenté des trajectoires dans le plan F1-F3 obtenues sur un modèle. Majid (1986) a étudié, dans le plan F1-F2 les effets de variations des paramètres du modèle de Maeda (1979).

Pour étudier les questions soulevées dans l'introduction, nous avons enregistré dans un studio, des énoncés de transitions voyelle-voyelle V1-V2, sans phrase porteuse, par 5 locuteurs masculins pour les combinaisons des 11 voyelles orales françaises. Ces réalisations sont en cours d'analyse à l'aide de techniques ceptrales pour extraire les valeurs des formants F1, F2 et F3. Chaque ensemble de valeurs est

représenté dans l'espace F1, F2, F3. A titre d'exemple, la figure 1 présente un exemple de trajectoires dans le plan F1-F2 d'un locuteur à partir d'une même voyelle.

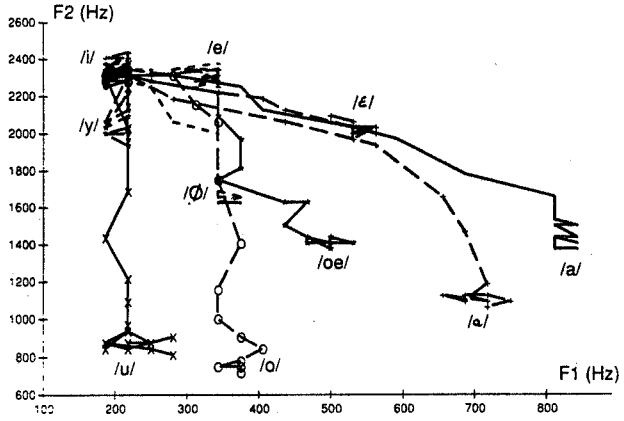


FIG. 1. Représentation dans l'espace F1-F2 de trajectoires obtenues à partir de la voyelle /i/ pour un même locuteur.

L'examen de ces trajectoires conduit aux premières remarques suivantes :
a) les trajectoires ne sont pas des lignes directes entre les voyelles (par exemple, sur la figure 1, on note que la transition /i oe/ passe par /e/ puis par /Ø/),
b) des chemins privilégiés apparaissent,
c) dans certains cas, on note une inversion du sens de variation d'un formant,
d) si le point de départ paraît relativement stable, en revanche, dans certains cas, la cible n'est pas atteinte,
e) en début et en fin de trajectoire, on constate des variations plus importantes qu'au cours du trajet.

Certaines de ces observations peuvent être déduites de données sur l'anglais rapportées par Repp (1989) lorsque nous représentons ces données dans l'espace F1-F2-F3.

En première conclusion, on peut supposer que les trajectoires naturelles que nous avons analysées suivent des chemins spécifiques que nous nous proposons de mieux comprendre.

II. MODELISATION DES TRAJECTOIRES A L'AIDE DU DRM. UNE NOUVELLE STRATEGIE DE COMMANDE.

Un modèle constitué de régions distinctives (modèle DRM) a été proposé (Mrayati, Carré, et Guérin, 1988). Sa définition et certaines de ses propriétés dynamiques ont été formulées et ont fait l'objet de plusieurs publications (Mrayati, Carré et Castelli, 1989 ; Carré et Mrayati, 1989; Mrayati et Carré, 1989). Rappelons que le modèle est déduit du comportement acoustique d'un tube uniforme et qu'il incorpore un découpage de la fonction d'aire en régions distintives de sections uniformes et de longueurs inégales. La figure 2 représente un modèle à quatre tubes correspondant à deux formants F1, et F2 et un modèle à huit tube correspondant à trois formants F1, F2 et F3.

Le modèle DRM effectue de manière pseudo-orthogonale la transformation des variations des aires des régions en effets acoustiques. On a représenté à titre d'exemple ce comportement dynamique dans le plan F1-F2 (figures 3 et 4).

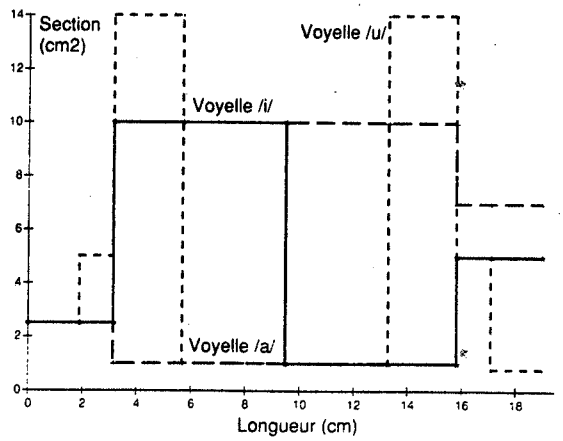


FIG. 2. Schéma d'un modèle à Régions constitué de quatre régions (configurations /a/ et /i/) et de huit régions (configuration /u/).

La figure 3a montre des trajectoires dans le plan F1-F2 obtenue pour le passage d'une configuration de type /a/ à une configuration de type /i/ du modèle constitué ici de 4 régions (R1, R2, R3, R4). Ce passage correspond à une variation inverse de R2 et R3. Ces trajectoires gardent leurs formes générales mais se déplacent dans le plan en fonction de l'aire de R4. La figure 3b montre des trajectoires pour des valeurs données de R2 et R3 lorsque R4 varie de 0.7 a 11 cm2 (corrélation avec l'ouverture aux lèvres). La figure 4 est la superposition des figures 3a et 3b. On retrouve aussi un comportement pseudo-orthogonal dans la modélisation effectuée par Ladefoged et Harshman (1979).

La contrainte sur la variation inverse des sections de R2 et R3 revient à conserver le volume de la langue constant. Une deuxième contrainte pourrait être ajoutée qui consiste à appliquer une certaine corrélation entre les mouvements de R3 et de R4 : en particulier, si R3 est grand, R4 est difficilement très petit. La zone correspondant à de basses valeurs de F1 et F2 est alors inaccessible.

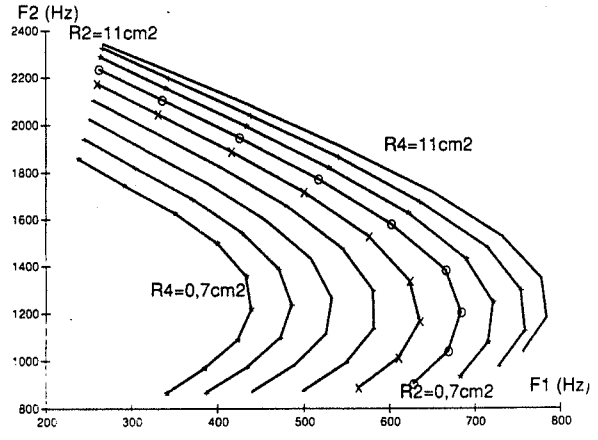


FIG. 3a. Représentation, dans le plan F1-F2, de neuf trajectoires de type /ai/ avec un modèle à quatre régions. Chacune des courbes représente l'effet de la variation de 0.7 à 11 cm2 de la région R2 associée à une variation

inverse de R3. R4 (corrélée avec l'ouverture aux lèvres) prend neuf valeurs comprises entre 0.7 et 11 cm2 pour des valeurs fixes de R2 et R3.

Les courbes précédentes ont été obtenues à partir de calculs systématiques sur un modèle constitué de 4 régions. La longueur du modèle est de 19cm intégrant approximativement l'effet inductif dû au rayonnement aux lèvres et l'allongement du conduit vocal lorsque la section aux lèvres diminue. La région R1 est conservée constante (1.4 cm2) approchant le cas naturel tandis que les trois autres régions R2, R3, R4 prennent chacune 9 valeurs comprises entre 0,7 à 11 cm2 et variant par pas de multiplication de $\sqrt{2}$. Sans tenir compte de contraintes de type articulatoire, on a 9x9x9 = 729 combinaisons représentant 729 configurations différentes du modèle. Pour ces configurations, les formants ont été calculés en utilisant les algorithmes proposés par Badin et Fant (1984). Les valeurs des trois formants correspondant à ces 729 configurations définissent l'espace vocalique du modèle.

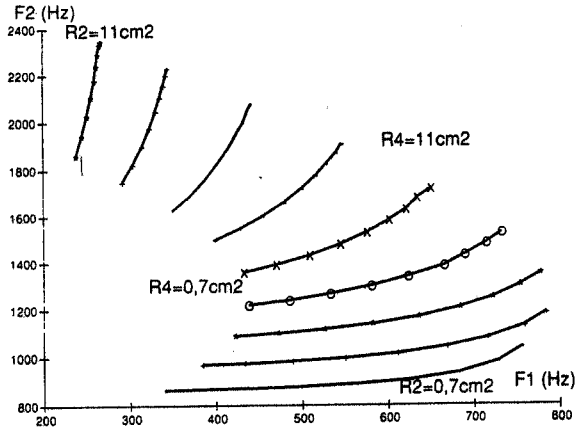


FIG. 3b. Mêmes données que pour la figure 3a mais ici, chacune des trajectoires correspond à une variation entre 0.7 et 11 cm2 de R4.

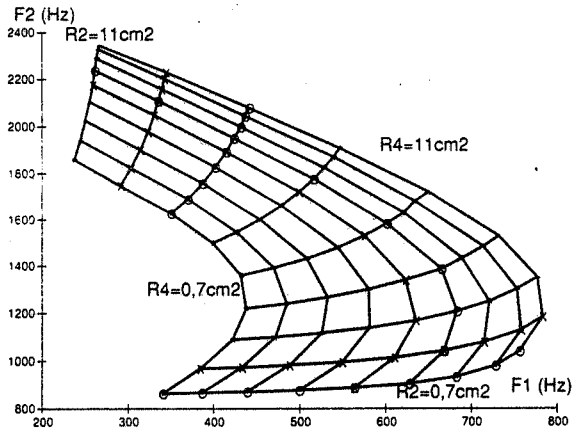


FIG. 4. Superposition des trajectoires représentées figures 3a et 3b. On note le comportement pseudo-orthogonal dynamique du modèle à Régions.

On peut montrer que le modèle à Régions est capable de produire un espace vocalique plus grand que tout modèle constitué d'un même nombre de tubes, pour une variation des aires entre des limites données. La figure 5 donne cet espace pour quatre régions. Dans le cas de huit régions, le triangle vocalique est plus étendu, en particulier pour les basses valeurs de F1 et F2.

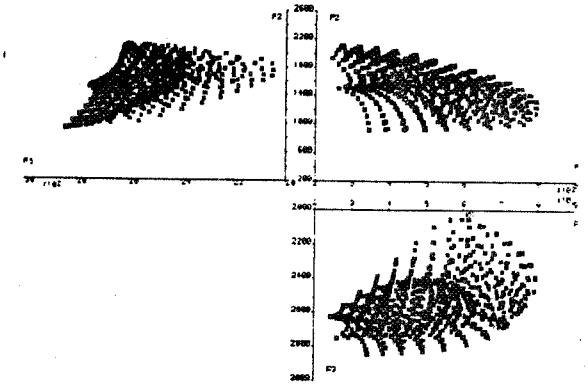


FIG. 5. Espace vocalique du modèle constitué de quatre régions dans les plans F1-F2, F1-F3, F2-F3.

Nous venons de présenter l'espace vocalique du modèle à Régions et certains de ses aspects dynamiques. Nous allons maintenant étudier plus systématiquement les relations entre des mouvements de régions et des déplacements dans l'espace vocalique de trajectoires.

III. LES GRANDS AXES DE DEPLACEMENT DES TRAJECTOIRES DANS L'ESPACE VOCALIQUE

Dans cette partie, la simulation a été effectuée sans introduction des pertes, l'impédance de rayonnement est nulle et la longueur du modèle est de 17,5 cm.

La figure 6a montre toutes les trajectoires obtenues par variation de R4 lorsque l'on modifie pas à pas les deux autres régions R2 et R3. Des remarques importantes peuvent être formulées :

a) l'ouverture de R4 (corrélée à l'ouverture aux lèvres) augmente toujours F1 et F2 sauf pour les zones à F2 élevé,
b) en général, les pentes des trajectoires augmentent avec F2,
c) la dynamique sur F1 diminue avec F2,
d) la dynamique sur F2 reste relativement constante,
e) les changements de pente à F2 élevée correspondent à des zones quantiques (Stevens, 1972),
f) les trajectoires sont translatées dans la direction de l'axe F2 avec moins de dynamique sur F1 et plus de pente lorsque R3 diminue,
g) les trajectoires sont translatées dans la direction de l'axe F2, F1 diminuant, lorsque R2 croît,
h) les trajectoires ont été calculées sans tenir compte des contraintes articulatoires.

La figure 6b représente toutes les trajectoires obtenues par variation de R3 (corrélation avec la dimension de la cavité buccale) lorsque l'on modifie pas à pas les deux autres régions R2 et R4. De nouvelles remarques sont à prendre en compte :

a) comme dans la figure précédente, les trajectoires ont été calculées sans tenir compte des contraintes articulatoires,

b) l'ouverture de R3 diminue toujours F2,
c) l'ouverture de R3 augmente généralement F1 sauf pour les zones à F1 et F2 bas ; cette dernière situation (R3 grand et R4 petit) n'est pas réaliste,
d) la dynamique sur F2 diminue avec F2,
e) la dynamique sur F1 est minimale dans la zone centrale de F2,
f) les trajectoires sont translatées dans la direction de F1 pour une ouverture de R4, et vers le bas et à droite avec une fermeture de R2,
g) les changements de pentes à fréquence F2 intermédiaire correspondent à des zones quantiques.

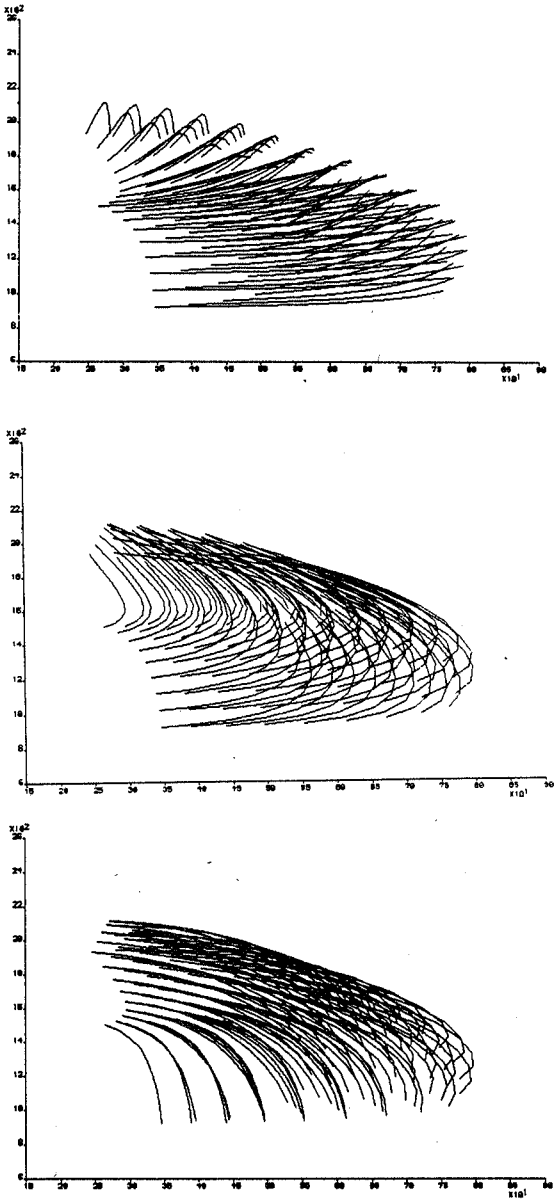


FIG. 6. a) Trajectoires obtenues dans l'espace F1-F2 par variation de R4, les deux autres régions R2 et R3 étant modifiées pas à pas.
b) Trajectoires obtenues dans l'espace F1-F2 par variation de R3, les deux autres régions R2 et R4 étant modifiées pas à pas.
c) Trajectoires obtenues dans l'espace F1-F2 par variation de R2, les deux autres régions R3 et R4 étant modifiées pas à pas.

La figure 6c représente toutes les trajectoires obtenues par variation de R2 (corrélation avec la cavité pharyngale) lorsque l'on modifie pas à pas les deux autres régions R3 et R4. Des conclusions de même type que les précédentes peuvent être formulées facilement.

On peut noter qu'en général, les trajectoires sont monotones du point de vue de leur direction dans le plan F1-F2, sauf pour les trajectoires de R3. Mais lorsque F1 et F2 sont bas, les configurations correspondantes du modèle à quatre régions ne sont pas naturelles. Si ces situations sont éliminées, les changements de monotonie se retrouvent alors seulement en extrémités de trajectoires.

Une trajectoire naturelle peut être projetée dans les trois plans précédents et analysée en référence à ces plans. Par exemple la trajectoire /ai/ se compare facilement avec une diminution de R3 et une augmentation de R2. La figure 7 montre des transitions obtenues avec le modèle par interpolation linéaire entre configurations vocaliques (Carré et Mrayati, 1990). En comparant les trajectoires obtenues naturellement et sur le modèle, on peut comprendre quelles sont les régions mises en jeu durant la production de ces V-V. On peut, en particulier, retrouver précisément sur le modèle une imitation de l'original. A titre d'exemple, notons la grande différence dans les trajectoires /iu/ naturelle et modélisée : on peut combler cette différence par une accélération du mouvement sur R4 (anticipation). Ce cas d'anticipation est effectivement connu (Lubker, 1981). Par déduction, on peut ainsi étudier les mouvements relatifs des régions les unes par rapport aux autres.

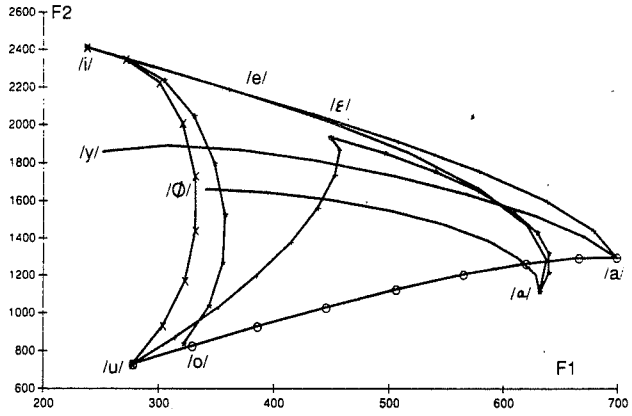


FIG. 7. Trajectoires obtenues sur le modèle par interpolations linéaires entre configurations vocaliques.

Il est important de rappeler que les trajectoires décrites plus haut résultent d'une commande **transversale** de l'aire des régions. Cette formalisation nouvelle de la commande de la forme de la fonction d'aire se retrouve implicitement dans différents travaux sur la parole naturelle (Shigenaga, 1969 ; Ariizumi et Shigenaga, 1968 ; Kiritani et Fujimura, 1975 ; Harshman, Ladefoged et Goldstein, 1974 ; Maeda, 1979). Cette commande transversale liée à l'existence même de régions fixes implique de manière implicite de réaliser des **constrictions** principales à l'emplacement de ces régions. Cette quantification des lieux de

constriction apparaît justifiée et par l'existence des lieux spécifiques d'articulation des **consonnes** correspondant aux régions d'une part et aux lieux privilégies de constriction des **voyelles** mis en évidence, en particulier, par Wood (1979), d'autre part. Enfin, cette commande transversale incorpore aussi implicitement le déplacement **longitudinal** de la constriction classiquement adopté (Carré et Mrayati, 1990).

En résumé, le Modèle à Régions (modèle DRM) est capable de produire toutes les cibles vocaliques possibles et aussi de produire simplement des trajectoires réalistes dans l'espace vocalique. La formalisation de ces trajectoires tenant compte des contraintes articulatoires est un pas significatif pour la compréhension des stratégies de réalisation d'une trajectoire et des facteurs à mettre en jeu pour déplacer cette trajectoire dans l'espace vocalique.

IV. ROLE DU DYNAMIQUE DANS LA DISTINCTION DES CONFIGURATIONS VOCALIQUES DIFFERENTES MAIS AYANT MEMES FORMANTS

Il est bien connu que l'on peut produire un même son au moyen de configurations vocaliques différentes. Atal et al. (1976) ont introduit la notion de fibre de fonctions d'aire possédant les mêmes fréquences et bandes passantes de formants. Statiquement, aucune différence acoustique ne peut donc être relevée pour les sons produits par ces configurations. Toutefois, un problème soumis par L.J. Boë, nous a conduit à approfondir cette affirmation. En effet, dans un dialecte africain, deux configurations vocaliques ayant une structure formantique très proche, y représentent deux voyelles différentes dans le système vocalique. Nous avons simulé ces configurations au moyen du modèle à Régions et nous avons alors constaté, dans l'espace F1-F2-F3 que les trajectoires intervocaliques entre ces deux configurations (voyelles V1 et V2) ne se réduisent pas à un seul point (figure 8).

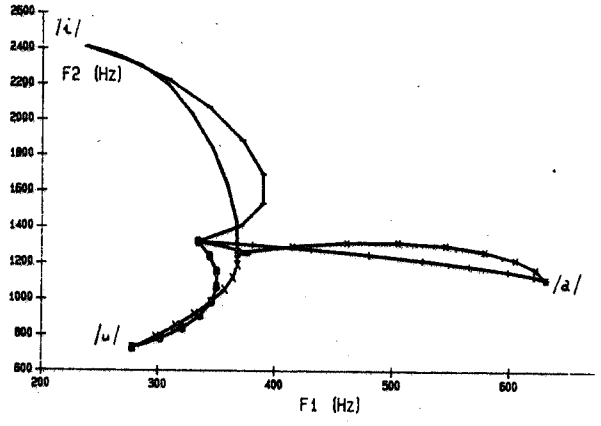


FIG. 8. Transitions réalisées sur le modèle à partir de configurations différentes ayant une structure formantique très proche.

Partant du point V1 une trajectoire spécifique V1V2 est suivie qui reboucle sur V2 superposé à V1. Nous avons ensuite simulé au moyen du même modèle des trajectoires impliquant les trois voyelles extrêmes /a, i, u/ avec V1 et V2 sont différents. Les transitions formantiques entre une voyelle cardinale et V1 ou V2 sont donc différentes. Par exemple, la transition du premier formant de /iV1/ passe par un changement de sens ce qui n'est pas le cas pour /iV2/. Cette constatation nous mène à supposer que la dynamique peut jouer un rôle pour la distinction des voyelles comme il le fait pour les plosives par exemple. Cette mise en évidence du rôle de l'angle d'attaque des voyelles mérite naturellement un sérieux approfondissement.

CONCLUSIONS.

Dans cet article, nous avons essayé de répondre à plusieurs questions relatives à l'étude des trajectoires formantiques. Ces trajectoires peuvent être groupées dans des "conduits" privilégiés dans l'espace F1-F2-F3. Certains aspects dynamiques du modèle à Régions ont été mis en évidence et tout particulièrement l'orthogonalité des trajectoires produites par ses différentes régions et la stratégie transversale produisant ces trajectoires. Les grands axes de déplacement des trajectoires formantiques dans l'espace Fn du modèle ont été mis en évidence avec les facteurs qui sont à l'origine de ces déplacements. Enfin, dans les trajectoires voyelle-voyelle, l'angle d'attaque dynamique peut être important pour la distinction des voyelles.

Le modèle constitué de régions permet une étude au-delà des contraintes articulatoires qui peuvent être mises en oeuvre ultérieurement. Il permet de mettre en évidence des limites (stabilités formantiques par exemple) non atteintes par des modèles articulatoires.

Dans des applications de synthèse par règles de transitions voyelle-voyelle par exemple, on peut envisager une stratégie de commande à partir du modèle à Régions avec d'une part un contrôle temporel de gestes liés à ces régions correspondant à un objectif spécifique de trajectoires formantiques, et la mise en oeuvre des tâches à effectuer par un modèle du conduit vocal lequel a ses propres contraintes, d'autre part.

BIBLIOGRAPHIE

Ariizumi, H., and Shigenaga, M. (1969). "On the movement of articulators", Abstract of the report of Committee on Speech of Acoust. Soc. Japan.

Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting techniques", J. Acoust. Soc. Am., 63, 1535-1555.

Badin, P. et Fant, G. (1984). "Notes on vocal tract computations", STL-QPSR, 2-3, 53-107.

ten Bosh, L.F.M. (1987). "About diphthongs. An implementation into vowel dispersion theory", Proc. of the Institute of Phonetic Sciences, Univ. of Amsterdam, 1-14.

Carré, R. (1971). "Identification des locuteurs : exploitation des données relatives aux fréquences des formants", Proc. 7ème Congrès International d'Acoustique, Budapest.

Carré, R., and Mrayati, M. (1990). "Articulatory-acoustic-

phonetic relations and modeling, regions and modes", à paraître dans Speech Production and Speech Modeling, W.J. Hardcastle and A. Marchal (eds), Kluwer Academic Publishers.

Fant, G. (1959), Acoustic analysis and synthesis of speech with applications to Swedish, Ericsson Technics, no 1.

Gay, T. (1968). "Effect of speaking rate on diphthong formant movements", J. Acoust. Soc. Am., 44, 1570-1573.

Gottfried, M. (1989). "Some acoustical properties of diphthongs", J. Acoust. Soc. Am., 86, S123.

Harshman, R., Ladefoged, P., and Goldstein, L. (1977). "Factor analysis of tongue shapes", J. Acous. Soc. Am., 62, 693-707.

Holbrook, A., and Fairbanks, G. (1962). "Diphthong formants and their movements", J. Speech Hear. Res., 5, 33-58.

Ladefoged, P., and Harshman, R. (1979). "Formant frequencies and movements of the tongue", UCLA Working Paper in Phonetics.

Lubker, J.F. (1981). "Temporal aspects of speech production : anticipatory labial coarticulation", Phonetica, 38, 51-65.

Maeda, S (1979). Un modèle articulatoire basé sur une étude acoustique, Bulletin de l'Institut de Phonétique de Grenoble, 8, 35-55.

Majid, R. (1986). "Modélisation articulatoire du conduit vocal. Exploration et exploitation. Fonctions de macro-sensibilité paramétriques et voyelles du Français", Thèse de docteur ingénieur, Grenoble.

Menon, K.M.N., Rao, P.V.S., and Thosar, R.B.T. (1971). "Perception of stop consonants",
Proc. of Int. Cong. on Acoustics, Budapest paper 19C4.

Mermelstein, P. (1973). "Articulatory model for the study of speech production", J. of Acoust. Soc. of Am., 53, 1070-1082.

Mrayati, M., Carré, R., and Guérin, B. (1988). "Distinctive regions and modes : a new theory of speech production", Speech Communication, 7, 257-286.

Mrayati, M., Carré, R., and Castelli, E. (1989). "A new approach for speech dynamic studies", J. Acoust. Soc. Am., 85, S144.

Mrayati, M., and Carré, R. (1989). "Speech synthesis based on vocal tract region theory", Proc. of Eurospeech 89, 172-175.

Peeters, W.J.M. (1987). "Acoustic structure and perceptual relevance of 'steady states' and 'glides' within formant trajectories of diphthongs, complex vowels, and vowel clusters", Proc. of European Conference on Speech Technology, CEP Consultants, 42-45.

Peterson, G.E., and Barney, H.L. (1952). "Control method used in a study of the vowels", J. Acoust. Soc. Am., 24, 175-184.

Pols, L. (1977). "Spectral analysis and identification of Dutch vowels in monosyllabic words", Doct. Diss. Free University Amsterdam.

Repp, B. (1989). "Traversing upper vowel space : a smooth or a bumpy ride", Speech Communication, 8, 1-15.

Shigenaga, M., Ariizumi, H., and Tanaka, K. (1968). "A model for transitions of area functions", Proc. of the 6th Int. Cong. of Acoustics, Tokyo.

Stevens, K.N. (1972). "The quantal nature of speech : evidence from articulatory-acoustic data", in Human Communication : a unified view (Mac Graw-Hill), 51-66.

Wood, S. (1979). "A radiographic analysis of constriction locations for vowels", J. of Phonetics, 7, 25-43.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

PREMIERES MODELISATIONS SUR LE TIMING DES PICS DE VITESSE DE LA MANDIBULE.

ABRY C., PERRIER P., & JOMAA M.

INSTITUT DE LA COMMUNICATION PARLEE (URA CNRS n°368)
INPG/ENSERG & Université Stendhal
Domaine Universitaire - 38040 Grenoble Cédex - France

ABSTRACT.

In the general framework of motor control theory, the study of velocity profiles for articulatory trajectories seems to be an efficient tool. We propose here preliminary simulations of some velocity profiles for jaw movements for VVC versus VCC productions in Tunisian Arabic. Our stiffness (2nd-order) control model allows us to generate appropriate profiles for consonant-vowel timings, inasmuch as selective plateau strategies and specific undershoot phenomena are taken into account.

I. INTRODUCTION.

Pour l'élaboration de théories sur le contrôle du mouvement, la représentation de l'évolution de la vitesse du mouvement en fonction du temps (son profil de vitesse), et plus particulièrement la description de l'allure de ces profils en termes de symétrie ou d'asymétrie par rapport à l'occurence du maximum de vitesse (pic de vitesse), constitue un domaine d'observation privilégié. Nelson (1983) a ainsi différencié les mouvements minimisant la vitesse, l'énergie, ou le **jerk**, par les caractéristiques de leurs profils de vitesse, et a proposé, sur ce critère, une modélisation par un système du second ordre des mouvements appris. Hogan (1984) s'est basé sur l'allure en forme de cloche des profils de vitesse des mouvements volontaires chez les singes pour suggérer que le contrôle de ces mouvements repose sur une minimisation du **jerk**. Bullock & Grossberg (1988) ont testé la validité de leur modèle neural VITE de génération de trajectoires articulatoires essentiellement sur des critères ayant trait au profil de vitesse. Massone & Bizzi (1989), désireux de modéliser les trajectoires des membres avec un réseau de neurones, ont choisi les profils de vitesse comme référence pour l'apprentissage de leur réseau. (Cf. aussi Marteniuk et al. (1987), pour une bonne illustration de la distinction de profils de vitesse, en fonction de gestes de saisie à finalités différenciées).

Nous nous proposons ici de recourir à une modélisation d'un articulateur sous la forme d'un système distribué du second ordre (Perrier et al., 1989) pour générer les profils de vitesse très différents observés en arabe tunisien dans des séquences VVC versus VCC, et retrouver les variations mesurées lors de modifications du débit.

II. DESCRIPTION DES DONNEES.

Le contrôle du timing en arabe tunisien s'exerce sur les voyelles et les consonnes, ce dialecte comme d'autres variétés de l'arabe possédant des distinctions de quantité phonologiques. La paire que nous avons choisie pour illustrer ce point est [zɛɛza] "il a tondu le mouton" vs [zɛzza] "il a récompensé".

Les items, insérés en phrase porteuse, ont été lus en chambre sourde, chaque item étant répété 12 fois en ordre aléatoire. Une première série a été enregistrée au débit choisi par le locuteur ; la seconde avec consigne de débit rapide.

L'articulateur choisi est la mandibule, qui régule le timing des articulateurs portés - les lèvres et la langue. Son déplacement est suivi par le kinésiographe mandibulaire (K5AR), les signaux de position et de vitesse étant échantillonnés à 160Hz, pour édition en synchronie avec le signal acoustique numérisé à 8 Khz (Worley, 1989).

Deux cycles ont été déterminés :

1. Le cycle "vocalique", qui va du début d'abaissement de la mandibule (Vocalic Cycle Onset), pour produire la première voyelle ([ɛɛ] ou [ɛ]), au début d'abaissement pour la voyelle suivante [a] : dans ce cycle une phase allant de VCO au début d'élévation pour la consonne ([z] ou [zz]) a été mesurée.

2. Le cycle que nous appellerons "cycle de vitesse vocalique", et qui a pour bornes les pics de vitesse négatifs (Max. Vocalic Velocity, soit MVV) des gestes d'abaissement pour ces mêmes voyelles : là aussi, une phase "vocalique" MVV-MCV (Max. Consonantal Velocity) est mesurée en pourcentage du cycle MVV-MVV.

Le cycle de vitesse vocalique (MVV-MVV, Fig. 1) nous a permis de retrouver au mieux l'opposition phonologique (Delattre et al., 1989) : celle-ci se maintient en débit normal, et disparaît en débit rapide, ce qui correspond par ailleurs aux données du test de rime et aux mesures acoustiques (Jomaa & Abry, 1988).

III. MODELISATION (essais préliminaires).

1. Position du problème.

Les types de signaux mandibulaires produits dans le contrôle de la durée en arabe nous semblent suffisamment contraints pour offrir la possibilité de tester notre modèle que nous avons proposé en vue de générer la dynamique des articulateurs (Perrier et al., 1989).

Rappelons brièvement les points forts de nos résultats, qui constituent des enjeux pour une telle modélisation. La différence entre les voyelles longues et les voyelles brèves apparaît sur la courbe de variation de la vitesse, par un différence dans le délai d'apparition du pic de vitesse dans la transition voyelle-consonne, c'est-à-dire par la vitesse à laquelle la voyelle est "cachée" par le mouvement d'élévation de la mandibule, en fait plus rapidement pour la voyelle brève. Il est donc décisif de générer des trajectoires qui conservent les propriétés du profil de vitesse, le timing du pic de vitesse étant un paramètre efficace pour différencier des signaux linguistiquement pertinents.

Nous devrons donc être particulièrement prudents dans le choix des commandes utilisées pour rendre compte de tels profils de vitesse.

2. Notre modèle.

Notre modèle est un modèle fonctionnel, contrôlant la raideur (k) sur une paire de ressorts, un pour les effecteurs agonistes (k1) et l'autre pour les antagonistes (k2). Les cibles sous-jacentes sont définies par un rapport k1/k2 (définissant un point d'équilibre). Leur position au cours du temps est celle que l'on peut déterminer à partir du signal réel. Ainsi, seuls les **undershoots/overshoots** spatiaux sont pris en compte. (Nous laissons de côté, pour l'instant, les **undershoots/overshoots** temporels, c'est-à-dire, la persévération et l'anticipation : ces dernières ne sont pas hors de portée du modèle, mais elles n'ont pas encore été explorées). On suppose que, de cible à cible, la raideur varie sinusoïdalement dans le temps. Des commandes de tenue sont disponibles, pour éviter ce qui constitue, à notre avis, un usage inadéquat de l'amortissement, proposé par Browman & Goldstein (1984) pour modéliser les plateaux. Le contrôle du niveau général de co-contraction (k1+k2) est d'une importance primordiale pour l'atteinte des cibles. Trois modes d'action restent possibles : le premier maintient le niveau de co-contraction constant, les raideurs de chacun des ressorts variant de manière antisymétrique ; le second diminue le niveau général de co-contraction et cette diminution est équirépartie sur les deux ressorts ; le troisième agit d'abord sur l'agoniste, tandis que l'antagoniste réagit (en **feedback**) pour freiner, si et seulement si la cible programmée risque d'être dépassée.

3. Les simulations.

[zɛɛza] débit normal (Fig.2.) - Pour un niveau de co-contraction suffisant, les cibles sont effectivement atteintes au bon moment. Les trajectoires se révèlent être optimisées suivant le principe de minimisation du **jerk** (Nelson, 1983) : les profils de vitesse sont donc typiquement en forme de cloche. Tout comme sur le signal mesuré, le pic de vitesse positive dans cette phase est loin du pic de vitesse négative dans le même cycle : l'intervalle entre ces deux pics représente 72% de la phase allant du pic négatif à la fin du cycle*.

[zɛɛza] débit rapide (Fig. 3) - En gardant les valeurs mécaniques utilisées pour la condition de débit normal et en ne changeant que la position des cibles au cours du temps (cibles déterminées à partir du signal choisi), nous reproduisons les effets suivants (en comparaison avec le débit normal): (1) le changement de l'instant d'apparition du pic de vitesse positive (64%, pour 62% observé); et (2) les tendances à un **undershoot** spatial des cibles vocaliques [ɛɛ] et consonantiques [z].

[zɛzza] débit normal (Fig. 4) - Nous avons tout d'abord pu constater qu'il était impossible de reproduire, avec notre modèle, le patron observé, si nous gardions les valeurs mécaniques choisies pour la voyelle longue (avec un séquencement temporel des cibles repris, bien sûr, du signal observé pour la brève [ɛ]). Ensuite, il nous a été également impossible de positionner le pic de vitesse positive, à une place correcte dans le cycle, et ceci jusqu'à ce que nous prenions en compte - parmi les possibilités et combinaisons offertes par le modèle** - une commande de tenue de 45ms, correspondant à l'effet de **plateau** observé, comme tendance générale sur les signaux [zz]. Une fois que cette commande est acceptée, on note qu'il n'y a qu'une légère différence entre les deux solutions restantes. L'équirépartition de la co-contraction, comme le mode rétroactif, nous livrent un bon espacement des deux pics de vitesse (52%; observé 51%); seule subsiste une petite différence dans le phasage du geste d'élévation (56% vs. 58%; observé 60%). Cependant, le mode rétroactif semble donner une pente plus raide en allant vers le pic de vitesse positive; après celui-ci, il présente une meilleure adéquation avec l'évolution de la vitesse de part et d'autre du passage par zéro : cette inflexion de la vitesse reflète mieux la tendance que l'on observe, à savoir un plateau.

[zɛzza] débit rapide (Fig. 5) - Dans ce cas, contrairement à ce qu'on avait obtenu pour la voyelle longue, on ne peut pas toujours générer un pic de vitesse bien placé, en accélérant simplement le timing de ces cibles (avec les mêmes commandes mécaniques qu'en débit normal). Dans certains cas, comme dans le résultat de simulation présenté en Fig. 5, nous avons dû enlever la commande de tenue, pour reproduire les tendances observées dans le placement du pic de vitesse positive (proche des valeurs obtenues pour [zɛɛza] rapide, allant jusqu'à 61%); le phasage du geste d'élévation étant toujours autour de 58%-60%.

Ces quatre simulations nous fournissent deux stratégies différentes pour allonger les voyelles et les consonnes. Lorsqu'on essaie, conformément à ce qui a été observé tout particulièrement en arabe tunisien, de faire disparaître (sous l'effet du débit) les différences importantes qui caractérisent les profils de vitesse, nous pouvons mettre en évidence que : 1. l'**undershoot** est facilement obtenu par une simple accélération de la structure mécanique établie pour les voyelles longues; 2. au contraire, les consonnes longues doivent renoncer à leur commande de tenue spécifique.

* Dans les simulations suivantes, tous les phasages du pic de vitesse positive, dans l'élévation de la mandibule, seront - pour des raisons de commodité - exprimés par rapport à cette base. Les mesures observées sont celles des exemples de signaux choisis pour la modélisation.

** Parmi lesquelles : 1. diminuer la raideur globale sur la voyelle brève; 2. augmenter la raideur sur la consonne géminée qui suit, et ceci selon les deux modes suivants : 2.1.avec équirépartition sur les deux protagonistes; 2.2. avec rétroaction sur l'antagoniste (mode feedback).

IV. CONCLUSIONS.

Il est probablement prématuré d'établir des implications générales à partir de ces simulations préliminaires. Par conséquent, nous nous limiterons à suggérer quelques propositions pour la commande dynamique d'un synthétiseur articulatoire (Maeda, 1988) avec des commandes mécaniques (Perrier et al., 1989).

1. Concernant la modélisation dynamique de leurs trajectoires spécifiques, les voyelles et les consonnes semblent se comporter de manière plutôt différentes, comme cela a pu être mis en évidence par la manipulation de leur allongement/compression, dans les deux débits. Nous avons eu l'occasion de raffiner la modélisation du profil de vitesse de ces tâches différentes. Notre conclusion préliminaire - qui ne tient pas seulement au modèle utilisé ici - consiste à dire que tout modèle doit pouvoir traiter la tendance qu'ont les consonnes (au contraire des voyelles) à produire des effets de plateaux. Ces effets pourraient être en réalité plus importants que nous n'avons déjà pu le mettre en évidence ici, en montrant la nécessité de disposer déjà pour la mandibule de véritables commandes de tenue. En effet les cibles des articulateurs portés (la lèvre inférieure, la lame ou le dos de la langue), présentent des possibilités de plateaux encore plus grandes que l'articulateur porteur, la mandibule.

2. Une étude complémentaire s'imposera pour explorer d'autres dialectes arabes; puis pour généraliser à d'autres types de langues possédant de telles oppositions de quantité. En particulier, les consonnes longues, phonologiquement hétéromorphémiques - produites par une agrégation de deux gestes (Munhall & Löfqvist, 1987) dans le cas des consonnes homorganiques - semblent présenter actuellement les cas de plateaux les plus clairs (Delattre, 1988).

REMERCIEMENTS : à Rudolph SOCK pour ses remarques pertinentes sur ce texte. A Christine Delattre pour son aide dans la collecte des données.

REFERENCES

Browman C.P. & Goldstein L.M. (1984), "Dynamic modeling of phonetic structure.", Haskins Labs. Status Rept. Speech Res. SR-79/80, 1-17.

Bullock D. & Grossberg S. (1988), "The VITE model : a neural command circuit for generating arm and articulatory trajectories.", In Dynamic patterns in complex systems, Eds. J.A.S. KELSO A. J. MANDELL M. F. SHLESINGER, Singapore : World Scientific Publishers.

Delattre C. (1988), "Etude des patrons de phases sur les mouvements mandibulaires en français dans les cycles vocaliques et consonantiques.", Mémoire de D.E.A. de Sciences du Langage (dir. C. ABRY), Université Stendhal, Grenoble.

Delattre C., Jomaa M., Worley & C. Abry C. (1989), "The phasing of the jaw in consonant and vowel lengthening : Arabic and French patterns.", European Conference on Speech Communication and Technology, Vol. 2, 416-419.

Hogan N. (1984), "An organizing principle for a class of voluntary movements.", The Journal of Neurosciences, Vol. 4, N°11, 2746-2754.

Jomaa M. & Abry C. (1988), "La résistivité de la quantité vocalique aux variations de la vitesse d'élocution : le cas de l'arabe tunisien.", 17èmes Journées d'Etudes sur la Parole (Groupe Communication Parlée - S.F.A.), Nancy, 231-236.

Maeda S. (1988), "Improved articulatory model.", J. of Acoust. Soc. Am., 84, Supp. 1, S1 46.

Marteniuk R., G. Mackenzie C.L., Jeannerod M., Athenes S. & Dugas C. (1987), "Constraints on human arm movement trajectories.", Canadian Journal of Psychology, 4, 365-378.

Massone L. & Bizzi E. (1989) , "A neural network model for limb trajectory formation.", Biol. Cybern., II, 1-9.

Munhall K. & Löfqvist A. (1987) , "Gestural aggregation in Speech.", PAW review, 2, 13-15.

Nelson W. L. (1983), "Physical principles for economy of skilled movements.", Biol. Cybern, 46, 135-147.

Perrier P., Abry C. & Keller E. (1989), "Vers une modélisation des mouvements du dos de la langue.", Journal d'Acoustique 2, 69-77.

Worley C. (1989), "*Organisation temporelle articulatori-acoustique des gestes vocaliques et consonantiques. Signaux kinésiographiques labiaux et mandibulaires.* " Thèse I.N.P., Grenoble.

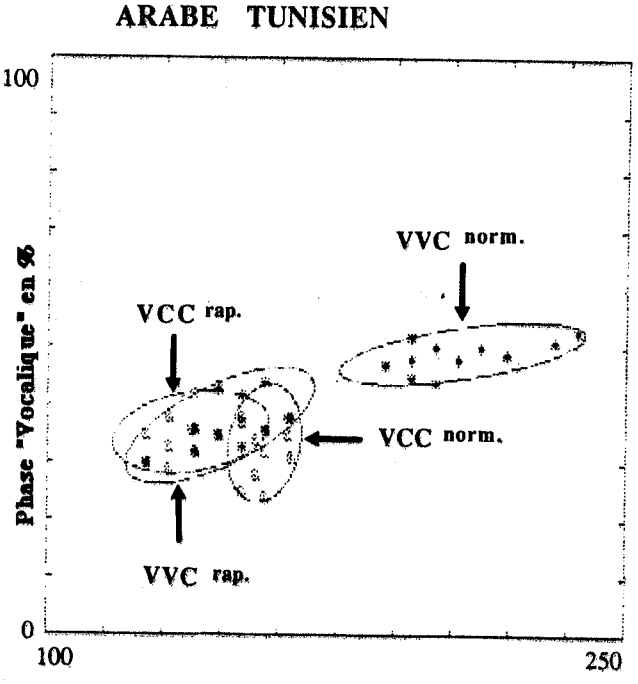


Figure 1. L'opposition VVC/VCC en arabe tunisien dans le cycle de vitesse "vocalique" (entre les pics de vitesse MVV-MVV). En ordonnée : la phase dite "vocalique" (MVV-MCV) déterminée par l'apparition du pic de vitesse consonantique (MCV). Locuteur masculin. Deux conditions de vitesse d'élocution : normale et rapide.

Figures 2 et 3. Simulations des signaux de position mandibulaire pour [zɛɛza] (en arabe tunisien). Au centre la vitesse instantanée; en bas les commandes de raideur.

Figure 2 : vitesse d'élocution normale

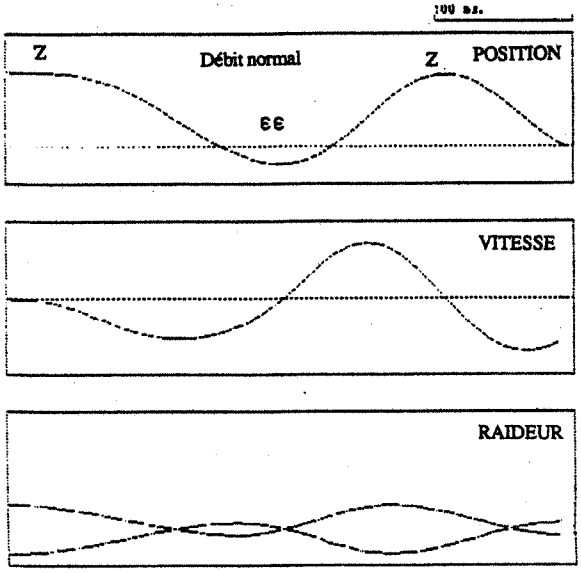
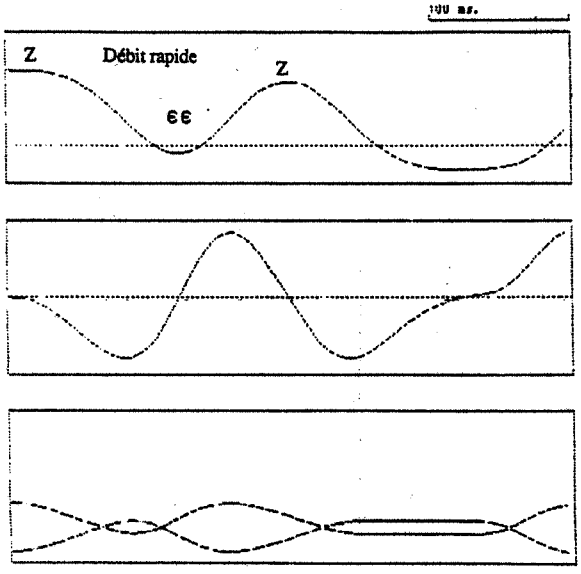


Figure 3 : vitesse d'élocution rapide.



Figures 4 et 5. Simulations des signaux de position mandibulaire pour [zɛzza] (en arabe tunisien). Au centre la vitesse instantanée; en bas les commandes de raideur pour deux conditions de commande (avec et sans feedback).

Figure 4 : vitesse d'élocution normale

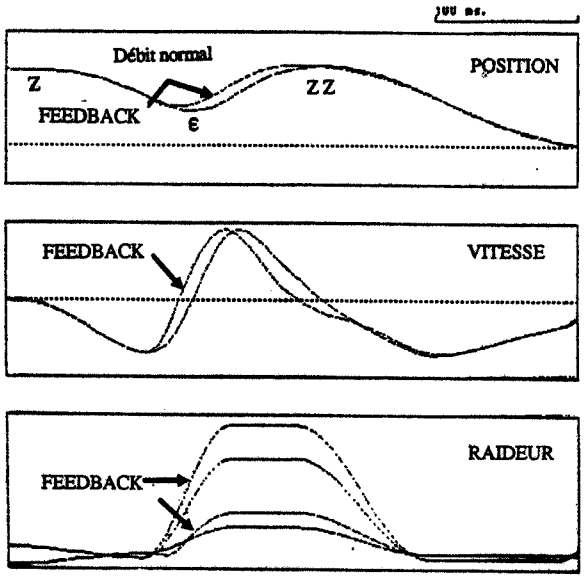
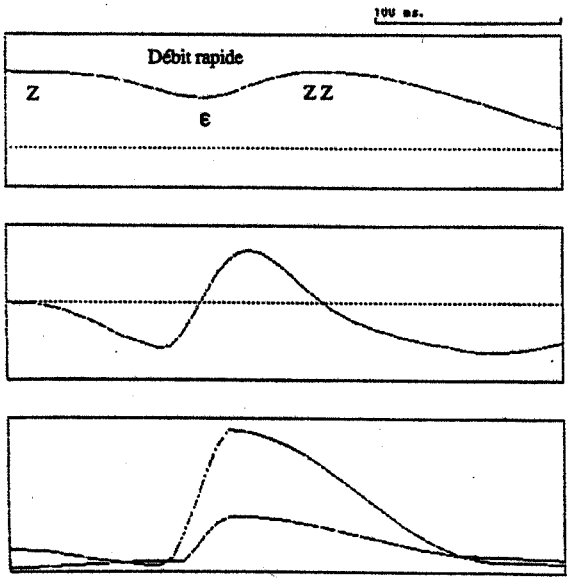


Figure 5 : vitesse d'élocution rapide.



XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

LES DIMENSIONS "CACHEES" DES CONTOURS LABIAUX INTERNE ET EXTERNE ET LEURS RELATIONS AVEC LA MANDIBULE.
Une révision des données de PLANT (1980).

Argyro TSEVA

INSTITUT DE LA COMMUNICATION PARLEE
INPG / ENSERG - UNIVERSITE STENDHAL, URA CNRS n° 368
Domaine Universitaire, B.P. 25 x, 38040 GRENOBLE Cedex, FRANCE

RESUME

Plant (1980), effectuant des mesures labiales sur les voyelles de l'anglais australien, s'est intéressé à l'apport respectif des contours interne et externe. Il en conclut que les paramètres du contour externe ont une faible dynamique ce qui l'amène à rejeter la distinctivité de ces mesures. Il en est de même pour le paramètre qui définit l'excursion mandibulaire. La contribution de cet article est de répondre à la question concernant l'existence d'une distinction physique des formes vocaliques en dépit de cette faible dynamique des paramètres. Pour ceci, nous avons été amenée à re-présenter les données de Plant pour une lisibilité meilleure. Nous avons, par la même occasion, exploité systématiquement ces paramètres pour faire apparaître leurs relations significatives. Nous pouvons ainsi retenir pour l'étirement externe et la mandibule, un pouvoir séparateur qui avait été nié par Plant. La contribution de l'étirement externe est d'autant plus remarquable que c'est en le mettant en relation (par différence) avec l'étirement interne que nous avons montré ses possibilités de discrimination : ainsi la forme du vermillon (ses contours interne et externe) peut être un corrélat important de l'arrondissement.

1. INTRODUCTION

Les mesures physiques (articulatoires) traditionnellement retenues pour définir la forme frontale des lèvres lors de la réalisation des voyelles se limitent généralement aux paramètres liés au contour interne, tels que l'écartement horizontal (A), la séparation verticale (B) et l'aire intérolabiale (S) (et ceci depuis Fromkin, 1964 ; parmi ses successeurs, citons entre autres : Abry & al., 1980 ; Linker, 1982 ; Zerling, 1988). Il faut mentionner cependant quelques exceptions notables*. C'est le cas de Brooke & Summerfield (1983) qui, pour les besoins de la synthèse de mouvements articulatoires visibles, ont utilisé plusieurs points pour circonscrire le vermillon. De même, c'est à Summerfield (1979 ; cf. aussi Summerfield & al., 1989) que l'on doit aussi les tests de perception présentant le vermillon seul, dans le but d'établir son apport dans la communication visuelle (et ceci par rapport aux dents et au visage complet). Cette contribution du vermillon seul par rapport au visage complet est importante puisque les résultats de ces expériences donnaient respectivement 54% ou 50.3% de réponses correctes, contre 65.3% ou 77.8% pour le visage complet. Néanmoins, il faut noter qu'il ne semble pas y avoir, à notre connaissance, des mesures et/ou des expériences perceptives qui mettent en évidence la contribution du bord interne des lèvres (soit la fente labiale) par rapport au bord externe, puisque les présentations du vermillon (avec ou sans les dents) donnent bien entendu simultanément ces deux contours.

Seul Plant (1980), effectuant des mesures labiales sur les voyelles de l'anglais australien, s'est intéressé à cet apport respectif des deux contours. Il en conclut (p. 86) : "... it would appear unlikely that external lip dimensions are [an] important cue in visual vowel identification". Cet argument est repris tel quel par Montgomery & Jackson (1983) pour justifier leurs mesures qui ne prennent en compte que le bord interne. Ils notent à ce propos que Plant n'a effectué aucune analyse systématique des relations entre les dimensions internes (pas plus qu'externes d'ailleurs) et les confusions perceptives qu'il obtient, (analyse que mènent sur leurs propres données Montgomery & Jackson, 1983).

Il semble donc qu'il ne resterait même plus, dans l'état actuel de la littérature, à se poser la question de la validation perceptive des informations sur le bord externe, puisque Plant aurait montré que les conditions nécessaires - les distinctions géométriques - ne sont pas disponibles sur ce contour. Or, l'argument avancé par Plant pour rejeter la distinctivité des mesures externes est pour le moins surprenant : "The differences between the external dimensions of the vowel are very small, the vertical measures being distributed within 14 mm and the horizontal measures within 6 mm" (p. 86). Cet argument sur la faible dynamique des paramètres soulève deux questions :

1. En dépit de cette faible dynamique existe-t-il une distinction physique des formes vocaliques ? Et dans ce cas, ce ne sont certes pas les représentations graphiques données par Plant qui permettent de répondre à cette question - il utilise en effet la même échelle pour les mesures internes et externes, aboutissant ainsi à confondre toutes les données externes sur un petit espace du quadrant droit de sa figure 3. Nous serons donc amenée en premier lieu à re-présenter ses données pour une lisibilité meilleure.

2. Dans le cas où nous trouverions des distinctions physiques, la perception visuelle de la parole utilise-t-elle les possibilités de l'oeil pour d'aussi petites distinctions ? Rappelons que certaines dimensions des lèvres, comme la protrusion, évoluent sur quelques millimètres (cf. Lallouache, 1990). Ce qui n'empêche pas le locuteur d'être particulièrement sensible à cette dimension, puisqu'on a pu montrer qu'en français, par exemple, la frontière perceptive entre les catégories [i] et [y] pouvait basculer sur une variation de l'ordre du millimètre (Cathiard, 1988, p.58).

---

* Notamment le programme sur la perception des voyelles entrepris par Erber et al. (1979) qui inclut aussi des mesures de l'épaisseur des vermillons, des dents et de la mandibule.

Cette deuxième question ayant trait à la validation perceptive des mesures de Plant, restera ouverte. Non parce qu'il serait impossible de soumettre ses données articulatoires et perceptives à une analyse "systématique" appropriée (comme l'ont fait pour leurs données Montgomery & Jackson, 1983). Mais parce que l'objet du présent article est seulement de montrer - comme on l'a déjà deviné - que, sans reprendre d'autres mesures que celles de Plant même (ce qui aurait pour effet d'entacher le débat du soupçon d'un changement sur les conditions des mesures), il est possible de donner la contribution géométrique du bord externe dans la distinction arrondie / non arrondie des voyelles.

Nous exploiterons aussi, par la même occasion, les paramètres donnés par Plant pour faire apparaître leurs relations significatives. Ce faisant, nous pourrons montrer peut-être que c'est la prise en compte à la fois du contour externe et du contour interne qui pourrait fournir les meilleures possibilités de discrimination.

2. METHODE ET MESURES ARTICULATOIRES

2.1. Le corpus de PLANT (1980)

Rappelons tout d'abord les conditions du corpus utilisé par Plant. Les 12 voyelles de l'anglais australien [ɪ, i, ɛ, æ, ɑ, ʌ, ɒ, ʊ, u, ɜ, ə, ɔ] ont été situées dans un contexte [b-b], contexte qui délimite bien le début et la fin du noyau vocalique. Chaque segment était inséré dans la phrase porteuse : "Please say ...". Les phrases retenues ont été enregistrées par un sujet féminin, parlant l'anglais australien dit "général". Le corpus a été filmé en 16 mm, noir et blanc (mais aussi en couleur), à 24 images par seconde avec une caméra Bolex. Seule une zone comprenant le menton, les lèvres et la pointe du nez du locuteur a été cadrée. Pour assurer une distance constante entre la caméra et le locuteur, ce dernier avait la tête maintenue immobile selon une méthode conçue par Abbs & Stivers (1978 ; cité par Plant). Pour des raisons d'étalonnage, un micromètre a été placé devant les lèvres du locuteur et a été filmé avant l'enregistrement du test.

2.2. Mesures

Le film couleur était destiné au test d'identification visuelle que nous n'évoquerons pas ici. Le film noir et blanc a été utilisé pour les mesures articulatoires. Pour ceci, une série de photos a été reproduite en taille réelle, pour chaque syllabe, à partir de la position de la fermeture labiale du [b] initial jusqu'à la fermeture labiale du [b] final. Les mesures articulatoires effectuées visent à définir, d'une part, les contours labiaux (contour interne et contour externe) pour chaque réalisation vocalique et, d'autre part, le degré du mouvement mandibulaire. Parmi la série de photos produite pour chaque syllabe, une seule a été retenue pour les mesures : celle qui correspondait à la partie centrale de chaque noyau vocalique, et ceci aussi bien pour les mesures labiales que mandibulaire.

2.2.1. Dimensions labiales

Cinq paramètres des lèvres ont été mesurés : trois pour la définition du contour interne et deux pour la défintion du contour externe. Nous les donnons ici en reprenant les symboles alphabétiques utilisés depuis Fromkin (1964) ou en créant au besoin de simples dérivés.

Pour le contour interne (Fig. 1, schéma I)
- L'étirement A ("the horizontal width of the lip opening from the internal lip corners")
- La séparation B ("the vertical height of the lip opening measured at the mid-point of the lips")
- Les deux hémi-composantes de la séparation B, soit B1 et B2 ("the vertical distance from the mid-point to the upper and lower lips)

Pour le contour externe (Fig. 1, schéma II)
- L'étirement A' ("the horizontal distance between the outer-most points of the lip corners")
- La hauteur B' ("the vertical distance measured at the mid-point from the vermilion margin of the upper lip to that of the lower lip")

2.2.2. Dimension mandibulaire (Fig. 1, schéma III)

Pour définir l'excursion mandibulaire, l'auteur a mesuré la distance existante entre un point fixé immédiatement au-dessous du nez et l'extrémité du menton.

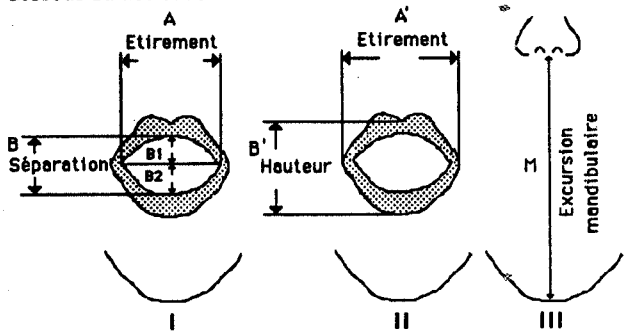


Fig. 1 : Paramètres retenus par Plant (1980) pour la définition du contour interne (schéma I), du contour externe (schéma II) et de l'excursion mandibulaire (schéma III).

2.2.3. Extension

Afin d'obtenir les différences entre deux variétés de l'anglais australien, un deuxième sujet féminin, parlant l'anglais australien dit "large" (broad), a réalisé les mêmes phrases du corpus. Pour cette étude de contrôle, l'auteur a effectué les mesures des paramètres A, B, A' et B' pour chaque voyelle, en suivant la procédure exposée précédemment. Notons, à titre indicatif, que l'anglais australien "général" et l'anglais australien "large" sont utilisés respectivement par 55 % et 34 % des locuteurs (Mitchelle & Delbridge, 1965 ; cité par Plant).

2.2.4. Données

Les données des mesures articulatoires sur les paramètres retenus par Plant sont présentées dans le tableau 1.

	1er Locuteur Anglais Australien "Général"							2ème Locuteur Anglais Australien "Large"			
	A	B	A'	B'	B1	B2	M	A	B	A'	B'
i	33.25	9.80	45.30	24.03	2.98	6.82	65.32	31.14	9.88	45.00	25.35
ɪ	31.30	5.82	44.30	22.66	1.50	4.32	65.32	30.00	6.88	44.04	22.28
ɛ	35.41	10.77	45.80	25.65	1.98	8.79	68.27	33.98	8.78	45.00	23.90
æ	44.40	16.88	45.30	27.15	4.96	11.92	71.36	39.09	10.35	44.04	24.41
ə	40.91	12.84	45.30	32.08	2.98	9.86	69.32	39.32	14.89	44.04	29.36
ʌ	38.41	9.92	44.30	25.00	1.00	8.92	67.84	39.09	11.03	45.00	27.38
ɒ	26.20	9.88	40.35	25.20	2.98	6.90	68.08	25.66	5.42	47.00	22.28
ʊ	10.05	1.04	40.25	18.08	0.52	0.52	62.64	14.94	2.70	43.03	17.33
u	15.00	2.98	40.25	20.00	1.98	1.00	64.78	16.84	3.85	44.04	20.42
ɜ	18.18	4.00	41.30	19.98	2.00	2.00	64.39	20.73	3.85	43.03	21.57
ə	20.62	2.90	45.30	18.08	1.00	1.90	64.19	27.05	2.94	43.03	20.42
ɔ	15.57	1.22	40.25	19.01	0.61	0.61		17.99	3.85	40.00	22.95

Tableau 1 : Mesures articulatoires (en mm) sur les 12 voyelles de l'australien réalisées par deux sujets (résultats de Plant, 1980).

Précisons que ces données ont été extraites à partir des graphiques donnés par l'auteur (ses figures 3, 4, 5 et 7 pour le premier locuteur ; 9 et 10 pour le second). Pour ce faire, ne disposant pas des tableaux de l'auteur, nous avons dû procéder, à un agrandissement (X 4) de chaque figure. Notre marge d'erreur ne dépasse jamais 0.5 mm : cette vérification s'ensuit de l'équation B1 + B2 = B (paramètres pour lesquels l'auteur fournit des résultats pour le premier locuteur). L'erreur relative étant plus petite pour B1 et B2 sur les graphiques donnés, nous avons choisi, une fois cette estimation d'erreur contrôlée, de donner les valeurs de B par la somme de B1 + B2 mesurés.

3. ANALYSE DES DONNEES

Nous nous bornerons dans le traitement du tableau donné ci-dessus à utiliser les représentations graphiques, corrélations et régressions des paramètres deux à deux, dans l'espace de chaque locuteur, sans avoir recours à des traitements

multidimensionnels (que nous avons utilisé par ailleurs : Abry & Benoit, 1985 ; Tseva, 1989). La raison en est que nous voulons garder une interprétabilité physique directe aux discriminations et relations (Abry & Boë, 1986).

3.1.Relations entre les contours interne et externe

L'étude de toutes les relations entre les paramètres communs aux deux locuteurs, soit A, A' et B, B', nous donne l'ensemble des constatations qui vont suivre (cf. Annexe, Tableau 2).

Les paramètres B et B' sont fortement corrélés et ceci pour les 2 locuteurs ($r = 0.90$ et 0.93 avec $r = 0.70$ à $p < 0.01$). En revanche, les corrélations ne sont pas bonnes pour A et A'. Notons que la corrélation A-A' du premier locuteur n'est relativement élevée ($r = 0.80$) qu'à cause de l'éloignement de deux groupes de données (visibles sur la Fig. 2), sans qu'il existe de corrélation interne significative pour chacun des groupes.

Les paramètres A et B sont eux aussi fortement corrélés pour les deux locuteurs ($r = 0.94$ et $r = 0.89$). Cela était déjà visible intuitivement sur les figures 4 et 10 de Plant (p. 86 et 89). Cette situation est le cas le plus général (Fromkin, 1964 ; Montgomery & Jackson, 1983 ; Zerling, 1988) lorsqu'on ne prend pas en compte l'influence des consonnes labialisantes (cf. Abry & Boë, 1986).

L'étirement du contour interne A contribue à la séparation des voyelles en quatre classes [ʊ, u, ɔ, ɜ] / [ɒ, ə] / [ɪ, i, ɛ] / [ʌ, a, æ] aussi bien pour l'anglais australien "général" que pour l'anglais australien "large" (Fig. 2). Cet échelonnement est déjà donné par Plant (p. 86) : "The mid-low front and low central vowels [æ, a, ʌ] have the largest lip opening with the mid-high back, high central and central vowels [ʊ, u, ɜ, ɔ] having an extremely narrow opening".

L'étirement du contour externe A' permet la séparation des voyelles de l'anglais australien "général" (1er locuteur) en deux classes [ʊ, u, ɔ, ɜ, ɒ] / [ə, ɪ, i, ɛ, ʌ, a, æ]. Malgré l'argument avancé par Plant pour rejeter la distinctivité des mesures externes en se basant seulement sur la faible dynamique des paramètres (voir supra), la figure 2 met en évidence que le paramètre A' permet la distinction physique des voyelles, en voyelles arrondies et non arrondies. Rappelons que selon Plant les descriptions classiques de l'anglais australien donnent [ɜ] comme arrondie (p. 86 ; en dépit de l'A.P.I. : centrale ou centrale non arrondie, cf. Pullum & Ladusaw, 1986). Cette distinctivité de A' ne fonctionne pas, en revanche, pour les voyelles de l'anglais australien dit "large" (Fig. 2, deuxième locuteur). Bien que l'échelonnement des voyelles soit le même sur A que pour le premier locuteur, la protrusion n'est pas manifestée ici par l'étirement du contour externe A'. La clé de cette différence dialectale semble être donnée par Dinning (1938, cité par Plant, p. 88), qui caractérise les locuteurs parlant l'anglais australien "large" comme suit : "The Australian often speaks without obviously moving his lips at all, [through] an immobile slit, and in extreme cases through closed teeth". Bernard (1970, cité par Plant, p. 88) effectuant une étude radiologique a trouvé que le locuteur parlant l'anglais australien "large" : "show smaller average lip and teeth apertures than the General subjects". En guise de conclusion provisoire, on peut considérer que A' permet de séparer surtout les deux dialectes de l'anglais australien.

On peut dire ainsi que A remplit une condition de grande généralité, celle que satisfont certains corrélats qui sont des spécificateurs privilégiés, quelles que soient les langues, d'un trait phonologique. De ce point de vue, les valeurs du trait [rond] semblent tout à fait repérables sur le continuum A (Ladefoged, 1975). Par contre, les exigences posées par Ladefoged (p. ex., 1979) de décrire non seulement les spécificités propres à chaque langue dans l'utilisation d'un trait, mais aussi les différences dialectales à l'intérieur d'une même langue, nous rendent le paramètre A' précieux pour une telle spécification.

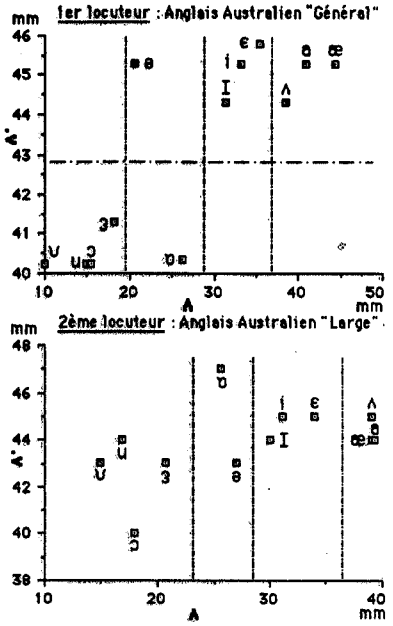


Fig. 2 : Projection sur le plan A-A' des résultats de Plant (1980).

Sur le contour interne, la séparation B, qui est un paramètre généralement moins discriminant que l'étirement A, permet tout de même la distinction des voyelles en deux catégories et ceci pour les deux variétés de l'anglais australien (Fig. 3) : [ʊ, u, ɔ, ɜ, ə] / [ɒ, ɪ, i, ɛ, ʌ, a, æ]. On remarquera la cohérence obtenue par fusion de classes déjà dégagées par A (à l'exception de [ə] et [ɒ] qui étaient déjà placées par A dans une classe intermédiaire entre les arrondies et les non arrondies).

Sur le contour externe, la hauteur B' ne permet pas d'opérer de distinction valable pour les 2 locuteurs.

Notons, comme on pouvait s'y attendre d'après les corrélations A-B et B-B', que les paramètres A et B' sont bien corrélés pour les deux locuteurs ($r = 0.88$ et $r = 0.83$), alors que les paramètres B-A' et A'-B' ne le sont pas.

Nous avons tenté, au vu des combinaisons des paramètres précédents (Fig. 2), une composition A' moins A. Ainsi qu'on peut le constater (Fig. 4) cette différence entre les étirements externe et interne donne une assez bonne séparation entre les voyelles arrondies [ʊ, u, ɔ, ɜ] et les non arrondies [ɪ, i, ɛ, ʌ, a, æ], les deux voyelles [ə] et [ɒ] du groupe intermédiaire déjà dégagé sur A hésitant, selon le locuteur, entre les arrondies et les non arrondies (cf. supra). Notons que ces résultats n'étaient pas obtenus sur A' pour le second locuteur. Ainsi, contrairement à ce qu'annonce Plant, l'efficacité de A' est à prendre en compte, d'autant plus qu'elle se révèle au mieux, tous locuteurs confondus, par la mise en relation des contours interne et externe. A ce titre, une telle mise en relation peut prétendre à autant de généralité pour fournir un corrélat au trait [rond] que peut le faire un paramètre simple comme A.

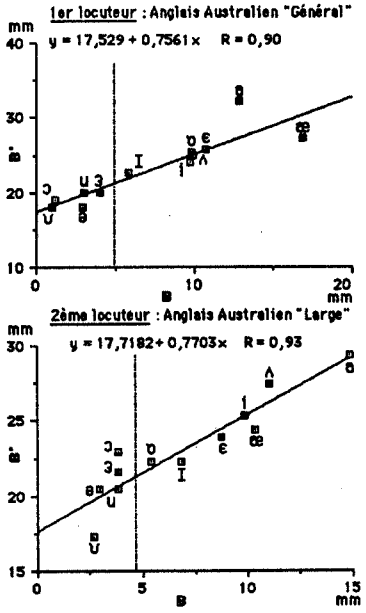


Fig. 3 : Projection sur le plan B-B' des résultats de Plant (1980).

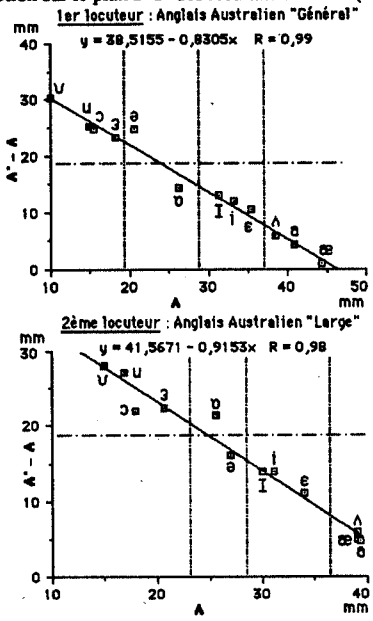


Fig. 4 : Projection sur le plan A - (A' moins A) des résultats de Plant.

3.2. Relations des deux hémi-composantes verticales de la séparation des lèvres avec les dimensions des contours interne et externe

La séparation attribuée à B ([ʊ, u, ɔ, ɜ, ə] / [ʊ, ɪ, i, ɛ, ʌ, ɑ, æ]) se retrouve bien pour B2 mais pas sur B1. La corrélation entre les paramètres B1 et B2, quoique significative, (r = 0.72) présente deux groupes relativement distancés, sans corrélation interne [ʊ, u, ɔ, ɜ, ə, ɪ] et [ʊ, i, ɛ, ʌ, ɑ, æ].

Des deux corrélations avec B, c'est celle avec B2 qui est la meilleure : ce qui confirme ce qui a déjà été dit sur la variation de B qui tient essentiellement à B2 (Fig. 5, en haut). Ceci donne une base quantitative aux remarques de Plant : "... it is the lower lip which contributes most to the lip shapes of the vowels" (p. 86) ; ou encore : "the movement of the lower lip appears to be the most important visible factor in vowel [..] articulation" (p. 91) . Notons que des résultats quantitatifs sur ce point sont intéressants pour un synthétiseur des lèvres, à cause de l'asymétrie verticale qui est la principale composante de B à contrôler (pour l'asymétrie horizontale, cf. Wolf & Goodale, 1987). De ce qui précède, on déduit aisément que la bonne corrélation de B' avec B2 (r = 0.91) et la mauvaise de B' avec B1 étaient toutes deux prévisibles connaissant la corrélation B-B'.

De toutes les relations avec l'étirement interne ou externe, c'est aussi une relation avec B2, celle de l'étirement interne A, qui est la plus nette et cela se comprend puisque l'aperture, principalement due à B2, va de pair avec un plus grand écartement de la fente labiale (Fig. 5, en bas).

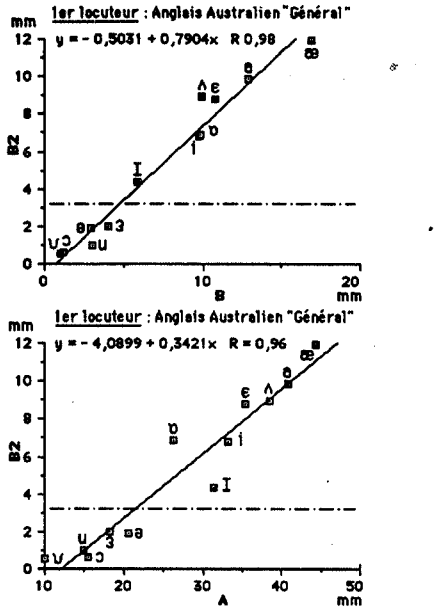


Fig. 5 : Projection sur les plans B-B2 et A-B2 des résultats de Plant.

3.3. Relation des dimensions interne et externe avec les positions mandibulaires

La constatation de Plant (p. 88), à savoir "jaw movement plays a relatively minor role in vowel production" ne peut tenir, car l'argument qu'il réemploie ici est toujours celui de la faible dynamique de l'articulateur : "the difference in the nose tip to chin measure is only 9mm". Contrairement à ce que dit Plant, le paramètre M permet très bien la distinction entre les voyelles à mandibule basse et les autres (Fig. 6, page suivante). Notons que cela range le [ɜ] à proximité de [ə], ce qui est dans l'usage de plusieurs transcriptions des dialectes de l'anglais (cf. Pullum & Ladusaw, 1986).

La corrélation de M avec B est bonne (r = 0.95 ; de même avec B' : r = 0.87) ; et cela tient davantage à B2 (r = 0.94) qu'à B1 (tout juste significatif : r = 0.75). L'étirement de la fente labiale A croissant directement avec B (surtout avec B2) et non directement avec A' (cf. supra), la bonne corrélation de M avec A (r = 0.87) et la mauvaise de M avec A' (r = 0.48) étaient toutes deux prévisibles.

4. CONCLUSION

Les données de Plant, on l'avouera, méritaient un réexamen parce qu'elles étaient à la fois plutôt mal interprétées par leur propre auteur et par ceux qui les citent et les utilisent. Ceci étant d'autant plus grave que ce type de mesures reste relativement rare dans la littérature.

En reprenant ses mesures, nous avons pu quantifier tous

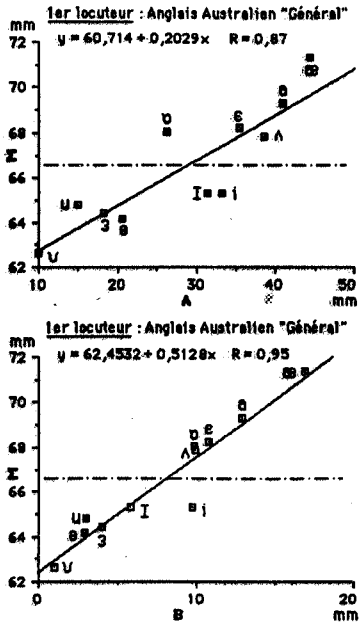


Fig. 6 : Projection sur les plans A-M et B-M des résultats de Plant (1980).

les cas où Plant donnait une simple suggestion sur ses données. La plupart de nos conclusions vont dans le même sens que les siennes tout en les développant. Et là où nous différons de ses conclusions c'est partout où il a utilisé un argument tenant à la trop faible dynamique de tel ou tel paramètre. Il en va ainsi pour l'étirement externe A' et la mandibule M, pour lesquels nous rétablissons un pouvoir séparateur qui avait été nié par Plant. La contribution de l'étirement externe A' est d'autant plus remarquable que c'est en le mettant en relation (par différence) avec l'étirement interne A que nous avons montré ses possibilités de discrimination : ainsi la forme du vermillon (ses contours interne et externe) peut être un corrélat important de l'arrondissement.

Il reste bien entendu à tester perceptivement la contribution de ces contours externe et interne du vermillon : cela est l'objet d'un travail en cours. De tels résultats seront nécessaires pour évaluer les performances d'une synthèse articulatoire à visage humain (pour ce programme cf. Abry & Schwartz in Cathiard, 1988/89), avec un "réalisme" du vermillon qui préserve - au delà d'un simple mimétisme - notre intelligence des mécanismes du système articulatoire (Parke, 1975). Mais il fallait d'abord clarifier des données géométriques comme celles de Plant, pour qu'elles n'apparaissent pas comme contradictoires avec de nouvelles données, alors qu'elles ne le sont pas, dans un domaine où précisément ce type de données est rare.

REMERCIEMENTS

Mes remerciements vont à Christian ABRY, qui a suivi ce travail de près, pour ses commentaires utiles sur l'analyse des données et à Marie-Agnès CATHIARD pour sa lecture critique.

REFERENCES BIBLIOGRAPHIQUES

ABRY C., BOE L.-J., CORSI P., DESCOUT R., GENTIL M. & GRAILLOT P. (1980) Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales. Publications de l'Université des Langues et Lettres de Grenoble.

ABRY C. & BENOIT C. (1985) Quelques exigences venues d'on ne sait haut pour la lecture des plans factoriels en phonétique. Bull. Inst. Phonétique Grenoble 14, 11-23.

ABRY C. & BOE L.-J. (1986) "Laws" for lips.Speech Communication 5, 97-104.

BROOKE N.M. & SUMMERFIELD Q. (1983) Analysis, synthesis and perception of visible articulatory movements. Journal of Phonetics 11, 63-76.

CATHIARD M.-A. (1988) Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français Mémoire de maîtrise, Départ. de Psychologie, Université Grenoble II.

CATHIARD M.-A. (1988/1989) La perception visuelle de la parole : aperçu de l'état des connaissances. Bull. Inst. Phonétique Grenoble 17–18, 109-193.

ERBER N.P., SACHS R.M. & DE FILIPPO C.L. (1979) Labiometrics I : analysis of articulatory dynamics in relation to perception of vowels through lipreading. J. Acoust. Soc. Am., 65 (suppl. 1).

FROMKIN V.A. (1964) Lip positions in American English vowels. Language and Speech 7, 215-225.

JACKSON P.L., MONTGOMERY A.A. & BINNIE C.A. (1976) Perceptual dimensions underlying vowel lipreading performance. J. Speech and Hearing Research 19, 796-812.

LADEFOGED P. (1975) A course in phonetics. Harcourt, Brace, Jovanovich. New York.

LADEFOGED P. (1979) What are linguistic sounds made of ? UCLA Working Papers in Phonetics 45, 1-24.

LALLOUACHE M.T. (1990) Un poste "Visage-Parole". Acquisition et traitement de contours labiaux. 18èmes JEP du G.C.P. de la S.F.A., Montréal.

LINKER W. (1982) Articulatory and acoustic correlates of labial activity in vowels : a cross-linguistic study. UCLA Working Papers in Phonetics 56. 124 p.

MONTGOMERY A.A. & JACKSON P.L. (1983) Physical caracteristics of the lips underlying vowel lipreading performance. J. Acoust. Soc. Am. 73, n° 6, 2134-2144.

PARKE F.I. (1975) A model for human faces that allows speech synchronized animation. Computer and Graphics 1, 3-4.

PLANT G.L. (1980) Visual identification of Australian vowels and diphthongs. Australian Journal of Audiology 2, 83-91.

PULLUM G.K. & LADUSAW W.A. (1986) Phonetic symbol guide. The University of Chicago Press, Chicago. 266 p.

SUMMERFIELD Q. (1979) Use of visual information for phonetic perception. Phonetica 36, 314-331.

SUMMERFIELD Q., MacLEOD A., McGRATH M. & BROOKE M. (1989) Lips, teeth and the benefits of lipreading. In : A.W. YOUNG & H.D. ELLIS (Eds.), Handbook of Research on Face Processing. Elsevier Science Publishers B.V., North-Holland, 223-233.

TSEVA A. (1989) L'arrondissement dans l'identification visuelle des voyelles du français. Premiers acquis. Bull. Laboratoire de la Communication Parlée 3, Grenoble.

WOLF M.E. & GOODALE M.A. (1987) Oral asymmetries during verbal and non-verbal movements of the mouth. Neuropsychologia 25, n° 2, 375-396.

ZERLING J.P. (1988) Les trois degrés de labialisation des voyelles tenues en français. Premiers résultats. 17èmes JEP du G.C.P. de la S.F.A., Nancy. 183-188.

ANNEXE (Etude des corrélations)

Les valeurs du r de Bravais-Pearson pour les paramètres du tableau 1 sont présentées, en ordre décroissant, séparément pour les deux locuteurs. En colonnes de gauche : les valeurs supérieures à $r = 0.70$ ($p < 0.01$).

1er Locuteur
Anglais Australien "Général"

Param.	r	Param.	r
B B2	0.98	A' B2	0.68
A B2	0.96	B' B1	0.67
B M	0.95	A B1	0.65
A B	0.94	A' B	0.64
M B2	0.94	A' B'	0.55
B' B2	0.91	A' M	0.48
B B'	0.90	A' B1	0.36
A B'	0.88		
A M	0.87		
B' M	0.87		
B B1	0.83		
A A'	0.80		
M B1	0.75		
B1 B2	0.72		

2ème Locuteur
Anglais Australien "Large"

Param.	r	Param.	r
B B'	0.93	A A'	0.45
A B	0.89	A' B	0.39
A B'	0.83	A' B'	0.26

Tableau 2 : Valeurs du r de Bravais-Pearson.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

L'EFFICACITE DES CYCLES ACOUSTIQUES DANS LA DISTINCTION DES QUANTITES VOCALIQUE ET CONSONANTIQUE
en arabe marocain

RHARDISSE N., SOCK R. & ABRY C.

INSTITUT DE LA COMMUNICATION PARLEE
U.R.A. - C.N.R.S. N° 368
Université Stendhal - Domaine Universitaire
38040 GRENOBLE CEDEX B.P. 25 X

ABSTRACT

In this article, we examine the behaviour of different phonetic structures as a function of speaking rate in Moroccan Arabic. The importance of timing in speech is demonstrated by languages such as Arabic, where a precise temporal control is required for semiotic goals related to quantity contrasts.

The relative timing approach adopted here, provides us not only with an experimental paradigm, together with associated concepts and statistical tools comparable with those used by psychomotricians working on movement control, it also allows us to evaluate the efficiency of cycles retained in separating specific phonetic classes.

Implications for certain phonological changes are discussed, as portrayed by the displacement of the different phonetic classes along the cycles investigated under rate conditions.

INTRODUCTION

Nous examinons dans ce travail le comportement de différentes structures phonétiques en arabe marocain, face à la variation de la vitesse d'élocution. L'importance de l'organisation temporelle (*timing*) en parole n'est plus à démontrer. Il suffit de se référer aux cas linguistiques qui demandent un contrôle temporel précis à des fins sémiotiques : c'est le cas de l'arabe où le contrôle de la durée vocalique et/ou consonantique est phonologique. De manière plus générale, le décours temporel de la parole doit être contrôlé - que ce soit au niveau segmental ou suprasegmental - pour préserver son intelligibilité et son naturel. Le choix d'un paradigme de variation de la vitesse d'élocution nous permettra ainsi d'apprécier l'élasticité temporelle (GAITENBY, 1965) de nos classes phonétiques, en particulier les facteurs de compression dans le changement de débit, facteurs qui constituent des données fondamentales pour les systèmes de synthèse.

Pour observer le comportement de la quantité, l'approche relative de la durée (LEHISTE, 1970), systématisée par la définition de cycles et de phases (KELSO et al., 1986), ne nous fournit pas seulement un paradigme expérimental, comparable à celui des psychomotriciens spécialistes en contrôle du mouvement, avec les concepts associés, comme le **phasage** (par exemple SHAPIRO et al., 1981) et des outils statistiques pour tester l'**invariance relative** (GENTNER, 1987). Mais en plus, elle nous permet d'évaluer l'**efficacité des cycles** retenus dans la séparation de tâches phonétiques déterminées (SOCK et al., 1987), comme ici les oppositions de quantité de l'arabe.

Enfin un cas, rencontré dans notre corpus d'arabe marocain, d'une **voyelle ultra-brève**, ainsi que le phénomène observé d'une **convergence** des différentes classes (VC, VVC, VCC et VVCC) **vers une structure simple VC**, lorsqu'on augmente la vitesse d'élocution, nous permettra de rappeler les processus linguistiques auxquels participent les phénomènes examinés.

MATERIAUX LINGUISTIQUES

CORPUS

Notre corpus est constitué de quatre items, quatre verbes du lexique de l'arabe marocain. Ces verbes constituent de véritables paires minimales permettant de tester les effets de gémination consonantique sur la longueur vocalique. Nous avons inséré chaque item dans la phrase porteuse interrogative suivante : [**qulha** (item) **dɛ:ba**] "Tu lui dis : ' _____ ', maintenant ?"

Les quatre paires minimales choisies sont les suivantes : [**qadu**] "il l'a conduit", [**qaadu**] "ils ont dirigé", [**qaddu**] "il est à sa mesure", [**qaaddu**] " il l'a égalisé"; soit VC, VVC, VCC et VVCC.

ENREGISTREMENT

Nous avons obtenu 12 répétitions de chaque item pour un locuteur marocain (étudiant originaire de Oujda) sous deux variations de vitesse d'élocution, normale et rapide. La première série d'enregistrement a été faite, en ordre aléatoire, à la vitesse d'élocution qui a paru la plus normale au locuteur. La deuxième série a été ensuite enregistrée en exigeant un débit rapide (le résultat obtenu est un facteur d'accélération de 1.40 en moyenne, mesuré après l'enregistrement par ajustement auditif d'un métronome sur le rythme syllabique).

MESURES

Le signal acoustique numérisé (à 16 KHz sur 12 bits) a été étiqueté manuellement en événements (ABRY et al., 1985) à l'aide d'un éditeur de signal (BENOIT, 1984).

Nous avons retenu à partir d'événements articulatori-acoustiques répétitifs sur le signal acoustique (CFO, VVO et VVT) les trois cycles suivants (fig. 1) : le **cycle détente** (de CFO à CFO); le **cycle vocalique** (de VVO à VVO); et le **cycle closion** (de VVT à VVT).

- Le **cycle détente** est déterminé par la répétition de l'événement CFO (début de la friction consonantique), soit d'un relâchement supraglottique à un autre.

- Le **cycle vocalique** est déterminé par la réitération de l'événement VVO (début du voisement vocalique), soit de l'établissement d'une structure formantique définie à une autre.
- Le **cycle closion** est déterminé par la récurrence de l'événement VVT (fin de voisement vocalique), soit de la disparition d'une structure formantique définie à une autre disparition de cette structure.

Nous observerons le comportement de l'opposition de quantité dans ces cycles selon la variation de débit (condition normale et condition rapide), en nous focalisant plus particulièrement sur les phases vocaliques et les tenues consonantiques (voir fig.1), qui sont, respectivement, les corrélats physiques privilégiés de la quantité vocalique et consonantique.

- La **phase** vocalique **D.VOC.** est la relation temporelle entre VVO et VVT, soit la partie de la voyelle présentant une structure formantique clairement définie.
- La **phase consonantique** est la relation temporelle entre VVT et CFO, soit la tenue consonantique.
- Les **phases VOT** (de CFO au VVO suivant , ce qui est la définition du VOT au sens de KLATT, 1975) - la première (détente de [q]) incluse seulement dans le cycle CFO-CFO, la seconde (détente des dentales voisées) comprise dans les deux autres cycles - n'ont pas été représentées graphiquement : leurs proportions se déduisent des phases vocaliques et consonantiques du cycle qui les contient.

RESULTATS

L'opposition des tâches vocaliques dans le cycle détente (CFO-CFO) (fig.2)

En débit normal, la structure VC se différencie de la structure VVC surtout sur le plan du cycle, mais avec une différence moyenne faible, d'environ 23 ms (significative t=5.44; seuil de t=1.75 à $p \leq 0.10$). Par contre les structures VCC et VVCC s'opposent en débit normal aussi bien par la phase (environ 22% de différence, t=17.82) que par le cycle (environ 109 ms de différence, t=14.50). En passant au débit rapide, les différences de phase et de cycle entre les structures VC et VVC se maintiennent, structurellement parlant. On constate par contre un rapprochement entre les structures VCC et VVCC dû à une compression remarquable du cycle pour la classe VVCC (il n'y a plus qu'une différence de 46 ms entre les deux tâches phonétiques). Il faut noter que la classe VCC, elle, ne peut plus être comprimée davantage que par sa partie consonantique (de 168 ms en débit normal à 121 ms en débit rapide), étant donné que la voyelle est de l'ordre de 19 ms déjà en débit normal. Ceci n'empêche pas les différences de phase d'être non significatives entre les deux débits : une phase vocalique moyenne de 17 ms en débit rapide donnant, dans le cycle, des pourcentages moyens de l'ordre de 10%, dont la variation est relativement importante. (Le cas de cette voyelle ultra-brève sera discuté plus loin).

Les relations de phase en débit rapide restent donc, dans l'ensemble, comparables à celles du débit normal (différences non significatives), sauf pour VVCC qui connaît une augmentation légère (avec t=3.19).

L'opposition des tâches consonantiques dans le cycle détente (CFO-CFO) (fig. 3)

En débit normal, la structure VC se différencie de la structure VCC sur le plan de la phase (environ 45% de différence, t=35.71), une séparation qui n'est pas possible au niveau du cycle. L'opposition VVC et VVCC se fait aussi bien sur le plan de la phase (environ 28% de différence, t=19.27) que sur le plan du cycle (avec une différence d'environ 57 ms, t=6.79). En passant au débit rapide, l'opposition entre VC et VCC se maintient, avec une différence comparable au niveau de la phase. Par contre la différence entre les structures VVC et VVCC ne tient plus décisivement qu'à une opposition de phase en débit rapide.

Les différences de phase entre débit ne sont significatives, encore une fois, que pour la classe VVCC qui accuse une légère diminution (avec t=4.51), complémentaire de l'augmentation de la phase vocalique (le VOT restant à peu près constant en proportion).

L'opposition des tâches vocaliques dans le cycle vocalique (VVO-VVO) (fig. 4)

On n'observe pas de différence significative sur la phase vocalique entre la classe VC et VVC en débit normal, la séparation entre ces deux classes se faisant sur le plan du cycle (avec environ 43 ms de différence, t=6.36). Par contre la séparation entre les classes VCC et VVCC en débit normal se réalise parfaitement au niveau de la phase vocalique (environ 23% de différence, t=16.63) et du cycle (104 ms, t= 16.93). La structure des oppositions reste la même, lorsqu'on passe au débit rapide, avec une possibilité de séparer les classes VC de VVC uniquement par le cycle. En ce qui concerne les classes VCC et VVCC, la séparation au niveau de la phase se maintient, mais le rapprochement entre ces classes se fait par la compression du cycle de la classe VVCC (la séparation entre ces deux classes ne tient plus qu'à une différence de 34 ms, t=5.50).

On notera que dans ce cycle (VVO-VVO) - légèrement décalé dans le temps, c.-à-d. du VOT de [q], par rapport au précédent (CFO-CFO) - les différences sur la phase vocalique sont tout juste significatives (sauf encore une fois, pour VVCC : t=4.65), la tendance étant là aussi, dans l'ensemble, plutôt à la stabilité.

L'opposition des tâches consonantiques dans le cycle vocalique (VVO-VVO) (fig. 5)

La séparation entre VC et VCC en débit normal se fait au niveau de la phase (environ 42% de différence, t=28.03). Au niveau du cycle la séparation est tout juste significative. Par contre pour séparer la classe VVC de la classe VVCC, la phase se révèle aussi efficace que le cycle (avec, respectivement, 26%, t=16.75 et 82 ms environ de différence, t=9.79). En passant au débit rapide, les relations structurelles entre VC et VCC restent semblables à celles du débit normal. Entre les structures VVC et VVCC, la différence de phase se maintient alors qu'on observe une réduction sur la différence de cycle (43 ms, t=4.15).

Entre les débits, les différences de tenue consonantique ne sont tout juste significatives que pour VCC (t=2.59) et VVCC (2.29). Ceci rejoint, avec quelques variations de la tenue consonantique complémentaires de celles de la phase vocalique dans ce même cycle, l'observation générale sur la stabilité des phases.

L'opposition des tâches vocaliques dans le cycle closion (VVT-VVT) (fig. 6)

Ce cycle n'est, bien entendu, pas le cadre privilégié pour séparer les tâches phonétiques **vocaliques**, étant donné que l'opposition de quantité vocalique ne se réalise pas essentiellement dans ce domaine temporel. Seuls des **effets** sont donc visibles, sur une différence de phase vocalique entre VCC et VVCC, tout juste significative en débit normal (t=2.01), et sur le cycle, en débit rapide (t=5.54).

Les différences de phase entre les débits sont, elles, tout juste significatives pour VC (t=2) et VVCC (t=3.42), donnant à l'ensemble une allure plutôt stable à travers les variations de débit.

L'opposition des tâches consonantiques dans le cycle closion (VVT - VVT) (fig. 7)

La séparation entre VC et VCC se fait en débit normal au niveau de la phase (avec environ 19% de différence, t=13.12) et au niveau du cycle (environ 97 ms de différence, t=8.96). Le scénario est le même pour l'opposition des structures VCC et VVCC. En passant au débit rapide, si la séparation en cycle est réduite, l'opposition en phase, elle, se maintient entre toutes les

tâches phonétiques.

Aucune différence de phase significative ne se manifeste entre les débits pour les mêmes classes.

Le lecteur aura bien entendu remarqué, au fil de nos commentaires, que de manière générale, les phases consonantiques dans les trois cycles sont, au VOT près, complémentaires des phases vocaliques.

* *
*

Nous avons observé que l'**efficacité des cycles** n'est pas identique pour les différentes tâches linguistiques à distinguer. L'avantage de notre lecture du signal en événements récurrents est de nous permettre de définir plusieurs cycles chevauchants - décalés chaque fois d'une phase - comme champs d'observation des maneuvres de la quantité vocalique et consonantique. On peut ainsi choisir le domaine temporel qui nous permet au mieux de distinguer les classes linguistiques en jeu.

Le cycle vocalique (VVO-VVO) semble être privilégié pour la mise en évidence de l'opposition de quantité vocalique en arabe marocain. Par rapport au cycle détente (CFO-CFO) - dont il diffère seulement par le fait qu'il ne prend pas en compte le VOT de [q], mais celui de [d] ou de [dd] -, il sépare mieux les classes phonétiques, en réduisant légèrement les dispersions intraclasses: ce qui lui permet d'être globalement plus discriminant. Notons tout de même que pour ces deux cycles, assez semblables de par leur domaine acoustique, les structurations des patrons de phase de la quantité vocalique se ressemblent fortement.

En se déplaçant dans le signal, on remarque que le cycle closion (VVT-VVT) nous offre un meilleur domaine pour l'opposition de la quantité consonantique : il reflète au mieux la partition consonnes simples/consonnes doubles, sans ajouter trop d'effets qui soient dûs à la quantité vocalique choisie sur la première voyelle [a] ou [aa].

DISCUSSION

A regarder les quatre patrons de phases (VC, VVC, VCC et VVCC) se transformer sous la variation de la vitesse d'élocution (du débit normal au débit rapide), on peut établir plusieurs observations qui concernent au premier chef le déplacement des classes. Ces phénomènes peuvent nous aider à comprendre les changements phonétiques, lesquels sont souvent dûs, parmi d'autres facteurs, à des conditions qui tiennent à la vitesse d'élocution (McCARTHY, 1986; BROWMAN & GOLDSTEIN, 1989).

On remarque tout d'abord, pour les cycles privilégiant les oppositions vocaliques - et ceci tout particulièrement dans le cycle détente (fig. 2) - une tendance à la **convergence** des structures complexes vers la structure la plus simple VC : impression générale qui est due essentiellement à l'évolution significative en cycle et en phase de VVCC, dans le passage au débit rapide. Ce phénomène de translation des structures complexes vers une structure simple a déjà été signalé pour d'autres dialectes de l'arabe - pour lesquels nous possédons des observations sur les effets du débit - comme le koweïtien (AL-DOSSARI, 1989, in DELATTRE et al., 1990). Le point de convergence des structures en timing relatif vers le phasage de VC (ici, env. 50%) est, à peu de choses près, le même pour ce dernier dialecte (env. 45%). Dans notre cas, on s'aperçoit que c'est l'élément **consonantique long** de VVCC qui a le plus accusé la compression : le rapport VV/CC passe en moyenne de 102/184 en débit normal à 80/93 en débit rapide (durées en ms).

Les tendances de VVC et VCC de l'arabe koweitien vers VC montrent bien que ce sont les éléments **longs** - qu'ils soient vocaliques **ou** consonantiques - qui sont les plus touchés (VV/C = 152/107 --> 116/88 et V/CC = 88/167 --> 72/107); mais il nous manque dans ce cas des informations sur ce qui se passe quand les deux éléments longs sont présents (VVCC).

On remarquera que, sur VVC, notre dialecte ne subit pas de changement de phasage, en dépit de la compression de la voyelle longue. Cela est tout à fait différent de ce qui se passe en arabe tunisien (JOMAA & ABRY, 1988 : VV/C = 113/85 --> 63/75 et DELATTRE et al., 1989) et ainsi que nous venons de le voir en koweïtien.

Mais ce qui semble différencier le plus, au premier abord, le comportement de la classe VCC de notre dialecte par rapport aux autres, est la valeur d'emblée ultra-brève de la voyelle (19 à 17ms). Cette voyelle ultra-brève (comparable aux éléments vocaliques dits épenthétiques ou svarabhactique) a une identité en timbre pratiquement inexistante par rapport à son entourage consonantique et vocalique. Rappelons que SERNICLAES et WAJSKOP (1971) ont trouvé que l'identification des voyelles s'optimisait plutôt à partir de 40 ms (en dessous de ce seuil, les voyelles réduites sont mal reconnues). Rappelons encore que, pour l'arabe libanais, ABOU-HAIDAR (1988) trouve tout de même 90 ms environ pour le [a] bref devant consonnes géminées ou dédoublées.

Il semble, dans le cas de l'arabe marocain, que l'on ait affaire à une séquence du type CCCV [qddu], avec insertion d'un élément vocalique épenthétique par resyllabification, donnant la structure $C^{V}CCV$ [$q^{a}ddu$]. De tels phénomènes traités plus généralement sous le nom de "Loi de Dorsey" (Dorsey's Law) sont intégrés à la fois dans les préoccupations récentes de la phonologie articulatoire (BROWMAN & GOLDSTEIN, 1986) et dans le cadre de la phonologie non linéaire (CLEMENTS ,1987). STERIADE (1988) prend ainsi en compte et discute les propositions de BROWMAN & GOLDSTEIN (1986), qui expliquent - par des variations dans le timing des gestes consonantiques surimposés aux gestes vocaliques - ces "insertions-apparitions" de voyelles. Le travail précurseur de PRICE (1980) sur sonorité et syllabicité, avait déjà pu montrer que la perception de différences - comme "prayed"/"parade", "plight"/"polite" ou encore "round"/"around" - repose sur des corrélats acoustiques autres que la durée vocalique (épenthétique), lorsque celle-ci est expérimentalement neutralisée, et ceci grâce à un phénomène de relais.

Quoiqu'il en soit, nous avons déjà signalé que l'absence de différence significative entre les phasages des débits normal et rapide, tenait à la très faible valeur de la voyelle dans le rapport V/CC. Celui-ci évolue de 19/168 à 17/121. C'est donc bien finalement l'élément **long** de VCC qui subit toute la compression due au débit, en arabe marocain; ce qui rend le phénomène comparable à ce qui se passe en koweïtien (mais pas en tunisien où V/CC garde son phasage : cf. JOMAA & ABRY, 1988 et DELATTRE et al., 1990).

CONCLUSION

L'observation des patrons de phase et de leur résistivité face aux variations de la vitesse d'élocution, dans la séparation des tâches linguistiques faisant appel à des contrôles de durée différenciés, nous donne, par la prise en compte des comportements linguistiques de dialectes apparentés - comme les variétés marocaine, tunisienne et koweïtienne de l'arabe - une leçon sur les contraintes propre à la parole, qui sont trop souvent oubliées dans les travaux qui recherchent une invariance généralisée en timing relatif (ou angle de phase, KELSO et al., 1986), sur le modèle des études en contrôle du

mouvement (SHAPIRO et al., 1981). Dans plusieurs situations que nous venons d'évoquer, en particulier lorsqu'un segment s'avère incompressible, comme le sont parfois les voyelles **brèves** (SOCK, 1983; ABRY et al., 1989) et à plus forte raison les **ultra-brèves** (épenthétiques), on voit mal comment la parole maintiendrait une invariance temporelle relative, tout en se comprimant au niveau segmental. De plus, on connaît depuis fort longtemps le caractère non-linéaire des compressions en débit rapide (GAITENBY, 1965), même si l'on ne sait pas toujours exactement quel segment **long** - vocalique **ou** consonantique - sera le plus fortement comprimé. Le fait que les segments **longs** soient atteints en priorité par le débit n'est certes pas pour nous surprendre, mais l'on sera sans doute étonné par deux choses. En premier lieu par le fait, maintenant attesté par l'arabe marocain, que les voyelles longues ne sont pas les plus atteintes par rapport aux consonnes longues, lorsque les deux sont en présence dans les structures VVCC. Les consonnes longues subissent ainsi des compressions importantes dans VVCC et dans VCC. Mais le fait, qui n'est sans doute pas le moins important, et dont on doit tenir compte, est que les compressions les plus classiquement connues peuvent ne pas donner lieu à une perte d'invariance relative, pour un milieu linguistique donné : ainsi tout le monde ne fait pas tomber ses voyelles longues de VVC en VC...

Par conséquent, l'amélioration de nos connaissances sur le contrôle temporel des gestes dans la parole, ne pourra se faire sans une réflexion sérieuse sur les contraintes spécifiquement linguistiques de cette organisation temporelle, telles qu'elles nous sont accessibles à travers la variété des systèmes réalisés.

REFERENCES

ABOU-HAIDAR L. (1988) La Gémination en Arabe : Consonne Double et Consonne Géminée. Etudes Contrastives - Locuteur Libanais et Locuteur Marocain.
in "Rencontres Régionales", Actes du 3ème Colloque Régional de Linguistique, 5-55.

ABRY C. BENOIT C. BOE L.J. SOCK R. (1985) Un Choix d'Evénements pour l'Organisation Temporelle du Signal de parole.
14ème JEP du GCP du GALF, 133-137.

ABRY C. ORLIAGUET J.P. SOCK R. (1989) Patterns of Speech Phasing. Their Resistivity in the Production of a Linguistic Timed Task : Single versus Double (Abutted) Consonants in French.
Actes du Premier Workshop Régional de Sciences Cognitives, LASCO 3, Chichilianne 19-21 mars, 134-166.

AL-DOSSARI A. (1989) "Le Phasage des Gestes Mandibulaires" Vocaliques et Consonantiques en Arabe Koweïtien.
Mémoire de D.E.A de Sciences du Langage (dir. C. ABRY), Université Stendhal, Grenoble.

BENOIT C. (1984) EDISIG : Encore un Editeur de Signal ?!
13ème JEP du GCP du GALF, 211-213.

BROWMAN C.P. GOLDSTEIN L.M. (1986) Towards an Articulatory Phonology.
Haskins Labs, Status Rept., Speech Res. 85, 219-249. (aussi : Phonology Yearbook 3, 219-252).

BROWMAN C.P. GOLDSTEIN L.M. (1989) Gestural Structures and Phonological Patterns.
Haskins Labs, Status Rept. Speech Res. 97/98, 1-23.

CLEMENTS G.N. (1987) Phonological Feature Representation and the Description of Intrusive Stops.
CLS 23, Parasession on Metrical and Autosegmental Phonology, 30 - 50.

DELATTRE C. JOMAA M. AL-DOSSARI A. WORLEY C. (1989) The Phasing of the Jaw in Consonant and Vowel Lengthening : Arabic a n d French Patterns.
European Conference on Speech Communication and Technology, vol. 2, 416-419.

DELATTRE C. JOMAA M. AL-DOSSARI A. WORLEY C. SOCK R. (1990) Comparaisons Articulatori-Acoustiques des Structures Temporelles en Arabe et en Français ou "peut-on séparer les classes dans les VC ?"
à paraître in 18ème JEP du GCP de la SFA (Montréal).

GAITENBY J.H. (1965) The Elastic Word.
Haskins Labs., Status Rept., Speech Res. 2,1-12.

GENTNER D.R. (1987) Timing of Skilled Motor Performances. Tests of the Proportional Duration Model.
Psychological Review 82, 225-260.

JOMAA M. ABRY C. (1988) La Résistivité de la Quantité Vocalique aux Variations de la Vitesse d'Elocution. Le cas de l'Arabe Tunisien.
17ème J.E.P du G.C.P de la S.F.A, 231-236.

KLATT D. (1975) Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters.
J. Speech Hearing Research 18, 686-706.

KELSO F.A.S . SALTZMAN E.L . TULLER. B. (1986) The Dynamical Perspective on Speech Production : Data and Theory.
Journal of Phonetics 14, 29-56.

LEHISTE I. (1970) Suprasegmentals.
The M.I.T Press, Cambridge Mass.

McCARTHY J. (1986) OCP Effects : Gemination and Antigemination.
Linguistic Inquiry 17, 207-260.

PRICE P. J. (1980) Sonority and Syllabicity : Acoustic Correlates of Perception.
Phonetica 37, 327-343.

RHARDISSE N. (1989) La Résistivité des Oppositions de Quantité Vocalique et Consonantique de l'Arabe Marocain aux Variations de la Vitesse d'Elocution.
Travail d'Etude et de Recherche de Maîtrise en Sciences du langage (dir. C. ABRY), Université Stendhal, Grenoble.

SERNICLAES W. WAJSKOP M. (1971) L'Identification Vocalique en Fonction de la Fréquence Fondamentale et de la Durée de Présentation.
Rapport d'Activités de l'Institut de Phonétique de Bruxelles 4, 54-70.

SHAPIRO D.C ZERNICKE R.F GREGOR R.J DIESTEL J.D. (1981) Evidence for Generalized Motor Programs Using Gait Pattern Analysis.
J. Motor Behav.13/1, 33-47.

SOCK R. (1983) L'Organisation Temporelle de l'Opposition de Quantité Vocalique en Wolof de Gambie. Sa Résistivité aux Conditions de Durée Segmentales et Suprasegmentales.
Thèse de Doctorat de 3ème Cycle, Université Stendhal, Grenoble.

SOCK R. OLLILA L. DELATTRE C. ZILLIOX C. ZOHAIR L. (1987) Timing Intersegmental et Intrasegmental en Français.
16ème J.E.P du G.C.P de S.F.A, 233-236.

STERIADE D. (1988) Gestures and Autosegments : Comments on Browman and Goldstein's "Gestures in Articulatory Phonology".
A paraître in J. Kingstone and M. Beckman (eds).
Papers in Speech. Cambridge University Press. M.I.T.

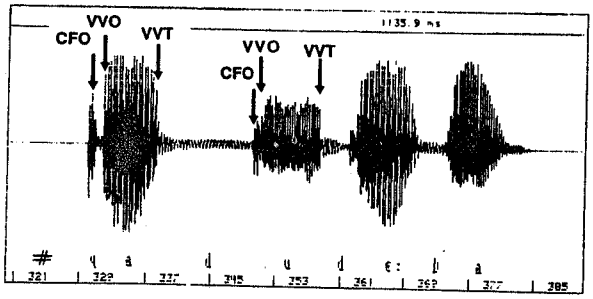
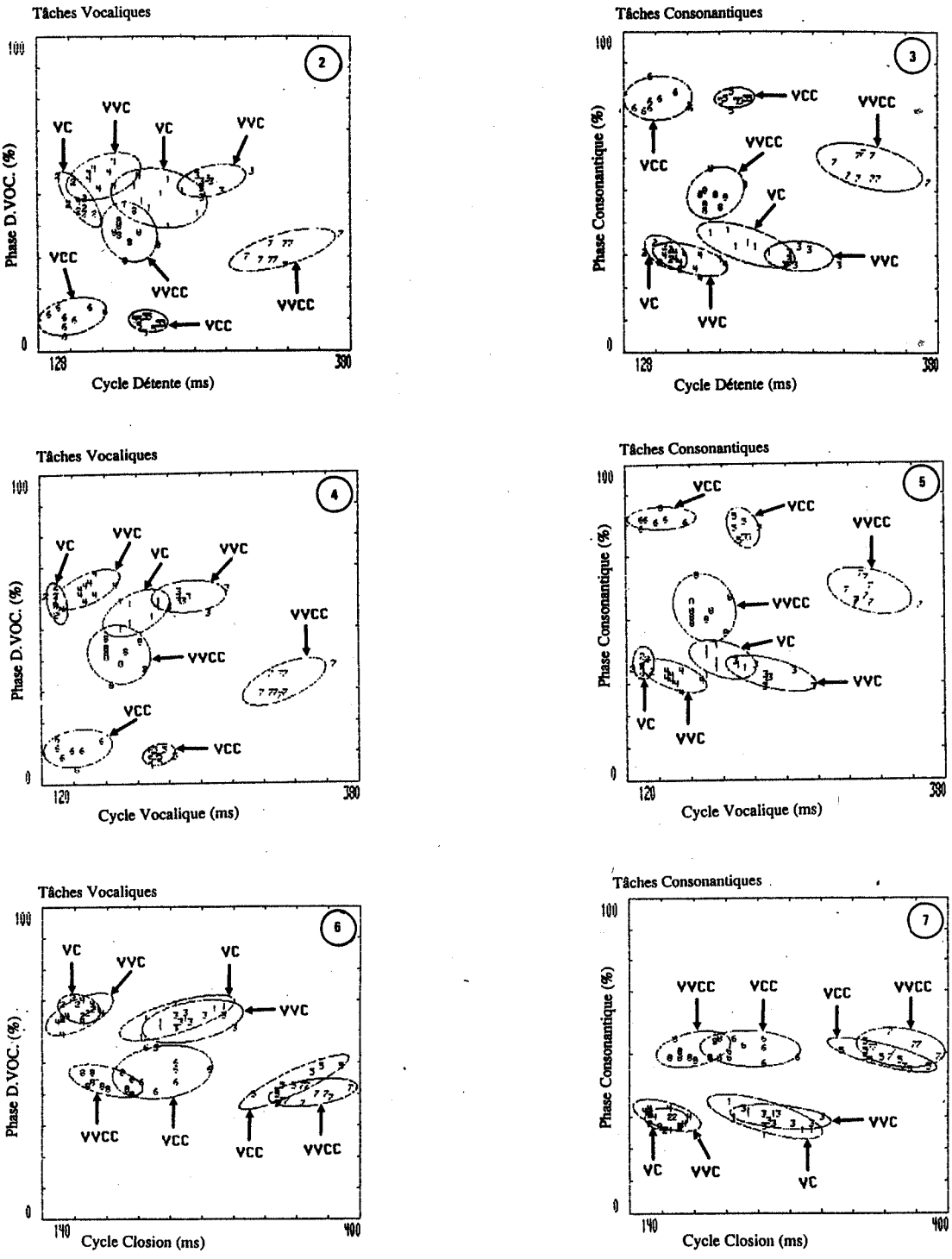


Figure 1. Les Cycles et les Phases mesurés sur la séquence [# qadu deba].

Les Cycles : **Le Cycle Détente** : d'un relâchement à un autre (CFO-CFO) ;
Le Cycle Vocalique : de l'établissement d'une structure formantique vocalique à un autre (VVO-VVO) ;
Le Cycle Closion : de la disparition d'une structure formantique vocalique à une autre (VVT-VVT).

Les Phases : **VOT** (*Voice Onset Time* ou délai d'établissement du voisement) : de CFO à VVO pour [q] et [d] ou [dd] ;
D.VOC. (partie formantiquement définie de la voyelle) : de VVO à VVT pour les voyelles [a] ou [aa] et [u] ;
Consonantique (ou tenue consonantique) : de VVT à CFO pour la deuxième consonne [d] ou [dd].



Figures 2 - 7.
Ellipses de dispersion pour l'ensemble des classes phonétiques dans les deux débits (1, 3, 5, 7 = débit normal ; 2, 4, 6, 8 = débit rapide) pour les Phases (%) D.VOC. et Consonantiques en fonction des Cycles (ms) : Détente (Fig. 2 & 3) ; Vocalique (Fig. 4 & 5) et Closion (Fig. 6 & 7).

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

COMPARAISONS ARTICULATORI-ACOUSTIQUES DES STRUCTURES TEMPORELLES EN ARABE ET EN FRANCAIS
ou "peut-on séparer les classes dans les VC ?"

DELATTRE C. JOMAA M. AL-DOSSARI A. WORLEY C. SOCK R.

INSTITUT DE LA COMMUNICATION PARLEE - U.R.A. C.N.R.S. N° 368
Université STENDHAL - Domaine Universitaire
38040 GRENOBLE CEDEX B.P. 25 X

Résumé

Cet article a pour but de mettre en évidence de possibles correspondances articulatori-acoustiques pour deux tâches motrices à finalité sémiotique différente : le contrôle du timing de la quantité vocalique et celui de la quantité consonantique, par l'intermédiaire d'un articulateur "porteur", la mandibule. Et celà à travers deux conditions de vitesse d'élocution (normale et rapide), en français et pour deux dialectes arabes (le koweïtien et le tunisien). Les correspondances articulatori-acoustiques mises en évidence dans les trois langues sont remarquables. Mais, il faut bien noter qu'il ne s'agit là que de correspondances structurales et non terme à terme, les domaines (les cycles) déterminés par l'acoustique et l'articulatoire n'étant bien entendu pas strictement en phase.

1. SITUATION THEORIQUE

Dans le cadre des études portant sur l'organisation temporelle ou *timing* de la parole, l'approche relative s'est avérée rentable pour la spécification des tâches linguistiques dans un cycle VC comme domaine élémentaire d'investigation. Le phénomène quasi-cyclique de la parole, qui comprend sous une forme plus ou moins régulière cette composante rythmique de base, nous permet ainsi d'adopter un paradigme, proche de celui des psychomotriciens, pour étudier les différents phasages (*phasing*), associés à différentes tâches phonétiques. Cette étude se focalisera sur deux cycles, articulatoire et acoustique, autant que possible en correspondance, comme champs d'observation de deux tâches motrices à finalité sémiotique : à savoir la production des quantités vocalique et consonantique. La variation de la vitesse d'élocution comme facteur perturbateur du contrôle de la durée vocalique et/ou consonantique - examiné ici en français et en arabe -, nous permettra d'évaluer les degrés de résistivité de nos différentes structures syllabiques face aux conditions destructrices de la quantité. On pourra dès lors comparer le pouvoir séparateur du milieu articulatoire avec celui de sa conséquence, le milieu acoustique, dans la distinction des classes linguistiques.

En outre, l'analyse des possibles réorganisations gestuelles, - dues au changement de débit -, qui peuvent apparaître dans ce paradigme, devrait nous offrir les moyens d'appréhender les mécanismes des **changements phonétiques diachroniques** [1]; les conditions de débit rapide étant considérés comme défavorables pour le maintien de structures **bigestuelles**. L'orientation de tels processus évolutifs est bien entendu régie, non seulement par les contraintes prises en compte par la **théorie du contrôle** en contexte d'exécution difficile, mais aussi par des exigences de haut niveau pour le maintien des différences sémiotiques.

Cette étude des variations de la quantité pourrait avoir pour cadre la phonologie autosegmentale qui fournit un continuum représentationnel pour les phénomènes de gémination et de longueur, basé essentiellement sur le nombre d'unités sur la couche CV de timing (voir, par exemple, [2]; ... ; [3]; et pour un test acoustique - négatif - concernant la manifestation en timing des différences représentationnelles des géminées, voir récemment [4]). Dans ce cadre une question reste pourtant en suspens, celle des règles d'implémentation phonétique : on demeure ici sur un échec, tout particulièrement en ce qui concerne le rendu du caractère **graduel** des processus et leur **dépendance** au débit (cf.[3], pp. 249-253). Les propositions récentes de la phonologie articulatoire [1] sont de ce point de vue bien plus directes : "[...] the [...] gestural structures have an inherent physical meaning : they directly (without mediation of any "implementation" rules) characterize the articulatory movements [...]". Il reste, ainsi que nous le verrons en conclusion, à prouver que les mécanismes qui rendent compte des changements dans la structure des mouvements sont bien les mêmes que ceux qui se produisent pour ces mouvements bien particuliers qui forment les sons du langage.

2. LES PARADIGMES LINGUISTIQUES

2.1. Corpus

Les oppositions de durée sont, aussi bien en français qu'en arabe, des tâches linguistiques plutôt discontinues sur la dimension temporelle. Cependant, le statut phonologique du contrôle temporel est relativement différent dans ces deux langues.

En arabe, les combinaisons intramorphémiques VC, VVC et VCC sont des séquences linguistiques généralisées. Ainsi en arabe koweïtien, les différences de durée consonantique et vocalique sont phonologiquement pertinentes (par exemple : [faṣel] "la saison", [faaṣel] "l'extrait", [faṣṣel] "faire faire un vêtement"). Tout comme en arabe tunisien, où [kifu] "comme lui", s'oppose à [kiifu] "son plaisir" et à [kiffu] "arrêtez ça".

Par contre en français, les oppositions linguistiques intramorphémiques ne sont régies pratiquement que par le contrôle de la durée vocalique (VC et VVC) et sont limitées à des aires linguistiques restreintes. Ainsi dans certains parlers non méridionaux, on fera l'opposition entre [ãpatɔ̃] "nous empattons la grue" et [ãpaːtɔ̃] "nous empâtons le pain". D'autres cas d'allongement vocalique proviennent plus souvent d'effets cointrinsèques consonantiques : les consonnes qui "allongent" une voyelle précédente [v, z, ʒ, ʁ] sont bien connues des linguistes. En ce qui concerne les patterns de gémination consonantique VCC en français, elles sont essentiellement le résultat de fusion de deux

consonnes hétéromorphémiques, lorsqu'il n'y a pas de relâchement de la première consonne (par exemple [õtuʃ] "on touche", VC vs. [õt⁻tuʃ] "on t'touche", VCC).

Pour l'investigation des deux tâches phonétiques - vocalique et consonantique - cette étude des phasages utilisera la **mandibule**, un articulateur fondamental dans la régulation d'ouverture et de fermeture du tractus vocal, ses déplacements étant fortement couplés à ceux de la langue et des lèvres. Ces couplages avec la langue et les lèvres, à des fins acoustiques, présentent des coordinations interarticulateurs différentes suivant la tâche linguistique à accomplir. Par conséquent, nous avons choisi des corpus qui assurent un bon couplage entre : la lèvre inférieure [f], ou la lame de la langue [s, ʂ, z, dʒ], ou encore sa masse [ɛ, a], et la mandibule. Ainsi les mesures recueillies sur la mandibule seraient révélatrices - au prix d'un peu de "bruit" - des signaux de mouvements pour la langue et les lèvres. C'est ce que nous testerons par la cohérence des phasages entre les milieux articulatoire et acoustique.

Les items choisis sont les paires minimales suivantes :

- pour l'**arabe koweïtien** :

Corpus 1 : [faʂel] "la saison", [faaʂel] "l'extrait", [faʂʂel] "faire faire un vêtement", soit VC, VVC et VCC.

Corpus 2 : [dʒezza] "il a tondu le mouton", [dʒeeza] "il a récompensé", soit VCC et VVC, comparables aux items tunisiens.

- pour l'**arabe tunisien** : [zezza] "il a tondu le mouton", [zeeza] "il a récompensé", soit les variantes dialectales correspondant au Corpus 2 du koweïtien.

- pour le **français** : [sez#a] "seize a", [sez#sa] "seize ça" (où # marque une jointure possible), soit VC et VCC, aussi proches que possible des structures arabes précédentes.

La comparaison des deux langues, possédant deux comportements linguistiques différents, l'arabe (avec une tâche vocalique et une tâche consonantique) et le français (avec une tâche consonantique), doit nous permettre de préciser de quelle nature est la tâche linguistique française : majoritairement vocalique ou consonantique ? En effet, grâce aux travaux préliminaires que nous avons menés au niveau acoustique [5], nous savons qu'en français la présence d'une consonne double a pour effet de réduire la voyelle précédente à une durée minimale (de l'ordre de 40-45 ms), cette durée ne changeant pratiquement pas sous l'effet de la vitesse d'élocution. Il y a donc certainement autre chose qu'un simple **effet** de la manoeuvre consonantique sur la voyelle.

2.2. Acquisition

Tous les items ont été insérés dans des phrases porteuses interrogatives. Ces phrases ont été lues douze fois en chambre anéchoïque, en ordre aléatoire, par un locuteur pour chaque langue, à deux vitesses d'élocution (normale et rapide). La première série d'enregistrement a été faite à la vitesse d'élocution qui a paru la plus normale au locuteur. La deuxième série a été ensuite obtenue en exigeant un débit rapide.

Les signaux de déplacement vertical de la mandibule ont été recueillis à l'aide d'un kinésiographe mandibulaire (K5AR) et échantillonnés à 160 Hz pour édition en synchronie avec le signal acoustique numérisé à 8 KHz [6].

2.3. Mesures

A l'aide des événements qui peuvent être repérés sur les signaux acoustiques [7] et articulatoires [8] des items choisis, nous avons pu retenir deux cycles - **détente** et **vitesse "vocalique"** (cf. **infra**) - qui se sont révélés être des domaines privilégiés pour la programmation de la syllabe VC. Deux autres cycles possibles, présentant seulement (*sic*) une invariance [8], ne seront pas considérés ici, puisqu'ils ne nous permettraient pas d'**opposer** les deux tâches linguistiques.

Les deux cycles retenus sont les suivants :

- **Sur le plan acoustique**, le **cycle détente** (correspondant au champ VC), peut être repéré par la reproduction de l'événement VVO (Vocalic Voice Onset), soit l'établissement d'une structure formantique clairement définie pour les réalisations de type vocalique. Dans ce cycle, l'arrivée de l'événement VVT (Vocalic Voice Termination) - correspondant à la fin de la structure formantique -, nous donne la durée de la **phase "vocalique"**.

- **Au niveau articulatoire**, le **cycle de vitesse "vocalique"** est défini comme la reproduction de l'événement MVV (Maximum Vocalic Velocity), soit le pic de vitesse dansl'abaissement de la mandibule pour produire la voyelle. Dans ce cycle, l'arrivée de l'événement MCV (Maximum Consonantal Velocity) - soit le pic de vitesse dans l'élévation de la mandibule pour produire la consonne -, nous donne la **phase "vocalique"**.

3. RESULTATS

Les figures 1, 2, 3 et 4 donnent les variations des phases en fonction des cycles (pour cette représentation dans l'étude du phasage, cf.[9].

3.1. Phasages articulatoires.

En arabe koweïtien (Corpus 1), l'opposition des tâches VC/VCC et VC/VVC se réalise, en vitesse d'élocution normale (fig. 2a), par une différence très nette en cycle, avec seulement une différence de phasage significative pour VVC/VCC. Mais cette dernière opposition disparaît avec l'augmentation de la vitesse d'élocution, seule VC se distinguant encore des autres classes par une différence significative sur le cycle (fig. 2b).

En arabe koweïtien (Corpus 2), seule une différence de phase permet d'opposer les classes VVC/VCC (fig. 3a). En débit rapide les deux tâches tendent à se rapprocher tout en gardant une nette différence de phase (fig. 3b).

En arabe tunisien, une différence de cycle **et** de phase permet d'opposer les classes VVC/VCC, en débit normal (fig. 4a). Mais cette opposition disparaît en débit rapide, les deux classes se confondant (fig. 4b).

En français, l'opposition entre les deux classes phonétiques VC/VCC repose, en vitesse d'élocution normale, à la fois sur une différence de phasage et de cycle (fig. 1a). Cette séparation des classes ne se maintient plus véritablement que par la phase, en vitesse d'élocution rapide (fig. 1b).

3.2. Phasages acoustiques.

En arabe koweïtien (Corpus 1), pour le débit normal (fig. 2c), l'opposition VC/VCC repose surtout, comme en articulatoire, sur une différence de cycle. La classe phonétique VVC s'oppose à la classe VC, à la fois en cycle **et** en phase, et uniquement par une différence de phasage significative avec la classe VCC . Le changement de débit provoque une convergence des classes vers la structure VC (fig. 2d).

En arabe koweïtien (Corpus 2), l'opposition des classes VVC/VCC n'est due qu'à une séparation en phase (fig. 3c), qui tend à s'affaiblir avec le changement de débit (fig. 3d).

En arabe tunisien, l'opposition des classes VVC/VCC se réalise par une différence de phase **et** de cycle en débit normal (fig. 4c). L'augmentation de la vitesse d'élocution provoque une fusion de la classe VVC avec la classe VCC (fig. 4d).

En français, les structures acoustiques présentent le même comportement que sur le plan articulatoire (fig. 1c), avec une légère tendance des classes à converger sous l'effet du changement de débit (fig. 1d).

4. COMPARAISONS INTERLANGUES DES STRUCTURES.

4.1. Les deux cas de l'arabe koweïtien.

- **Au niveau acoustique.** Dans les deux cas, l'opposition des tâches VVC/VCC ne se réalise en débit normal (fig. 2c et 3c) que par une différence de phase, qui se maintient en débit rapide (fig. 2d et 3d).

- **Sur le plan articulatoire.** En débit normal (fig. 2a et 3a) le comportement des deux tâches VVC/VCC est le même qu'au niveau acoustique, mais pour le cas du Corpus 1 (fig. 2b), l'augmentation de la vitesse d'élocution provoque une fusion très nette des classes VVC et VCC, alors que ce n'est qu'une tendance pour le Corpus 2 (fig. 3b).

L'explication de cet état de choses doit se trouver dans l'utilisation (pour ce Corpus 1) de deux couplages successifs différents, labio-mandibulaire [f] puis linguo-mandibulaire [ʂ]. En effet le changement de type consonantique de [z]tunisien (fig.4) à [dʒ] koweïtien (fig. 3) ne semble pas

introduire autant de "bruit" dans les structures articulatoires par rapport aux structures acoustiques, simplement par le fait que les successions de consonnes utilisées sont toutes produites avec la lame de la langue. On conçoit, par cet exemple, que la régularité des cycles - et leur régulation temporelle - puissent être fortement influencées par la nature des couplages avec la mandibule, base de la régulation syllabique.

4.2 L'arabe tunisien et l'arabe koweïtien

- Au niveau acoustique. En débit normal (fig. 4c et 3c), la structure de l'arabe tunisien, pour des items linguitiques correspondants, diverge de celle de l'arabe koweïtien, puisque les deux tâches VVC/VCC s'opposent à la fois en phase et en cycle. En arabe tunisien, le changement de débit (fig. 4d) provoque un rapprochement très net des deux classes, alors qu'en arabe koweïtien (fig. 3d), l'opposition tend à se maintenir.

- Sur le plan articulatoire. Le comportement des structures en débit normal (fig. 4a et 3a) est le même qu'au niveau acoustique (fig. 4c et 3c); cependant en arabe tunisien, la fusion des classes VVC/VCC en débit rapide (fig. 4b) est beaucoup plus nette.

4.3. Le français et l'arabe koweïtien.

Les différences de structures observées en français - à savoir une nette séparation en phase et en cycle des deux tâches linguistiques VC/VCC - divergent des structures comparables du koweïtien (Corpus 1), où l'opposition VC/VCC repose essentiellement sur une différence de cycle. Par contre la structure de l'arabe VC/VVC peut être plus directement comparée à la structure VC/VCC (*sic*) du français, si l'on prend la peine d'exprimer nos résultats en phases **consonantiques** comparables aux phases vocaliques. Ainsi la figure 1c nous donne une phase consonantique (la complémentaire de la phase vocalique) naturellement faible pour VC et forte pour VCC : ce qui est proprement la structure observée pour les phases vocaliques VC et VVC de l'arabe sur la figure 2c. Il semblerait donc que la réalisation de l'opposition VC/VCC en français corresponde le plus à une **manoeuvre vocalique**, pour la structure linguistique la plus comparable, celle du koweïtien.

5. CONCLUSION

1) Les cycles choisis nous permettent de mettre en évidence une correspondance articulatori-acoustique remarquable, dans les trois langues. On notera que la structure des oppositions est dans tous les cas, moins nette sur le plan articulatoire : ce qui est bien naturel puisque la mandibule ne nous donne qu'une vue partielle sur l'action des effecteurs produisant finalement le timing du signal. Il reste donc remarquable qu'avec ce seul articulateur porteur on obtienne d'aussi bonnes correspondances articulatori-acoustiques.

2) La nette séparation des tâches VC, VCC, ou VVC en débit normal (dans les trois langues) tend à s'affaiblir, en débit rapide. On remarque que les effets "destructeurs" du débit sur les structures sont beaucoup plus prononcés en arabe qu'en français, et tout particulièrement en arabe tunisien, pour VVC. Les tendances à la confusion se produisent au profit de la tâche linguistique la plus simple VC. Cette translation de structures complexes VCC et VVC vers une structure simple VC, quand la tâche devient plus difficile à contrôler, peut être primordiale pour comprendre les changements phonétiques.

Récemment, ces problèmes ont été correctement reposés en termes moteurs [10], rappelant que le véritable précurseur en ce paradigme expérimental était STETSON [11]. La tâche que ce dernier avait requise de ses sujets - soit la répétition, en accélérant la vitesse d'élocution, de syllabes CVC, afin de produire un continuum , de consonnes géminées à consonnes simples, en passant par des consonnes doubles (CVC+CVC -> CVCCVC -> CVCVC) - soutenait l'idée que les modifications actuelles, induites par le débit, étaient les causes principales des changements historiques permanents (voir **cuppa** du latin qui a donné **coupe** en français). C'est dans cette optique que KELSO et al.[10] ont repris une autre expérience de STETSON, qu'ils ont analysée dans le cadre explicatif des transitions de phase. La préférence donnée à des relations **en phase**, entre voyelles et consonnes, pourrait expliquer, selon eux, l'universalité de syllabes CV par rapport aux VC. Le fait que [ip+ip -> ipip...] aboutit en débit accéléré à [...pipi], mais jamais le contraire, semble coïncider assez bien avec une expérience sur le contrôle des doigts, dans laquelle les index étaient, au début, actionnés **en opposition de phase** (soit antisymétriquement en flexion et en extension), puis basculaient "irrésistiblement" vers un mode **en phase**, avec l'augmentation de la vitesse d'exécution[12]. Sur ce point, TULLER [13] a finalement présenté des données qui, de manière générale, n'ont pas démontré la transition de phase attendue (sauf pour un seul sujet, phonéticien). En fait, bien des difficultés se sont présentées pour réaliser la tâche fondamentale : un premier problème tient au phénomène de l'effacement de la **jointure** [ip#ip → ipip] ; un second à **l'absence d'un véritable geste vocalique**, de voyelle à voyelle (puisqu'à l'intérieur de ce paradigme nous avons affaire à la même voyelle) ; enfin, plus particulièrement, le fait d'avoir à contrôler deux gestes **consonantiques**, la fermeture des lèvres et l'ouverture de la glotte, donne en anglais un phénomène idiosyncratique : on finit par produire [p^hi] aspiré, soit une **opposition de phase.**

Cette approche semble pourtant prometteuse et - une fois qu'elle aura résolu un transfert conceptuel en termes corrects vers la parole - (ce qui ne nous semble pas encore être le cas de la phonologie articulatoire de BROWMAN et GOLDSTEIN [1], qui repose sur plusieurs bases contestées de la **Théorie de l'Action**) - elle pourra être sans doute utilisée avec profit dans l'étude de nos données sur le contrôle de la quantité vocalique et consonantique.

A Christian ABRY pour ses commentaires utiles sur ce travail.

REFERENCES

[1] BROWMAN C.P. GOLDSTEIN C.M. (1989) Gestural Structures and Phonological Patterns. Haskins Labs. Status Rept. SR- 97/98, 1-23.

[2] LEBEN W. (1980) A Metrical Analysis of Length. Linguistic Inquiry 11, 497-509.

[3] McCARTHY J. (1986) OCP Effects : Gemination and Antigemination. Linguistic Inquiry 17, 207-263.

[4] LAHIRI A. HANKAMER J. (1988) The Timing of Geminate Consonants. J. of Phonetics 16, 327-338.

[5] ABRY C. ORLIAGUET J.P. SOCK R. (1989) Patterns of Speech Phasing. Their Resistivity in the Production of a Linguistic Timed Task : Single versus Double (Abutted) Consonants in French. Actes du 1er Workshop Régional de Sc. Cognitives, Lasco III, Chichilianne 19-21mars, 134-166.

[6] WORLEY C. (1989) Organisation temporelle articulatori-acoustique des gestes vocaliques et consonantiques. Signaux kinésiographiques labiaux et mandibulaires. Thèse I.N.P.G, Grenoble.

[7] ABRY C.BENOIT C. BOE L.J. SOCK R. (1985) Un choix d'événements pour l'organisation temporelle du signal de parole. 14èmes J.E.P du GCP du GALF, 113-137.

[8] DELATTRE C. JOMAA M. WORLEY C. ABRY C. (1989) The Phasing of the Jaw in Consonant and Vowel Lengthening : Arabic and French Patterns; Europ. Conf. on Speech Comm. and Techn. 2, 416-419.

[9] GENTNER D. R. (1987) Timing of Skilled Motor Performances. Tests of the Proportional Duration Model. Psy. Review 82, 225-260.

[10] KELSO J.A.S. SALTZMAN E.L. TULLER B. (1986) The Perspective on Speech Production : Data and Theory. J. of Phonetics 14, 29-56.

[11] STETSON R.H. (1928) Motor Phonetics : a Study of Speech Movements in Action. Archives Néerlandaises de Phonétique Expérimentale 3, 1-216. (Nouvelle édition 1988: J.A.S Kelso & K.G Munhall, Boston : Little, Brown & Company).

[12] TULLER B. (1988) Phase Transitions in Speech Production : Empirical Observations. In M. Jeannerod (Ed). Attention and Performance XIII : Motor Representations and Control, (in press).

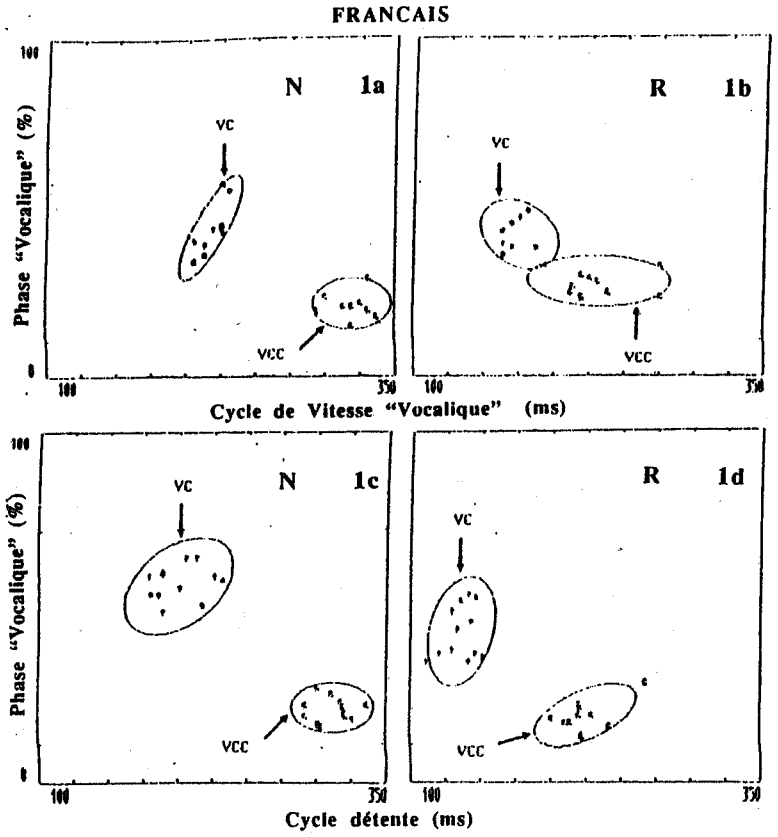


Figure 1 Gémination consonantique (VC vs. VCC) en français. Correspondances entre phasages articulatoires (en haut) et phasages acoustiques (en bas) dans deux cycles : vitesse "vocalique" (pics de vitesse) et détente (établissements d'une structure formantique claire pour la voyelle). Locuteur masculin. Deux conditions de vitesse d'élocution : normale (N) et rapide (R).

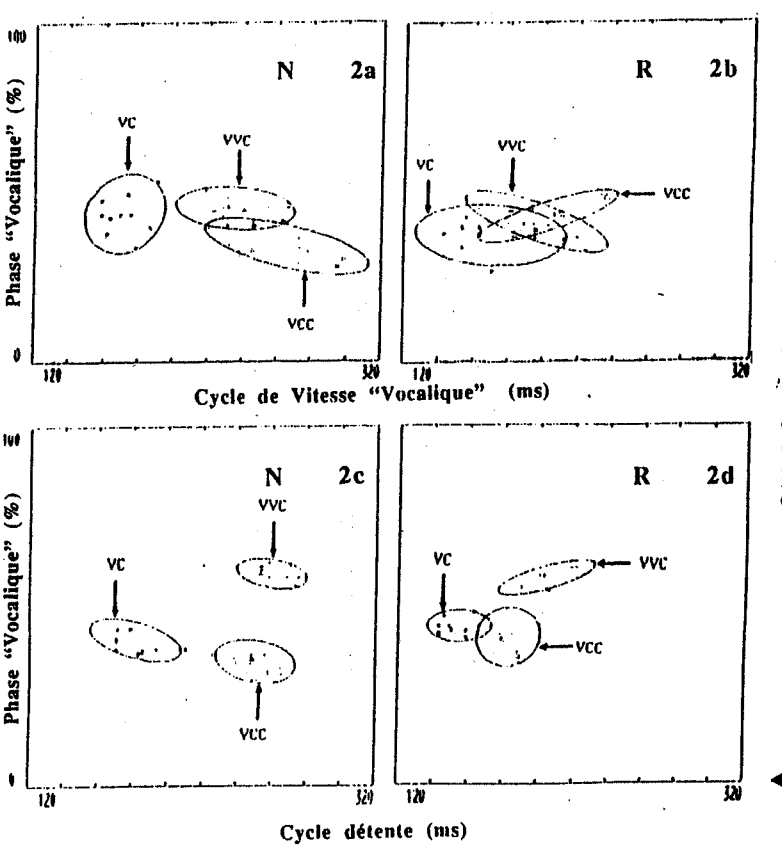


Figure 2 Quantité vocalique (VC vs. VVC) et consonantique (VC vs. VCC) en arabe koweïtien. Correspondances entre phasages articulatoires (en haut) et phasages acoustiques (en bas) dans deux cycles : vitesse "vocalique" (pics de vitesse) et détente (établissements d'une structure formantique claire pour la voyelle). Locuteur masculin. Deux conditions de vitesse d'élocution : normale (N) et rapide (R).

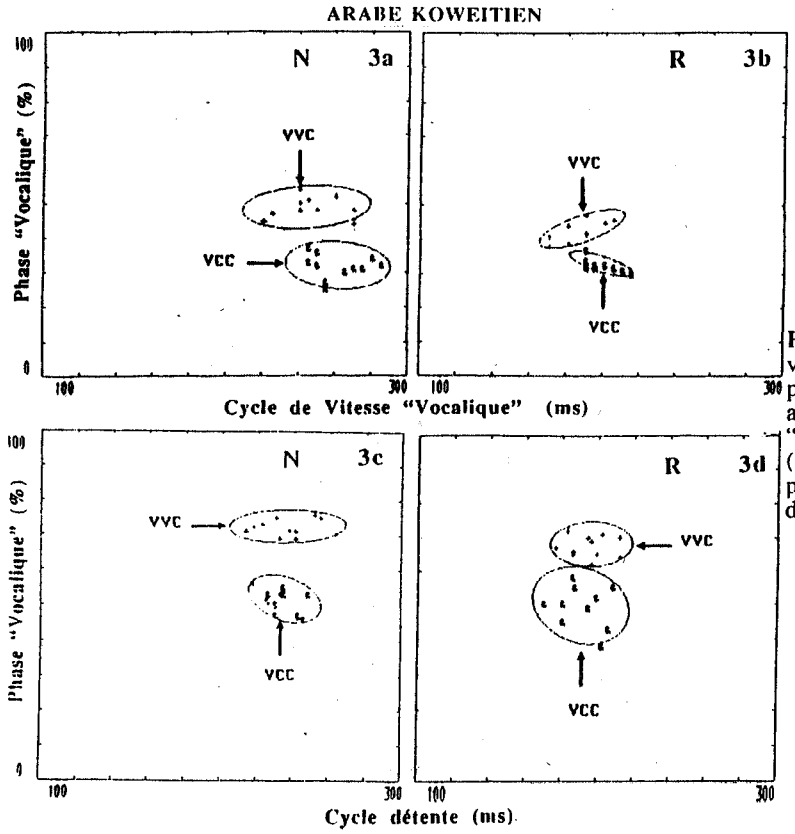


Figure 3 Quantité vocalique et consonantique (VVC vs. VCC) en arabe koweïtien. Correspondances entre phasages articulatoires (en haut) et phasages acoustiques (en bas) dans deux cycles : vitesse "vocalique" (pics de vitesse) et détente (établissements d'une structure formantique claire pour la voyelle). Locuteur masculin. Deux conditions de vitesse d'élocution : normale (N) et rapide (R).

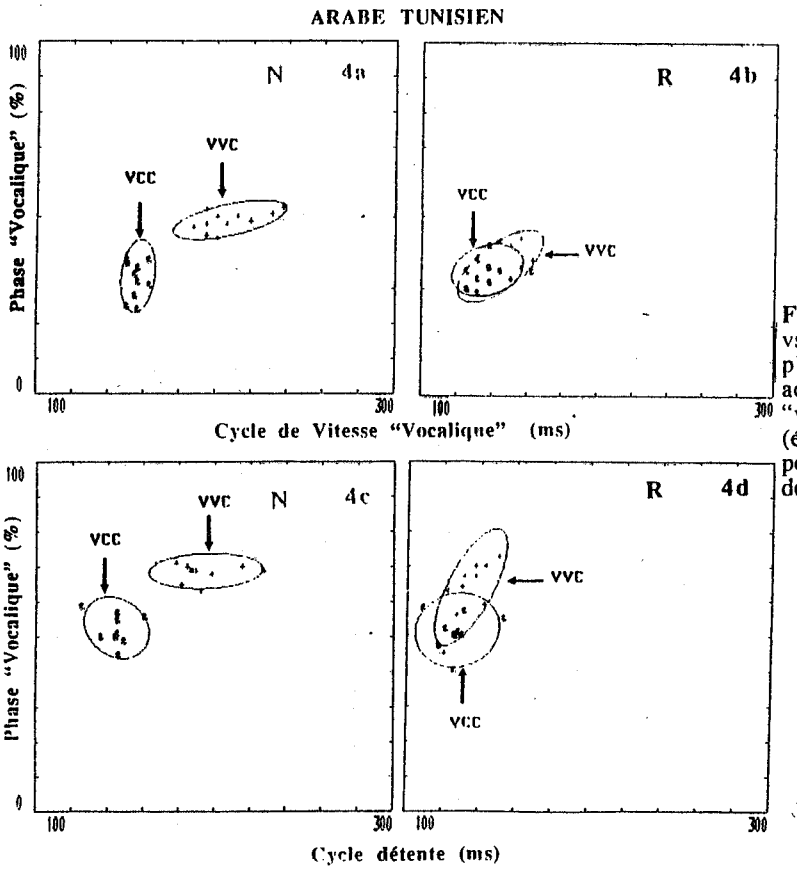


Figure 4 Quantité vocalique et consonantique (VVC vs. VCC) en arabe tunisien. Correspondances entre phasages articulatoires (en haut) et phasages acoustiques (en bas) dans deux cycles : vitesse "vocalique" (pics de vitesse) et détente (établissements d'une structure formantique claire pour la voyelle). Locuteur masculin. Deux conditions de vitesse d'élocution : normale (N) et rapide (R).

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

MANIPULATION DE PARAMETRES ISSUS D'UNE ANALYSE EN FORMES D'ONDES: TESTS PRELIMINAIRES

Lori LAMEL et Maxine ESKENAZI

LIMSI - CNRS BP 133 91403 - ORSAY CEDEX

Recently short-time waveform analysis has been used to analyse, manipulate, and synthesise speech [e.g., Lienard, ICASSP-87]. In order to better understand the parameters of this analysis, and to determine which ones carry pitch information in waveform-analysed speech, perceptual experiments (ABX test) were performed using synthetic stimuli of 300 ms (formed by repeating 10ms waveforms). These experiments, with 5 normal-hearing subjects, verified that the parameters affecting the perceived pitch were internal frequency, offset, and change in phase in successive waveforms; the remaining parameters primarily affected timbre. The parameters affecting pitch were varied individually and jointly to explore their interaction. Both frequency and offset were found to dominate perceived pitch.

I. Introduction

Dans les recherches sur la variabilité de la parole, on a souvent recours à des manipulations de signaux divers dans le but d'étudier la perception humaine.

L'idée relativement récente (RODET 84, LIENARD 87) d'employer une analyse du signal par formes d'onde pour analyser, manipuler, et/ou synthétiser des objets de parole est séduisante. Cependant, avant d'aborder la manipulation d'objets de parole pour prendre en compte la variabilité (que ce soit en synthèse ou en reconnaissance automatique), il est nécessaire de comprendre le comportement de ces formes ondes (GROVEL 89). Nous décrivons des études menées dans ce but qui manipulent les paramètres définissant les formes d'onde.

Une tâche relativement simple a été choisie comme cadre de départ pour la manipulation de ces paramètres: la perception de la fréquence fondamentale du signal (appelée "pitch" ci-après). Ce cadre a été choisi pour plusieurs raisons: l'interêt, pour la synthèse de bonne qualité, de pouvoir bien maîtriser les paramètres concernant le pitch; pour des regroupements éventuels d'objets comme des formes d'onde, des formants ... de mieux comprendre un des paramètres, le pitch, qui doit aider à leur regroupement; pour créer une tâche rélativement simple quant au nombre de paramètres que le sujet doit évaluer - une comparaison de hauteur seulement; pour éviter des tâches d'identification de type "trait phonétique" dont nous pensons ne pas avoir une compréhension suffisante pour pouvoir bien interpréter les résultats obtenus.

Les sons synthétisés pour ces tests sont dérivés d'objets que nous avons extraits d'analyses de parole réelle (voir III).

Nous avons donc essayé de déterminer le rôle des différents paramètres des formes d'onde dans le perception du pitch. D'abord il s'agissait de déterminer quels paramètres contribuent plutôt à définir le timbre des sons que leur pitch. Ensuite en testant ensemble et séparément les paramètres qui semblent avoir un rôle à jouer vis à vis du pitch, on les manipule de manière à obtenir la perception d'un son précis.

Ce papier donnera une définition succincte des grains (le lecteur peut se référer à la bibliographie pour de plus amples détails), décrira ensuite le protocole de test, les tests, et enfin présentera quelques premières conclusions qui peuvent en être tirées.

II. L'Analyse en Formes d'Onde

Ce type d'analyse vise à décomposer le signal en une somme de fonctions élémentaires, ou formes d'onde. Une telle décomposition peut s'envisager sous plusieurs aspects: théorie du signal, modèle de production vocale, analyse suivant des critères auditifs (D'ALESSANDRO 89). Le signal $x(n)$ est représenté par une somme de fonctions élémentaires, par la formule d'expansion:

$$x_m = \sum_n \sum_k c^{n,k} f_m^{n,k}$$

où le coefficient, c, et la fonction élémentaire, f, dépendent de l'instant (n), de la fréquence (k), et du temps (m) pour f.

Les fonctions, f, ou formes d'ondes élémentaires à court terme et à une largeur de bande limitée, apparaissent comme des sinusoides modulées par une enveloppe. De nombreuses variantes de l'enveloppe existent dans la littérature. Les formes d'ondes utilisées ici se définissent par six paramètres: l'instant de référence IR, la durée de l'attaque de l'enveloppe ATT, la durée de décroissance de l'enveloppe DEC (qui peut être un segment exponentiel ou sinusoïdal), l'amplitude maximum AMPL, et deux paramètres de la porteuse, la fréquence porteuse FREQ, et la phase PHI par rapport à l'instant de référence de l'enveloppe.

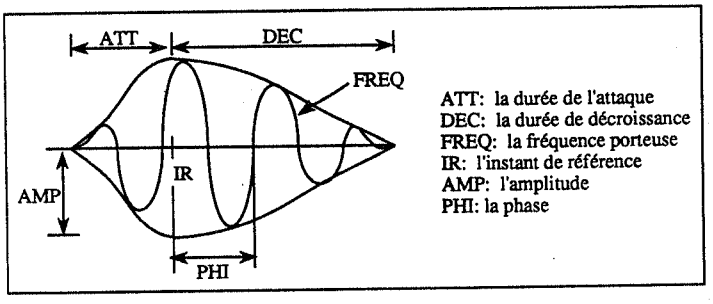


Figure 1. Paramètres d'une forme d'onde.

III. Protocole de test

Stimuli inconnus: Des stimuli inconnus d'une durée de 300 ms ont été synthétisés en utilisant un logiciel de synthèse par forme d'onde développé au LIMSI. Les valeurs par défaut des paramètres ont été choisies à partir d'une analyse en formes d'onde de plusieurs énoncés. Toutes les combinaisons de paramètres employées pour ces tests peuvent ne pas se rencontrer dans de la parole réelle, mais les valeurs employées sont toutes choisies à l'intérieur de limites provenant de segments voisés de parole prononcés par quatre locuteurs différents. Les paramètres des stimuli de base ont été choisis d'après l'analyse d'un /o/ prononcé par une locutrice ayant un premier formant à environ 500 Hz et un fondamental de 250 Hz. Des formes d'onde ayant des fréquences (FREQ) allant de 480 à 530 Hz, des durées (ATT + DEC) de 8 à 12 ms. et des attaques (ATT) durant de 30 à 70% de leur durée totale ont été observées à la sortie d'un filtre d'une largeur de bande de 2 Barks centrée à 500 Hz. La période de répétition (distance entre IR d'une forme d'onde et IR de la suivante - nous l'appelerons IRD) est comprise entre 4,3 et 3,8 ms., ce qui correspond à une fréquence fondamentale de 233 à 263 Hz. Le stimulus de base a donc une durée de forme d'onde de 10 ms., une fréquence de 500 Hz, et une période de répétition (IRD) de 4 ms. (250 Hz fréquence de répétition).

Des sons de 300 ms ont été synthétisés et stockés sur le disque dur d'un PC-AT compatible. Afin de minimiser le rôle de l'intensité dans le jugement des sujets, chaque son a été normalisé en suivant une courbe d'isosonie (Fletcher 53) afin que les sons purs de référence de fréquence différente aient une intensité perçue similaire. Les 20 ms. du début et les 20 ms. de la fin de chaque son ont été pondérés par une demie fenêtre de Hanning afin de ramener doucement l'amplitude du signal à zéro.

Présentation du test: Afin de déterminer l'effet de la manipulation des paramètres, les stimuli inconnus ont été présentés dans un test ABX (Cutting 74). Rapellons que dans ce type de test les sujets identifient l'inconnu, X, comme une des références, A ou B. Les deux sons de référence sont des sons purs de 250 et 500 Hz, correspondant respectivement à un fondamental et un premier formant plausibles pour le /o/ précédant. Une analyse spectrale (DFT) de plusieurs des stimuli inconnus préliminaires montre des composantes fréquentielles à 250 et à 500 Hz. Ceci rendait impossible l'emploi d'un test de type, "identique/différent", ou "ajustement absolu à un ton de référence", car dans ce type de test les sujets rapprocheraient le stimulus inconnu des deux références alors que le but est cependant de déterminer quelle fréquence est perçue comme dominante. L'emploi du test ABX fournit une tâche qui consiste en un choix entre les deux sons de référence et n'oblige pas le sujet à effectuer des jugements absolus sur la fréquence.

Les stimuli de test ont été présentés aux sujets à l'aide d'un logiciel interactif développé au LIMSI qui envoit les sons à l'auditeur et enregistre ses réponses transmises à l'aide d'une souris pointant sur des icônes affichées à l'écran. Les sujets utilisent un casque Beyer Dynamic DT-48. Le signal est échantillonné à 10 kHz et des filtres antirepliement à 4,8 kHz sont employés.

Le triplet de test, AXB, est présenté une fois au sujet. Ensuite il lui est possible de rejouer les triplets, ou l'un quelconque des sons, et ceci un nombre illimité de fois et dans n'importe quel ordre en désignant les icônes présentés à l'écran. Un silence de 190 ms sépare chacun des sons à l'intérieur d'un triplet et il y a une pause minimum de 500 ms entre les répétitions éventuelles d'un triplet AXB. Un triplet de test consiste donc en: A (300 ms) silence (190 ms) X (300 ms) silence (190 ms) B (300 ms).

Chaque stimulus inconnu est présenté cinq fois pour le Test TIM et 10 fois pour les Tests VAR. L'ordre de présentation des stimuli est automatiquemment rendu aléatoire par le programme de test ainsi que l'ordre de présentation des sons de référence (comme A ou B). Chaque stimulus inconnu est aussi présenté pour la moitié des exemples avec le son de référence le plus bas en "A" et pour l'autre moitié avec ce son en "B" afin d'éliminer un biais dans les réponses. Les sons de référence ont également été présentés en tant que sons de test afin de confirmer que ces tons ont été correctement mis en correspondance. Après avoir choisi une icône réponse, le triplet suivant est automatiquemment présenté au sujet. Un dispositif de correction d'erreur est prévu et permet au sujet de rejouer le triplet précédent.

Sujets: Il y avait cinq sujets, chercheurs au LIMSI. Un audiogramme a été réalisé pour chaque sujet afin de confirmer qu'il n'y avait aucune attenuation de l'ouïe dans la région fréquentielle qui nous intéresse.

Instructions: Des instructions écrites ont été fournies aux sujets. Celles-ci comprenaient une explication de la démarche du test ainsi qu'une description de la manière dont ils pouvaient interagir avec la machine. Il fallait choisir la référence, A ou B, qui était la plus proche en fréquence au stimulus inconnu X. Les sujets ont été forcé à choisir entre ces deux options, mais ils pouvaient également spécifier qu'X ne se rapprochait ni d'A, ni de B, mais avec un choix plutôt A / plutôt B..

Les sujets ont d'abord subi une session d'entrainement afin de les accoutumer au déroulement du test. Au début de chaque test les sujets ont ajusté le niveau de sortie du signal à un niveau qui était confortable. La durée de chaque test était d'environ 10 à 15 minutes (selon le sujet) et les sujets ont fait des pauses entre les tests afin d'éviter une fatigue auditive. Les tests ont été ordonnés des tâches les plus "faciles" (stimuli inconnus contenant typiquemment un ton spectralement simple) aux tests ayant des tâches plus "difficiles" (souvent contenant des stimuli inconnus composés de tons complexes spectralement). La difficulté relative a été déterminée lors de tests préliminaires.

Tableau 1. Contenu des tests

Test	n° présentations	n° stimuli	n° sujets	totale
TIM				
PH	5	10	5	250
AT	5	20	5	500
VAR				
FR/IRD	10	10	5	500
FR	10	10	5	500
IRD	10	10	5	500
DPH	20	10	5	1000

IV. Tests TIM (relatifs au timbre)

Cette série de tests a été conçu pour confirmer que la modification de certains des paramètres des formes d'onde, telle la durée de l'attaque, n'affecte pas la perception de la *fréquence fondamentale perçue* (pitch). Les paramètres de phase interne de la porteuse et durée d'attaque ont donc été variés ici. A priori on s'attend ici à un changement au niveau du timbre perçu, mais pas au niveau de la fréquence dominante.

IV.A Test PH: Phase Interne

Cinq stimuli inconnus, S0 à S5, ont été générés à une fréquence (FREQ) de 500 Hz et cinq autres stimuli, S6 à S11 ont été générés à une fréquence de 250 Hz. La phase interne de la porteuse (PHI), relative à l'instant de référence, IR, variait linéairement de 0 à 360 degrés en 4 pas de 72 degrés. Toutes les formes d'onde d'un même stimulus sont générés en phase (voir la Figure 2).

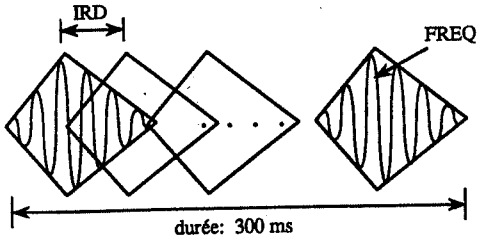


Figure 2. Un stimulus inconnu. IRD = la distance entre deux formes d'onde consécutifs, FREQ = la fréquence porteuse.

IV.B Test AT: Durée d'attaque

Vingt stimuli inconnus ont été générés, S0 à S9 à une fréquence (FREQ) de 500 Hz et S10 à S19 à 250 Hz. La durée d'attaque (ATT) a été variée (8 pas à une distance de 0,89 ms) de manière linéaire de 10 à 90% de la durée totale (ATT + DEC) de la forme d'onde (10 ms - voir la Figure 3).

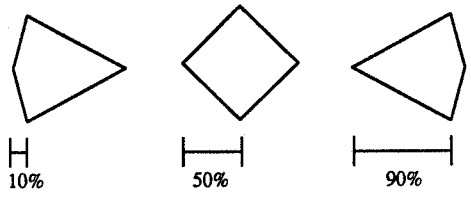


Figure 3.Valeurs d'attaque (ATT) - trois valeurs utilisées, présentées en ordre croissant (pour une durée totale (ATT + DEC) invariante).

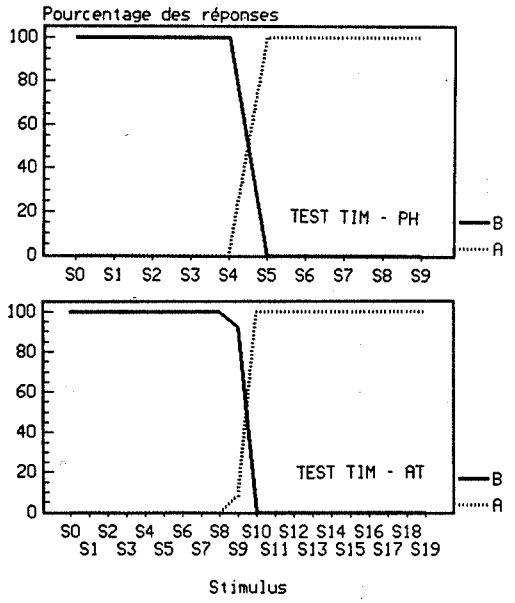


Figure 4. Résultats pour les tests TIM.

Résultats: Les résultats pour les cinq sujets se trouvent sur la Figure 4 pour ces deux tests. Il est évident, à la vue de ces graphiques, que les stimuli inconnus d'une fréquence de 250 Hz ont été rapprochés à la référence de 250 Hz, et ceux de 500 Hz ont été rapprochés de la référence de 500 Hz.

Discussion: Aucun de ces deux paramètres a affecté la perception de la fréquence dominante, cependant plusieurs des sujets ont noté, pour le Test A1 (où l'attaque variait), que les deux fréquences de référence étaient présentes dans le stimulus inconnu. L'intensité des autres composantes du son augmente pour des attaques très "brusques", résultant en un changement de la qualité du son. Mais, comme les instructions demandaient un rapprochement en fréquence à un des sons de référence, les résultats sont relativement non-ambigus.

V. Tests VAR (variation de paramètres)

Dans cette série d'expériences les tests sont menés sur la variation d'un ou de plusieurs paramètres qui sont soupçonnés d'avoir un effet sur la fréquence perçue. Le Test FR/IRD contient des stimuli dont la fréquence porteuse (FREQ) et la fréquence de répétition des formes d'onde IRD (définie comme le temps entre deux IR consécutifs) varient en même temps. Les tests FR, IRD, et DPH varient les paramètres de fréquence porteuse (FREQ), décalage (IRD), et delta-phase (différence de phase (PHI) entre deux formes d'onde consécutifs, les phases sont mesurées à IR), respectivement. Les stimuli inconnus S0 et S9 ont été définis dans le but d'obtenir une perception non-ambigue de 250 et 500 Hz respectivement, et les valeurs des paramètres ont été interpolés linéairement entre ces valeurs en 8 pas intermédiaires à distance égales (voir la Figure 2).

V.A Test FR/IRD: fréquence et décalage

Dans ce test la fréquence porteuse et la fréquence de répétition des formes d'ondes ont été ajustés aux valeurs nécessaires pour fournir un son à une fréquence de 250 Hz et à celles nécessaires pour un son de 500 Hz. Le stimulus inconnu, S0 a donc une fréquence porteuse (FREQ) de 250 Hz et une distance (IRD) entre formes d'onde de 4ms (250 Hz fréquence de répétition). L'autre stimulus extrême, S9, a une fréquence porteuse de 500 Hz et une distance (IRD) de 2 ms entre les formes d'onde (500 Hz fréquence de répétition). Huit stimuli intermédiaires ont été générés.

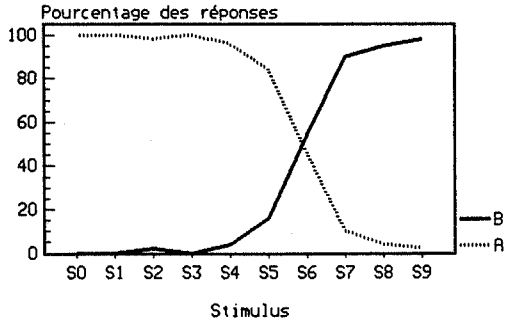


Figure 5. Résultats pour le test VAR - FR/IRD

La Figure 5 montre les résultats pour ce test, tous sujets confondus. On peut remarquer qu'un croisement entre les perceptions des deux stimuli existe vers S6, qui a une fréquence porteuse de 417 Hz et un distance entre formes d'onde de 2,67 ms. (378 Hz fréquence de répétition). La fréquence porteuse est

plus élevée que le milieu perceptif entre les deux extrèmes (prévu par une approximation géométrique en échelle de Mels -368 Hz), mais la fréquence correspondant à la distance entre formes d'onde consécutifs y est proche.

V.B Test FR: fréquence

En gardant une distance entre formes d'onde consécutifs (IRD) fixe à 4 ms. (250 Hz fréquence de répétition), la fréquence porteuse du stimulus inconnu (FREQ) a été variée de 250 Hz (S0) à 500 Hz (S10) avec un pas de décalage de 27,8 Hz.

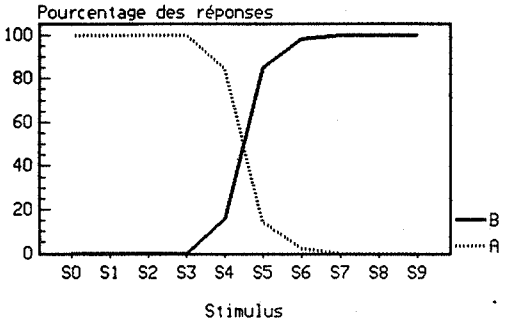


Figure 6. Résultats pour le test VAR - FR

Les résultats tous sujets confondus se trouvent sur la Figure 6. Les sujets ont un croisement assez nettement marqué entre S4 et S5. S4 a une fréquence porteuse de 361 Hz et S5 est à 389 Hz. (n.b. Il n'y a pas de stimulus inconnu à une fréquence porteuse de 368 Hz, ce qui correspond à une fréquence à mi-chemin entre les deux stimuli extrèmes, selon une approximation géométrique en échelle de Mels. Ce croisement a lieu à une fréquence porteuse moins élévée que celle du test FR/IRD, mais correspond grossièrement à la même fréquence que la fréquence de répétition trouvée dans le test FR/IRD.

Une analyse spectrale (DFT ci-dessus) montre que tous les stimuli inconnus ont des composantes à 250 Hz et à 500 Hz, et quelques uns ont une composante à 750 Hz. Le croisement a lieu où les composants de 250 et de 500 Hz sont d'amplitude à peu près égale, à S4. S0 à S3 ont une composante à 250 Hz qui est plus intense que celle de 750 Hz et S7 à S9 ont une composante de à 750 Hz qui est plus intense que celle de 250 Hz.

Des tests préliminaires utilisant des conditions complémentaires (i.e. une distance entre formes d'onde consécutifs de 500 Hz et une fréquence porteuse d'entre 250 Hz et 500 Hz) ont montré que l'inconnu a toujours été perçu comme étant plus proche du son de référence élevé (500 Hz). Il faudrait cependant noter que cette combinaison particulière de paramètres n'a pas été retrouvée dans les analyses effectués sur de la parole réelle voisée dans les régions fréquentielles comparables.

V.C. Test IRD: période de répétition

En gardant la fréquence fixée à 250 Hz, la période de répétition (IRD) entre formes d'onde successives varie de 2 ms (500 Hz), pour S0, à 4 ms (250 Hz), pour S9 avec un pas de 0,22 ms.

Les résultats de ce test se trouvent sur la Figure 7. Il y a un croisement entre S3 et S4. La distance entre deux formes d'onde consécutifs est de 2,67 ms (378 Hz fréquence de répétition) pour S3 et de 2,89 ms (346 Hz) pour S4. Le croisement a lieu près de la fréquence à mi-chemin entre les deux références (367,7 Hz), définie par la fréquence de répétition des formes d'onde. Une analyse spectrale des stimuli inconnus montre que la composante dominante s'approche de la fréquence de répétition.

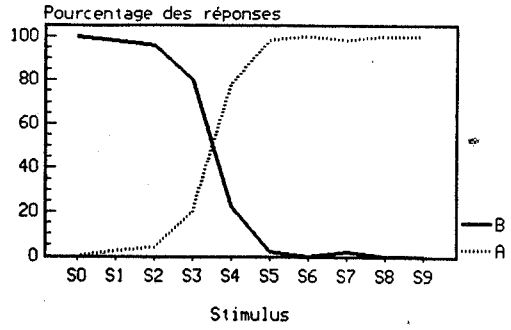


Figure 7. Résultats pour le test VAR - IRD.

Les combinaisons de paramètres ci-dessus n'ont également pas été retrouvés lors d'analyses en formes d'onde de la parole réelle. Les paramètres complémentaires (c'est à dire une fréquence porteuse (FREQ) de 500 Hz et une fréquence de répétition entre 250 et 500 Hz) pourraient exister. Des investigations préliminaires de ces paramètres indiquent que ces stimuli inconnus sont toujours rapprochés du son de référence de 500 Hz.

V.D. Test DPH: Delta-Phase

Dans ce test la fréquence (DPH) et la fréquence de répétition sont maintenus à une valeur constante et la valeur de la phase (PHI) d'une forme d'onde à la prochaine est variée (voir la Figure 8). La distance entre formes d'onde (IRD) a été fixée à 4 ms (250 Hz - fréquence de répétition), mais la moitié des stimuli (S0 à S9) ont été fixés à une fréquence porteuse de 250 Hz et l'autre moitié (S10 à S19) à une fréquence porteuse de 500 Hz. La différence de phase entre une forme d'onde et la suivante (appelée delta-phase ci-après) a été variée de 0,05 périodes à 0,95 périodes, soit de 18 à 342 degrés. Si on présume que la delta-phase s'entend comme un décalage temporel par rapport à l'instant de référence, la période donnée par la gamme des delta-phases est entre 4,2 ms. et 7,8 ms. à 250 Hz et entre 4,1 ms. et 5,9 ms. à 500 Hz. On peut aussi envisager le contraire, c'est-à-dire que la delta-phase étant entendue comme une avance, donne une période plus courte que la fréquence de répétition des formes d'onde.

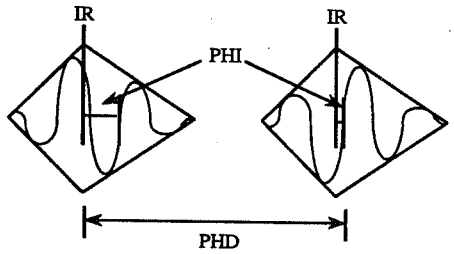


Figure 8. Variation de phase entre deux formes d'onde successives pour les stimuli inconnus du test VAR - DPH.

Résultats: Les résultats sont montrés sur la Figure 9. Il est à noter que tous les stimuli inconnus à 500 Hz de fréquence porteuse (S10 à S19) ont été rapprochés au son de référence de 500 Hz. Pour les stimuli inconnus à 250 Hz, S0 à S2 et S7 à S9 ont été rapprochés au son de référence de 250 Hz. Les stimuli intermédiaires, S3 à S6 ont des résultats plus ambigus. Ces stimuli contiennent deux composantes ayant des amplitudes à moins de 10 dB de différence, l'un en dessous de 250 Hz et l'autre entre 250 et 500 Hz. Les résultats ambigus peuvent indiquer que les sujets se focalisaient sur l'un ou l'autre des deux composantes. Cependant, il n'y a jamais de cas de réponses ambigues à 500 Hz. (Une analyse plus poussée des résultats individuels peut aider à mieux comprendre ces resultats.)

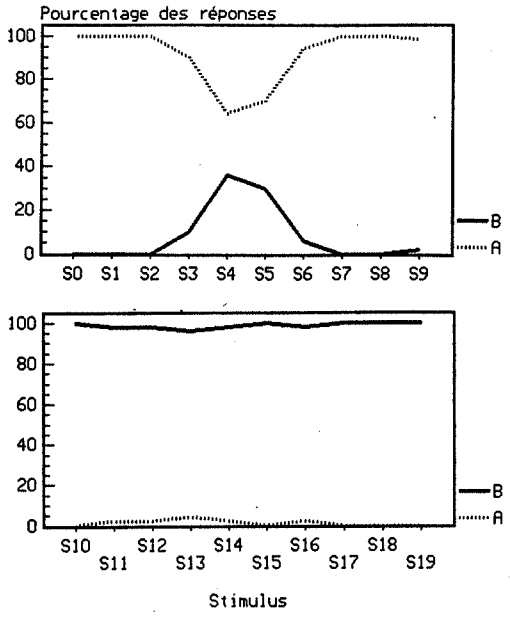


Figure 9. Résultats VAR - DPH. En haut, pour les stimuli inconnus de 250 Hz; en bas, pour les stimuli inconnus de 500 Hz.

V.E Discussion des TESTS VAR

Cette série de tests montre qu'il existe différentes manières de modifier la fréquence perçue des stimuli inconnus. Les cas les plus nets sont FR/IRD, FR, et IRD. La fréquence perçue a changé dans DPH, mais les changements ne sont pas suffisamment importants pour changer le son de référence choisi.

Les cas complémentaires mentionnés brièvement lors de la description des tests FR et IRD montrent que pour certaines valeurs des paramètres, le changement des autres paramètres a peu, voire aucun, effet sur la fréquence perçue. Il est interessant de noter que le même resultat a été obtenu en variant la distance entre formes d'onde successives qu'en variant la fréquence porteuse. Cette forme de test n'indique cependant pas s'il y a d'autres différences entre les stimuli inconnus. Des expériences en cours d'élaboration actuellement sont destinées à examiner ce phénomène.

Est-ce que les tests FR et IRD indiquent que c'est toujours la fréquence la plus élevée qui domine? Le resultat n'est pas clair car, dans le test FR/IRD, le croisement a eu lieu à l'endroit où la fréquence de répétition est près de la valeur perceptuellement intermédiaire entre les deux sons de référence malgré le fait que la valeur de la fréquence porteuse l'ait déjà dépassé.

La variation de la delta-phase a crée des sons de fréquence moins élevée et d'autres de fréquence plus élevée que les deux extrèmes de référence. Il a été noté qu'avec une fréquence porteuse de 500 Hz., les modifications de delta-phase n'ont pas suffisamment affecté la fréquence pour obtenir un jugement de 250 Hz. Ceci peut vouloir dire que la possibilité de modifier la delta-phase pour obtenir un changement en fréquence est limitée. Il semble qu'il soit bien plus facile, pour changer la fréquence, de modifier la fréquence de répétition ou la fréquence porteuse. Des études futures devraient donc essayer d'évaluer la contribution du paramètre de la phase. Les changements de phase d'une forme d'onde à la suivante peuvent s'avérer plus pertinentes en ce qui concerne la qualité (le "naturel") du son que le "pitch".

VI. Conclusions et perspectives

Deux séries de tests sur la perception du "pitch" de stimuli inconnus créés par une synthèse par forme d'onde ont été présentées. La première série de tests montre que la variation des paramètres de phase interne de la porteuse relative à l'instant de référence et de durée de l'attaque n'affectent pas la perception de la hauteur. La seconde série de tests montre que les paramètres de fréquence porteuse, fréquence rélative, et delta-phase peuvent affecter la perception du "pitch".

Les trois paramètres peuvent tous être manipulés pour obtenir un changement dans le "pitch". Cependant, la manipulation de chaque paramètre dépend des valeurs des autres paramètres. Un troisième ensemble de tests qui examine l'importance relative de ce trois paramètres est actuellement en cours.

Des études futures concerneront des manipulations et tests de perception de la pertinence de ces paramètres sur de la parole réelle, et la manipulation des valeurs des paramètres afin d'obtenir une son de hauteur voulue. Elles devront nous aider à relier nos observations à de la parole réelle en utilisant la synthèse, et la manipulation des paramètres de voyelles synthétiques.

REMERCIEMENTS

Nous souhaitons remercier Christophe D'Alessandro pour ses commentaires sur le contenu scientifique de cet article, Sophie Grovel pour les programmes de synthèse et Jean-Luc Gauvain pour le mécanisme de playback en temps réel. Et tous nos remerciements aux chers sujets.

BIBLIOGRAPHIE

Carhart, R., and Jerger, J.F. (1959), "Preferred method for clinical determination of pure-tone thresholds", *Journal of Speech and Hearing Disorders*, 24, p.330-345.

Cutting,J.E., and Rosner, B.S. (1974), "Categories and boundaries in speech and music", *Perception and Psychophysics*, 16 (3), p.564-570.

D'Alessandro, C. (1989), *Représentation du signal de parole par une somme de fonctions élémentaires*, These de Doctorat de l'Université Paris 6 en Informatique.

Demany, L. (1988),"Perception de la hauteur tonale," Chapitre 2 dans *Psychoacoustique et Perception Auditive*, Paris: INSERM.

Fletcher, H. (1953), *Speech and Hearing in Communications*, New York: Van Nostrands.

Gabor, D. (1946), "Theory of communication," *Journal of the IEE*, No. 93-1946, Londres.

Grovel, S., Liénard, J.S., and D'Alessandro, C. (1989), "Representation of the Speech Signal with Elementary Waveforms: A Preliminary Perceptive Study," *Proc. of the 13th International Congress on Acoustics*, p.369-372.

Kewley-Port, D., Watson, C.S., Foyle, D.C. (1988), "Auditory temporal acuity in relation to category boundaries; speech and nonspeech stimuli," *Journal of the Acoustical Society of America*, 83, p.1133-1145.

Liénard, J.S. (1987), "Speech Analysis and Reconstruction Using Short-time, Elementary Waveforms," *Proc. IEEE ICASSP-87*, p.948-951.

Rodet, X. (1984), "Time-Domain Formant-Wave-Function Synthesis," *Computer Music Journal*, 8 (3).

Silverman, H.F., and Lee, Y.T. (1987), "On the spectrographic representation of rapidly time-varying speech," *Computer Speech and Language*, 2, p.63-86.

4 PERCEPTION

Président: J.M. DOLMAZON
ICP-Grenoble,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

MISE EN CORRESPONDANCE D'ESPACES ACOUSTIQUE ET PERCEPTIF PAR L'ANALYSE FACTORIELLE MULTIPLE : APPLICATION A L'ETUDE D'UN CORPUS CVCV MULTILOCUTEUR.

PASCAL D.

CNET LAA/TSS/CMC
ROUTE DE TREGASTEL BP 40 22300 LANNION

Résumé

L'analyse factorielle multiple a été conçue spécialement pour analyser des tableaux de données structurées. Sur un même ensemble d'individus décrits par plusieurs groupes de variables ($J \geqslant 2$), cette analyse permet la comparaison systématique des différentes typologies définies par chacun d'eux, par exemple deux typologies acoustique et perceptive. Son originalité est de mettre en évidence les structures communes et spécifiques de ces groupes. L'AFM produit à la fois une représentation graphique des individus caractérisés par l'ensemble des variables (ou points moyens) et des représentations superposées de ces mêmes individus caractérisés chacun par un groupe de variables (ou points partiels). Elle autorise ainsi une lecture des ressemblances globales entre individus et une lecture des ressemblances partielles en fonction de chaque groupe de caractéristiques. C'est un outil de synthèse remarquable pour toute étude dont l'objectif essentiel est la caractérisation. Nous avons appliqué cette analyse factorielle multiple aux données acoustiques et perceptives extraites d'une expérience de similarité de voix (Pascal & col., 89).

Introduction

Le problème de la mise en correpondance de descriptions distinctes - tant aux niveaux acoustique, articulatoire que perceptif - relativement à un même ensemble de stimuli, se retrouve constamment dans bon nombre d'études sur la parole. Cette mise en correspondance permet, bien évidemment, d'éclairer et d'expliciter les unes par les autres les informations complexes recueillies dans chaque champ de connaissances. Ceux-ci étant généralement multidimensionnels, la problèmatique est de corréler globalement des espaces distincts. L'analyse factorielle multiple ou AFM (Escofier & Pagès, 88) est une nouvelle procédure, developpée spécialement pour le traitement de tableaux de données structurées, qui nous semble idéale pour analyser globalement la corrélation d'espaces différents.

Les analyses factorielles ont, à l'origine, été conçues pour étudier un tableau de données unique. Or, la plupart du temps, l'expérimentateur dispose, pour décrire ses stimuli, d'une série de données distinctes (acoustiques, articulatoires, perceptives, quantitatives, qualitatives, etc...) ou d'une suite de descriptions identiques évoluant en fonction d'un paramètre quelconque, dont il aimerait faire une analyse simultanée. C'est la problèmatique introduite par les analyses canonique ($J=2$) et multicanonique ($J>2$) formulées par Hotelling et plus tard par Carroll. L'AFM peut d'ailleurs être considérée comme une analyse multicanonique particulière et une alternative au programme INDSCAL (Carroll & Chang, 70) qui présente, sur celui-ci, de nombreux avantages (dont l'absence de problème de convergence et la garantie de poids positifs* (Pascal, 84)).

Dans le passé, ces données multiples ont, le plus souvent, été analysées comme un tableau complexe formé par la juxtaposition des différents tableaux. Si cette solution reste techniquement possible, en dépit d'une perte notable d'une partie de la structure interne des données, elle n'autorise aucunement la prise en compte simultanée dans une même analyse de données de type mixte : à la fois quantitatives et qualitatives. C'est le second avantage de l'analyse factorielle multiple que de permettre l'analyse conjointe, en tant qu'éléments actifs, de données mixtes. Les données soumises à l'AFM sont ainsi réparties en plusieurs groupes dont chacun peut jouer soit un rôle actif, soit un simple rôle explicatif s'il est introduit en tant que groupe supplémentaire; l'unique contrainte est qu'à l'intérieur d'un même groupe les données soient toutes de même nature. (fig 1a : tableau de données X).

I. Aspects théoriques de l'analyse factorielle multiple

L'AFM procède, séparèment sur chacun des groupes de variables, à des ACP normées (analyse en composantes principales) dont les facteurs spécifiques sont introduits, ensuite, en variables supplémentaires dans l'ACP globale de l'ensemble des variables (fig 1b : schéma du déroulement de l'analyse). C'est ce qui permet la comparaison effective des groupes dans le cadre d'un référentiel commun, contrairement aux résultats de séries d'analyses indépendantes et, donc, incomparables par nature (sauf à entreprendre de fastudieuses rotations d' axes pour tenter de réveler l'éventuelle identité de sous-espaces factoriels).

I.0. Rappels sur l'ACP

Comme dans toute ACP, les données sont préalablement centrées et réduites pour s'affranchir de l'arbitraire des unités de mesure. Le tableau analysé a pour terme général $(x_{ik} - \bar{x}_k) / s_k$. L'ensemble des individus caractérisés par les variables du groupe j constitue le nuage N_J dont le centre de gravité . est confondu avec l'origine des axes du fait du centrage. La ressemblance entre deux individus i et l est représentée par leur distance d(i,l) définie par :

* Avec la collaboration de B. Escofier, nous envisageons de comparer systématiquement les deux analyses (AFM et INDSCAL).

$$d^2(i,l) = \sum_{k \in Kj} m_k (x_{ik} - x_{lk})^2$$

(distance euclidienne usuelle dans l'espace Kj, pour des poids m_k des variables tous égaux à 1). La liaison entre deux variables k et h est mesurée par le coefficient de corrélation :

$$r(k,h) = \sum_i p_i ((x_{ik} - \bar{x}_k) / s_k) ((x_{ih} - \bar{x}_h) / s_h)$$

(les poids p_i des individus sont pris tous égaux à 1/I, $\forall$ i$\in$I)

L'ACP fournit des images planes approchées des nuages des individus : ici N^j_I situé dans l'espace R^{Kj} . Les axes factoriels Grjf1, Grjf2, ... ou composantes principales sont des directions privilégiées de cet espace, orthogonales par construction, maximisant l'inertie par rapport à l'origine de la projection sur ces axes du nuage N^j_I. Pour les individus, les distances ne sont que peu déformées par les projections. Pour les variables, représentées par des vecteurs de l'espace à I dimensions noté R^I et dont le nuage N^j_K est situé sur une hypershère de rayon 1, ce sont les angles entre les vecteurs qui sont peu déformés par les projections : le cosinus de l'angle formé par les vecteurs représentant deux variables h et k étant égal au coefficient de corrélation de ces deux variables (r(h,k)).

Les deux nuages N^j_I des individus et N^j_K des variables sont liés par des relations dites de dualité. Leur inertie totale est la même :

$$\text{Inertie tot. } (N^j_I \text{ ou } N^j_K) = 1/I \sum_{k \in K_j} \sum_i ((x_{ik} - x_k) / s_k)^2 = Kj$$

et la décomposition de cette inertie totale selon les deux séries d'axes factoriels (facteurs sur les individus f_s et facteurs sur les variables g_s) sont identiques. On peut, en effet montrer (Escofier & Pagès, 88) que les inerties projetées sur les axes de même rang s sont égales et représentées par la valeur propre λ_j^s de rang s de l'analyse séparée du groupe j. Les relations de transition entre les facteurs de rang s s'écrivent :

$$f_s(i) = 1/\sqrt{\lambda^s} \sum_k (x_{ik} - \bar{x}_k) / s_k)\, g_s(k)$$

$$g_s(k) = 1/\sqrt{\lambda^s} \sum_i (x_{ik} - \bar{x}_k) / s_k)\, f_s(i)$$

La projection $f_s(i)$ d'un individu i est une combinaison linéaire des projections $g_s(k)$ de toutes les variables.

I.1. Equilibre des groupes

La solution technique à la prise en compte simultanée, dans une même analyse, de points de vue différents exprimés chacun par un groupe de variables est d'appliquer à ces groupes une pondération qui en équilibre l'influence. La solution retenue est ici de diviser par λ_j^1 (première valeur propre de l'ACP du seul groupe j) le poids initial de chaque variable d'un même groupe j. L'inertie du nuage associé N^j_I , nuage des individus vus au travers du groupe de variables j, est alors multiplié par ce poids dans chaque direction de l'espace. Dans l'ACP normée du groupe de variables ainsi pondérées, le premier axe de chaque groupe j est alors associé à une valeur propre de 1. Le rôle des groupes est équilibré "en ce sens qu'aucun groupe ne peut influencer à lui seul la première composante principale de l'ensemble (qui maximise l'inertie projetée de toutes les variables)" (Escofier & Pagès, 88). Le carré de la distance entre deux points i et l du nuage global N_I est la somme des carrés de leur distance dans les N^j_I. Si i^j est le point représentant i dans le nuage N^j_I :

$$d^2(i,l) = \sum_{k \in K} m_k (x_{ik} - x_{lk})^2 = \sum_{j \in J} \sum_{k \in Kj} m_k (x_{ik} - x_{lk})^2 = \sum_{j \in J} d^2(i^j, l^j)$$

Toutefois, l'influence d'un groupe croît avec sa dimension propre et reste fonction de sa structure interne.

I.2. Représentation simultanée des J nuages N^j_I associés aux groupes

Cette représentation simultanée est obtenue en projetant chaque nuage N^j_I sur un même sous-espace. Le choix de ce sous-espace doit satisfaire deux conditions essentielles :

- chaque nuage N^j_I doit être bien représenté ---> on cherche des projections des N^j_I d'inertie importante,
- les points homologues (représentant le même individu) doivent être les plus proches possible ---> on cherche à minimiser l'inertie intra des nuages N^J_i d'un même individu i autour de leur centre de gravité i*. (cf récapitulatif des nuages en présence). Le compromis entre ces deux conditions est donné par le théorème de Huyguens (inertie inter = inertie totale - inertie intra). Ainsi "le sous-espace de R^K sur lequel la projection de N^j_I a une inertie inter maximum est engendré par les premiers axes d'inertie, notés u_s, du nuage N^*_I des centres de gravité. Ce nuage étant homothétique au nuage N_I associé à l'ensemble des variables, le sous-espace cherché s'obtient par une ACP du tableau X tout entier." (Escofier & Pagès, 88). Pour obtenir la représentation simultanée des N^j_I , on introduit, en supplémentaires dans l'ACP globale, les tableaux X^j (de dimensions (I,K) dans lequel Xj est complétée par des zéros). Dans ce cas, l'indicateur inertie projetée / inertie totale est toujours très faible, même si la forme du nuage est bien respectée (2 projections succéssives). A la place des X^j, il revient au même d'introduire le tableau des composantes principales de Xj en éléments supplémentaires. Il est alors possible de comparer entre elles et avec les variables les premières composantes Grjf1, Grjf2 ... de chaque groupe.

N_I	= (i, i$\in$I) nuage des individus associé à X
N^j_I	= (i^j, i$\in$I) nuage dans R^{Kj} associé à Xj
	= projection de N_I sur R^{Kj}
N^J_i	= (i^j,j$\in$J) images du même individu i
N*I	= (i*, i$\in$I) nuage des centres de gravité des N^J_i, homothétique de N_I (rapport 1/I)

Tableau 1 : Récapitulatif des nuages en présence dans R^K

I.3. Recherche de facteurs communs et Aides à l'interprétation

Sur un même axe, le coefficient de corrélation entre le facteur de l'ensemble des groupes et ses représentants dans chaque groupe constitue la mesure de la similitude des projections de chacun des nuages "partiels" N_I^j et du nuage global N_I. Ces représentants, encore appelés variables canoniques, sont des combinaisons linéaires des variables du groupe. Si, pour un groupe, la corrélation vaut 1, le facteur global est une combinaison linéaire des facteurs de ce groupe. Le tableau des corrélations entre variables canoniques et variables générales constitue un véritable indice de la structure commune partagée par l'ensemble des groupes ou seulement par certains d'entre eux. Des valeurs élevées indiquent qu'il existe une direction de dispersion presque analogue dans les groupes j concernés et le nuage global. Par ailleurs, le tableau d'aide à l'interprétation des groupes actifs rassemble les contributions des variables de chaque groupe aux différents facteurs globaux. Ces contributions définissent la coordonnée des groupes sur un axe. Lorsqu'un facteur commun existe, son importance dans chaque groupe se lit directement dans ce tableau. La somme par facteur de ces contributions est égale à la valeur propre λ^s associée.

Afin d'aider à l'interprétation, comme en ACP, sont fournies en sortie* et exprimés en fonction des facteurs de l'analyse globale, les coordonnées, les qualités de représentation par l'axe (COR) et les contributions (CTR) des individus partiels -c'est-à-dire caractérisés par un seul groupe de variables- et moyens -ou caractérisés par l'ensemble des données- , ainsi que celles des centres de gravités -partiels et moyens- des classes d'individus définies par les variables qualitatives. Les tableaux concernant les variables et les composantes principales de chaque analyse partielle permettent d'étiqueter les axes de l'analyse globale et de comparer les axes partiels de chaque groupe aux facteurs communs. Rappelons que, dans le cas des variables, comme en ACP simple, les coordonnées représentent le coefficient de corrélation entre la variable k et le facteur s. Toutes ces informations sont visualisées dans des graphiques .

I.4. Analyse de tableaux de données mixtes :

Dans une ACP simple, on ne peut qu' introduire en éléments supplémentaires des variables qualitatives (en calculant les centres de gravité des classes d'individus définies par les modalités de ces variables). L'originalité de l'AFM est permettre l'analyse simultanée en tant qu'élément actif de variables de deux types, quantitatif et qualitatif. D'une façon classique, l'analyse de variables qualitatives est réalisée par l'application de l'analyse des correspondances multiples (ou ACM**) au tableau disjonctif complet créé à partir des modalités prises par ces variables. Les modalités d'une même variable sont orthogonales entre elles et leur direction n'est pas affectée par une quelconque pondération. On montre que l'ACM est équivalente à une ACP normée appliquée aux variables indicatrices *** pondérées par le rapport $(I-I_k)/I$, I_k représentant le nombre d'individus possèdant la modalité k. Le programme réalise lui-même cette pondération. Pour pouvoir appliquer l'AFM à l'analyse simultanée de groupes de variables qualitatives, il suffit que chaque groupe soit constitué de l'ensemble des indicatrices associées à une seule et même variable. Ainsi l'AFM est une méthode qui généralise l'ACM comme d'ailleurs l'ACP (cas où les groupes sont réduits à une seule variable quantitative).

* outre les histogrammes des valeurs propres pour chaque analyse partielle et pour l'analyse globale

** ACM = AFC d'un tableau disjonctif complet

** *indicatrices = modalités de la variable qualitative en AFM

II. Application à l'étude d'un corpus CVCV multilocuteur

Nous avons appliqué l'analyse factorielle multiple aux données acoustiques et perceptives extraites d'une expérience de similarité de voix (Pascal & col., 89). Le corpus étudié était composé de 9 items CVCV (ufu, sisi, bubu, kuku, kiki, baba, papa, kaka et tata) prononçés chacun par dix locuteurs masculins(LJ, RD, MS, GM, AG, FC, PD, PC, CG et MG). Des auditeurs, auxquels était présenté un même item prononcé par deux locuteurs différents, devaient évaluer la distance perceptive entre les deux réalisations. Les résultats de l'expérience ont été analysés par INDSCAL (Carroll & Chang, 70); ils nous ont fourni une description perceptive du corpus en 3 dimensions que , pour l'heure, nous ne remettons pas en cause (Pascal & col., 89). Par ailleurs, parmi un choix de mesures acoustiques, treize variables ont été retenues:

- durées des différents segments : voyelle interconsonnantique (DV1), consonne intervocalique (DC2), voyelle finale (DV2), segment VC (DVC) et durée totale (TOT)
- fréquence fondamentale (FO), sa dynamique (DFO), assorties d'une mesure de l'écart de fréquence (DIFO) entre les fondamentales des 2 voyelles V1 et V2,
- formants (F1, F2, F3), ainsi que leurs dérivés (F01) et (F12), écarts entre le premier formant et la fréquence fondamentale, d'une part, les deux premiers formants, d'autre part. L'ensemble de ces données constitue ainsi un tableau rassemblant 4 groupes de variables (Gr1:variables de durées, Gr2:variables relatives à FO, Gr3:variables formantiques et Gr4:variables perceptives) dont l'analyse par l'AFM est particulièrement indiquée. La démonstration concerne ici l'analyse des données moyennées sur les différents contextes consonantiques des voyelles extrêmes i, a, u. (Fig. 2 : Décomposition des tableaux de données, la variable locuteur a été mise en supplémentaire)

	1=F	2=F	3=F	4=F	5=F	6=F
Gr1	858	514	725	357	529	295
Gr2	789	869	508	661	546	287
Gr3	224	540	725	500	442	789
Gr4	948	831	597	863	293	209

Tableau 2 : Correlations variables canoniques et variables générales

Les corrélations entre variables canoniques et variables générales sont rassemblées dans le tableau 2. Ce tableau indique une forte communalité du premier facteur entre le groupe perceptif (Gr4), et les groupes définis par les mesures de durées (Gr1) d'une part et du fondamental (Gr2) d'autre part. La valeur propre correspondante à ce premier facteur est relativement forte 2.20. Le premier axe correspond à une direction de dispersion importante pour ces trois groupes. Les premiers axes partiels 101, 201 et 401 devraient être relativementt proches entre eux. Le second facteur est une dimension commune des groupes 2 et 4, le troisième facteur, lui, est lié aux données formantiques et au groupe des durées, enfin le quatrième facteur est une exclusivité du groupe perceptif (Gr4). Les quatre premiers facteurs expliquent 79 % de la variance totale contenue dans l'ensemble des données acoustiques et perceptives. On a poussé l'analyse jusqu'au facteur 6 pour tenter de mieux déceler l'influence des formants dont on sait le lien déterminant avec les items (cf Pascal & col., 89), mais la valeur propre associée à ce facteur est extrêmement réduite. Le groupe des formants est pratiquement orthogonal au plan formé par les deux premiers facteurs. L'absence de corrélations entre les axes perceptifs issus d'INDSCAL et les variables formantiques (hormi, justement, pour les données moyennées sur l'ensemble des items) indique que les

caractéristiques formantiques des locuteurs ne sont pas perçues comme telles par les auditeurs*.

	IN1	IN2	IN3
TOT	.37	-.05	.71
DV1	.63	-.01	.43
DC2	-.23	-.19	.65
DVC	.13	-.18	.85
DV2	.59	.17	.30
FO	-.02	-.73	-.40
DIFO	-.32	.51	-.36
DFO	-.54	-.19	-.32
F1	.08	.04	.02
F2	.14	.03	.00
F3	.02	-.11	-.02

Tableau 3 : Corrélations des dimensions perceptives et des variables acoustiques.

Par ailleurs, c'est le groupe perceptif qui témoigne de la plus forte contribution de ses variables aux facteurs globaux. L'importance des données perceptives, par rapport au choix de la description acoustique, se trouve ainsi validée par l'analyse générale de l'AFM. Concernant le traitement perceptif, l'analyse en composantes principales séparée des dimensions fournies par INDSCAL révèle des facteurs quasiment identiques à ceux de l'analyse générale (soit 1f, 2f et 4f) mais cependant distincts des dimensions extraites du modèle. Ce fait suggère de mettre en doute la capacité d'INDSCAL à distinguer des dimensions réellement indépendantes (IN1 et IN3 sont d'ailleurs liées entre elles dans le plan (1,4)). Une comparaison systèmatique des résultats extraits d'une analyse INDSCAL et d'une analyse factorielle multiple sera entreprise car l'AFM peut être considérée comme une alternative plus rigoureuse (au sens mathématique) à l'analyse INDSCAL fondée, elle, sur un modèle psychométrique empirique.

L'interprétation des facteurs se fait comme en ACP classique : on recherche les variables les plus liées aux facteurs au sens de leur contribution relative (CTR). Sur la figure 3a, on constate que le premier facteur, facteur commun à 3 des groupes, synthétise les variables concernant le débit (TOT et DVC très corrélées entre elles et avec le premier axe partiel 101 de Gr1) et la dynamique de FO (DFO constituant l'axe partiel 201). Ces deux informations se retrouvent, sans doute, au niveau de la troisième dimension IN3 extraite de l'analyse INDSCAL, fortement corrélée à l'axe 1 ainsi qu'aux deux variables précédentes de durées. Le second facteur, commun à Gr2 et Gr4 uniquement, représente la hauteur; il oppose IN2 et FO comme nous l'avons vu dans l'analyse INDSCAL et de plus correspond sensiblement aux secondes composantes principales 402 et 202 des groupes perceptif et fondamental. On constate au passage que les dimensions perceptives IN3 et IN2 sont bien orthogonales. Le troisième axe global est caractérisé par F2, seconde raisonnance formantique; il est pratiquement confondu avec la première composante du groupe 3. D'autre part, il est lié au groupe 1 par la durée de la consonne intervocalique DC2 qui correspond sensiblement, à une inversion près, au second axe partiel 102 du groupe 1. Le facteur 4 est uniquement déterminé par IN1.

* Le fait de traiter le groupe locuteur en élément actif et donc, d'admettre une certaine redondance de l'information locuteur, relègue encore plus loin l'influence des formants , c'est-à-dire sur le cinquième facteur.

La représentation simultanée (fig. 3b, plan des deux premiers facteurs) des locuteurs caractérisés par chacun des 4 groupes (ex: FC 1 = point partiel du locuteur FC défini par le groupe Gr1 des données temporelles, AG 2 = point partiel AG caractérisé par le groupe Gr2 des données relatives à FO...) fait clairement apparaître la spécificité des locuteurs "typés" déterminés dans un premier travail (Pascal & col., 89). Le locuteur AG, dont les principales caractéristiques sont un fondamental élevé (influence sur l'axe 2) et une dynamique associée importante (influence sur l'axe 1), est représenté par un point AG 2 extrême et un point partiel perceptif AG 4 très proche. Cette proximité des images des locuteurs au niveau perceptif et au niveau de leurs caratèristiques acoustiques les plus marquantes est une constante de cette analyse. Le locuteur MS, dont le débit est extrêmement rapide, est caractérisé par un point partiel MS 1 extême, mais relativement proche du point partiel perceptif MS 4. On observe, au passage, que les locuteurs LJ et MS, à l'écoute si proches, ont des images relativement semblables dans chacun des groupes. Ceci se traduit par la quasi identité de leur point moyen LJ et MS, ainsi que de leur image perceptive LJ 4 et MS 4, et par la proximité de leurs points partiels "durées" et "fondamental" (ce sont deux locuteurs rapides, ayant sensiblement la même fréquence fondamentale). Le locuteur FC présente une caractéristique marquante sur les données du groupe 2 (excentricité du point FC 2, sur 2f d'une part, du fait d'une voix particulièrement grave et, sur 1f d'autre part, du fait d'une dynamique importante DFO. Son image perceptive FC 4 se trouve donc relativement proche de son point partiel le plus caratéristique. Ces deux derniers locuteurs n'ont pas de caractéristiques temporelles marquées: d'où la position centrale de leur point image du groupe 1. Enfin, GM, dernier locuteur "typé", est essentiellement caractérisé aux niveaux acoustique et perceptif par des durées très longues de ses segments VC et de ses consonnes (proximité des points GM 1 et GM 4) ainsi que par une nette différence entre les fondamentaux de ses voyelles (point GM 2 extrême). Parmi les locuteurs sans caractéristiques marquées, on constate que les différentes images des locuteurs RD et CG sont analogues: ces deux locuteurs sont globalement semblables sur le plan (1,2).

Conclusion

Les procédures classiques le plus souvent utilisées pour la mise en correspondance d'espaces descriptifs distincts consistent à régresser les unes sur les autres les dimensions de ces espaces. L'inconvénient majeur de cette technique est d'éliminer des candidats explicatifs d'un premier ensemble dans le cas d'une corrélation négligeable avec les axes du second ensemble. Traité globalement, c'est-à-dire en association avec les autres élèments de son propre ensemble, la liaison avec les éléments du second espace peut toutefois exister. La solution à cet impasse consiste à recourir à l'analyse factorielle multiple ou AFM qui, seule, permet d'analyser globalement la corrélation des deux espaces. Cette démarche ne permet cependant pas de faire l'économie d'une régression multiple qui permettra d'exprimer les variables perceptives en fonction des mesures acoustiques. L'AFM, conçue spécialement pour analyser des tableaux de données structurées en groupes, permet la comparaison effective de ces groupes dans le cadre d'un référentiel commun. Sa seconde originalité est d'autoriser le traitement simultané en tant qu'éléments actifs de variables de deux types : 1) quantitatif et 2) qualitatif. L'analyse factorielle multiple est, ainsi, une méthode qui généralise à la fois l'ACP, applicable aux données de type 1 et l'ACM, applicable aux données de type 2.

L'auteur remercie B. ESCOFIER pour ses précieux conseils et pour la valeur pédagogique de son livre, écrit en collaboration avec J. PAGES, dont il s'est très largement inspiré. Merci à J. L. TAFFOU pour sa pratique des analyses et à R. CHEVREUL pour les illustrations schématiques.

REFERENCES

CARROLL J. D. & CHANG J.J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, N° 35, 283-319.

ESCOFIER B. & PAGES J. (1988) Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation. Dunod, Paris.

PASCAL D., THILL C. & BOYER M. (1989) Perceptive similarity of male voices: Correlation with acoustic measures. Eurospeech, 26-28 Sept., Paris.

PASCAL D. (1984) : Analyse de données perceptives par la méthode d'évaluation multidimensionnelle INDSCAL ou analyse individuelle des proximités. Séminaire Obtention et Analyse de données phonétiques, GALF, Groupe Communication Parlée, Grenoble, 17-21 Sept.

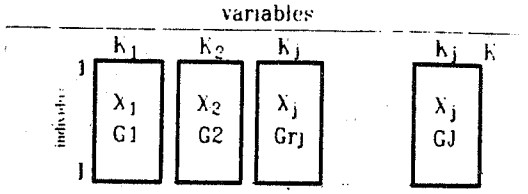


Fig. 1a : Le tableau des données : X

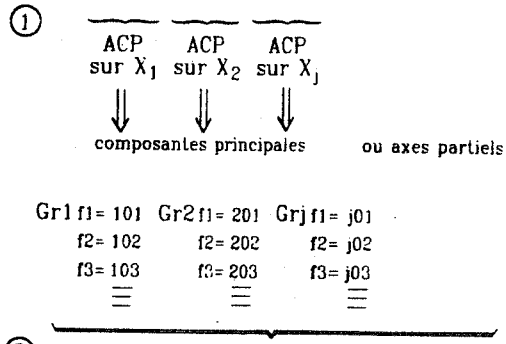


Fig. 1b : Déroulement de l'AFM

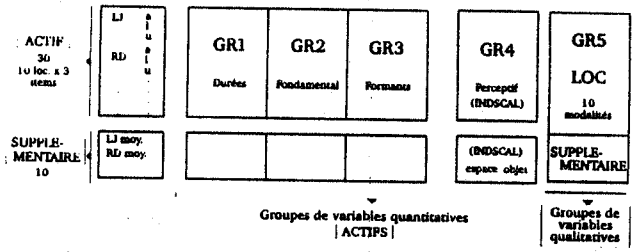


Fig. 2 : Décomposition des tableaux de données

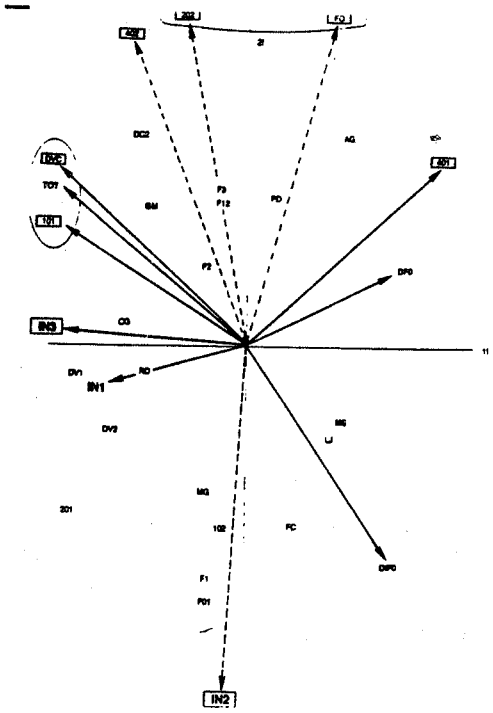


Fig. 3a : AFM (Groupe locuteur en supplémentaire)
Représentation des variables, des indicatrices de la variable qualitative et des axes partiels des ACP séparées des groupes. Plan des 2 premiers facteurs.

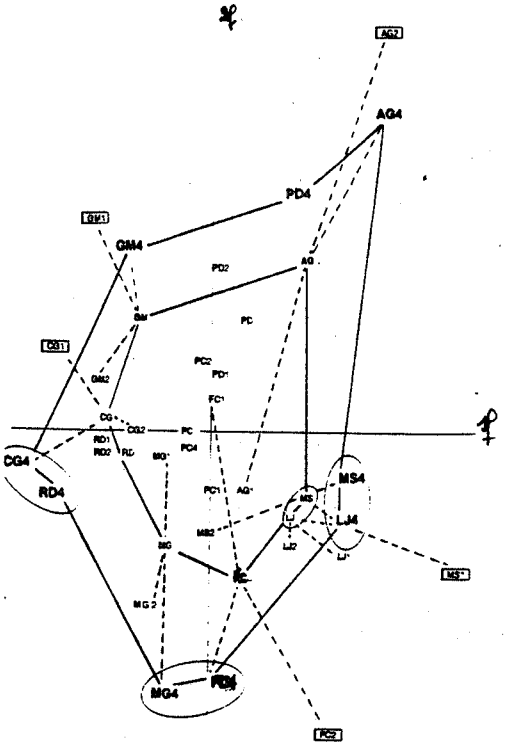


Fig. 3b : AFM (Groupe locuteur en supplémentaire)
Représentation des centres de gravités :

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Etude Perceptuelle de la Parole Bruitée

Jean-claude JUNQUA

Speech Technology Laboratory, Division of Panasonic Technologies Inc.

3888 State Street, Santa Barbara, California 93105, U.S.A

et C.R.I.N / I.N.R.I.A Campus scientifique BP 239, 54506 Vandoeuvre les Nancy, France

RESUME

Le but de cette étude est d'étudier l'intelligibilité de la parole produite dans du bruit par comparaison à celle de la parole normale. Sur une base de données de 49 mots isolés (produite par 10 locuteurs) contenant des mots acoustiquement distincts et similaires, des tests d'intelligibilité ont été effectués sur des auditeurs français, anglais et américains pour différents rapport signal/bruit. Il a été observé que pour des vocabulaires de mots acoustiquement similaires l'intelligibilité diminuait lorsque les mots avaient été prononcés en milieu bruité (sauf pour le sous-vocabulaire constitué des mots "n" et "m"). Ce résultat est valable pour les différentes classes d'auditeurs. Enfin, il a été observé une grande corrélation entre l'augmentation de l'effort vocal et la perte d'intelligibilité.

1 INTRODUCTION

Le bruit est un facteur parmi tant d'autres pouvant influencer l'intelligibilité de la parole. En effet les conditions d'écoute, le contexte, le locuteur ou l'auditeur peuvent aussi affecter l'intelligibilité. En présence de bruit, la parole est plus ou moins masquée et sa production est soumise à ce qui est couramment appelé l'effet Lombard. L'effet Lombard est le reflexe qu'ont les locuteurs à augmenter leur effort vocal lorsqu'ils parlent en présence de bruit. Cet effet fut tout d'abord décrit par Etienne Lombard en 1911 [5].

De nombreux travaux (essentiellement à partir de 1950) ont montré l'influence de divers types de bruit sur l'intelligibilité de la parole. Il s'agit essentiellement de travaux sur de la parole normale ou criée et des locuteurs et auditeurs de même langue maternelle. Plus particulièrement des études ont été consacrées à l'influence sur l'intelligibilté de la parole bruitée de certains paramètres comme l'effort vocal [6, 2], la fréquence d'occurence et la durée des mots et l'étendue du vocabulaire [3]. Howes a en particulier montré que l'intelligibilité de la parole augmentait avec la longueur des mots, leur fréquence d'occurence et lorsque l'on réduisait la taille du vocabulaire. Au niveau de l'effort vocal Dreher et O'Neill raportèrent que, à rapport signal/bruit égal et pour des mots isolés ou de la parole continue, la parole produite dans du bruit est plus intelligible que la parole produite dans un environnement non-bruité [2]. Toutefois, les travaux de Pikett montrèrent que lorsque le locuteur augmente son effort vocal jusqu'à un niveau correspondant à de la parole criée l'intelligibilité diminue [6]. En effet l'effort vocal augmente l'énergie acoustique mais diminue l'information phonétique [8]. Le changement du type de voix entraîne un glissement du niveau phonétique vers un niveau phonologique inhabituel, propre à la voix criée [9].

Le but de l'étude présentée dans cet article est d'étudier l'intelligibilité de la parole produite dans du bruit par comparaison à celle de la parole normale. En particulier, l'influence de l'effet Lombard sur l'intelligibilité de vocabulaires acoustiquement similaires a été étudiée. Cet aspect semble avoir totalement été négligé par les études précédentes. L'étude a été menée pour des locuteurs américains et différentes classes d'auditeurs (français, anglais et américains) afin de pouvoir mesurer l'intelligibilité pour des sujets utilisant différents degrés de redondance au niveau de l'identification du mot émis. Pour chaque classe d'auditeurs le rapport signal/bruit a été changé. Cette étude vient en complément d'une étude acoustique sur l'influence de l'effet Lombard pour la même base de données [1]. Le but à moyen terme de ces études est d'utiliser les résultats obtenus afin d'améliorer la robustesse des systèmes de reconnaissance automatique de parole, particulièrement en milieu bruité.

2 CONDITIONS EXPERIMENTALES

Cette étude a utilisé une base de données de 49 mots comprenant le vocabulaire alphanumérique plus quelques mots de contrôle (voir Table 1). Les mots ont été prononcés de façon isolée par 10 locuteurs (5 hommes et 5 femmes) originaires de différentes régions des Etats-Unis.

zero	enter	start	h	r
one	erase	stop	i	s
two	go	yes	j	t
three	help	a	k	u
four	no	b	l	v
five	off	c	m	w
six	on	d	n	x
seven	repeat	e	o	y
eight	right	f	p	z
nine	rubout	g	q	

Table 1 Vocabulaire étudié

Chaque locuteur a prononcé la liste de mots (présenté chaque fois dans un ordre aléatoire) quatre fois dans une chambre sourde, deux fois dans un contexte non-bruité et deux fois dans un contexte bruité correspondant à l'injection d'un bruit blanc de 85 dB spl dans un casque calibré (TDH 49). C'est cette base de données qui a été utilisée dans tous les tests présentés. La seul conseil donné au locuteur avant de lire la liste de mots fut de parler de façon à pouvoir être compris. L'intelligibilité a été mesurée de façon globale (c'est à dire pour toute la base de données) mais aussi pour les sous-vocabulaires présentés Table 2. Dans ce cas même si les tests ont toujours été effectués de façon globale, les erreurs pour chaque sous-vocabulaire ont été comptabilisées.

Sous-vocabulaires considérés	Enumération du contenu des sous-vocabulaires
V1	zero, one, five, six, seven, nine
V2	enter, erase, help, repeat, right, rubout,
V3	f, s, x, yes
V4	a, eight, h, k
V5	b, c, d, e, g, p, t,v, z, three
V6	n, m

Table 2 Les sous-vocabulaires considérés

Ces sous-vocabulaires ont été choisis de façon arbitraire. Ils correspondent à des sous-vocabulaires de mots acoustiquement similaires (V3, V4, V5, V6) et des sous-vocabulaires de mots acoustiquement distincts (V1, V2). Le mot "j" n'a pas été inclus dans le sous-vocabulaire V4 car il fut rarement confondu avec les autres mots du même sous-vocabulaire.

Afin de pouvoir effectuer la relation entre l'intensité de l'effort vocal et les taux d'intelligibilité obtenus, l'augmentation de l'effort vocal entre la parole produite dans du bruit et la parole normale a été calculée pour chaque locuteur. Celle-ci a été obtenue en calculant la différence du rapport signal/bruit (*SB*) pour chaque locuteur et chaque mot dans les deux conditions : parole normale et parole produite dans du bruit, et en moyennant les résultats obtenus sur les 49 mots. Ces résultats sont présentés Table 3. Dans cette Table l'augmentation de l'intensité de l'effort vocal est représentée par $SB_2 - SB_1$.

	SB_1 (parole normale)	SB_2 (parole produite dans du bruit)	$SB_2 - SB_1$
locuteur ac	38 dB	54 dB	16 dB
locuteur ak	39 dB	53 dB	14 dB
locuteur bah	32 dB	55 dB	23 dB
locuteur dz	33 dB	56 dB	23 dB
locuteur na	35 dB	53 dB	18 dB
locuteur pf	38 dB	42 dB	4 dB
locuteur sp	31 dB	48 dB	17 dB
locuteur ta	29 dB	53 dB	24 dB
locuteur th	38 dB	45 dB	7 dB
locuteur vk	43 dB	51 dB	8 dB

Table 3 Augmentation de l'intensité de l'effort vocal entre la parole normale et la parole produite dans du bruit pour chaque locuteur de la base de données étudiée.

Ces résultats nous amènent à penser que la parole produite par les locuteurs bah, dz et ta s'approche de la parole criée.

3 ANALYSE PERCEPTUELLE

3.1 Test A

Pour ce test il a été demandé aux auditeurs d'identifier des mots dérivés de quatre conditions différentes correspondant à de la parole :

- normale sans bruit additif,
- produite dans du bruit sans bruit additif,
- normale en présence de bruit additif (SB=10 dB),
- produite dans du bruit en présence de bruit additif (SB=10 dB).

Chaque condition a fait l'objet d'un test séparé. Ceci a été vérifié expérimentalement. Dans tous les cas il a été demandé aux auditeurs d'indiquer une réponse, celle-ci devant être un mot de la liste de mots possibles. Pour chaque locuteur chacun des 49 mots était susceptible d'apparaître plusieurs fois ce qui évitait l'élimination d'un mot déjà identifié (5 mots ont été rajoutés aléatoirement à chaque liste). L'ensemble des auditeurs de ce test furent des auditeurs français ayant une bonne compréhension de l'anglais. La première partie du test a servi à sélectionner ceux qui avaient une bonne compréhension de l'anglais. En effet ceux qui ont commis plus de 10 erreurs au niveau de l'écoute de la parole normale sans bruit additif ont été éliminés. Dans tous les tests une répétition de chaque mot a été utilisé ce qui correpond à 540 mots par test (avec les mots rajoutés). Sur 16 auditeurs qui ont subi le test 9 ont été retenus. Ce sont les résultats obtenus pour ces 9 auditeurs (et tous les locuteurs de la base de données) qui sont présentés Table 4. Tous les auditeurs n'avaient à priori aucun problème d'écoute au moment des tests. Chaque test individuellement a duré environ une heure.

Procédure Les mots furent présentés aux auditeurs par l'intermédiaire d'un ordinateur masscomp et du logiciel snorri réalisé au laboratoire CRIN/INRIA. Les mots ont été présentés dans un ordre aléatoire de même que chaque locuteur de la base de données. Pour chaque mot, présenté par l'intermédiaire d'un casque, l'auditeur était invité à identifier le mot reconnu par l'intermédiaire du message *"Quel est le mot reconnu"*. Toute réponse permettait de passer au mot suivant. Pour tous les tests chaque mot pouvait être réécouté autant de fois qu'il était désiré avant de l'identifier. Lorsque du bruit blanc fut ajouté au signal de parole celui-ci fut ajouté sur tout le fichier de façon à ce que le bruit apparaisse bien avant l'apparition du mot à identifier.

Résultats Les résultats furent enregistrés automatiquement à l'aide du masscomp. Ceux-ci sont présentés Table 4.

Auditeurs	Taux d'identification			
	Parole normale sans bruit additif	Parole produite dans du bruit sans bruit additif	Parole normale en présence de bruit additif (SB=10 dB)	Parole produite dans du bruit en présence de bruit additif (SB=10 dB)
bo	99.0 %	99.0 %	95.9 %	92.9 %
me	99.2 %	99.3 %	95.9 %	90.0 %
ga	98.8 %	98.4 %	93.3 %	85.9 %
ma	98.8 %	97.8 %	93.5 %	87.8 %
to	99.0 %	98.4 %	95.7 %	91.2 %
ha	98.6 %	98.2 %	93.7 %	88.2 %
mo	99.0 %	99.0 %	94.7 %	93.7 %
so	98.0 %	98.4 %	94.7 %	93.5 %
wr	98.2 %	98.0 %	94.3 %	91.6 %

Table 4 Taux d'identification obtenus pour 9 auditeurs français et différentes conditions

Au vu de ces résultats remarquons tout d'abord que le taux d'intelligibilité avoisine les 99 % pour de la parole normale (non soumise à l'effet Lombard), ce qui dénote une bonne compréhension de l'anglais. Il est tout d'abord intéressant de noter que l'effet Lombard ne gêne que très peu les auditeurs (la moyenne se situe environ autour de 98.5 %) ce qui n'est pas le cas en reconnaissance automatique de parole par ordinateur pour laquelle un score d'environ 70 % sur la même base de données a été obtenu [4]. Lorsque le rapport signal/bruit est fixé à 10 dB les résultats obtenus montrent qu'il y a une diminution notable de l'intelligibilité lorsque la parole est produite dans du bruit pour tous les auditeurs ayant effectué le test (seuls les résultats moyennés sont présentés ici). Ceci est contraire aux résultats de précédents travaux comme ceux de Dreher et al. [2] qui montrent, comme cela a été indiqué dans l'introduction, que l'intelligibilité de la parole produite dans du bruit est supérieure à celle de la parole produite dans le silence. Afin d'expliquer ces résultats nous avons décidé de regarder de façon plus détaillée les erreurs commises par les auditeurs. Celles-ci sont présentées Table 5 et Table 6 pour respectivement la parole produite dans le silence et la parole produite dans du bruit.

Auditeurs	V1	V2	V3	V4	V5	V6
bo	0	0	3	1	13	0
me	0	0	3	3	13	1
ga	0	0	3	2	14	4
ma	0	1	5	2	16	0
to	0	0	2	0	17	0
ha	0	0	4	4	19	1
mo	0	0	3	1	17	3
so	0	0	4	1	17	1
wr	0	0	6	3	13	0
Total des erreurs	0	1	33	17	139	10

Table 5 Erreurs des auditeurs français pour la parole produite dans du silence (SB=10 dB).

Auditeurs	V1	V2	V3	V4	V5	V6
bo	0	1	11	1	21	0
me	0	2	7	6	27	0
ga	0	0	8	10	37	0
ma	0	0	12	9	35	0
to	0	0	11	4	28	0
ha	0	0	9	12	34	0
mo	0	0	3	3	24	0
so	0	0	8	0	21	0
wr	0	0	8	5	25	0
Total des erreurs	0	3	77	50	252	0

Table 6 Erreurs des auditeurs français pour la parole produite dans du bruit (SB=10 dB).

Nous pouvons remarquer que les erreurs ne se produisent pas au niveau des sous-vocabulaires de mots acoustiquement distincts mais plutôt au niveau des sous-vocabulaires de mots acoustiquement similaires. Ceci explique les différences de nos résultats avec ceux provenant d'études antérieures. En fait il apparait que des vocabulaires de mots acoustiquement similaires sont moins intelligibles lorsqu'il sont prononcés dans du bruit que lorsqu'ils sont prononcés dans le silence. Ceci semble raisonnable si l'on réfère à des études acoustiques sur la parole produite dans du bruit qui ont montrées que la durée des consonnes avaient tendance à être diminuée et que l'augmentation de l'intensité de l'effort vocal renforce plus les voyelles que les consonnes [1, 7]. Hors, les consonnes sont beaucoup plus responsables de l'intelligibilité que les voyelles. Enfin notons qu'au niveau de la paire "n", "m" l'intelligibilité s'améliore lorsque celle-ci est produite en milieu bruité. Ce résultat sera discuté au niveau de la conclusion.

Par une étude précise des résultats nous avons aussi essayé de répondre aux questions suivantes :

- y a t-il des locuteurs pour laquelle la différence d'intelligibilité entre la parole produite dans le silence et la parole produite dans du bruit est plus importante que pour d'autres?
- peut-on corréler cette différence à l'augmentation de l'intensité de l'effort vocal?

Même s'il ne nous est pas possible de fournir dans cet article tous les résultats nous avons mis en évidence qu'il y avait des locuteurs de la base de données considérée qui étaient beaucoup moins intelligibles que d'autres lorsqu'ils parlaient en milieu bruité (ak, bah, dz, ta). Il y a de plus une corrélation évidente entre l'augmentation de l'intensité de l'effort vocal et la perte d'intelligibilité. Toutefois cette relation n'est pas linéaire.

3.2 Test B

Ce test est similaire au test précédent à l'exception des points suivants :

- les auditeurs sont maintenant des anglais. 10 ont effectué le test.

- la parole utilisé est de la parole produite dans du silence et dans du bruit chaque fois en présence de bruit additif. Le rapport signal/bruit est de 0 dB.

Le but de ce test était de répondre aux mêmes questions que celles posées au test précédent avec cette fois-ci des auditeurs anglais (utilisant une plus grande redondance initiale pour identifier les mots) et un rapport signal/bruit plus faible (0 dB au lieu de 10 dB).

Résultats

	Taux d'identification	
Auditeurs	parole normale en présence de bruit additif (SB=0 dB)	parole produite dans du bruit en présence de bruit additif (SB=0 dB)
an	79.6 %	76.5 %
ca	87.6 %	82.4 %
cu	82.2 %	79.0 %
do	80.6 %	79.8 %
hd	85.1 %	81.0 %
he	85.1 %	83.7 %
ho	84.7 %	83.5 %
ma	87.8 %	88.4 %
mc	84.5 %	77.8 %
pe	80.4 %	82.0 %

Table 7 Taux d'identification obtenus pour 10 auditeurs anglais.

De façon générale comme pour le test précédent il y a une baisse d'intelligibilité globale. Une analyse détaillée des erreurs est fournie Table 8 et Table 9 pour les différents sous-vocabulaires.

Auditeurs	V1	V2	V3	V4	V5	V6
an	0	1	17	9	45	7
ca	0	1	12	4	33	2
cu	0	1	10	7	41	7
do	3	2	13	6	46	3
hd	2	0	11	8	33	3
he	1	1	15	1	40	6
ho	2	1	10	6	39	4
ma	2	1	9	4	27	7
mc	2	1	10	3	40	4
pe	2	3	13	11	44	3
Total des erreurs	14	12	120	59	388	46

Table 8 Erreurs des auditeurs anglais pour de la parole produite dans du silence (SB=0 dB).

Auditeurs	V1	V2	V3	V4	V5	V6
an	0	1	18	9	66	0
ca	0	1	14	6	48	0
cu	3	3	16	11	54	1
do	2	0	14	9	35	2
hd	4	2	16	11	43	1
he	2	1	14	5	50	0
ho	1	2	8	9	49	0
ma	0	1	11	9	34	0
mc	0	1	16	11	57	1
pe	3	1	11	10	53	2
Total des erreurs	15	13	138	90	489	7

Table 9 Erreurs des auditeurs anglais pour de la parole produite dans du bruit (SB=0 dB).

Les conclusions qui peuvent être tirées de ces résultats sont similaires à celles énoncées pour le test précédent avec toutefois une augmentation très nette de l'intelligibilité pour le sous-vocabulaire V6. En ce qui concerne la corrélation entre la perte d'intelligibilité et l'augmentation de l'intensité de l'effort vocal les remarques précédentes sont aussi toujours valables. Les auditeurs du test A et du test B sont de nationalité différentes et utilisent donc un niveau de redondance différent pour identifier les mots. Cependant, il est intéressant de noter que les phénomènes observées sont identiques pour les deux tests (à des rapports signal/bruit différents). Quant aux résultats obtenus pour les vocabulaires V1 et V2 il est difficile de conclure car le nombre d'erreurs est faible.

3.3 Test C

Enfin, un test identique au test B a été mené pour quatre auditeurs américains avec un rapport signal/bruit de -10 dB. Les résultats présentés Table 10 indiquent aussi une dégradation de l'intelligibilité. La dégradation intervient, comme dans les précédents tests, principalement au niveau des mêmes sous-vocabulaires de mots acoustiquement similaires.

	Taux d'identification	
Auditeurs	Parole normale en présence de bruit additif (SB=-10 dB)	Parole produite dans du bruit en présence de bruit additif (SB=-10 dB)
an	56.5%	45.5%
ba	62.2%	56.1%
ez	55.1%	43.7%
pu	57.6%	47.6%

Table 10 Taux d'identification moyennés sur 4 auditeurs américains.

4 DISCUSSION ET CONCLUSIONS

Les précédents travaux qui se rapprochent du travail présenté dans cette article concement des études sur la différence d'intelligibilité entre la parole normale et la parole produite dans du bruit [2, 10] et l'intelligibilité de la parole en milieu bruyant pour des locuteurs étrangers [9]. Toutefois les premières études se sont intéressées à des vocabulaires de mots acoustiquement distincts alors que l'étude de Rostolland et Parant s'est surtout intéressée à la relation entre le niveau de connaissance de la langue et le niveau de bruit admissible pour obtenir une intelligibilité donnée dans le cas d'auditeurs étrangers.

Les particularités de notre étude sont :

- d'utiliser un vocabulaire suffisamment important pour pouvoir prendre en compte des sous-vocabulaires de mots acoustiquement distincts ou similaires,
- de comparer les différences d'intelligibilité existantes entre la parole normale et la parole produite dans du bruit pour différentes classes d'auditeurs,
- la procédure utilisée pour les tests où chaque mot pouvait être réécouté à volonté ce qui permettait d'utiliser tous les indices possibles pour l'identification en particulier ceux qui n'avaient pas été perçus lors de la première écoute.

Les résultats obtenus ont permis de mettre en évidence qu'il y a une diminution notable de l'intelligibilité lorsque la parole est produite dans du bruit pour des sous-vocabulaires de mots acoustiquement similaires. Ce résultat qui est contraire à ce qui a été observé pour des mots acoustiquement distincts est intéressant pour les incidences qu'il peut avoir en reconnaissance automatique de la parole. En effet, une des théorie avancée est qu'il doit être possible d'améliorer la performance des systèmes actuels de reconnaissance de parole en milieu bruité en apprenant aux locuteurs à parler clairement [10]. Même si la parole produite en milieu bruité n'est pas indentique à la parole énoncée clairement, la perte d'intelligibilité que nous avons obtenu pour des mots acoustiquement similaires montre que le paramètre *nature du vocabulaire* joue un rôle très important. Les solutions à adopter pour améliorer les performances d'un système de reconnaissance automatique peuvent varier d'un vocabulaire à l'autre.

Contrairement aux autres sous-vocabulaires étudiés, le sous-vocabulaire formé des mots "m" et "n" est plus intelligible lorsque les mots sont prononcés en milieu bruité. Si l'on se rapporte à l'analyse acoustique faite sur la même base de données [1], nous avons observé que pour ces mots, lorsqu'il sont prononcés en milieu bruité, les nasales avaient tendance à être allongées et qu'il y avait même très souvent insertion d'un "e" muet en fin de mot (2 fois plus souvent lorsque la parole est produite dans du bruit que dans le silence). Ce phénomène peut expliquer pourquoi nous obtenons une meilleure intelligibilité lorsque ces mots sont prononcés en milieu bruité.

Nous avons aussi montré que les résultats s'appliquent à différentes classes d'auditeurs. Il est en particulier intéressant de noter que la fréquence d'utilisation du vocabulaire ne semble pas être un paramètre important au niveau de cette étude.

Enfin, il a été observé qu'il y a une grande corrélation entre la perte d'intelligibilité et l'augmentation de l'intensité de l'effort vocal. Celle-ci provoque une diminution de la qualité du signal de parole. Une perte d'intelligibilité avait déjà été observée au niveau de la voix criée par rapport à la parole normale [6]. Cependant, le problème est complexe car il n'y a pas une relation biunivoque entre le niveau de parole et le type de voix.

Au niveau des perspectives une des extensions de cette étude est d'analyser la répartition des erreurs en fonction de l'intensité de l'effort vocal qui varie pour les différents locuteurs. Il serait aussi souhaitable d'effectuer le test C sur davantage d'auditeurs américains ce qui permettra notamment de relier certaines parties de notre étude à des travaux antérieurs similaires.

REMERCIEMENTS

L'auteur tient à remercier tous les auditeurs du laboratoire CRIN/INRIA ainsi que les lecteurs anglais ou américains des différentes Universités de Nancy qui ont aimablement et bénévolement accepté de participer aux tests.

Bibliographie

[1] Y. Anglade and J.C. Junqua. Etude acoustique de l'effet lombard sur des phonèmes de l'anglais américain dans le cadre de mots isolés. 1990. XVIIIème Jounées d'Etudes sur la Parole, Montreal.

[2] J.J. Dreher and J.J. O'Neill. Effects of ambient noise on speaker intelligibility for words and phrases. *J. Acoust. Soc. Am.*, 29:1320–1323, 1957.

[3] D. Howes. On the relation between the intelligibility and frequency of occurrence of english words. *J. Acoust. Soc. Am.*, 29(2):296–305, 1957.

[4] J.C. Junqua and H. Wakita. A Comparative Study of Cepstral Lifters and Distance Measures for All-pole Models of Speech in Noise. In *ICASSP-89*, 1989.

[5] E. Lombard. Le Signe de l'Elévation de la Voix. *Ann. Maladiers Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[6] J.M. Pickett. Effects of vocal force on the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 28(5):902–905, 1956.

[7] D. Rostolland. Influence de l'effort vocal sur les caractéristiques physiques de la voix et sur l'intelligibilité de la parole. In *Jounées d'Etude sur la Boucle Audio-Phonatoire, Bruxelles*, 1978.

[8] D. Rostolland and C. Parant. Distorsion and intelligibility of shouted voice. In *Symposium: Speech Intelligibility. Liège*, pages 293–304, 1973.

[9] D. Rostolland and C. Parant. The intelligibility of a foreign language in a noisy environment. In *Symposium Intelligibility of Speech, Nottingham*, 1975.

[10] W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.*, 84(3):917–928, 1988.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

IDENTIFICATION DE VOYELLES SYNTHETIQUES PROJETEES SUR LES AXES PRINCIPAUX D'UNE ANALYSE FACTORIELLE

N. Nguyen-Trong, S. Santi, et C. Cavé

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261
29 av. R. Schuman, 13621 Aix-en-Provence, France

ABSTRACT: The so-called principal-components analysis is often used to extract the main "features" of a vocalic system, represented as a cloud of points in a multidimensional acoustic space. But attention has rarely been paid to the psychological verisimilitude of such features. In this study, 10 subjects had to identify synthetic french oral vowels [i,y,e,Ø,E,oe,a,O,o,u], whose initial positions has been transformed through orthogonal projections on each of the principal axes of the F1-F2-F3 space. The confusion matrices show that axes 1 and 2 can be related as expected to acuteness and compactness. Nevertheless, the part of the initial variance "explained" by an axe and, more generally, by a factor space, doesn't seem to be in perfect adequation with the amount of "information" it provides for the listeners (if such an amount is considered as the percentage of the vowels correctly identified). This result leads us to criticize the usual rule stating that a factor space has a relevance in direct proportion to the explained variance.

1. INTRODUCTION: LES APPLICATIONS DE L'ACP EN PHONETIQUE

En phonétique acoustique, les problèmes que l'on veut résoudre par l'analyse en composantes principales (ACP) se rangent essentiellement sous deux catégories:

a) l'identification des dimensions phoniques à travers lesquelles se laisse décrire "au mieux" (en un sens statistique que nous rappelons ci-dessous) un ensemble de consonnes ou de voyelles donné.

Ce problème émerge dès lors que les unités phonétiques doivent âtre considérées comme des *vecteurs* possédant un nombre relativement important de composantes (supérieur à 3), ce que l'on sait se produire fréquemment, puisqu'il est ainsi devenu ordinaire de représenter une voyelle par x valeurs numériques constituant les réponses moyennes d'autant de filtres passe-bande. En effet, le nuage de points associé à de tels vecteurs, se trouve dans un espace multidimensionnel non visualisable. L'intérêt de l'ACP est de transformer ce nuage en une image géométrique, qui est sa projection orthogonale sur le plan le moins "déformant" possible, c'est-à-dire sur lequel les distances entre les points sont maximisées (cf. Diday et al., 1982). Ainsi peut-on alors le percevoir et l'interpréter.

Plus généralement, lorsqu'une description phonétique fait intervenir x variables initiales, l'ACP permet de remplacer celles-ci par de nouvelles variables dites composantes principales, dont le nombre est plus réduit, et qui sont construites de telle sorte que l'"information" perdue dans l'opération est minimale. Les travaux de Pols et de ses collaborateurs nous ont familiarisés avec cette méthode (Klein et al., 1970; Plomp et al., 1967; Pols et al., 1969, 1973; citons également Favella et al., 1969, et Grenié, 1987).

b) l'encodage d'un signal phonique devant être transmis au moindre coût.

Il est parfois souhaitable d'éliminer la redondance résultant du fait que, dans un spectre évolutif, l'énergie ne varie pas indépendamment d'une bande de fréquences à l'autre. L'ACP peut être utilisée si l'on admet une nouvelle fois l'hypothèse selon laquelle la forme du spectre est gouvernée par un petit nombre de combinaisons linéaires entre les différents filtres. Ces "composantes principales" sous-jacentes requièrent, pour être stockées, une place moins importante que les variables d'origine. Par la synthèse, elles sont transformables à leur tour en un nouveau signal dont la redondance a diminué mais dont l'intelligibilité reste élevée (voir Boehm et al., 1968; Li et al., 1969; Zahorian et al., 1981).

Dans ces derniers travaux, seule l'intelligibilité globale semble avoir été prise en considération. Ainsi reste-t-on sans savoir *quels types* d'erreurs ont lieu lorsque des sujets doivent identifier un signal synthétisé sous le contrôle d'une ACP, si certaines confusions se répètent plus fréquemment que d'autres, et si le recodage affecte ou non toutes les unités phonétiques d'une même manière. Ces questions revêtent une importance technique certaine, et entrent comme telles dans la catégorie b). Mais la solution qui leur serait apportée pourrait également nous informer sur les principaux aspects d'un système vocalique ou consonantique. Supposons en effet qu'une fois représentées par les énergies moyennes contenues dans p bandes de fréquences, des voyelles se répartissent dans RP suivant un axe interprétable comme une échelle d'acuité générale. L'axe se subdivise donc en deux parties coïncidant l'une avec les sons "graves", l'autre avec les sons "aigus" (aucun véritable rapport n'étant à établir entre ces termes et les traits distinctifs de Jakobson). Supposons encore que les points soient projetés à angle droit sur cet axe; l'encodage est maximal, puisque le nuage est rendu unidimensionnel; il en résulte une composante principale unique. Si l'on synthétise, par son intermédiaire, les voyelles dont on a ainsi modifié les positions, nous pouvons selon toute vraisemblance prédire que:

- la distinction aigu-grave se conservera;

- les voyelles aiguës seront mutuellement confondues;
- les voyelles graves seront mutuellement confondues.

Les réponses des sujets prendront donc la forme suivante:

		SORTIE					
		← aig. →← grav. →					
	↑	x	x	x			
	aig.	x	x	x			
ENTREE	↓	x	x	x			
	↑				x	x	x
	grav.				x	x	x
	↓				x	x	x

Tableau 1: Matrice de confusion théorique entre voyelles réduites à une seule dimension d'acuité.

Mais l'important est que, réciproquement, la matrice de confusion nous permettra d'attribuer à l'axe principal une signification phonétique. Cette caractéristique doit être considérée avec attention, car elle rend possible une nouvelle manière d'interpréter une ACP. Par une épreuve de perception portant sur des stimuli synthétiques, on reconstitue sans difficulté les classes au sein desquelles des confusions sont produites. Dès lors, l'axe (et, en général, l'espace de projection) s'identifie immédiatement: il est cette dimension acoustique sur laquelle repose la distinction entre classes, la seule à s'être conservée. On voit donc ce que l'opération d'encodage définie en b) peut avoir d'utile pour la description phonétique elle-même (point a).

Dans ce qui suit, nous exposons les premiers résultats d'une épreuve qui semblent montrer que notre méthode ouvre des perspectives intéressantes.

2. UNE EPREUVE DE PERCEPTION

2.1. Stimuli

Les données initiales consistent en 10 voyelles orales françaises ([i,y,e,Ø,E,oe,a,O,o,u], où [O] désigne la voyelle prononcée dans le mot "motte") représentées par 6 points chacune (soit 60 points au total) dans l'espace des formants F1, F2, F3. Elles sont obtenues grâce à un processus aléatoire qui est tel que, pour chaque voyelle et chaque formant, les valeurs respectent (lorsque leur nombre tend vers l'infini), une distribution gaussienne dont la moyenne et l'écart type ont été fixés *a priori*, en accord avec le tableau ci-après:

	i	y	e	ø	ɛ
F1	290(20)	240(20)	360(20)	340(20)	540(20)
F2	2170(40)	1720(40)	2050(40)	1530(40)	1930(40)
F3	3070(200)	2300(200)	2580(200)	2170(200)	2580(200)

	oe	a	O	o	u
F1	570(20)	720(20)	480(20)	380(20)	260(20)
F2	1530(40)	1290(40)	860(40)	680(40)	810(40)
F3	2440(200)	2580(200)	2300(200)	2300(200)	2170(200)

Tableau 2: Fréquences initiales moyennes des formants pour les 10 voyelles. Source: Durand, 1985. Les écarts types figurent entre parenthèses et sont invariants d'une voyelle à l'autre.

Les Hz sont convertis en demi-tons, et les formants centrés en 0 (ainsi que le demande une ACP). Le calcul puis la diagonalisation de la matrice de covariance V nous donnent les 3 valeurs propres $\lambda_1, \lambda_2, \lambda_3$ et les 3 vecteurs propres associés $\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}$. Les composantes principales $\mathbf{y^1}, \mathbf{y^2}, \mathbf{y^3}$ en dérivent comme suit:

$$\underset{(60x1)}{\mathbf{y^j}} = \underset{(60x3)}{X} \quad \underset{(3x1)}{\mathbf{u_j}} \qquad j=1,2,3$$

où X désigne le tableau des valeurs d'origine. Les coordonnées des voyelles projetées sur l'axe $[\mathbf{u_1}]$ (ou "axe 1") dans l'espace des formants rapporté au repère initial sont alors déterminées selon la formule:

$$\underset{(60x3)}{X'} = \underset{(60x1)}{\mathbf{y^1}} \cdot \underset{(1x3)}{{}^t\mathbf{u_1}} \qquad (1)$$

où X désigne le tableau des valeurs terminales, et ${}^t\mathbf{u_1}$ la transposée de $\mathbf{u_1}$. Enfin, on recentre les formants dont les fréquences sont de nouveau exprimées en Hz.

Le calcul (1) est répété pour:
- l'axe $[\mathbf{u_2}]$
- le plan $[\mathbf{u_1},\mathbf{u_3}]$
- le plan $[\mathbf{u_2},\mathbf{u_3}]$
- le plan $[\mathbf{u_1},\mathbf{u_2}]$

Il en résulte un ensemble total de 360 points transformés en voyelles synthétiques grâce au système de Klatt (1980), implanté sur un mini-ordinateur MASSCOMP 5400 et muni d'un environnement interactif (multi-fenêtrage, menus déroulants, graphisme...) conçu par R. Espesser (laboratoire de phonétique d'Aix-en-Provence). Rappelons que ce système simule une sortie acoustique considérée comme le produit du spectre de source et de la fonction de transfert du conduit vocal. Les stimuli possèdent une durée constante de 250 ms; les courbes d'intensité globale et de F0 sont également invariantes.

2.2. Sujets

Dix sujets ont répondu au test. Tous sont droitiers et ne souffrent d'aucun trouble auditif connu. Leur tâche est d'identifier un à un les 350 stimuli. Les réponses sont inscrites sur des tables comportant chacune 10 lignes sur lesquelles figurent les 10 symboles i,y,e,Ø,E,oe,a,O,o,u dans cet ordre. Les stimuli se succèdent à 4 secondes d'intervalle, en séries de 10 séparées par un bip. Les sujets doivent leur attribuer un symbole respectif et un seul. Aucune difficulté particulière n'a été rencontrée.

2.3. Résultats

2.3.1. L'identification des voyelles initiales

	i	y	e	ø	ɛ	oe	a	O	o	u	
i	51		9								60
y		60									60
e			58		2						60
ø		5		52		3					60
ɛ			3		57						60
oe			2	2	10	46					60
a							60				60
O								57	3		60
o								1	59		60
u										60	60
	51	65	72	54	69	49	60	58	62	60	600

Tableau 3: Matrice de confusion entre les 60 voyelles non transformées; 10 sujets ont été soumis à l'épreuve, et ont donc produit, pour ces stimuli, un ensemble total de 360 réponses.

Parmi les 360 stimuli figurent les 60 voyelles initiales, synthétisées sans modification de leurs positions dans l'espace des formants. On vérifie ainsi qu'elles sont identifiables de manière non équivoque, en accord avec notre propre jugement, et que l'alphabet phonétique est correctement interprété par tous les sujets. La matrice ci-dessus nous sert de référence dans l'analyse des suivantes. Les voyelles ont été bien reconnues pour 94% d'entre elles.

2.3.2. L'identification des voyelles projetées sur le plan des axes 1 et 2

Cette deuxième condition engendre davantage d'erreurs, qui laissent supposer que les voyelles ont été centralisées ([i] devient [e], [y] devient [Ø], [o] devient [O]...) L'origine d'un tel phénomène est facile à établir: la projection orthogonale réduit les distances autour du barycentre du nuage; or celui-ci constitue une "moyenne" entre toutes les voyelles initiales, dont le F1 aurait pour fréquence 380 hz, le F2 1290 hz, et le F3 2435 Hz. Ainsi les sujets ont-ils plus fréquemment répondu [Ø] (82 fois) et [oe] (98 fois), toutes choses égales par ailleurs. Cependant, les identifications correctes restent dominantes (84.5%). Elles confirment l'idée déjà bien admise, selon laquelle deux paramètres suffiraient à déterminer entièrement un système vocalique (ces paramètres étant ici des combinaisons linéaires de F1, F2 et F3).

	i	y	e	ø	ɛ	oe	a	O	o	u	
i	45		15								60
y		42		16						2	60
e			58	2							60
ø				57		2			1		60
ɛ			1	3	20	36					60
oe				3		57					60
a							60				60
O				1		2		57			60
o						1		8	51		60
u										60	60
	45	42	74	82	20	98	60	65	52	62	600

Tableau 4: Matrice des confusions entre les voyelles projetées sur le plan [u_1,u_2].

2.3.3. L'identification des voyelles projetées sur l'axe 1

	i	y	e	ø	ɛ	oe	a	O	o	u	
i			57		3						60
y				55		5					60
e			49	3	8						60
ø				56		4					60
ɛ			17	34	4	5					60
oe				55		5					60
a				43	2	15					60
O				36		4		4	15	1	60
o								2	58		60
u								1	59		60
	0	0	123	282	17	38	0	17	132	1	600

Tableau 5: Matrice des confusions entre les voyelles projetées sur l'axe [u_1].

Les réponses révèlent clairement un effet de centralisation en ce sens que les voyelles périphériques [i] et [u] n'ont jamais été mentionnées à une exception près. Il se dégage de la matrice trois "points fixes" [e], [Ø], et [o], qui semblent chacun "absorber" les voyelles voisines: [i] est identifiée comme [e], [y] comme [Ø]... Notons que les sujets perçoivent maintenant mal la distinction labial/non labial (ou plutôt son corrélat acoustique): ainsi [E] est-elle assimilée à [Ø]. On délimite deux zones principales de confusion, l'une contenant [i,y,e,Ø,E,oe,a,O], l'autre [o,u], qui suggèrent que l'axe 1 coïncide avec une dimension d'*acuité* relative.

2.3.4. L'identification des voyelles projetées sur l'axe 2

	i	y	e	ø	ɛ	oe	a	O	o	u	
i		17		40		3					60
y		41		19							60
e				60							60
ø		1		59							60
ɛ				11		49					60
oe				7	1	52					60
a							60				60
O				5		55					60
o				42		18					60
u		11		49							60
	0	70	0	292	1	177	60	0	0	0	600

Tableau 6: Matrice des confusions entre les voyelles projetées sur l'axe [u_2].

Dans cette dernière condition, les réponses [Ø] et [oe] prennent une importance majeure (292 et 177 occurrences respectives) et, corrélativement, les confusions s'étendent à des voyelles plus éloignées ([i] devient [Ø] par exemple). Mais il faut surtout noter que la distinction aigu/grave est ici supprimée: [u] se transforme en [y], [o] en [Ø]. Le tableau 6 semble lui aussi définir deux classes parmi les voyelles: [i,y,e,Ø,o,u] et [E,oe,a,O]. On peut en conclure que ces voyelles se distribuent sur l'axe 2 selon leur *compacité* relative.

3. PERSPECTIVES

L'analyse de nos résultats a fait apparaître un phénomène supplémentaire. On sait que le nuage des voyelles initiales possède une inertie dont l'axe 1, par définition, "explique" une proportion supérieure à celle de l'axe 2. Mais les taux d'identification correcte se rangent dans l'ordre inverse, comme le montrent leurs valeurs suivantes:

	axe 1	axe 2
PIE	54%	34%
PIC	29%	35%

... PIE désignant le pourcentage d'inertie expliquée par l'axe, et PIC le pourcentage d'identification correcte des voyelles projetées sur le même axe. Or, une équivalence est souvent établie entre inertie et quantité d'"information" au sens large; on juge également naturel de dire que les premiers axes principaux sont les plus "pertinents". Pourtant, dans notre cas, cette pertinence statistique est en relation de *non coïncidence* avec la pertinence telle qu'elle est définie en phonologie (voir sur ce point Nguyen-Trong, N., 1988). Un problème important se pose ici, que de nouvelles épreuves de perception devront élucider.

4. REMERCIEMENTS

Nous remercions vivement Robert Espesser et Pierre Durand pour leurs contributions respectives à ce travail.

5. BIBLIOGRAPHIE

Boehm, J.F., and Wright, R.D.
(1968). "Dimensional analysis and display of speech spectra", *J.A.S.A.*, 44(1), 386.

Diday, E., Lemaire, J., Pouget, J., et Testu, F.
(1982). *Eléments d'analyse de données* (Bordas/Dunod, Paris).

Durand, P.
(1985). *Variabilité acoustique et invariance en français: consonnes occlusives et voyelles* (CNRS, "Sons et Parole", Paris).

Favella, L.F., Reineri, M.T., and Righini, G.U.
(1969). "On a mathematical procedure for detecting significant parameters in the classification of a statistical ensemble of phenomena and its application", *Kybernetik*, 5(5), 187-194.

Grenié, M.
(1987). "Nature et hiérarchie d'indices acoustiques indépendants du locuteur", Thèse de IIIème cycle, Phonétique (Univ. Aix-Marseille I, Aix-en-Provence).

Klatt, D.
(1980). "Software for a cascade/parallel synthetizer", *J.A.S.A.*, 67(3), 971-995.

Klein, W., Plomp, R., and Pols, L.C.W.
(1970). "Vowel spectra, vowel spaces, and vowel identification", *J.A.S.A.*, 48 (4/2), 999-1009.

Li, K.-P., Hughes, G.W., and House, A.S.
(1969). "Correlation characteristics and dimensionality of speech spectra", *J.A.S.A.*, 46 (4/2), 1019-1025.

Nguyen-Trong, N.
(1988). "Analyse en composantes principales et traits acoustiques",in Proceedings of *Speech '88*, 7th FASE Symposium, Edinburgh, 22-26 August 1988, pp.691-696.

Plomp, R., Pols, L.C.W., and Geer, J.P. van der
(1967). "Dimensional analysis of vowel spectra", *J.A.S.A.*, 41(3), 707-712.

Pols, L.C.W., Kamp, J. Th. van der, and Plomp, R.
(1969). "Perceptual and physical space of vowel sounds", *J.A.S.A.*, 46 (2/2), 458-467.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R.
(1973). "Frequency analysis of Dutch vowels from 50 male speakers", *J.A.S.A.*, 53 (4), 1093-1101.

Santi, S.
(1990). "Extraction et modélisation de paramètres acoustiques pour la synthèse du français"
Travaux de l'Institut de Phonétique d'Aix-en-Provence, n° 13.

Zahorian, S.A., and Rothenberg, M.
(1981). "Principal-components analysis for low-redundancy encoding of speech spectra", *J.A.S.A.*, 69(3), 832-845.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

ETUDE ACOUSTIQUE DE L'EFFET LOMBARD SUR DES PHONEMES DE L'ANGLAIS AMERICAIN DANS LE CADRE DE MOTS ISOLES

Yolande ANGLADE et Jean-Claude JUNQUA *

C.R.I.N / I.N.R.I.A LORRAINE BP 239 F-54506 Vandoeuvre les Nancy cedex.
* et SPEECH TECHNOLOGY LABORATORY, Santa Barbara, California.

RESUME

Le but de cette étude est de déterminer, au niveau phonétique, les différences acoustiques entre la parole normale et la parole prononcée en milieu bruité (effet Lombard). Sur une base de données de 49 mots isolés de l'anglais américain, prononcés par 10 locuteurs, nous avons effectué une analyse statistique contextuelle d'environ 40 paramètres, présentés pour tous les locuteurs confondus et séparément pour les hommes et les femmes. D'un point de vue prosodique, nos résultats confirment ceux des précédentes études sur le sujet : la fréquence fondamentale est en hausse (principalement pour les hommes) ainsi que la durée des voyelles. Les paramètres spectraux sont également touchés par l'effet Lombard : baisse de l'énergie entre 0 et 500Hz pour les voyelles, générale pour les nasales, fricatives et occlusives, montée du centre de gravité énergétique, du premier formant, de la limite inférieure de bruit fricatif et du nombre de passages par zéro (chez les locutrices), diminution de la norme cepstrale, et modification de certains paramètres calculés sur la barre d'explosion. Nous avons également observé une augmentation du nombre d'omissions de certaines occlusives et fricatives placées en fin de mot

PRESENTATION

En dépit d'une importante littérature sur l'effet Lombard, peu d'efforts ont été faits pour définir les changements acoustiques et phonétiques liés à l'effort vocal effectué par les locuteurs en présence de bruit. De plus, l'influence de cet effet a été négligé dans les systèmes de reconnaissance automatique de la parole. Pourtant, récemment [Raj86] et [Jun89], il a été montré que la variation de la production de la parole due à la présence de bruit engendre beaucoup plus de dégradations que le bruit ambiant lui-même.

Plusieurs études ont été consacrées à l'étude des différences entre la parole Lombard et la parole normale [Sum88] [Sta88] [Bon88] [Gre89]. L'objectif principal de ces études était d'améliorer les performances des systèmes de reconnaissance automatique grâce à la prise en compte des résultats concernant les changements acoustiques dus à l'effet Lombard. Cependant les résultats obtenus en reconnaissance automatique de la parole ne sont pas encore satisfaisants.

Pour exprimer les différences acoustiques entre la parole Lombard et la parole normale, nous avons calculé sur une base de données de l'anglais américain, des paramètres choisis du fait de leur importance en reconnaissance automatique de la parole. Nous avons choisi d'effectuer cette étude au niveau phonétique pour disposer de résultats assez précis de façon à pouvoir les appliquer à plusieurs domaines de la reconnaissance automatique, comme la reconnaissance à l'aide d'indices acoustiques ou la reconnaissance globale. Pour ce faire, un étiquetage du corpus a été effectué. L'ensemble des résultats obtenus a été analysé à l'aide de tests statistiques afin d'en dégager les différences significatives. Les résultats ont été présentés pour chaque phonème et chaque paramètre. Nous en donnerons ici une synthèse.

Arp.	IY	IH	EY	EH	AA	AO	UW
IPA	i	I	e^{y}	ϵ	a	$\supset$	u
Arp.	OW	AH	AW	AX	UR	AY	EL
IPA	o^{w}	$\wedge$	$\underline{a}^{w}$	aϖ	∂	a^{y}	∂l

Figure 1 : voyelles et diphtongues de l'anglais-américain représentées à l'aide de la notation Arpabet et de la notation IPA.

Arp.	Y	W	L	R	HH	F	TH	S	V	Z
IPA	j	w	l	r	h	f	θ	s	v	z
	semi-consonnes		liquides		fricatives					
Arp.	M	N	CH	JH	P	T	K	B	D	G
IPA	m	n	$\check{c}$	j	p	t	k	b	d	g
	nasales		affriquées		occlusives					

Figure 2 : consonnes de l'anglais américain,representées à l'aide de la notation Arpabet et de la notation IPA.

PROCEDURE EXPERIMENTALE

LA BASE DE DONNEES

Nous avons travaillé sur une base de données comportant 49 mots (chiffres, alphabet et mots techniques couramment utilisés) issus de l'anglais américain. Ces mots ont été prononcés de façon isolée par 10 locuteurs — 5 hommes et 5 femmes — originaires de différentes régions des Etats-Unis. Chaque locuteur a prononcé quatre fois la liste de mots dans une chambre sourde, deux fois dans une ambiance non bruitée et deux fois dans une ambiance bruitée : un bruit blanc de 85 DB SPL était injecté dans un casque calibré (TDH49). L'ordre des mots a été chaque fois interverti afin que le locuteur ne s'y habitue pas.

Le corpus contient la majorité des phonèmes de l'anglais américain. Nous avons pu travailler de façon satisfaisante sur 25 d'entre eux et de façon plus approximative sur 9 autres pour lesquels nous possédions peu d'occurences. Les phonèmes ont

été classés en différentes classes phonétiques, ils sont représentés figures 1 et 2 à l'aide de la notation arpabet et de la notation IPA.

L'ensemble du corpus comporte 5280 phonèmes. La bande de fréquence étudiée va de 0 à 5000Hz car les données ont été digitalisées à 10kHz.

LES PARAMETRES CALCULES

Nous avons calculé environ 40 paramètres sur le corpus. La majorité d'entre eux a été calculée en trois points de chaque phonème répartis de façon à négliger au maximum les influences des contextes gauche et droit (au premier quart, à la moitié et au dernier quart). Après élimination des valeurs aberrantes éventuelles, nous avons calculé une moyenne de ces trois points pour obtenir la valeur finale qui a été utilisée pour les tests de comparaison. Voici les paramètres retenus :

- l'énergie dans sept bandes de fréquences : 0–250Hz, 250–500Hz, 500–1000Hz, 1–2kHz, 2–3kHz, 3–4kHz, 4–5kHz;
- le centre de gravité énergétique : il permet de synthétiser en un seul paramètre le déplacement fréquentiel de l'énergie;
- les quatre premiers formants et leurs largeurs de bande;
- les 20 premiers coefficients cepstraux : nous nous sommes dans un premier temps intéressés à leur norme utilisée dans les mesures de distance;
- les trois premiers pics du spectre de fréquences calculé à l'aide de la méthode d'analyse perceptivement fondée PLP (en anglais "Perceptually-based Linear Prediction), développée par Hermansky [HHW85b]. Cette méthode qui modélise le comportement de l'appareil auditif périphérique fait notamment appel à une analyse en bandes critiques et à un modèle tout pôle d'ordre réduit;
- les coefficients cepstraux calculés à partir des coefficients PLP;
- La durée : nous avons étudié la durée de chaque phonème ainsi que celle de chaque mot;
- La pente spectrale : elle a été calculée en basses et hautes fréquences,
  a. Pente BF : nous avons sélectionné les énergies des quatre premières bandes de fréquences (de 0 à 2kHz) et cherché la droite de régression s'approchant de ces quatre points. C'est la différence angulaire entre les pentes BF calculées sur la parole normale et la parole Lombard qui a été analysée;
  b. Pente HF : la même analyse a été effectuée avec les trois autres bandes de fréquences (de 2 à 5kHz);
- La fréquence fondamentale;
- Les passages par zéro du signal temporel;
- La limite inférieure de bruit fricatif;
- La barre d'exposion (burst) : elle a été analysée, pour les occlusives et les affriquées, à l'aide de sept paramètres :
  a. les trois principales concentrations d'énergie dans la barre d'explosion;
  b. les fréquences extrêmes apparaissant dans le burst : fréquence basse et fréquence haute;
  c. "force" du burst : évalue le rapport entre l'énergie la plus intense du burst et les énergies extrêmes du spectrogramme;
  d. "compacité" du burst : évalue le rapport "énergie de la fréquence la plus intense du burst / énergie moyenne du burst";

L'ETIQUETAGE

Nous avons étiqueté l'ensemble du corpus manuellement à l'aide de "Snorri" [Lap88], logiciel d'analyse et de traitement de la parole développé au CRIN/INRIA LORRAINE. Les débuts et fins de segments ont été enregistrés, ainsi que l'étiquette de chaque phonème, à l'aide de la notation arpabet. Les occlusives et les affriquées ont été étiquetées en deux parties correspondant d'une part au silence et d'autre part à la barre d'explosion suivie du bruit de friction. A l'aide de cet étiquetage, nous avons pu effectuer le calcul des différents paramètres sur chacun des phonèmes, ainsi que le recensement des omissions et des insertions de phonèmes.

LES TESTS STATISTIQUES

Nous avons utilisé des tests statistiques pour analyser les résultats obtenus pour chacun des paramètres calculés. L'objectif de ces tests est de fournir les tendances générales concernant les différences entre les populations "parole normale" et "parole Lombard" à partir des échantillons de valeurs que nous possédons. Nous avons comparé ces deux populations en formant des paires d'échantillons contenant la valeur d'un paramètre pour la parole normale et pour la parole Lombard. Si l'une des répétitions ne donne lieu à aucune valeur du paramètre (phonème non prononcé ...) la répétition correspondante est elle-aussi supprimée, afin de comparer des populations homogènes de même dimension.

Nous avons utilisé deux tests classiques de comparaison [AR77] pour analyser les différences : le T-test et une analyse de variance à deux dimensions. Il était intéressant de disposer de ces analyses couramment utilisées et de comparer leurs résultats pour confirmer les conclusions de notre étude. Nous avons relevé très peu de différences entre ces tests qui ont été appliqués pour chaque paramètre et chaque phonème dans les configurations suivantes :

1. pour tous les locuteurs dans tous les contextes (où apparaît le phonème).
2. pour chaque locuteur dans tous les contextes.
3. pour tous les locuteurs dans chacun des contextes.

Les résultats de ces tests sont donnés sous forme d'une probabilité (p) que l'on compare à un seuil de décision égal à 0,05 (seuil généralement employé). Pour quantifier les différences existantes, nous avons calculé les moyennes des deux ensembles d'échantillons et leur différence en pourcentage.

LES RESULTATS

Les résultats sont présentés pour chaque phonème et chaque paramètre de façon systématique. Les références qui ont été prises sont les valeurs des paramètres calculées en parole normale. Nous avons essayé de répondre dans chaque cas aux questions suivantes (explication des symboles utilisés dans les tableaux de résultats) :

1. peut-on dégager une tendance concernant le comportement du paramètre ?
2. cette tendance va-t-elle dans le sens :
   - d'une augmentation : symbole 
   - d'une diminution : symbole 
   - d'une égalité : symbole 
3. cette tendance dépend-elle du locuteur ? (traduite par le côté gauche de chaque signe, prenons le cas d'un) :

- pas du tout ou très peu : le nombre n de locuteurs suivant la tendance est tel que n>= 8 ; le symbole est noirci du côté gauche ◐
- sensiblement : le nombre n de locuteurs suivant la tendance est tel que 8>n>=5 ; le symbole est rayé du côté gauche ⊘
- beaucoup : le nombre n de locuteurs suivant la tendance est tel que 5>n>=3 ; le symbole est blanc du côté gauche ◯

4. cette tendance dépend-elle du contexte dans lequel est placé le phonème ? (traduite par le côté droit de chaque signe) :

- pas du tout ou très peu : le symbole est noirci du côté droit ◑
- sensiblement : le symbole est rayé du côté droit ⊘
- beaucoup : le symbole est blanc du côté droit ◯

Nous avons présenté les résultats pour tous les locuteurs confondus et pour chaque sexe séparément.

Des demi-symboles apparaissent dans les tableaux de résultats. Ils sont utilisés dans le cadre de résultats par sexe et correspondent à des tendances que l'on ne retrouve pas de façon nette globalement et pour lesquelles on ne peut définir l'influence du contexte.

Nous allons maintenant présenter des tableaux graphiques de résultats concernant des paramètres pour lesquels on observe des changements significatifs, puis une synthèse de l'ensemble des résultats obtenus.

Tableaux graphiques de résultats

Ces tableaux présentés figures 3 et 4, respectivement pour les femmes et pour les hommes, concernent l'énergie dans les deux premières bandes de fréquences étudiées (de 0 à 500Hz), le premier formant, la fréquence fondamentale ainsi que les passages par zéro. Il est intéressant de noter que des différences interviennent entre les deux sexes. Les résultats sont commentés dans la synthèse générale.

Synthèse des résultats

Des tableaux similaires ont été établis pour l'ensemble des paramètres étudiés. La taille restreinte de cet article ne nous permet pas de les présenter ici, nous allons simplement en donner les conclusions essentielles qui correspondent à des changements statistiquement significatifs.

- Energie : L'énergie baisse de façon significative (de 17 a 37%) entre 0 et 500Hz pour les voyelles chez tous les locuteurs, et plus faiblement jusqu'à 1000Hz. Cette baisse est probablement liée à l'augmentation de la fréquence fondamentale. Une augmentation moins sensible apparaît uniquement chez les femmes entre 4 et 5kHz. Les nasales, fricatives, affriquées et occlusives sont affaiblies dans toutes les fréquences.
- Le centre de gravité énergétique : il s'élève pour l'ensemble des locuteurs et des phonèmes (particulièrement voyelles).
- Formants : il y a une montée générale du premier formant pour les voyelles, semi-consonnes, glides, liquides et nasales, allant de 40 à 115 Hz. Cette tendance est plus marquée chez les locutrices que chez les locuteurs, mais ne dépend pas du contexte. Le second formant suit cette même tendance, uniquement chez les femmes.
- Norme des coefficients cepstraux : celle des voyelles baisse de façon indépendante du locuteur et du contexte (ordre de grandeur : de 15 à 30 %).
- La durée : elle augmente unanimement pour les voyelles, entraînant ainsi une augmentation générale de la durée des mots. On retrouve la tendance inverse, beaucoup moins marquée pour les consonnes.
- La fréquence fondamentale : on observe une augmentation très nette de ce paramètre chez les hommes, alors que la tendance est beaucoup moins nette et plus faible chez les femmes. Le contexte n'intervient pas dans ce changement.
- Le nombre de passages par zéro : il augmente de façon significative pour les voyelles, semi-consonnes, glides, liquides et nasales, chez les locutrices uniquement (entre 2 et 25 passages supplémentaires toutes les 10ms, selon le phonème).
- La limite inférieure de bruit fricatif : elle est en hausse pour les phonèmes /S/ et /Z/ ainsi que pour le phonème /F/ chez les hommes.
- La barre d'explosion : les omissions ainsi que les problèmes de détection de burst nous ont amené à analyser ces différents paramètres sur un nombre plus limité de phonèmes (entre 23 et 70% du corpus selon les occlusives) qui ne nous a pas permis d'analyser la dépendance vis à vis des locuteurs.

  a. la première concentration d'énergie du burst est plus haute en fréquence pour les phonèmes /T/ et /G/ (environ 25%). Elles est en baisse pour le phonème /K/ (9%).
  b. la seconde est en hausse pour les phonèmes /K/ et /G/, ainsi que pour le /T/ lorsqu'il est placé en début de mot. Elle baisse pour le phonème /P/ placé en fin de mot.
  c. la troisième monte pour les phonèmes /G/ et /P/ lorsqu'il est en fin de mot. On observe la tendance inverse pour le phonème /D/.
  d. la fréquence haute du burst ne varie pas.
  e. la fréquence basse du burst évolue de façon diverse pour les différents phonèmes. Elle est en augmentation pour le /P/ placé en début et surtout en fin de mot, pour le /T/ placé en début de mot (entre 77 et 104%), et pour le /B/ (environ 80%). En revanche, elle baisse pour le /D/ (environ 30%) et pour le /G/ (17%).

- Les omissions de phonèmes : elles touchent les consonnes placées en fin de mot et tout particulièrement les phonèmes /T/ et /P/, le /F/ étant lui aussi concerné de façon moindre par ce phénomène. Les omissions interviennent en parole normale et en parole Lombard où elles sont entre trois et cinq fois plus nombreuses selon le phonème, atteignant jusqu'à 15% du corpus pour le phonème /P/.
- Les insertions de phonèmes : elles interviennent sous forme de "e muet" placés en fin des mots M et N. Elles sont deux fois plus nombreuses en parole Lombard, mais restent cependant limitées.

Nous continuons l'analyse pour les autres paramètres présentés dans l'étude.

CONCLUSION / DISCUSSION

L'étude que nous menons actuellement est proche notamment de quatre autres études récentes effectuées sur l'effet Lombard..

Pisoni, Bernacki, Nusbaum et Yuchtman [Pis85], ont analysé

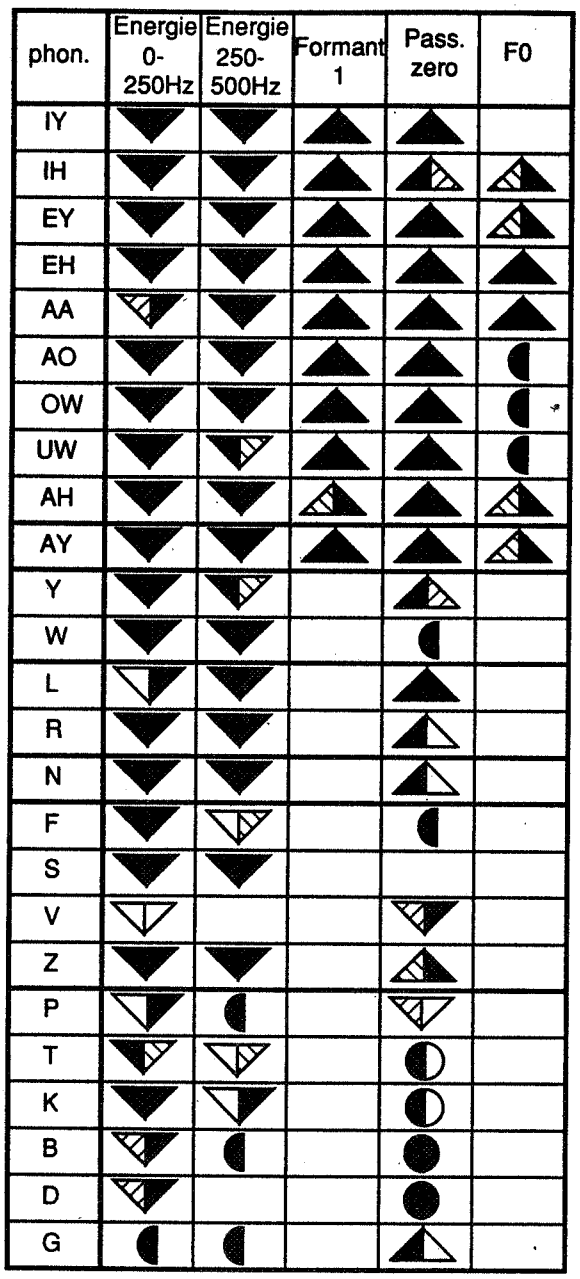


Figure 3 : variations concernant l'énergie dans les deux premières bandes de fréquences étudiées, le premier formant, la fréquence fondamentale et les passages par zéro pour les femmes.

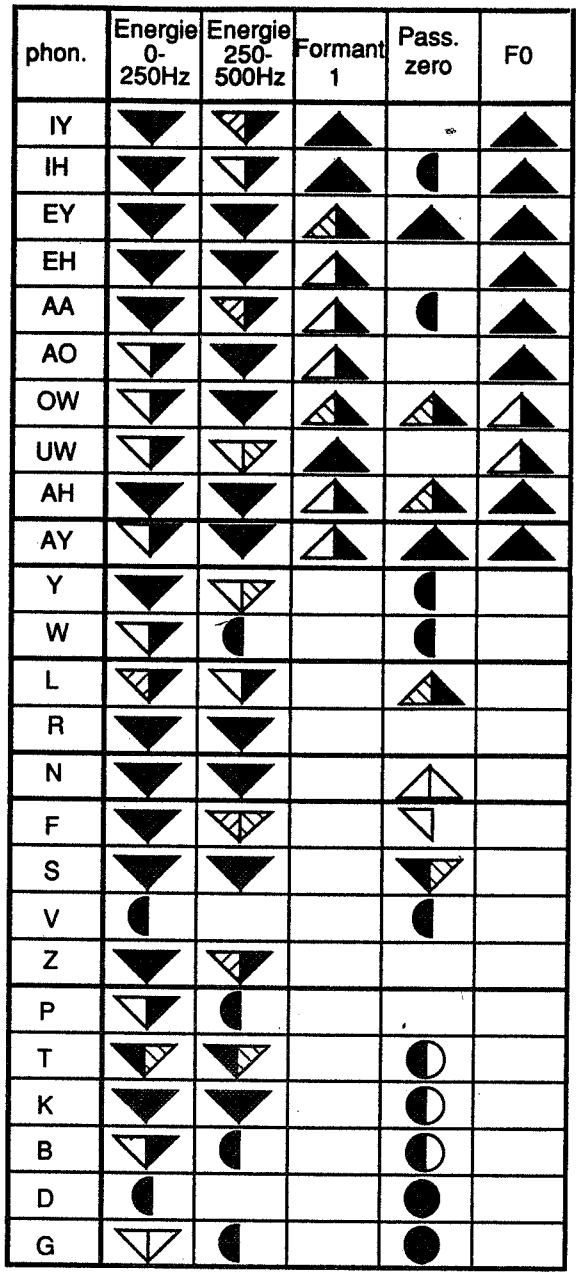


Figure 4 : variations concernant l'énergie dans les deux premières bandes de fréquences étudiées, le premier formant, la fréquence fondamentale et les passages par zéro pour les hommes.

une base de données comportant les chiffres et quatre mots techniques de l'anglais américain, prononcés de façon isolée par deux locuteurs masculins, dans quatre configurations distinctes : ambiance calme et ambiance bruitée (avec trois niveaux de bruit différents). Certains paramètres — durée totale, fréquence fondamentale et pente spectrale — ont été calculés sur les mots ; ils y observèrent des changements déjà signalés dans quelques études préalables. D'autres — formants et énergie — ont été analysés sur les voyelles ; ils aboutirent à une montée du premier formant et à une baisse du second, dues à l'effet Lombard.

Une étude a été pratiquée sur des phonèmes français par Grenié et Del Negro [Gre89]. Dix-neuf mots ont été prononcés par deux hommes et une femme dans une ambiance calme et dans une ambiance bruitée (avec trois niveaux de bruit différents). Les conclusions obtenues concernent l'augmentation de la durée, de la fréquence fondamentale, de l'amplitude totale ainsi que du premier formant.

Des recherches ont également été menées par Bond, Moore et McCreight [Bon88] sur de la parole continue : quatre locuteurs masculins ont prononcé un vocabulaire contenant vingt phrases

courtes en anglais américain, deux fois en ambiance calme et deux fois en ambiance bruitée. La segmentation a été effectuée au niveau phonétique, les paramètres calculés sont les suivants : énergie totale, énergie en basses et hautes fréquences, formants, fréquence fondamentale et limite inférieure de bruit fricatif. Les conclusions fournissent des résultats précis concernant neuf voyelles (/IY/, /IH/, /EY/, /EH/, /UW/, /OW/, /AA/, /AH/, /EA/), où une augmentation de la fréquence fondamentale atteignant 40Hz, du premier formant entre 20 et 70Hz ainsi qu'une diminution du second formant de 20 à 100Hz ont été mises en évidence.

Enfin Stanton, Jamieson et Allen [Sta88] ont mené une étude sur de la parole continue prononcée par cinq hommes, dans un milieu calme, dans un milieu bruité (90DB SPL) et en parole criée. Les paramètres suivants ont été calculés sur environ 11000 phonèmes segmentés : durée, énergie dans différentes bandes de fréquences (0 à 8kHz), centre de gravité énergétique, pente spectrale basses fréquences (0 à 3kHz) et hautes fréquences (3kHz à 8kHz), fréquence fondamentale, trois premiers formants. Cette étude aboutit à plusieurs conclusions : diminution de l'énergie entre 0–500Hz et 4–8kHz avec une augmentation correspondante entre 500Hz-4kHz pour les voyelles et semi-voyelles, augmentation sensible de la fréquence fondamentale (30Hz), du premier formant (35Hz) alors que le second et le troisième formants ont des comportements plus variables. La durée des voyelles tend à augmenter alors que celle des fricatives et des plosives suit la tendance inverse.

Nos résultats peuvent être comparés à ceux de ces études concernant les paramètres communs. Il en ressort des résultats similaires pour le premier formant, la fréquence fondamentale, la durée, le centre de gravité énergétique. Les résultats sont partiellement similaires pour l'énergie, non pas concernant la baisse entre 0 et 500Hz mais pour les bandes de fréquences intermédiaires où Stanton a noté une hausse d'énergie significative. Cette différence peut provenir de l'hétérogénéité de notre base de données comportant des voix d'hommes et de femmes, ainsi que de la variabilité interlocuteurs. Nous avons également donné des résultats concernant cette variabilité dans tous nos tableaux récapitulatifs (tels ceux des figures 3 et 4donnés précédemment). Cette variabilité intervient par exemple pour la fréquence fondamentale qui varie beaucoup plus pour les hommes que pour les femmes. Si certains paramètres sont affectés de façon unanime par l'effet Lombard, il est intéressant de remarquer que le degré de variation évolue d'un locuteur à un autre jusqu'à un facteur 3. Nous avons mené une étude sur l'énergie et le premier formant qui tend à montrer qu'un locuteur chez lequel on observe une forte évolution pour un paramètre, la conserve pour les autres paramètres touchés et inversement. Certains locuteurs sont donc plus sensibles que d'autres à l'effet Lombard. Il serait toutefois intéressant de prolonger cette étude par une analyse plus approfondie de l'influence de l'effet Lombard sur les variabilités intra et inter-locuteurs. Parmi les nouveaux paramètres observés, il est intéressant de noter la diminution de la norme des coefficients cepstraux; ce résultat pourra être utilisé pour les mesures de distance des systèmes de reconnaissance automatique.

Nous avons également analysé l'influence du contexte dans les différents résultats. Les tableaux récapitulatifs nous indiquent également si son influence est significative ou non. Dans les premiers cas, nous avons tenté de dégager les différentes configurations qui restent cependant limitées à notre base de données.

Les conclusions que nous avons obtenues, sur un nombre de locuteurs plus important, permettent en outre de mettre en évidence des différences intéressantes entre les locuteurs et les locutrices. Il s'agit, à notre connaissance, de la première étude effectuée sur une base de données contenant plusieurs voix féminines.

Enfin, il est également intéressant d'effectuer un parallèle avec des travaux effectuées sur de la voix criée [Sch89] [Tra85], où une corrélation entre l'augmentation de la fréquence fondamentale et du premier formant a été faite (paramètres également observés en augmentation dans la parole Lombard), et où il est aussi fait mention d'une tendance à l'augmentation de la durée des voyelles et d'une diminution de celle des consonnes.

L'ensemble des résultats obtenus témoignent de l'importance de l'influence de l'effet Lombard sur les paramètres acoustiques utilisés dans les systèmes de reconnaissance automatique de la parole et de la nécessité d'en tenir compte pour améliorer leur robustesse au bruit. L'intégration de ces résultats constitue l'objectif futur de ce travail, lorsque l'analyse de tous les paramètres aura été effectuée.

BIBLIOGRAPHIE

[AR77]H.L. Alder and E.B. Roessler, editors. Introduction to Probability and Statistics. W.H. Freeman and Compagny, 1977.

[Bon88]Z.S. Bond, T.J. Moore and K. MacCreight. "Some phonetic characteristics of sentences produced in noise". Spring meeting of the Acoustical Society of America, mai 1988, Seattle, Washington.

[Gre89]M.Grenié, and S. Del Negro. "Acoustic-phonetic analysis of speech produced under noise and auditory feedback". Proceedings EUROSPEECH 89, Vol. 2, p. 681, Paris, Septembre 1989.

[HHW85b]H. Hermansky, B.A. Hanson, and H Wakita. "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain". Speech Communication, (4):181–187, 1985.

[Jun89]J.C. Junqua, and H. Wakita. "A comparative study of cepstral lifters and distance mesures for all pole models of speech in noise". Proceedings ICASSP 89, p 476–479, mai 1989.

[Lap88]Y. Laprie. "Snorri,un système d'étude interactif de la parole". CRIN/INRIA Nancy, 1988.

[Pis85]D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman. "Some Acoustic-Phonetic Correlates of Speech produced in noise". Proceedings ICASSP 85, March 1985.

[Raj86]P.K. Rajasekaran and G.R. Doddington. "Recognition under Stress and in Noise". Proceedings ICASSP 86, April 1986.

[Sch89]R. Schulman. "Articulatory dynamics of loud and normal speech". J. Acoust. Soc. Am., 85 (1), 295–312, January 1989.

[Sta88]B.J. Stanton, L.H. Jamieson, and G.D. Allen. "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions". Proceedings ICASSP 1988.

[Sum88]W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stockes. "Effects of noise on speech production : Acoustic and perceptual analyses". J. Acoust. Soc. Am., 84 (3), 917–927, Septembre 1988.

[Tra85]H. Traunmüller. "The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness". PERILUS, No 4, (Inst. Linguist., Univ. Stockolm), p 92–102, 1985.

5 SYNTHÈSE ET CODAGE

Président: R. CARRÉ
ENST-Paris,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

L'ANALYSE DE LA RELATION LANGUE-PAROLE POUR UN SYSTEME DE SYNTHESE ARTICULATOIRE

Danièle Archambault*, Gilles Boulianne**, Henrietta Cedergren**

*INRS-Télécommunications **Université du Québec à Montréal

RESUME

Nous avons voulu mettre au point un système de synthèse qui permette l'examen des relations complexes entre la connaissance phonologique, la parole et la structure physiologique sous-jacente. Dans ce but, toutes les étapes de simulation de la production du français parlé, des plus abstraites jusqu'aux articulateur eux-mêmes, ont été formulées de façon à être facilement accessibles et manipulables. Au cours de l'élaboration d'un tel système, nous avons dû délimiter les niveaux de représentation phonologique, phonétique et physiologique, décrire leur structure interne, et définir leur interface avec les autres niveaux, en raccordant des approches différentes pour faire un tout opérationnel. Le système obtenu permet de vérifier expérimentalement des hypothèses qui concernent plusieurs niveaux, sur l'interface entre la phonologie et la phonétique, par exemple.

1. INTRODUCTION

Les développements récents en synthèse de parole par ordinateur ont permis l'apparition sur le marché de systèmes variés mettant à profit différentes techniques. Cette possibilité de produire de la parole à l'aide de l'ordinateur présente beaucoup d'intérêt pour les linguistes et les phonéticiens qui y voient un outil d'investigation permettant d'examiner de plus près les processus de production de la parole [1,13]. Nous avons donc voulu mettre au point un système de synthèse qui se veut avant tout un instrument d'analyse scientifique pour les chercheurs en parole.

Notre principal objectif est d'enrichir nos connaissances sur le rapport existant entre la connaissance phonologique, la parole et la structure physiologique sous-jacente. Cet examen des relations complexes entre différents niveaux se fait par l'entremise d'un synthétiseur articulatoire simulant toutes les étapes de la production du français parlé, des plus abstraites jusqu'aux articulateurs eux-mêmes.

Nous donnerons d'abord une description générale du système. Nous examinerons ensuite la question de la représentation des connaissances phonologique, phonétique et physiologique, et nous discuterons du problème de l'interface entre les différents types de représentation.

2. DESCRIPTION DU SYSTEME

Ce système se veut avant tout un outil d'analyse scientifique pour les chercheurs en langue et parole. En ce sens, il se dissocie de plusieurs systèmes de synthèse qui s'adressent au public en général. En effet, la plupart des systèmes texte-à-parole ont pour but la lecture orale par l'ordinateur de n'importe quel texte écrit et ont comme applications, par exemple, la lecture de textes pour les aveugles ou encore l'accès à de grandes banques de données via un système téléphonique [2]. Comme ils s'adressent à des utilisateurs naïfs, les différentes phases d'analyse telles la conversion graphémo-phonétique, la définition des paramètres acoustiques et les règles de conversion et d'évolution de ces paramètres, par exemple, sont généralement cachées à l'utilisateur.

Les paramètres de contrôle de notre système sont accessibles, afin de permettre aux chercheurs de saisir plus concrètement les mécanismes multidimensionnels de contrôle de la parole, de tester des théories et d'exploiter les relations entre la représentation phonologique et le contrôle des articulateurs. L'usager entre au terminal la transcription phonétique large

d'un mot ou d'une phrase[1]: à partir de cette entrée, le système donne une représentation phonologique, puis phonétique, et enfin articulatoire, avant de produire la synthèse. L'usager peut alors modifier les représentations et redemander le calcul de celles qui en découlent.

Le système comprend trois grandes composantes (voir figure 1). D'abord, une série de modules où sont représentées les différentes connaissances phonologiques, phonétiques et physiologiques. Ensuite, un algorithme de transformation des positions articulatoires en sons parlés (le synthétiseur articulatoire proprement dit). Enfin, une structure de représentation centrale par l'intermédiaire de laquelle les modules communiquent entre eux. L'usager peut examiner et modifier cette structure via l'interface usager.

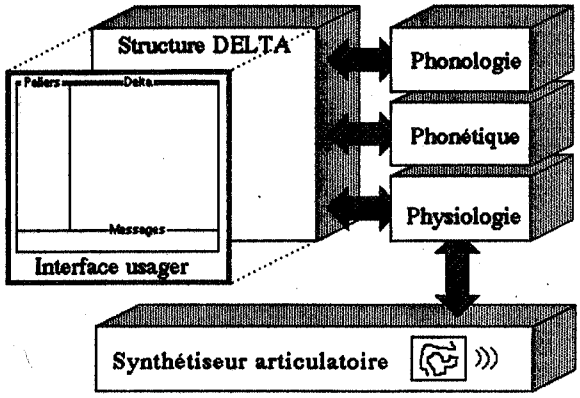


Figure 1. Composantes du système

La structure centrale ne contient, au départ, que la représentation incomplète fournie par l'usager. Les modules la consultent, la manipulent et la remplissent jusqu'à ce qu'elle soit complète à tous les niveaux. L'organisation de cette structure centrale, sous forme d'un delta multidimensionnel [3], a été choisie de manière à permettre le maximum de souplesse dans la représentation d'informations hétéroclites et de relations temporelles et hiérarchiques.

Chaque module regroupe des règles de transformation. Ainsi des règles du module phonologique agissent sur les paliers phonologiques pour permettre l'apparition des variantes contextuelles alors que d'autres règles du module phonétique vont intervenir pour assurer, par exemple, les différents types d'assimilation. Nous avons choisi comme langage de programmation, à l'intérieur des modules, le langage Delta [3] qui a été spécifiquement conçu pour permettre aux linguistes de formuler et de vérifier relativement facilement des théories phonologiques et phonétiques.

3. MODULES DE CONNAISSANCE ET PALIERS

Une première partie de notre travail a consisté à délimiter, d'une part, les niveaux de représentation dont nous voulions rendre compte dans le système, et d'autre part, la représentation propre à chaque niveau. Certaines hypothèses, tant sur l'organisation du composant phonologique de la grammaire que sur les contraintes particulières du système biomécanique de production de la parole, ont été déterminantes dans nos choix.

3.1 Représentation phonologique

Dans la foulée des propositions récentes en phonologie, nous supposons que la représentation phonologique est hiérarchique et multidimensionnelle (voir figure 2). Elle comporte des distinctions sur les plans prosodique, rythmique et segmental. Chacun de ces plans est contraint par des principes particuliers. Le plan prosodique règle l'organisation suprasegmentale de la chaîne phonétique [4,5,6]. Le plan rythmique rend compte des rapports de proéminence accentuelle [5,7]. Le plan segmental concerne la structure interne des segments. Dans le plan prosodique, nous distinguons cinq paliers; sur le plan rythmique nous distinguons l'accent; finalement, sur le plan segmental, nous distinguons les phonèmes avec leurs attributs classificatoires binaires.

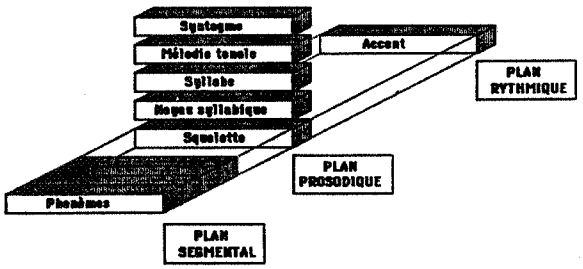


Figure 2. Paliers de la représentation phonologique

[1] Dans un premier temps, nous avons décidé de nous limiter au niveau du mot, étant donnée la quantité d'information à traiter déjà à ce niveau (syllabe, accent, ton, phonème, squelette, noyau, segment, etc.). Cependant, le système devrait dans un avenir rapproché prendre comme entrée la phrase entière. Le système se limite également à des variantes internes du français québécois et ne traite pas les anglicismes.

Les distinctions de "syllabicité" ou "d'accent" sont exclues du plan segmental, car nous en rendons compte sur les plans prosodique et rythmique respectivement. La syllabicité est décrite par une association avec le "noyau" de la syllabe, tandis que le statut d'accentué dérive d'une association avec le trait [accent] sur le plan rythmique.

Nous avons pour l'instant concentré nos efforts à la description du plan segmental, à sa structure interne et à sa relation avec la description phonétique. Chaque phonème est défini par un ensemble de traits binaires qui définissent des classes fonctionnelles naturelles [8]. Ainsi les phonèmes /p/, /v/ et /e/ partagent la même valeur du trait d'arrondissement [-rond] (figure 3), mais peuvent être distingués au moyen des traits [sonant] et [continu].

	/p/	/v/	/e/
rond	-	-	-
sonant	-	-	+
continu	-	+	+

Figure 3. Traits phonologiques de /p,v,e/

3.2 Représentation phonétique

Parallèlement à cette structure abstraite, nous avons établi une structure de représentation phonétique dont les propriétés classificatoires se rapprochent de l'organisation hiérarchique du système phonatoire. Les traits phonétiques ne sont pas nécessairement binaires; ils servent à spécifier les dimensions articulatoires majeures [9] (voir figure 4).

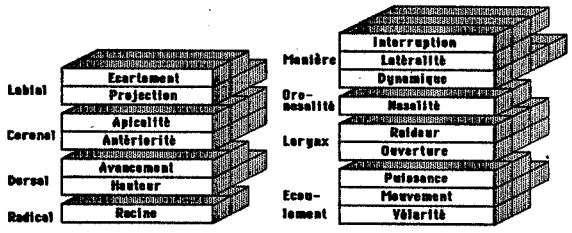


Figure 4. Paliers de la représentation phonétique

Dans ce système, l'organisation des traits reflète celle des composants du système phonatoire. Les distinctions les plus complexes ont lieu dans le résonateur buccal; on y distingue les classifications qui découlent des actions des lèvres, de l'apex, du dos et de la racine de la langue.

Ainsi les segments /p/, /v/ et /e/, qui partagent le trait [-rond] au niveau phonologique, se distinguent phonétiquement par deux paramètres de labialité: la projection et l'écartement. En ce qui concerne la projection labiale, /p/ et /e/ sont [neutre], alors que /v/ est [rétracté]; tandis que pour l'écartement, le /p/ [neutre] se distingue du /e/ [écarté]. L'absence commune du trait phonologique [rond] se traduit par des spécifications phonétiques différentes, qui se distinguent par la position des lèvres.

3.3 Paliers physiologiques

Nous disposions déjà d'un prototype de synthétiseur articulatoire [10] contrôlé par les articulateurs de la figure 5. Chaque articulateur est assigné à un palier. A partir des neuf premiers paliers, la configuration du conduit vocal est calculée à l'aide d'un modèle articulatoire [11]. Trois autres paliers (pression pulmonaire, tension et ouverture glottiques) contrôlent un modèle à deux masses des cordes vocales [12].

Plusieurs autres ensembles de paramètres articulatoires ont été proposés [13,14]. Les paramètres choisis ont l'avantage de contrôler individuellement chaque articulateur et incorporent leurs influences mutuelles dues aux contraintes physiologiques.

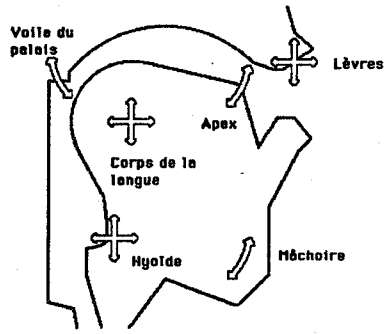


Figure 5. Paliers physiologiques

Ces paliers physiologiques permettent de réaliser une même forme du conduit vocal de plusieurs manières. Ainsi la fermeture des lèvres peut être obtenue en fermant les lèvres

elles-mêmes, ou en fermant la mâchoire sans bouger les lèvres. De même, la position de la pointe de la langue dépend de la position du corps de la langue, lui-même solidaire de la mâchoire, de sorte qu'un même lieu de constriction apicale peut être obtenu par des combinaisons différentes de ces trois articulateurs.

4. INTERFACES ENTRE LES REPRESENTATIONS

Une grande partie des règles des modules agissent sur des paliers situés au même niveau de représentation, mais certaines sont consacrées au passage d'un niveau à l'autre. Lorsque le calcul des paliers phonologiques est terminé, par exemple, des règles particulières sont activées pour obtenir une représentation phonétique préliminaire, afin que les autres règles du module phonétique puissent s'appliquer.

Ces règles d'interface ne font pas qu'une simple traduction d'un type de représentation à l'autre. En effet, lors du passage d'un niveau d'abstraction plus élevé à un moins élevé, d'un palier phonologique à un palier phonétique, par exemple, le degré de spécification doit augmenter. Comme plusieurs descriptions phonétiques différentes correspondent à la même unité phonologique, il faut ajouter de l'information aux paliers phonologiques pour remplir les paliers phonétiques.

La figure 6 représente géométriquement les divers niveaux de représentation. Chaque trait binaire du niveau phonologique, par exemple, correspond à un axe portant deux graduations. L'ensemble des axes ainsi défini sous-tend un espace phonologique. Un phonème correspond à un point de cet espace, situé sur un sommet d'un cube. La figure 6a illustre un espace phonologique sous-tendu par trois traits.

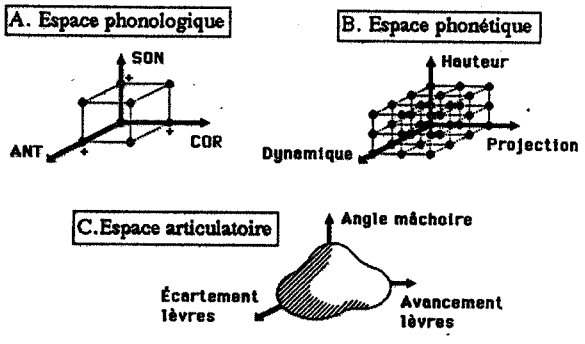


Figure 6. Géométrie des niveaux de représentation

Le niveau phonétique, qui utilise des traits à valeurs discrètes (mais pas uniquement binaires), peut être représenté de la même manière. Dans cet espace, les éléments phonétiques individuels occupent les intersections d'une grille multidimensionnelle comme illustré à la figure 6b.

Dans l'espace physiologique, obtenu lorsque chaque axe représente l'étendue de déplacement d'un articulateur, les positions articulatoires sont infiniment rapprochées, puisque les articulateurs se déplacent de manière continue. L'ensemble des articulations possibles est représenté par un volume, par exemple celui en forme de poire de la figure 6c.

Considérons un point isolé de l'espace phonologique, qui représente un phonème particulier, le /b/, par exemple. Dans l'espace phonétique, ce point unique ne correspond plus à un seul, mais plutôt à un ensemble de points (réalisations). Le point phonologique /b/ désigne, par exemple, le [b] de "bile", le [b] de "bulle", etc. La description phonologique ne spécifie pas entièrement la phonétique, mais la contraint en déterminant des régions dans l'espace phonétique. De la même façon, un point unique de l'espace phonétique détermine des volumes dans l'espace articulatoire.

Les règles internes du module phonétique (règles d'assimilation pour l'arrondissement des lèvres, par exemple) permettront de choisir une trajectoire particulière passant par les régions, donc une suite de points phonétiques particuliers.

Cette suite de points phonétiques correspond à des volumes successifs dans l'espace articulatoire. La trajectoire articulatoire retenue, parmi toutes celles qui passent par ces volumes, sera celle qui satisfait le mieux aux règles internes du niveau physiologique.

4.1 Un exemple: de la phonétique à la physiologie

Une spécification phonétique ne fait donc qu'apporter des contraintes à la représentation articulatoire, sans la spécifier complètement. L'exemple suivant illustre la forme que peuvent prendre ces contraintes.

La valeur [neutre] d'un trait phonétique tel que l'écartement correspond à une région articulatoire: on peut déterminer sa forme en faisant bouger les articulateurs du synthétiseur et en comparant les configurations obtenues avec les données radiographiques pour les voyelles neutres [15]. On obtient la région de la figure 7 pour les articulateurs "angle mâchoire" et "écartement lèvres".

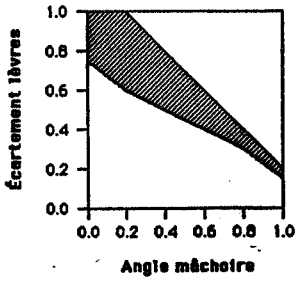


Figure 7. Région correspondant au trait écartement [neutre].

On constate que les limites sur l'écartement des lèvres changent selon la position de la mâchoire. La simple traduction du trait [écartement] en limites absolues sur l'écartement des lèvres n'est pas possible. En fait, pour tous les traits phonétiques, une traduction en limites sur la position de chaque articulateur est impensable, puisque la position d'un articulateur influence les mouvements de plusieurs autres. Les traits phonétiques doivent donc être transformés en régions complexes dans l'espace physiologique.

Lors du calcul d'une production, une fois la suite de traits phonétiques traduite en une série de régions dans l'espace physiologique, il reste à déterminer la trajectoire spécifique qui passe par ces régions. L'information nécessaire est fournie par les contraintes internes de précision, d'inertie et d'économie des mouvements des articulateurs.

5. CONCLUSION

L'élaboration d'un tel système, qui oblige à considérer ensemble théories phonologiques, phonétiques et modèles articulatoires, cristallise un des problèmes de fond en linguistique: la difficulté de raccorder ces approches dissemblables pour un faire un tout opérationnel. Nous ne prétendons pas avoir résolu entièrement le problème. Nous avons plutôt fait certains choix théoriques afin d'obtenir un système fonctionnel. Il est maintenant possible, grâce au système obtenu, de vérifier expérimentalement la validité de ces choix, de tester des hypothèses sur l'interface entre la phonologie et la phonétique, par exemple, et de faire des regroupements et des généralisations qui touchent plusieurs niveaux à la fois.

6. BIBLIOGRAPHIE

[1] Browman, C.P., L. Goldstein, J.A.S. Kelso, P. Rubin, E. Saltzman, "Articulatory synthesis from underlying dynamics", *J. Acoust. Soc. Am.*, vol. 75, 1984, pp. S22-S23.

[2] J. Allen, M.S. Hunnicutt, D.H. Klatt, *From text to speech: the MITalk system*, Cambridge, Cambridge University Press, 1987.

[3] Susan R. Hertz, "The Delta Programming Language: An Integrated Approach to Non-Linear Phonology, Phonetics, and Speech Synthesis", Phonetics Lab of Cornell University, mai 1988, à paraître.

[4] M. Nespor, I. Vogel, *Prosodic Phonology*, Dordrecht, Foris Publications, 1986.

[5] E.O. Selkirk, *Phonology and Syntax: the relation between sound and structure*, Cambridge, MIT Press, 1984.

[6] E.O. Selkirk, "On derived domains in sentence phonology", *Phonology Yearbook*, no. 3, Cambridge, Cambridge University Press, 1986.

[7] M. Liberman, A. Prince, "On stress and linguistic rhythm", *Linguistic Inquiry*, vol.8, 1977, pp.249-336.

[8] N. Chomsky, M. Halle, *The Sound Pattern of English*, New York, Harper & Row Publishers, 1968.

[9] P. Ladefoged, "The many interfaces between phonetics and phonology", *UCLA Working Papers in Phonetics*, no. 70, juillet 1988, pp.13-23.

[10] G. Boulianne, "Simulation d'un modèle articulatoire en vue de la synthèse: applications linguistiques", *Annales de l'ACFAS*, vol. 57, 1989, p.199.

[11] P. Mermelstein, "Articulatory model for the study of speech production", *J. Acoust. Soc. Am.*, vol.53, no.4, 1973, pp.1070-1082.

[12] K. Ishizaka et J.L. Flanagan, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords", *The Bell System Technical Journal*, vol.51, no.6, juillet-août 1972, pp. 1233-1268.

[13] C.H. Coker, "A Model of Articulatory Dynamics and Control", *Proceedings of the IEEE*, vol.64, no.4, avril 1976, pp.452-460.

[14] S. Maeda, "Un modèle articulatoire basé sur une étude acoustique", *Bulletin de l'Institut Phonétique de Grenoble*, vol.8, 1979, pp.35-55.

[15] R. Majid, L.J. Boë, P. Perrier, "Fonctions de sensibilité, modèle articulatoire et voyelles du français", *15ièmes JEP*, Aix-en-Provence, mai 1986, pp.59-63.

Recherche financée par le Fond de développement académique du réseau de l'Université du Québec.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

POSITIONNEMENT AUTOMATIQUE DE L'ACCENT LEXICAL EN ITALIEN EN VUE DE LA SYNTHESE

Philippe Martin

University of Toronto

Résumé

La synthèse de l'italien à partir du texte requiert le positionnement automatique de l'accent lexical pour la génération des durées vocaliques et la courbe de fréquence fondamentale. Or, l'italien est une langue à accent libre (la position de l'accent dans le mot n'est pas fixe comme en français), et il n'existe pas de règles simples permettant de déterminer la place de l'accent de mot. La solution retenue fait appel à la propriété d'accentuablilité des composantes morphologiques du mot, préfixe, lexème, morphème et flexion. Chacune de ces composantes est caractérisée par une propriété d'accentuablité (une syllabe est susceptible de recevoir un accent dans un contexte morphologique spécifique), et c'est la dernière syllabe accentuable qui determine la position de l'accent du mot. Le programme de positionnement automatique applique cette propriété en analysant le mot en ses composantes morphologiques, dont l'accentuabilité est déterminée par consultation de 5 lexiques (environ 60.000 lexèmes, 300 suffixes, 300 flexions, 100 préfixes et 100 séquences de pronoms atones). La levée d'ambiguïté accentuelle des homographes est basée sur les fréquences d'occurence de groupes de trois catégories syntaxiques successives.

1. Introduction

La synthèse de l'italien à partir du texte requiert le positionnement automatique de l'accent lexical à la fois pour la génération des durées vocaliques et de la courbe de fréquence fondamentale [5]. La difficulté vient de ce qu'il n'existe pas de règles simples comme en français pour déterminer la position de la syllabe accentuée dans le mot qui dépend de facteurs morphologiques complexes.

Ainsi, dans les substantifs, les adjectifs, les pronoms et les adverbes, l'accent peut tomber sur la dernière syllabe (*virt'ù, caff'è*), sur l'avant dernière (*acc'ento, cont'ino*), l'antépénultième (*tel'efono, c'elebre*), voire même, mais rarement, sur la quatrième avant la fin (*c'austico*). Ce n'est qu'en position finale que l'accent est transcrit dans l'orthographe.

Pour les verbes, du fait de l'adjonction de pronoms personnels, l'accent peut se placer de la première à la cinquième syllabe à partir de la fin (Certains grammairiens proposent même des cas d'accentuation de la 6ème syllabe à partir de la dernière). On peut avoir par exemple:

cas oxyton: *amer'ò* (1ère pers futur simple)
cas paroxyton: *am'are* (infinitif)
cas proparoxyton: *pr'endilo* (impératif, "prends-le")
cas 4ème/fin: *f'abbricalo* (impératif, "fabrique-le")
cas 5ème/ fin: *f'abbricamelo*(impératif, "fabrique-le-moi")

Les choses se compliquent encore par l'existence d'homographes, appartenant soit à des catégories grammaticales distinctes, soit aux mêmes catégories.

Dans *Sono cose che capitano, capitano*
("ce sont des choses qui arrivent, capitaine")

la première occurence de *capitano* est un verbe et s'accentue sur la 4ème syllabe à partir de la fin (*c'apitano*), alors que le second est un substantif et est paroxyton (*capit'ano*).

Par contre, les homographes *pr'incipi* ("princes"), *princ'ipi* ("principes"), ou encore *t'urbine* ("tourbillon") et *turb'ine* ("turbines") appartiennent aux mêmes catégories grammaticales. En l'absence de contexte, la levée d'ambiguïté ne pourra se faire que par une analyse ad-hoc de l'environnement morphologique (accords en genre, nombre, personne: *il t'urbine* vs. *le turb'ine*, pluriel de *la turbina*), ou syntaxique (type de catégorie adjacentes: séquence verbe, ponctuation, nom dans *c'apitano, capit'ano*).

2. Une solution statistique

Les applications de synthèse de la parole étant souvent élaborées par des spécialistes du traitement du signal plutôt que par des linguistes, on peut s'attendre a priori à ce que les solutions adoptées dans ce cadre pour la localisation de l'accent tournent le dos à la morphologie. C'est effectivement le cas dans le système de synthèse de l'italien du CSELT [2],[3].

Une première observation statistique porte sur le nombre de mots accentués sur l'avant dernière syllabe. Sur un corpus d'environ 8.000 mots, 78% sont accentués sur la pénultième [2]. Avec cette seule règle, on peut donc déjà obtenir un taux d'erreur dans le positionnement automatique de 22%.

L'approche retenue au CSELT se base sur les corrélations observables entre des trigrammes orthographiques (séquences de trois graphèmes) et la place de l'accent dans le mot. Ces corrélations sont utilisées sous forme de règles (environ 250) mises en oeuvre par un automate fini explorant la composition des mots à partir de la fin.

Cette solution, totalement dépourvue de données morphologiques ou mêmes phonétiques explicites (on procède directement sur le texte d'entrée), permet d'obtenir sur des textes standards un taux de positionnement correct d'environ 97%, en dehors des formes verbales. Ce taux se dégrade évidemment si le texte traité comporte des formes verbales demandant une analyse morphologique, même sommaire, ou encore s'il présente des homographes.

3. Une solution phonologique-phonétique

D'un point de vue linguistique, il peut paraître naturel de penser que la qualité accentuée d'une syllabe est liée à sa structure phonologique. Il serait alors possible d'établir des règles contextuelles déterminant le caractère accentué des syllabes selon leur composition phonétique et/ou phonologique.

Cette approche a été proposée par Delmonte [1], et implantée dans un système de synthèse. Le problème de ce type de solution vient du grand nombre de règles (et du grand nombre d'exceptions) nécessaires à l'obtention d'un taux de performances satisfaisant.

4. Une approche morpho-phonétique

Dans Profili [4], les règles phonétiques sont assorties de règles morphologiques portant sur l'accentuabilité ou l'inaccentuabilité des suffixes (considérés dans leur forme phonologique). Ainsi, les suffixes *-'illa,'-esse* sont toujours accentuables (*dist'illo, profet'essa*), alors que les suffixes *ido, -bile* sont inaccentuables (*t'imido, sens'ibile*).

Là aussi il faut adjoindre des listes d'exceptions pour réaliser une localisation correcte.

5. Une solution morphologique

On conçoit que la construction et la maintenance d'un système basé sur un ensemble de règles (essentiellement phonétiques dans [2],) est à la fois difficile et délicate [7]. L'inconvénient majeur réside dans leur interdépendance, la modification de l'une d'elles entrainant le plus souvent la révision des autres et de leurs listes d'exceptions. De plus, les performances de ce type de système est limité.

L'augmentation exponentielle des capacités de mémoire des machines informatiques permet d'envisager une solution de "force brute", consistant à mettre en mémoire toutes (ou en tous cas un très grand nombre) les formes fléchies, ainsi que, outre la place de l'accent, l'indication de la catégorie grammaticale.

Ce genre d'approche, bien qu'elle ait été effectivement utilisée pour d'autres langues, en particulier pour le français, apparaît cependant comme trop lourde. Aux quelques 50.000 formes de base [10], s'ajouterait, pour chaque forme, 3 entrées supplémentaires pour les flexions de l'adjectif, 50 formes fléchies pour chaque verbe, etc.

Entre l'emploi d'un ensemble de règles, relativement compact mais difficile à modifier, et celui d'un lexique pouvant dépasser 250.000 entrées, on proposera donc une solution différente utilisant une propriété, du reste peu utilisée, de l'accent lexical en italien. La méthode procède par analyse morphologique de chaque mot, en accédant à des lexiques séparés de racines (lexèmes) et de morphèmes (préfixes, suffixes, flexions, pronoms atones).

Cette méthode se base sur le mécanisme d'accentuation en italien, plutôt que sur propriétés accentuelles des syllabes ou des suffixes. Ce mécanisme (cf. [12], [13], considère le mot décomposé en ses éléments morphologiques, un lexème, suivi d'un ou plusieurs suffixes et d'une flexion et éventuellement précédé de préfixe(s).

Les suffixes et flexions peuvent être soit accentuables, soit inaccentuables. Ainsi, comme on l'a vu plus haut, *-'ill-* est un suffixe accentuable, et *-id-* est inaccentuable. (par accentuable, on entend susceptible de recevoir l'accent, mais non nécessairement accentué).

Selon leur nature, c'est-à-dire le type de lexème qu'ils déterminent, des suffixes homographes pourront se révéler soit accentuables, soit inaccentuables. Par exemple, *-in-*, suffixe diminutif suivi de la flexion inaccentuable *-o* du masculin singulier, est accentuable (*piccol'ino*), mais *-ino*, flexion 3ème personne pluriel du subjonctif est inaccentuable (*'amino*). De même, *-o* inaccentuable est la flexion verbale de la 1ère personne du présent de l'indicatif, alors que *-'o* accentuable est la flexion verbale de la 3ème personne du singulier du passé simple.

Les lexèmes sont toujours accentuables, soit sur la dernière (*ar'en-* , dans *ar'ena*), soit sur l'avant dernière syllabe (*f'abbric-*, dans *f'abbrica*). D'après [13], 82% des lexèmes sont accentuables sur la dernière syllabe, et seulement 18% le sont sur l'avant dernière.

Le principe d'accentuation spécifie alors simplement que, dans une séquence

(préfixe)+ (lexème) + (suffixe) +...+ (suffixe) + (flexion)

dont un ou plusieurs éléments sont accentuables, c'est le dernier élément, lexème ou morphème, qui portera l'accent du mot.

Ainsi, les dérivés de *'oper-*, accentuable sur la pénultième, sont accentués comme suit:

'opera : 'oper + -a (flexion inaccentuable) "opéra"
oper'oso : 'oper + -'os- (suffixe accentuable) + -o (flexion inaccentuable) "travailleur"
oper'etta : 'oper + -'ett- (suffixe accentuable) + -a (flexion inaccentuable) "operette"
operosit'à : 'oper + -'os- + -it'a (suffixe accentuable) "le fait d'être travailleur"

De même, on a:

turb'ina : turb'in + -a "turbine"
t'urbine : t'urbin + -e "tourbillon"

la position de l'accent dans le lexème permet de différentier les deux dérivés d'un même mot latin.

Le mot peut également s'analyser en deux lexèmes, interdépendants ou indépendants l'un par rapport à l'autre. Dans le premier cas, c'est-à-dire si aucun des deux lexèmes ne peut apparaître seul, c'est le premier lexème qui attirera l'accent [18]:

tel'efono : tel'e + f'ono "téléphone"
s'incrono : 'sin + cr'ono "synchrone"

Si le dernier ou si les deux lexèmes sont indépendants, la règle de dominance du dernier élément s'applique, et c'est le dernier élément accentuable (lexème ou suffixe) qui déterminera la position de l'accent:

aeron'ave : a'ero + n'av + -e "aéronef"
alisc'afo : 'ali + sc'af + -o "hydroglisseur"

Le système traite ces mots composés comme lexèmes à part entière.

Le problème de localisation de l'accent qui, sur des formes complètes, peut se placer entre la dernière et la 5ème syllabe à partir de la fin, peut donc être traité par une analyse du mot en ses composantes morphologiques, et par l'identification des syllabes accentuables des éléments dégagés par cette analyse.

La consultation d'un lexique permet alors de déterminer la position de la syllabe accentuable du ou des lexèmes identifiés par l'analyse, ainsi que celle des suffixes accentuables.

Cette identification demande, comme on l'a vu, d'appareiller les catégories de suffixes aux catégories de lexèmes.

Schématiquement, le processus est basé sur la constitution de plusieurs lexiques:

- un lexique de lexèmes avec indication de la catégorie syntaxique et de la position de la syllabe accentuable;

- un lexique (inverse) de flexions verbales, nominales et adjectivales avec indication de la syllabe accentuable.

- un lexique de pronoms atones pour l'analyse des formes verbales 'accouplées';

- un lexique de suffixe avec indication de la syllabe accentuable pour l'aide à la saisie des lexèmes;

Dans la phase d'analyse, la consultation du premier lexique permet la construction d'une liste d'hypothèses d'analyse. Dans le cas général, plusieurs racines différentes pourront être associées au mot analysé:

capitano: capit'an- (substantif) reliquat -o

c'apit- (verbe) reliquat -ano

L'identification des reliquats résultant de la première opération se fait par la consultation du lexique de flexions:

-o : flexion nominale du masculin singulier, inaccentuable

-a-: flexion verbale du pluriel, inaccentuable

-no: flexion verbale (3ème pers.), inaccentuable

L'application de la règle de dominance du dernier élément accentuable détermine ensuite la place de l'accent dans le mot analysé.

Les deux analyses possibles de l'exemple sont donc:

capit'an-o :substantif ("capitaine")
c'apit-an-o :forme verbale ("se passent", "arrivent")

Il reste alors, par examen du contexte, de lever l'ambiguïté dans le cadre de la phrase. Dans

Sono cose che c'apitano, capit'ano

("Ce sont des choses qui arrivent, capitaine")

l'existence d'un signe de ponctuation immédiatement avant et après le deuxième *capitano* permet d'en établir le caractère de substantif.

Lorsqu' aucun lexème ne peut être trouvé dans le lexique, le programme procède à une analyse basée sur l'identification des seuls suffixes, flexions et pronoms atones pour déterminer la place de l'accent. Si cette procédure échoue, c'est la syllabe pénultième qui reçoit l'accent.

6. Implantation

Le processus de localisation de l'accent par analyse morphologique a demandé l'élaboration

- d'un lexique de lexèmes (60.000 entrées environ);

- d'un lexique de suffixes (300 formes environ);

- d'un lexique de flexions des noms, adjextifs, verbes (600 formes environ);

- d'un lexique de pronoms atones entrant dans les formes verbales (120 formes environ); Les formes irrégulières, et particulièrement les formes complètes de quelque 200 verbes, sont incluses dans le lexique de lexème.

- d'un logiciel de saisie rapide des lexiques, d'analyse morphologique, et d'analyse syntaxique contextuelle. Ce logiciel est de plus capable de générer toutes les formes verbales, nominales et adjectivales régulières d'un lexème donné.

La mise en oeuvre finale du système de positionnement implique, outre la consultation et l'analyse morphologique, l'application de règles de désambiguation basée sur l'occurence de chaines syntaxiques et de concordances de genre, personne et nombre, ainsi que la mise en place de règles de collision d'accent [6],[11]. Ce

processus utilise entre autre un tableau de groupes de 1, 2 et 3 catégories syntaxiques ordonnées par ordre de probabilité d'apparition décroissant.

Outre la position accentuelle, le programme génère un étiquetage morpho-syntaxique pour chaque mot, qui peut être utilisé pour une analyse syntaxique de la phrase.

7. Résultats

EXEMPLE: Traitement du fichier de test CNET

Le fichier cnet.txt contient 29 phrases test utilisées pour l'évaluation des systèmes de postionnement automatique de l'accent. Il présente un certain nombre d'homographes à accentuation différentes, dans certains appartiennent à la même catégorie syntaxique.

1. *I ragazzi studiano il periodo contemporaneo.*
2. *La ventesima pantomima era fallita.*
3. *Era, tra parentesi, una cosa ardita.*
4. *Sono cose che capitano, capitano.*
5. *Ludovico non era un bravo politico.*
6. *Sono giochi da villano questi.*
7. *I vulcani non esplodono spesso.*
8. *Soffrivano di sinusite o di elefantiasi?*
9. *Il sagrestano super'o il limite indicato dal semaforo.*
10. *Curati, se non andrai a finire male.*
11. *La lezione è stata capita da tutti.*
12. *Capita sempre per l'ora di pranzo.*
13. *Ci siamo curati la salute.*
14. *Nell'ambito più ambito della fonetica, il professor Pontormo è un capo.*
15. *Il contino si è permesso un sorriso.*
16. *Finchè li contino tutti, noi riusciremo a scappare.*
17. *Ci pigliamo un tramezzino all'uscita.*
18. *Era un vero carnivoro sin dalla nascita.*
19. *E in vendita l'appartamento di Giacomo a Rimini.*
20. *Paolo è sempre molto caustico.*
21. *Hanno detto che lo spedivano allo scrivano.*
22. *Troverai il prezzo della medicina in pagine tre.*
23. *La polizza scade domani.*
24. *I principi sono spesso senza scrupoli e senza principi.*
25. *La chimica non piaceva alla tua amica, vero?*
26. *La tua visita ci è sempre gradita.*
27. *Il prestito e il credito sono operazioni di banca.*
28. *Bisognerebbe partire immediatamente.*
29. *Ma quando non ti dimentichi di fare il bucato?*

Le temps de calcul est d'environ 10 sec sur une machine opérant à 33 MHz (soit 350 ms par phrase en moyenne).

Ce fichier présente de nombreux pièges de positionnement de l'accent, et en particulier des homographes distingués par l'accentuation:

1. *per'iodo vs peri'odo*
2. *c'apitano vs capit'ano*
3. *c'urati vs cur'ati*
4. *c'apita vs cap'ita*
5. *'ambito vs amb'ito*
6. *c'ontino vs con'tino*
7. *tram'ezzino vs tramezz'ino*
8. *scr'ivano vs scriv'ano*
9. *pr'incipi vs princ'ipi*

Le programme de désambiguation syntaxique a pu sélectionner la forme correcte toutes les fois que les catégories des homographes étaient différentes (cas 1, 2, 6, 7, 8). Par contre, une forme incorrecte a été retenue dans les cas où les homographes appartiennent à la même catégorie syntaxique (cas 3, 4, 5, et 9). Seule une analyse syntaxique complète de la phrase permettrait alors de lever l'ambiguité (des solutions ad-hoc pourraient convenir dans des exemples très particuliers, mais s'avéreraient inopérants dans le cas général).

(Ce travail a fait l'objet d'un contrat CNET No 88 1B 108)

Références

[1] R. DELMONTE (1981) "L'accento di parola nella prosodia dell'enunciato dell'italiano standard", Studi di Grammatica Italiana", Vol. X, 351-394.

[2] S. SANDRI et E. VIVALDA (1981) "Automatic Stress Assignement for Italian Text-to-Speech Synthesis", CSELT Rapporti Tecnici, Vol. VIII, No 3, juin 1981, 213-216.

[3] S. QUAZZA et E. VIVALDA (1987) "Contextual Syntactic Analysis for Text-to-Speech Conversion", Proc. ICASSP 87, 389-392.

[4] O. PROFILI (1987) "L'accent et sa prévisibilité", Rapport Syntalit/ Italien, CNET-Lannion.

[5] Ph. MARTIN (1978) "L'intonation de la phrase en italien", Studi di Grammatica Italiana, VIII, Acc. della Crusca, Firenze, 395-417.

[6] M. NESPOR et I. VOGEL (1979) "Clash avoidance in Italian", Linguistic Inquiry, X, 3, 467-482.

[7] E. LAPORTE (1986) "Application de la morpho-phonologie à la production automatique de textes phonétiques", Actes GALF Symposium Lexique, 215-227.

[8] G. MALAGOLI (1946) "L'accentazione italiana", Sansoni, Firenze.

[9] U. BORTOLINI, C. TAGLIAVINI et A. ZAMPOLLI (1971) "Lessico di frequenza della lingua italiana contemporanea", Garzanti, Milano.

[10] A. BATINTI et W. TRENTA (1982) "Ricerche sul lessico di base dell'italiano contemporaneo", Guerra, Perugia.

[11] Ph. MARTIN et O. PROFILI (1987) "Accent de mot et structure syntaxique en italien", "Information-Communication", U. of Toronto, 15-26.

[12] P. GARDE (1968) "L'accent", PUF, Paris.

[13] P. ANTONETTI et M. ROSSI (1970) "Précis de Phonétique Italienne, Synchronie et Diachronie", Aix-en-Provence, 356p.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

OPTIMISATION D'UN ALGORITHME DE SYNTHESE DE PAROLE POUR SON IMPLANTATION TEMPS-REEL

L.Le Faucheur, E. Moulines

CNET Lannion, route de Trégastel, Dpt TSS/RCP

ABSTRACT

This paper describes an algorithm that have been proposed recently for speech synthesis using diphones. The algorithm is based on a pitch-synchronous overlap-add (PSOLA) approach for modifying the speech prosody and concatenating diphones waveforms. In order to have a cost-effective real-time implementation, the diphone data-base is compressed by using a variable rate multi-pulse LPC coder. We give the theoretical basis of the time-domain PSOLA technique and we describe some of the constraints of the implementation.

INTRODUCTION

Les systèmes de synthèse de parole à partir du texte commercialisés aujourd'hui sont intelligibles, mais la qualité de la parole produite est encore relativement limitée. Chaque étape du traitement - analyse linguistique, génération des contours prosodiques... - influence la qualité globale de la parole de synthèse. Nous concentrons notre attention dans cet article sur la dernière étape du traitement à savoir la production du signal de parole elle-même.

Dans le système que nous décrivons, la parole de synthèse est obtenue par concaténation d'unités acoustiques [Emerard,77]. Le répertoire dont nous disposons pour le français comporte 1500 unités, dont 1200 diphones et environ 300 diphones en contexte qui tiennent compte de variantes allophoniques de certains phonèmes (principalement les consonnes liquides, certaines nasales et les semi-voyelles). Un tel répertoire correspond à environ 3 minutes de parole, et le volume de stockage requis, en l'absence de codage est de plus de 5 Moctets [1]. Les langues possédant des voyelles accentuées peuvent nécessiter des inventaires d'unités acoustiques encore plus vastes: en anglais, par exemple, le répertoire d'unités acoustiques représente plus de 5 minutes de parole... soit plus de 10 Moctets dans les conditions décrites ci-dessus. Si dans un futur plus ou moins lointain de tels volumes pourront être aisément manipulés, ils sont aujourd'hui rédhibitoires pour la plupart des implantations. Le codage à bas-débit du répertoire d'unités acoustiques est donc, dans la plupart des cas, un pré-requis.

En tenant compte de cette contrainte, il apparaît qu'un algorithme de synthèse par concaténation d'unités doit effectuer 3 types d'opérations:

- une opération de décodage: le signal de parole constituant les unités acoustiques est reconstruit à partir des informations codées à bas débit;

- une opération d'ajustement des paramètres prosodiques: la prosodie intrinsèque des unités est corrigée de façon à reproduire la courbe intonative et les durées phonémiques spécifiées par le module de génération de la prosodie;

- une opération de concaténation: la suite des unités acoustiques est concaténée après l'ajustement de leur prosodie intrinsèque.

Les vocodeurs paramétriques à bas-débit (vocodeur à prédiction linéaire ou vocodeur à formants) sont particulièrement bien adaptés pour résoudre ce type de problème. Ils offrent la possibilité de réduire de façon importante le débit binaire (moins de 800 bits/sec sont parfois suffisants pour produire un signal de parole intelligible) et autorisent d'autre part des manipulations aisées des paramètres prosodiques: ces vocodeurs, qui sont la réplique de modèle de production simplifiés, utilisent une décomposition source-filtre et un signal d'excitation commandé par un nombre restreint de paramètres, dont la fréquence fondamentale elle-même, qui apparaît ici comme un paramètre explicite de synthèse.

---

1 à 16 kHz de fréquence d'échantillonnage, 16 éléments binaires par échantillon

Ces avantages multiples sont contre-balancés par le fait que, malgré de nombreux efforts pour améliorer la qualité de la déconvolution source-filtre ou pour utiliser des modèles plus réalistes de signaux d'excitation [Hedelin,86] [Bimbot,88], la parole produite par un vocodeur paramétrique souffre généralement d'altérations caractéristiques [2] : timbre métallique, mauvaise restitution de certains sons comme les constrictives ou encore les fricatives voisées [Klatt,87].

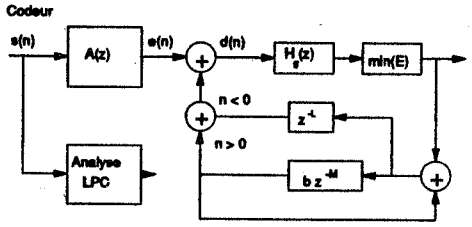
Il est donc apparu nécessaire d'utiliser des représentations plus complexes du signal vocal et de définir, sur celles-ci, des méthodes de modification des paramètres prosodiques. Des variantes du codage par harmonique utilisant une distribution du voisement en fréquence ou encore des méthodes de décomposition du signal de parole en formes d'ondes élémentaire sont évidemment des candidats sérieux pour réaliser ce type de tâche, car elles possèdent le double avantage de permettre une réduction sensible de la quantité d'informations à stocker pour synthétiser une parole de bonne qualité et de disposer de la fréquence fondamentale comme paramètre plus ou moins explicite du modèle [Rodet & al,87] [D'allessandro & Liénard 88].

L'approche que nous avons utilisée est plus indirecte. Elle combine un vocodeur prédictif à bas-débit à excitation multi-impulsionnelle et un algorithme de modification des paramètres prosodiques par addition-recouvrement de signaux élémentaires dans le domaine temporel (TD-PSOLA, ou Time-Domain Pitch Synchronous Overlap Add) [Moulines & Charpentier,88] [Hamon & al,89] [Charpentier & Moulines, 89] [Moulines,90]. Nous présentons dans une première partie l'algorithme de codage à bas-débit que nous avons étudié. Il peut être utilisé aussi bien en bande téléphonique qu'en bande élargie, et permet une réduction significative du débit binaire du signal (30 kbits / sec à 16 kHz de fréquence d'échantillonnage, 16 kbits / sec à 8 kHz) au prix d'une diminution minime de la qualité de synthèse. Nous décrirons dans une seconde partie succinctement l'algorithme TD-PSOLA en insistant sur son interprétation fréquentielle: celle-ci guidera nos choix pour l'implantation.

1 CODAGE PREDICTIF DE LA BASE DE DONNEES D'UNITES ACOUSTIQUES

L'essentiel des recherches en codage de parole portent aujourd'hui sur la réduction du débit binaire pour un niveau de qualité légèrement inférieur à la qualité téléphonique: le signal décodé est intelligible tout en présentant des distorsions audibles. Notre objectif a été, à l'opposé, de repartir des structures du codage bas-débit, de trouver des solutions permettant d'en améliorer la qualité et de les adapter à une utilisation sur des signaux en "bande élargie" (8 kHz de bande), réputés plus difficiles à coder du fait de leur grande dynamique spectrale et de leur sensibilité aux effets de souffle. Le codeur que nous avons étudié utilise une excitation de type multi-impulsionnelle. Une version à excitation par codes multi-impulsionnels et à débit variable adapté aux segments de parole à coder a été également étudiée [Moulines,90] [White & al,90].

La structure du codeur est illustrée dans la figure 1. L'excitation optimale est obtenue par analyse-synthèse [Atal & Remde,82]. Dans cette structure, deux signaux de prédiction sont utilisés.



On calcule tout d'abord le signal résidu de prédiction à court-terme $e(n)$ en filtrant le signal original $s(n)$ par le prédicteur à court-terme $A(z)$. Ce prédicteur est estimé au moyen d'un algorithme de covariance pondérée sur des fenêtres de 20 msec [Shingal & Atal,84]. La largeur des bandes passantes des résonnances est contrainte, suivant un seuil variable en fréquence, de façon à éviter l'obtention de filtres sur-tendus. La cadence de renouvellement de ce prédicteur est variable: elle est synchrone de la fréquence fondamentale sur les segments voisés et dépend de la non-stationnarité du signal sur les segments non-voisés et les transitions. La quantification de ce prédicteur utilise une méthode "Split Vector Code" (quantification de vecteurs éclatés), qui consiste à diviser le vecteur de paramètres représentant les coefficients prédicteur en un certain nombre de sous-vecteurs et à quantifier ces différents sous-vecteurs au moyen de quantificateurs vectoriels. On utilise comme paramètres pour la quantification les logarithmes des rapports d'aire et comme critère de quantification la distance euclidienne associée.

Le second signal d'erreur de prédiction est obtenu en soustrayant de $e(n)$ le signal $\hat{e}(n)$ égal au résiduel reconstruit $\bar{e}(n)$ pour $n<0$ et à la prédiction à long-terme du signal reconstruit pour $n=0,\ldots,L-1$ (L étant la longueur de la trame courante):

$$d(n)=e(n)-\hat{e}(n) \quad \hat{e}(n)=\begin{cases}\bar{e}(n) & n<0\\ b\bar{e}(n-M) & n=0..L-1\end{cases}$$

Pour estimer les amplitudes et les positions des impulsions $u(n)=\sum a(i)\delta(n-n_i)$, on minimise l'énergie du signal d'erreur de reconstruction filtré: $h_g(n)*(d(n)-u(n))$

$$\min_{\{a_i,n_i\}}\left(\sum_{n=0}^{L-1}\left(d(n)*h_g(n)-\sum_{i=0}^{P-1}a_i h_g(n-n_i)\right)^2\right)$$

2 Cette remarque s'applique uniquement pour des systèmes d'analyse-synthèse totalement automatiques. Il est en effet connu que certains synthétiseurs à formants, lorsque les paramètres sont corrigés manuellement par des experts, permettent d'obtenir une qualité sonore presque indiscernable du naturel.

$h_g(n)$ est la réponse impulsionnelle du filtre de pondération $1/A(\gamma(z)z)$. Le facteur γ, qui contrôle la bande passante des résonnances du filtre de pondération et donc l'atténuation de l'erreur dans les zones formantiques, est usuellement fixe. Dans cette implantation, nous avons choisi de le rendre variable en fréquence. Cette particularité permet de prendre en compte l'évolution de la largeur des bandes critiques, qui conditionne la largeur des bandes de fréquence "masquées" par une proéminence spectrale. Ce choix répartir de façon plus optimale l'erreur en fréquence, et s'avère particulièrement efficace pour le codage en bande élargie.

Le délai paramètres du prédicteur à long-terme sont obtenus dans la boucle d'analyse-synthèse, en minimisant l'erreur de reconstruction pondéré a priori c'est-à-dire en minimisant:

$$\min_{(b,M)}\left(\sum_{n=0}^{L-1}((e(n)-b\hat{e}(n-M))*h_g(n))^2\right)$$

Le gain du prédicteur à long-terme est réajusté après l'estimation de la séquence d'impulsions. Un algorithme séquentiel rapide, utilisant une récursion en temps et prenant en compte les valeurs de délai inférieurs à la longueur de la trame, a été développé pour accélérer le calcul de ces paramètres.

2 ALGORITHME TD-PSOLA DE MODIFICATION DES PARAMETRES PROSODIQUES

Pour modifier les paramètres prosodiques, nous avons mis en oeuvre l'algorithme TD-PSOLA, qui procède par addition/recouvrement dans le domaine temporel de signaux à court-terme extraits de façon synchrone de la fréquence fondamentale [Hamon & al, 89] [Charpentier & Moulines, 89] [Moulines,90].

Le signal de parole numérisé $x(m)$ est décomposé en une suite temporelle de signaux à court-terme $x_n(m)$ obtenus en multipliant le signal par des fenêtres d'analyse glissantes, centrées autour de l'origine et translatés de t_n échantillons depuis leur position initiale.

$$x(t_n,m)=x_n(m)=h_n(t_n-m)x(m)$$

Les instants successifs t_n sont apposés de façon synchrone de la fréquence fondamentale sur les segments voisés du signal de parole, et à une cadence uniforme sur les segments non-voisés. Cette opération d'analyse est à rapprocher des transformations de type Transformée de Fourier à Court-Terme, qui ont pour but de projeter un signal mono-dimensionnel $x(n)$ dans un espace bi-dimensionnel (généralement temps-fréquence). Le flux des signaux élémentaires de synthèse sert à produire un flux de signaux élémentaires modifiés $\hat{x}_q(m)$ qui sont synchronisés sur une nouvelle suite d'instants t_q.

L'algorithme détermine en fonction des facteurs de modification de pitch et de durée les différents instants t_q et la fonction $t_q \rightarrow t_n$ dont le rôle est de préciser pour chaque signal élémentaire de synthèse $\hat{x}_q(m)$ le signal élémentaire d'analyse $x_n(m)$ qui doit être utilisé pour l'engendrer. La suite des instants t_q définit les délais qui doivent être appliqués entre les signaux élémentaires de synthèse. Le traitement consiste donc à sélectionner certains signaux élémentaires d'analyse $x_n(m)$ et à les translater selon la suite des délais $\delta_q = t_q - t_n$:

$$\hat{x}(t_q,m)=\hat{x}_q(m)=x(m-\delta_q)=x_m(m+t_n-t_q)$$

On obtient enfin le signal de synthèse $\hat{x}(n)$ en superposant et additionnant le flux des signaux de synthèse, éventuellement multipliés par un facteur de normalisation énergétique α_q, en utilisant soit une formule de synthèse de type WOLA, soit une technique de type LS-OLA, plus correcte d'un point de vue théorique:

$$\hat{x}(m)=\frac{\sum_q \alpha_q \hat{x}_q(m)h_q(t_q-m)}{\sum_q h_q^2(t_q-m)}$$

Le terme apparaissant au dénominateur est une fonction temporelle de pondération prenant en compte la cadence non-uniforme du traitement. Les fenêtres d'analyse étant choisies de façon à admettre un certain recouvrement lors de l'étape de synthèse, ce terme sera non-nul. Cette étape de normalisation peut être évitée en pratique en faisant un choix judicieux des fenêtres d'analyse (quitte à s'affranchir de la symétrie) en fonction des facteurs de modification de la fréquence fondamentale et de la durée, de façon à respecter:

$$\sum_q h_q^2(t_q-m)=1 \quad \forall m$$

3 INTERPRETATION FREQUENTIELLE DES MODIFICATIONS TD-PSOLA

Nous allons, dans ce paragraphe, donner un fondement théorique à cette opération de synthèse et spécifier les conditions d'analyse permettant une qualité optimale de synthèse. Pour ne pas alourdir le traitement, nous utiliserons pour les développements théoriques un modèle de son voisé stationnaire, superposition d'un signal déterministe périodique et d'un signal aléatoire faiblement stationnaire de puissance moyenne finie. La composante déterministe périodique rend compte de la composante harmonique du signal de parole qui doit être affectée par l'algorithme de modification de la fréquence fondamentale. La composante aléatoire modélise les irrégularités que l'on observe dans les sons de parole d'une période à l'autre, dues, par exemple, aux fluctuations du mouvement des cordes vocales ou encore aux turbulences du flux d'air passant à travers la glotte durant la phase d'ouverture. Pour obtenir une qualité optimale, cette composante ne devrait pas être affectée de façon significative par les modifications TD-PSOLA. Nous intéressons tout

d'abord aux modifications de la composante déterministe de ce signal, l'impact de ces algorithmes sur la partie aléatoire étant traité dans le paragraphe suivant [3]. Dans les deux cas, nous formulons les hypothèses suivantes:

(1) la composante déterministe du signal est périodique de période P. Les instants d'analyse successifs t_n sont disposés de façon synchrone de la fréquence fondamentale et l'origine des temps est choisie de façon à ce que: $t_n = nP$,

(2) le facteur de modification est supposé constant, égal à β, et les instants de synthèse se déduisent des instants d'analyse par la relation: $t_n = n\beta P$

(3) les fenêtres d'analyse $h_n^2(m)$

sont identiques égales à $h(m)$ et vérifient la condition de normalisation. Nous supposons également que le facteur de normalisation énergétique est constant: $\alpha_n = 1$.

3.1 modifications de la partie déterministe

En nous appuyant sur l'ensemble des hypothèses formulées ci-dessus nous démontrons aisément que le signal obtenu par synthèse TD-PSOLA est égal à:

$$y(m) = \sum_s h(s\beta P - m)x(m - s\beta P)$$

$$= \sum_s w(s\beta P, m) \quad w(n,m) = h(n-m)x(m-n)$$

La fenêtre d'analyse étant de durée finie, le signal bi-dimensionnel est absolument sommable et admet des transformées de Fourier partielle par rapport à chaque variable de temps, que nous notons respectivement $W_1(\omega, m)$ et $W_2(n, \psi)$. La formule sommatoire de Poisson, qui relie la somme d'un signal sur un réseau périodique de points à la somme des coefficients de sa transformée de Fourier sur le réseau réciproque permet d'affirmer que :

$$y(m) = \sum_s w(s\beta P, m) = \frac{1}{\beta P} \sum_{k=0}^{\beta P - 1} W_1(2\pi k/\beta P, m)$$

Par construction, la valeur du signal $w(n,m)$ au point (n,m) est égale à: $w(n,m) = w(0, m-n)$. Cette propriété sur la translation de l'origine des temps permet de relier simplement les valeurs des transformées de Fourier partielles pour différentes valeurs de l'index temporel m: $W_1^*(\omega, m) = W_2(0,\omega)\exp(j\omega m)$. En substituant cette relation dans la formule de Poisson, il vient:

$$y(m) = \frac{1}{\beta P} \sum_{k=0}^{\beta P - 1} W_2(0, 2\pi k/\beta P)\exp(j2\pi k m/\beta P)$$

Le signal de synthèse est donc périodique de période βP, et l'expression ci-dessus donne son développement en série de Fourier. Cette expression permet, de plus, de relier les amplitudes complexes des harmoniques de synthèse aux valeurs de la transformée de Fourier du signal à court-terme d'analyse $h(-n)x(n)$, prélevées à la fréquence des harmoniques de synthèse. Ce résultat a deux conséquences pratiques importantes:

(1) la résolution spectrale de la fenêtre d'analyse influe directement sur le spectre du signal de synthèse et il est donc nécessaire d'utiliser une fenêtre d'analyse possédant de bonnes caractéristiques spectrales (type Hamming, Blackman..) [Harris,78]. De plus, il apparaît que le comportement de l'algorithme se modifie en fonction du rapport entre la résolution de la fenêtre $h(m)$ et l'espacement des harmoniques du signal de synthèse [Moulines & al, 89]. On peut, à cet égard, démontrer qu'il est en général plus favorable de se placer dans des conditions d'analyse en "large bande", correspondant au cas où la résolution fréquentielle de la fenêtre est telle qu'elle ne résout pas les harmoniques du signal d'analyse. Pour les fenêtres usuelles, cette condition est satisfaite lorsque la longueur de la fenêtre n'excède pas deux fois la période fondamentale locale.

(2) la position du centre de la fenêtre d'analyse par rapport au signal a une répercussion sur le signal de synthèse. Plus précisément, dans des conditions d'analyse "large bande" [4] nous avons démontré qu'il est optimal de synchroniser le centre de la fenêtre d'analyse avec l'instant d'excitation maximale du conduit vocal, à savoir l'instant de fermeture de la glotte [Moulines,90]. Cette synchronisation permet en effet de préserver les relations d'amplitudes entre les différents maxima du spectre du signal d'analyse et, dans une certaine mesure, la distribution de phase des harmoniques.

3.2 modifications de la partie stochastique

Le modèle de son voisé stationnaire que nous avons défini comporte une autre composante, à savoir un signal aléatoire stationnaire de puissance moyenne finie. Ce s.a. rend compte des fluctuations du signal d'une période à l'autre (qui peuvent dominer le spectre lors de la production de certains sons comme les fricatives voisées). Pour ne pas introduire d'artefacts, il est souhaitable de na pas trop altérer, lors d'une modification TD-PSOLA de la fréquence fondamentale, la densité spectrale de puissance de ce signal.

Nous étudions dans la suite les modifications d'un s.a. $x(m)$,De moyenne nulle et de densité spectrale de puissance $S_x(\tau)$. Sous les hypothèses que nous avons formulé ci-dessus, nous pouvons démontrer que la densité spectrale de puissance du signal synthétique après modification TD-PSOLA de la fréquence fondamentale est égale à [Moulines,90]:

3 L'ensemble des résultats que nous démontrons peuvent être généralisés sans difficulté aux signaux quasi-périodiques et quasi-stationnaires [Moulines,90].

4 la fenêtre est ici supposée symétrique

$$S_y(\omega) = \frac{1}{\beta P^2} \sum_{k=0}^{P-1} |H(\omega(1-\beta) + 2\pi k/P)|^2 S_x(\beta\omega - 2\pi k/P)$$

où $H(\omega)$ est la transformée de Fourier de la fenêtre d'analyse. Cette expression met en évidence le rôle du spectre de la fenêtre d'analyse. Elle intervient par le carré du module de sa transformée de Fourier, à la manière de la fonction de transfert du filtre dans le cas d'un s.a. Filtré. Notons toutefois, qu'à la différence d'un filtrage linéaire, la valeur de la D.S.P du signal de synthèse à une fréquence particulière ω ne dépend pas uniquement de la valeur de la D.S.P du signal d'analyse à cette même fréquence, ce qui souligne le caractère non linéaire de l'opération de transformation TD-PSOLA.

Comme dans le cas déterministe, le comportement de l'algorithme change en fonction de la résolution du filtre d'analyse vis-à-vis de la cadence $2\pi/P$ du traitement (qui est la fréquence du signal déterministe sous-jacent que l'on modifie). Dans des conditions d'analyse à "bande étroite" (fréquence de coupure du filtre $\omega_h \leq \pi/P$), au plus un seul terme de la somme définissant la D.S.P du signal de synthèse est simultanément non-nul. La D.S.P du signal de synthèse s'annule sur un réseau périodique de points, créant ainsi une structure pseudo-harmonique. La pseudo-période associée dépend à la fois de la période P et du facteur de modification β: $P_s = |1-\beta|P$. Cette propriété se manifeste en pratique par l'apparition, dans certaines circonstances, de bruits tonaux.

Dans le cas d'analyse "large bande", l'influence du filtre d'analyse se manifeste par l'introduction d'un étalement spectral. De façon un peu inhabituelle, ce n'est pas ici un produit de convolution qui intervient, mais une sommation discrète. Pour une fenêtre "classique", le nombre de termes non-nuls dépend de la longueur de la fenêtre: 3 termes pour une longueur égale à deux fois la période ($2P$), 4 pour une fenêtre de longueur $3/2P$. Si la fréquence fondamentale du signal qu'on modifie est égale à f_0 Hz, ceci signifie que des termes distants de f_0 voire $2f_0$ Hz vont intervenir dans le calcul de la D.S.P du signal synthétique à une fréquence particulière. Comme dans le cas du signal déterministe, il est clair que cet étalement est d'autant plus préjudiciable que le signal d'entrée a une dynamique spectrale importante, les zones les plus énergétiques venant masquer les autres détails du spectre.

4 CONTRAINTES D'IMPLANTATION

Les résultats que nous avons rappelés définissent un ensemble de conditions à satisfaire pour optimiser la qualité des modifications obtenues par l'algorithme TD-PSOLA. Celles-ci peuvent se classer en trois types:

(a) longueur de la fenêtre

Les fenêtres doivent d'une part bénéficier d'une résolution spectrale compatible avec la structure fine du spectre du signal à modifier et d'autre part vérifier une condition de normalisation énergétique. La première contrainte limite la taille de la fenêtre [5]: pour respecter les conditions d'analyse "large bande", la longueur de la fenêtre ne doit en effet pas excéder deux fois la période fondamentale d'analyse. La seconde contrainte impose une condition sur le recouvrement des fenêtres à la synthèse.

Lors d'une élévation de la fréquence fondamentale, ces contraintes sont simultanément satisfaites en choisissant des fenêtres de longueur égale à deux fois la période fondamentale de synthèse. Lors d'un abaissement de la fréquence fondamentale, la première contrainte est satisafaite en choisissant des fenêtres de longueur égale à deux fois la période fondamentale d'analyse. La condition de normalisation n'est alors plus exactement vérifiée: le signal de synthèse doit être alors multiplié par un facteur de compensation énergétique dépendant du recouvrement des fenêtres à la synthèse. Cette étape peut être omise dans le cadre d'une implantation temps-réel, pour des facteurs d'abaissement limités [Moulines,90].

(b) type de fenêtre

Pour éviter des distorsions spectrales importantes, les fenêtres doivent respecter un certain nombre de conditions en terme de largeur du lobe principal et d'amplitude des lobes secondaires [Harris,78]. Ces conditions sont remplies par des fenêtres d'analyse spectrale usuelle (Hanning, Backman).

Dans le cadre d'une implantation temps-réel, il est coûteux d'avoir à régénérer les fenêtres, dont la taille varie en fonction de la fréquence d'analyse et de synthèse, à chaque période. Nous avons donc choisi d'en tabuler un nombre fini. Pour limiter le nombre de données à stocker, nous avons utilisé des fenêtres a priori dissymétriques, constituées de deux demi-fenêtres de Hanning, la durée exacte de la fenêtre étant ajustée en intercalant entre ces deux demi-fenêtres un segment de fenêtre rectangulaire.

Le spectre de ce type de fenêtre [6] peut être évalué en remarquant qu'elle résulte d'un produit de convolution d'une arche de cosinus d'une largeur $(\alpha/2)N$ et d'une fenêtre rectangulaire de largeur $(1-\alpha/2)N$ (ce produit de convolution évolue du rectangle à la fenêtre de Hanning lorsque le paramètre α varie de 0 à 1) La transformée de Fourier de ce type fenêtre est donc le produit du spectre d'une fenêtre rectangulaire et d'une fenêtre de Hanning de longueurs différentes. Elle présente dans le cas général une structure compliquée de lobes secondaires. L'affaiblissement de ceux-ci est satisfaisant (-25 dB) dès que le segment de fenêtre rectangulaire n'excède pas 20 % de la longueur totale de la fenêtre. Une quinzaine de demi-fenêtres tabulées s'avèrent alors suffisantes pour permettre de régénérer toutes les fenêtres nécessaires.

(c) synchronisation précise du traitement

---

5 on suppose ici qu'on utilise une fenêtre d'analyse spectrale classique de type Hanning

6 dans le cas d'une fenêtre symétrique

Il est souhaitable de synchroniser le centre des fenêtres d'analyse avec l'instant d'excitation maximale du conduit vocal, à savoir l'instant de fermeture de glotte. Cette synchronisation, effectuée lors de l'analyse des unités acoustiques, conditionne beaucoup la qualité de synthèse obtenue. La solution la plus sûre pour effectuer cette synchronisation est de disposer de mesures physiologiques directes de l'activité glottique (par exemple, de mesures électro-glottographiques synchronisées avec le signal de parole). Ces mesures ne sont toutefois pas indispensables et nous avons mis au point un algorithme permettant d'obtenir une synchronisation satisfaisante à partir du signal de parole lui-même. Cet algorithme combine les résultats d'une estimation de la fréquence fondamentale (utilisant une mesure d'intercorrélation du signal de parole filtré passe-bas) et d'une détection séquentielle d'évènements fondée sur une mesure d'évolution rapide du spectre à très court-terme [Di Francesco & Moulines, 89]. Cette méthode donne des résultats satisfaisant, quelques corrections manuelles étant néanmoins à apporter (principalement pour les débuts et fins de voisement).

5 BIBLIOGRAPHIE

[Atal,Remde, 82] Atal B.S. & Remde J., "A new model of LPC excitation for producing natural-sounding speech at low bit rates", IEEE Int. Conf. ASSP, Paris,611-614.

[Bimbot,88] F. Bimbot, "Synthèse de la parole: des segments aux règles, avec utilisation de la décomposition temporelle", Thèse de Doctorat E.N.S.T, 1988.

[Charpentier & Moulines, 89] F. Charpentier, E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Int. Conf. Eurospeech, 1989.

[Emerard, 77] F. Emerard, "Synthèse par diphones et traitement de la prosodie", Thèse de 3 ième cycle, 1977, Université de Grenoble.

[D'allessandro & Liénard,88] C. D'allessandro, J.S. Liénard, "Decomposition of the speech signal into short-timewaveforms using spectral segmentation", Proc. ICASSP 88, New-York.

[Hamon & al, 89] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", Proc. Int. Conf. ASSP, Glasgow, 1989.

[Harris,78] F. J. Harris, " On the use of windows for harmonic analysis with the Discrete Fourier Transform", Proc. of the I.E.E.E., vol-66, n0 1, 1978.

[Hedelin,86] P. Hedelin, "High quality glottal LPC vocoding", Proc. of Int. Conf. on Acoust. Speech and Signal Proc., 1986.

[Klatt,87] Klatt D.H., "Review of text to speech conversion for English", J. Acoust. Soc. Am., 82(3),737-793.

[Moulines & Charpentier,88] E. Moulines, F. Charpentier, "Diphone synthesis using multipulse linear prediction", Proc FASE, Int. Conf. Edinburgh, 1988.

[Moulines & al,89] E. Moulines, C. Hamon, F. Charpentier, "High-quality prosodic modifications of speech using time-domain overlap-add synthesis", XII colloque GRETSI, 1989.

[Moulines,90] E. Moulines, "Contributions à la synthèse de parole à partir du texte", Thèse présentée en vue de l'obtention du doctorat E.N.S.T, Février 90.

[Rodet & al,87] X. Rodet, P. Depalle, G. Poirot, "Analyse et synthèse de voix parlée et chantiée par modélisation de l'enveloppe spectrale et de l'excitation", XVI ième JEP, 1987.

[Shingal & Atal,84] S. Shingal, B.S. Atal, "Improving performance of multi-pulse LPC coders at low bit rates", Proc. ICASSP 84.

[White & al,90] S. White, P. Mabilleau, E. Moulines, "Codeur CELP à débit variable: application au codage des diphones", XVIII ième JEP, 1990.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Mesure subjective de la redondance contextuelle : un indice pour quantifier la complexité linguistique

Christian BENOIT

Institut de la Communication Parlée
U.R.A. CNRS N°368 - INPG/ENSERG - Université Stendhal
BP 25 X - 38040 GRENOBLE Cedex FRANCE

RESUME

Un test d'intelligibilité a permis d'évaluer différents synthétiseurs français à partir du texte à l'aide de phrases "sémantiquement imprédictibles". Les distributions de réponses des auditeurs montrent une forte relation entre le pourcentage des phrases et celui des mots correctement retranscrits. Le rapport de leurs logarithmes apparaît comme un indice robuste de la complexité d'un message oral. Des données extraites de la littérature confirment l'hypothèse selon laquelle plus une phrase contient d'informations contextuelles (sémantiques, syntaxiques, etc.), plus faible est cet indice. Il pourrait être relié au nombre d'unités de décision traitées par un auditeur à l'écoute d'une phrase. Les synthétiseurs distordent ici la compréhension des phrases. Or, les erreurs ne suivent pas la loi binomiale qui régirait leur distribution si l'on considérait un modèle simplifié où les unités testées (mots ou syllabes) auraient la même probabilité d'être correctement identifiées. L'analyse des divergences entre l'observation et la théorie issue de ce modèle simpliste montre clairement que la redondance linguistique corrige des mots "*a priori* incompris" et entache d'erreur des mots "*a priori* compris". Ce phénomène de correction/distorsion dans les phrases dépend surtout de leur complexité linguistique, quantifiable par l'indice suggéré, lequel présente également des variations du second ordre dues à la compétence des sujets, à leur entraînement, ou au niveau de dégradation acoustique du message.

1. INTRODUCTION

Si les psycholinguistes ont montré depuis longtemps que l'intelligibilité d'un message oral dépend de la quantité d'information linguistique qu'il véhicule [Miller et al., 1951], la redondance contextuelle n'a cependant jamais été quantifiée.

Il est souvent légitime de considérer une phrase "bien formée" (au sens où l'entend Ruwet [1967], p. 20) comme un ensemble hiérarchisé de mots issus de différents lexiques grammaticaux et dont les positions dépendent des règles syntaxiques de la langue utilisée. Cette conception est opérante pour décrire les langues écrites. Elle est même incontournable à ce jour lorqu'il s'agit de synthétiser la parole à partir du texte. Mais la communication parlée fonctionne selon un schéma éloigné d'une telle combinaison linéaire idéale d'unités. C'est la raison pour laquelle Miller [1962] suggéra le terme d'"**unités de décision**" dans la perception de la parole. Même cachée, leur existence se manifeste quotidiennement dans le dialogue oral ; elle résulte de décisions successives prises au niveau phonétique, lexical, syntaxique, sémantique et pragmatique. Ces décisions, hiérarchisées, reflètent des stratégies ascendantes et descendantes dont le nombre dépend naturellement de la compétence linguistique de chaque auditeur, comme de la complexité linguistique du message émis - ou perçu, s'il n'a pas été correctement compris.

Il n'est malheureusement pas possible de préciser les frontières (à supposer qu'elles existent...) de telles unités de décision dans le cerveau - ou, plus précisément, dans la mémoire de travail - d'un auditeur dont les performances ne peuvent être évaluées qu'en terme de temps de réaction ou de pourcentages d'unités correctement identifiées. Malgré cette incapacité à les identifier, nous tentons ici d'estimer le nombre d'unités de décision traitées par des sujets au cours d'une tâche de transcription orthographique de phrases courtes, acoustiquement dégradées, et au contenu linguistique contrôlé. Des phrases sémantiquement imprédictibles ont été générées automatiquement à partir d'un petit nombre de structures syntaxiques simples. Ce corpus a ensuite été synthétisé sous différentes conditions acoustiques.

Les résultats de cette évaluation de synthétiseurs et la méthodologie adoptée ont été publiés ailleurs (Benoît, 1989). Des tests similaires ont été menés simultanément en Angleterre (Hazan & Grice, 1989) et en Allemagne à partir d'un corpus "multilingue" comparable (Grice, 1989) dans le cadre d'un projet ESPRIT (Benoît et al., 1989).

Nous présentons ici un deuxième niveau d'analyse des réponses des auditeurs pour lequel les synthétiseurs ont servi à dégrader le **support acoustique** de l'information verbale. L'étude des distributions d'erreurs ainsi observées dans la transcription de phrases synthétisées nous permet ainsi de suggérer un nouveau moyen d'en mesurer la **complexité linguistique**.

2. DESCRIPTION DE L'EXPERIENCE

2.1 Le corpus

Nous avons retenu cinq structures syntaxiques simples n'impliquant pas plus de huit mots (y compris les articles définis). Vingt phrases ont été générées aléatoirement, conformément à chaque structure, par concaténation des monosyllabiques français les plus fréquents extraits de lexiques correspondant à leur catégorie grammaticale. Ce test est une extension d'expériences psycholinguistiques classiques [Miller, 1962 ; Miller & Isard, 1963] et d'un test d'intelligibilité de synthétiseurs [Nye & Gaitneby, 1974].

Il faut souligner que les cinq structures ne comportent pas le même nombre de mots (voir Annexe) et que la morphologie du français introduit parfois des modifications du nombre initial de syllabes dues à l'élision de l'article défini devant une voyelle initiale ou de l'insertion d'un *schwa* entre des structures multiconsonantiques finales et initiales de mots. Le nombre moyen de syllabes par structure est en fait inférieur au nombre théorique.

2.2 Les stimulis

Deux synthétiseurs à partir du texte ont été employés dans cette expérience. Ils utilisent le même dictionnaire de diphones, chacun avec sa propre technique de codage. Les cent phrases ont été synthétisées par chacun sous deux conditions prosodiques : un schéma "constamment plat" ; et la modélisation prosodique automatique propre à chacun d'eux. En outre, le locuteur dont la voix avait servi à la constitution du dictionnaire de diphones a enregistré les mêmes cent phrases (légèrement dégradées par la suite) pour comparer la parole naturelle aux 400 stimulis synthétiques. Un ordre de type "carré latin" a été respecté pour aparier les cinq types de voix avec les cinq structures syntaxiques sur cinq bandes magnétiques contenant cent stimulis chacune (mélangés pseudo-aléatoirement).

2.3 Le test

Cinq groupes de quatre sujets ont écouté, à travers un casque, dans une chambre sourde, l'une des bandes magnétiques, à chacune des cinq sessions constituant le test. Le même carré latin que ci-dessus a servi à ordonner la présentation des cinq bandes magnétiques aux cinq groupes d'auditeurs au cours des cinq sessions, ce qui permettait de calculer n'importe quelle score moyen sur les synthétiseurs, les syntaxes, les sessions et les groupes d'auditeurs. Les auditeurs, rémunérés, étaient naïfs en parole synthétique et n'étaient pas informés du contenu linguistique du test. Tous étaient de jeunes étudiants français, sans déficience auditive, et compétents en orthographe comme en écriture manuscrite, de façon à homogénéiser les performances linguistiques. Après une courte phase d'adaptation aux types de phrases et de synthétiseurs, il leur a été demandé de retranscrire sur une feuille de réponse ce qu'ils avaient entendu... ou cru entendre.

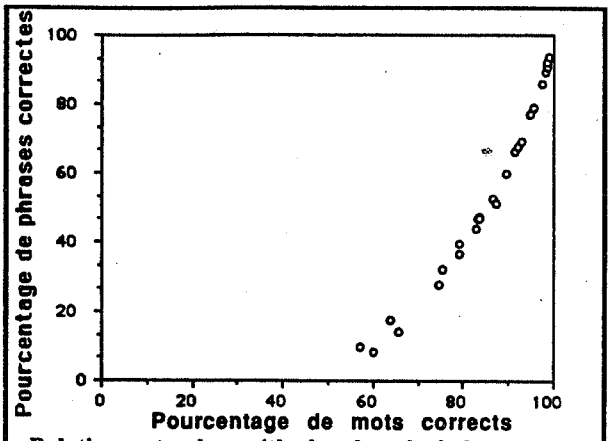
3. RESULTATS

Nous ne fournissons pas ici les résultats détaillés de cette expérience, publiés ailleurs (Benoît, 1989). Rappellons cependant que les cinq conditions acoustiques se répartissent d'une intelligibilité très pauvre à une clarté quasi-totale. Un effet d'apprentisage a été nettement observé, dû en particulier au fait que les cent phrases étaient représentées à chaque session, bien que sous des conditions acoustiques différentes.

3.1. Comparaison des méthodes de calcul des scores

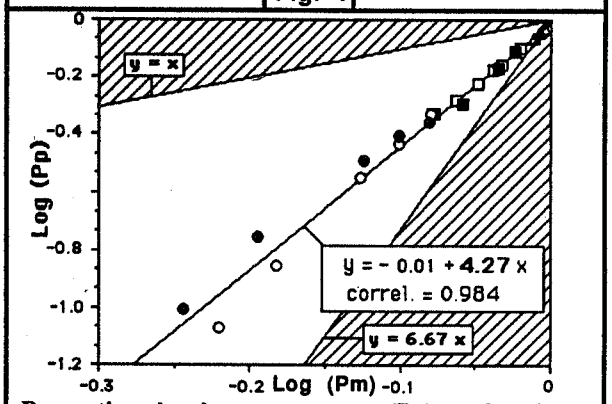
Pour l'ensemble des réponses des 20 auditeurs aux 5 sessions présentant 100 stimulis chacune, ont été calculés les nombres de mots corrects (66700 au total) et de phrases correctes (10000 au total). Tous les mots, articles y compris, ont été considérés. Ils devaient être retranscrits homophoniquement, appartenir à la bonne catégorie grammaticale et être à la bonne position dans la phrase (i.e. position correcte de la syllabe). Fautes, insertions et omissions ont été sanctionnées. Une phrase était correcte si elle ne comportait aucune erreur. Les proportions de mots corrects (Pm) et de phrases correctes (Pp) pour chaque conditon acoustique à chaque session, toutes syntaxes confondues, sont projetées Fig. 1. Une relation exponentielle apparaît clairement entre les deux modes de calcul des scores. Aussi avons-nous représenté Fig. 2 les logarithmes de ces proportions. Il en ressort une forte corrélation et une relation, linéaire cette fois-ci, entre les logarithmes des proportions de mots corrects et ceux des phrases correctes. Le résultat expérimental fournit la relation suivante :

$$\mathbf{Log\ (Pp) = r.Log\ (Pm) \Leftrightarrow Pp = Pm^r}\ \text{(avec r = 4.27)}$$



Relation entre les méthodes de calcul des scores
Chaque point représente 400 phrases ou 2700 mots (20 réponses par 20 auditeurs) sous 25 conditions (5 voix x 5 sessions), toutes structures syntaxiques confondues.

Fig. 1



Proportion de phrases correctes (Pp) en fonction de la proportion de mots corrects (Pm) en log.
5 types de voix x 5 sessions. Chaque point représente les réponses à 400 phrases.
△ Voix 1 □ Voix 2 ■ Voix 3 ○ Voix 4 ● Voix 5
Les lignes supérieure et inférieure montrent les limites théoriques des observations

Fig. 2

3.2 Prédiction des erreurs par la loi binomiale

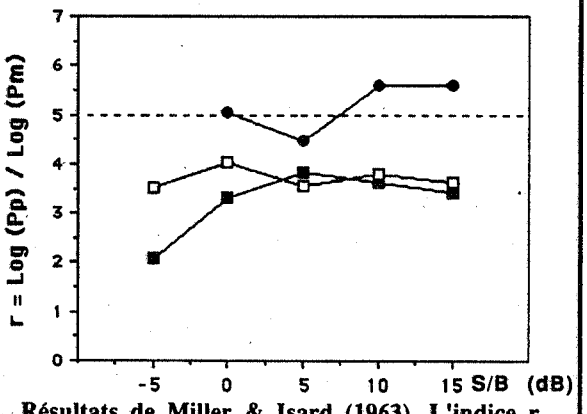
En faisant une approximation grossière, il est possible de considérer un modèle simplifié de phrase dont les unités - c'est-à-dire les mots ou les syllabes - quelles que soient leur position et leur catégorie grammaticale, auraient la même probabilité d'être correctement comprises. Dans un tel cas, une phrase se comporterait comme une suite d'unités dans laquelle une dégradation acoustique (le synthétiseur, ici) introduirait des erreurs de transmission. Si la probabilité qu'une unité soit correctement comprise est **pu**, la probabilité **p(k)** qu'une séquence de **n** unités contienne **k** unités correctes est alors donnée par la Loi Binomiale :

$$\mathbf{p(k) = pu^k \cdot (1 - pu)^{n-k} \cdot C[n,k]}$$

Si **k = n**, toutes les unités sont correctes. Notons **ps = p(n)** : probabilité que la séquence soit correcte. Donc, **ps = pu^n**. Il apparaît ainsi une divergence entre le résultat observé **Pp = Pm^r** et l'hypothèse simplificatrice **ps = pu^n**, puisque **r = 4.27 < n = 6.67** où **n** est le nombre moyen de mots par phrase dans notre corpus. Nous devons alors en conclure que le modèle binomial est trop simple quand il suppose que toutes les unités sont indépendantes, et/ou que les mots ne sont pas la bonne unité de perception à mesurer... Du fait des contraintes syntaxiques, et donc de l'information contextuelle, les mots ne sont pas indépendants, ni leurs erreurs de compréhension, tandis que le la loi binomiale suppose des unités aux probabilités indépendantes. Une question se pose alors : la dépendance entre les mots peut-elle être mesurée ? Nous essayons d'y répondre par le postulat suivant : la différence entre **n** et **r** peut être considérée comme un indice de la dépendance contextuelle des mots dans la phrase, où **r = Log (Pp) / Log (Pm)** représente la ***complexité linguistique d'une phrase***, en "équivalents - mots indépendants", ou en unités de décision...

3.3 Comparaison avec des données de la littérature

Miller et Isard (1963) ont évalué l'intelligibilité de phrases grammaticales (i.e. à contenu sémantique), de phrases sémantiquement imprédictibles et de chaînes agrammaticales (i.e. constituées de mots désordonnés). Le même vocabulaire était utilisé dans chaque type de phrases. Les pourcentages de mots corrects et de phrases correctes ont été mesurés sous cinq conditions de dégradation par addition de bruit. Nous avons calculé les rapports de leurs logarithmes qui sont présentés Fig. 3 en fonction du niveau de dégradation. L'indice **r** observé dans les deux cas de phrases syntaxiquement correctes est systématiquement inférieur à celui obtenu pour les chaînes agrammaticales, lequel est sensiblement égal au nombre initial de mots testés (cinq). De surcroît, la valeur de l'indice pour les phrases grammaticales est inférieur à celui observé pour les phrases sémantiquement imprédictibles dans les pires conditions de dégradation. Ce qui confirme l'idée que cet indice reflète la redondance contextuelle des phrases... sinon le nombre d'unités de décision nécessaire à leur compréhension !

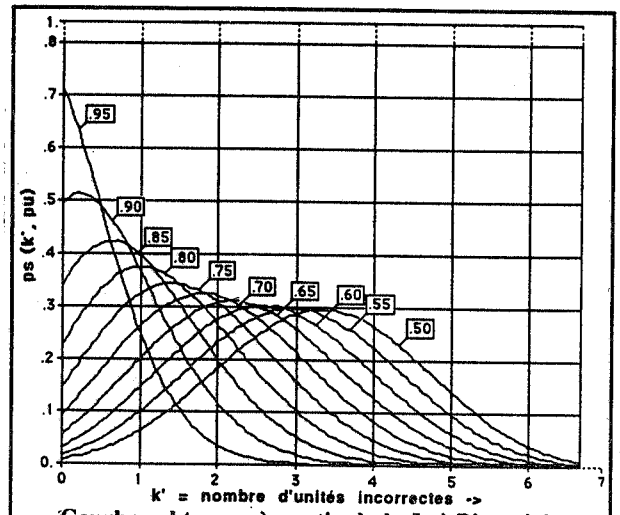


Résultats de Miller & Isard (1963). L'indice r a été recalculé en fonction de la dégradation.

Fig. 3

3.4 Analyse de la distribution des erreurs

Afin d'analyser nos résultats plus en détail, nous avons dû tout d'abord étendre la loi binomiale au cas où le nombre d'unités n'est pas entier, comme c'était le cas dans notre test où nous ne pouvions accéder qu'à un nombre moyen - et réel - de mots par phrase, quelle que soit la structure syntaxique considérée. Pour y parvenir, nous avons extrapolé la factorielle utilisée dans le calcul des combinaisons **C[n,k]** en la remplaçant par la fonction Gamma qui permet de traiter des valeurs réelles de **n** et de **k**. La Fig. 4 présente les courbes continues ainsi obtenues pour différentes valeurs de **pu**, avec **n = 6.67**.



Les mots n'ont pas été testés en présentation isolée. Des valeurs plus réalistes auraient ainsi pu être affectées à **Pm**, puisque le **Pm** mesuré ici tient déjà compte *per se* des différentes corrections et distorsions introduites par la redondance contextuelle. Nous n'avons donc pu qu'observer comment les erreurs *a posteriori* sur les mots affectent les erreurs *a posteriori* sur les phrases.

Nous avons calculé les lois binomiales théoriques **B(n',pm)** approximant au mieux les 25 distributions de réponses des 20 sujets sur 16 phrases sémantiquement imprédictibles, pour chaque structure syntaxique et chaque présentation acoustique. Deux exemples sont présentés Fig. 5. Dans trois conditions acoustiques sur cinq, **r** et **n'** ne diffèrent jamais de plus de 0.2 %, quelle que soit la structure syntaxique. Ceci prouve que le rapport des logarithmes correspond rigoureusement au nombre d'unités théoriques qui auraient fourni la distribution d'erreurs la plus proche de celle observée expérimentalement.

4. DISCUSSION

La distribution des erreurs dévie de plus en plus d'une distribution binomiale à mesure que la dégradation augmente. Cela peut signifier que le modèle est irréaliste, mais cela peut aussi être dû au fait qu'il n'est possible d'observer que les distributions de **n** mots incorrects alors qu'il nous faudrait observer les distributions de **n'** unités de décision incorrectes...

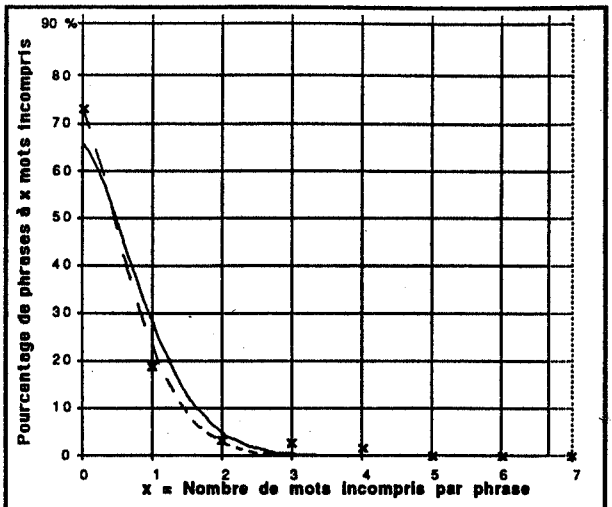


Fig. 5.a : voix synthétique N° 2, syntaxe N° 5

Cinq sessions moyennées ; 320 phrases (2140 mots) par point (X)
Pm = .940 ; Pp = .734 ; n = 6.69 m/p ; indice r = 4.97
——— : distribution théorique B(n,pm)
- - - - - : distribution approchée B'(n',pm) avec n' = 4.96
(9.3 % hors distr.)

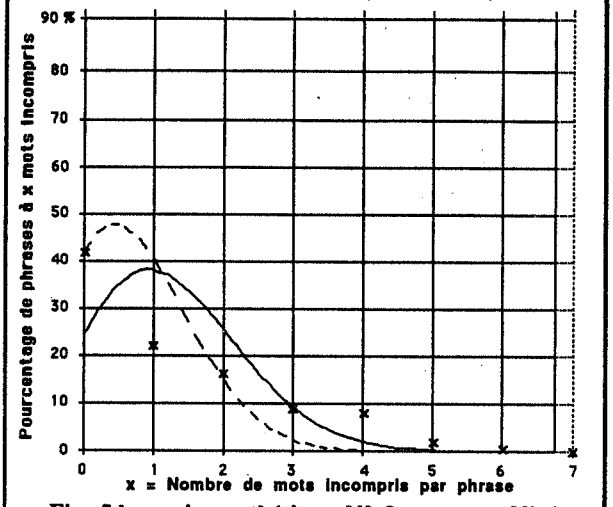


Fig. 5.b : voix synthétique N° 3, syntaxe N° 1

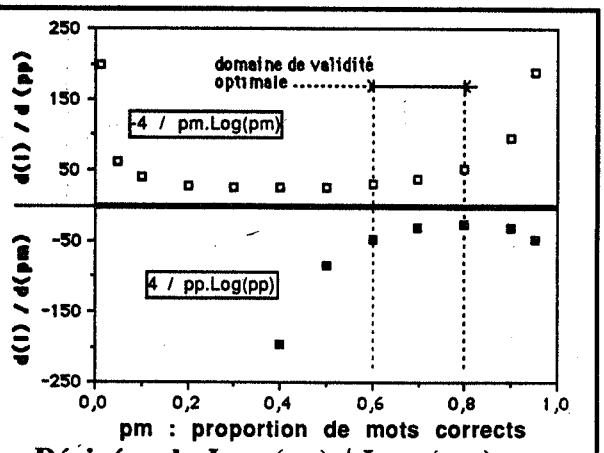
Cinq sessions moyennées ; 320 phrases (2160 mots) par point (X)
Pm = .811 ; Pp = .419 ; n = 6.75 m/p ; indice r = 4.16
——— : distribution théorique B(n,pm)
- - - - - : distribution approchée B'(n',pm) avec n' = 4.16
(37 % hors distr.)

Fig. 5

L'indice **r** tend vers la valeur de **n** quand la dégradation diminue. Ceci est naturellement dû au cas limite où la très faible proportion d'erreurs potentielles (ou observées) rend celles-ci strictement indépendantes les unes des autres. Lorsqu'un message est proche de l'intelligibilité totale, la distribution des erreurs résiduelles suit la loi binomiale, et l'observation rejoint la théorie : **r** est alors égal à **n**.

Il existe en réalité un niveau optimal de dégradation sous lequel il convient de mesurer l'indice **r**, comme le suggèrent les résultats, et comme le prouve la logique évidence : au-delà d'une trop forte dégradation, l'information contextuelle se dilue dans la pauvreté du flux acoustique ; au-dessus d'une trop grande clarté, la redondance n'est plus nécessaire à la compréhension d'un message suffisamment simple.

Afin de mieux cerner le domaine de validité de notre indice, nous présentons sur la Fig. 6 l'évolution de sa dérivée par rapport à **pm** et à **pp** en fonction de **pm** pour un nombre donné d'unités de décision théoriques (r = 4 sur la figure). On voit alors nettement qu'en-dessous de 50 % de mots corrects, l'indice **r** diminue fortement pour une faible augmentation de la proportion de phrases correctes ($\Delta r = 1$ pour une variation de 1% de phrases correctes, à 40 % de mots corrects observés), de même qu'il augmente fortement pour une faible augmentation de plus de 80 % de mots corrects ($\Delta r = 1$ pour un pourcentage de mots corrects observés variant de 90 à 91 %). Le niveau optimal se situe donc sur une plage limitée (entre 60 et 80 % de mots corrects pour un nombre d'unités de décision de l'ordre de 4) au-delà de laquelle la mesure de l'indice **r** nécessite un nombre élevé de tests subjectifs, en phrases, en répétitions, ou en nombre d'auditeurs pour être précise.



Dérivées de Log (pp) / Log (pm) par rapport à pm et à pp, en fonction de pm

Fig. 6

Dans l'analyse de nos résultats, nous avons comparé nos observations à un modèle dans lequel toutes les unités auraient eu la même probabilité d'être entachées d'erreurs. Nous avons affecté à cette probabilité supposée la valeur moyenne observée dans notre expérience sur un nombre de phrases dépendant de la distribution étudiée (quelques dizaines à quelques centaines). Or, cette valeur correspond à la moyenne arithmétique des erreurs observées. Il est plus réaliste d'imaginer des variations de ces proportions d'erreurs en fonction de la complexité phonotactique de chacun des mots, de leur densité de voisins phonétiquement "proches" dans le lexique, ou encore de leur catégorie grammaticale, voire de leur position dans la phrase. Si la proportion d'erreurs de chacun des constituants de chacune des phrases avait été mesurée, nous aurions alors pu comparer nos observations à des distributions probables d'erreurs calculées à partir des probabilités propres à chaque unité. La valeur particulière correspondant à la probabilité théorique **ps** qu'une séquence de **n** unités soit entièrement correcte nous aurait ainsi été fournie par le produit des **n** probabilités élémentaires : $ps_g = \Pi[i=1,n]\ pu_i$, alors que la loi binomiale utilisée ci-dessus calcule cette probabilité de séquence correcte à partir de la proportion moyenne de mots corrects dans la phrase : $ps_a = \{\Sigma[i=1,n]\ pu_i / n\}^n$. Les racines ènièmes de ces deux probabilités correspondent en fait respectivement aux moyennes **géométrique** et **arithmétique** des probabilités élémentaires. Or, *la moyenne géométrique est toujours inférieure à la moyenne arithmétique.*

D'une part, la différence entre la proportion **observée** de phrases entièrement correctes (**Pp**) et la probabilité **théorique** de leur apparition (ps_g) aurait ainsi été amplifiée. D'autre part, un calcul plus réaliste de notre indice de complexité (également basé sur la moyenne **géométrique** des proportions observées de mots corrects) aurait fourni une valeur r' (= Log (Pp) / Log (Pm_g), avec $Pm_g = \sqrt[n]{\{\Pi[i=1,n]\ pu_i\}}$) encore inférieure à celle que nous avons obtenue ici (r = Log (Pp) / Log (Pm_a), avec $Pm_a = \Sigma[i=1,n]\ pu_i / n$). Si les proportions intrinsèques de chaque mot avaient été utilisées dans cette expérience, un tel calcul aurait donc, non seulement confirmé, mais encore amplifié les différences entre les indices de complexité linguistique mesurés pour les phrases redondantes et pour les chaînes de mots indépendants.

5. CONCLUSION

Ce travail préliminaire suggère que la redondance contextuelle peut être quantifiée. Si sa validité était confirmée, l'indice proposé serait d'un grand intérêt, pour la psycholinguistique comme pour l'évaluation de synthétiseurs de parole, en tant qu'instrument de contrôle de la complexité linguistique. Mais avant cela, de nouvelles expériences doivent être menées afin d'observer l'évolution des pourcentages de mots corrects entre présentations isolées (proportions intrinsèques et *a priori*) et présentations en contexte (proportions *a posteriori*) sous différentes conditions de dégradation. Il est également souhaitable d'évaluer la robustesse de l'indice sur des phrases grammaticales et des chaînes agrammaticales dépassant le nombre "*magique*" de sept ± deux mots, ce qui permettrait de vérifier ou d'infirmer l'hypothèse selon laquelle cet indice serait relié au nombre d'unités de décision prises par les auditeurs. Il n'est en effet pas utopique d'envisager un seuil de complexité au-delà duquel l'absence de redondance linguistique rendrait toute tâche de retranscription correcte impossible, du fait du trop grand nombre d'unités (indépendantes) à traiter par les auditeurs, alors que la présence de la seule syntaxe fournirait suffisamment d'informations contextuelles aux sujets pour que ceux-ci puissent retranscrire correctement certaines phrases et nous permettent ainsi de mesurer le nombre d'unités de décisions auxquelles ils ont eu affaire...

Remerciements

Ce travail a été subventionné par le projet ESPRIT N°2589 et par le Centre National de la Recherche Scientifique. Nous tenons à remercier les partenaires européens du projet *SAM* pour les nombreux et fructueux échanges scientifiques qui nous ont permis d'avancer dans cette recherche, ainsi que Pascal PERRIER et Jean-Luc SCHWARTZ, pour leurs critiques et commentaires.

Références bibliographiques

BENOIT C. (1989), "Intelligibility test for the assessment of French synthesizers using Semantically Unpredictable Sentences", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.7.1.-1.7.4., Noordwijkerhout, Hollande.

BENOIT C., **ERP** A. van, **HAZAN** V., **GRICE** M. and **JEKOSH** U. (1989), "Multilingual synthesizer assessment using Semantically Unpredictable Sentences", Proceedings of EUROSPEECH'89 Conference, 633-636, Paris.

GRICE M. (1989), "Syntactic structures and lexicon requirements for Semantically Unpredictable Sentences in a number of languages", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.5.1.-1.5.4., Noordwijkerhout, Hollande.

HAZAN V. & **GRICE** M. (1989), "The assessment of synthetic speech intelligibility using Semantically Unpredictable Sentences", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.6.1.-1.6.4., Noordwijkerhout, Hollande.

MILLER G.A. (1962), "Decision units in the perception of speech", IEEE Tr. on Information Theory, 8, 81-83.

MILLER G.A., **HEISE** G.A. and **LICHTEN** W. (1951), " The intelligibility of speech as a function of the context of the test materials", J. Exp. Psychol., 41, 329-335.

MILLER G.A. and **ISARD** S. (1963), "Some perceptual consequences of linguistic rules", J. Verbal Learning and Verbal Behavior, 2, 217-228.

NYE P.W. and **GAITENBY** J. (1974), "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences", Haskins Lab. Status Rep. on Speech Research, 37/38, 169-190.

RUWET N. (1967), "Introduction à la grammaire générative", Plon, Paris.

ANNEXE : exemples des cinq structures syntaxiques utilisées dans l'expérience

Stx 1) *"La robe entre vers la science rouge."*
Stx 2) *"Le verre vrai ouvre le coin."*
Stx 3) *"Tourne peu la date et la main."*
Stx 4) *"Quand le texte pose-t-il la fille crue ?"*
Stx 5) *"La chose lance le train qui passe."*

6 OUTILS POUR LE TRAITEMENT DE LA PAROLE

Président: A. MARCHAL
Université de Provence,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Evaluation d'un détecteur de fréquence fondamentale du signal microphonique par comparaison à une référence laryngographique.

BARBE Thierry, BAILLY Gérard

Institut de la Communication Parlée
ENSERG/INPG - Université Stendhal
46, Avenue Félix Viallet - 38031 Grenoble Cedex

Abstract

This article is concerned with the evaluation of a new pitch detector in comparison with a reference contour extracted from the laryngograph. This robust pitch detector is speaker and noise-level independant, i.e. requires no manual thresholds nor frequency range adjustments. Dynamic Time Warping (DTW) is applied on a time-limited pitch candidates lattice at each voicing onset to determine a pitch anchor sequence. Simple dynamic frame-to-frame tracking is then applied to enable real-time implementation.

An algorithm to extract a reference pitch contour from a laryngograph is described and used to evaluate our pitch tracker. 98% of the pitch values are within 10% of the reference. 1% of the reference voiced pitch values have been devoiced and 10% of the reference unvoiced pitch values have been devoiced (70 % being located at the voicing ends due to acoustic wave propagation).

1. Introduction

Lorsqu'on réalise un détecteur de fréquence fondamentale, un des problèmes majeurs qui se pose à son concepteur est la mise en oeuvre de son évaluation. Pour cela, on se donne un corpus de test pour lequel on dispose de valeurs fiables de la fréquence fondamentale, et qu'on compare aux valeurs calculées par l'algorithme à évaluer. Le point crucial dans ce type de comparaison est bien sûr l'obtention de la courbe mélodique de référence.

On distingue trois grands types de procédures:

- Mesure manuelle ou semi-automatique de la courbe de référence. Cette méthode consiste à positionner à la main des marqueurs de période sur le signal, ou par autocorrélation, cepstre etc. [4,6,8] dans le cas d'une détection semi-automatique, l'inconvénient étant bien sûr, le temps nécessaire pour positionner ces marqueurs.
- Courbe de référence fixée et synthèse. Cette méthode consiste à s'imposer une courbe de référence et à synthétiser le signal de parole à partir de cette courbe mélodique, l'inconvénient étant l'impossibilité de traiter du signal de parole naturelle.
- Courbe de référence fournie par un autre algorithme. Cette méthode consiste à prendre comme courbe de référence, la courbe donnée par un autre détecteur de fréquence de fondamentale, supposé robuste et fiable.

Nous nous proposons ici d'évaluer un algorithme de détection de la fréquence fondamentale par comparaison avec les valeurs mesurées sur le signal issu d'un laryngographe [5]. Pour cela, on dispose de deux enregistrements, réalisés simultanément, l'un au niveau de la glotte à l'aide du laryngographe, fournissant le signal de référence, et l'autre au niveau de la bouche à l'aide d'un microphone, fournissant le signal à traiter (Fig. 1).

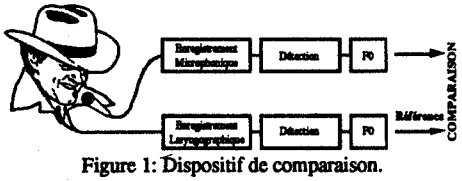


Figure 1: Dispositif de comparaison.

L'intérêt de cette méthode réside dans le fait que le signal analysé pour extraire la courbe de référence est un signal où seuls les cycles de vibrations des cordes vocales sont présents. Dans ce cas, la structure formantique qui complique la détection de la fréquence fondamentale est absente et les erreurs traditionnelles de détection sont quasiment inexistantes. Seuls les mouvements du larynx peuvent être source d'erreurs.

Pour effectuer cette comparaison, nous avons donc réalisé, en plus de l'algorithme de détection de la fréquence fondamentale à partir du signal microphonique, l'algorithme donnant la courbe de référence à partir de l'enregistrement laryngographique. Cet algorithme a été conçu, d'une par pour être implémenté ultérieurement en temps réel et d'autre part pour fournir une valeur de la courbe de référence sans aucun réglage de paramètres particulier.

2. Détection de la fréquence fondamentale à partir du signal issu d'un laryngographe.

Le laryngographe est un appareil qui fournit une mesure de l'impédance de la glotte. Pour cela, on applique deux électrodes, au niveau de la glotte, au travers desquelles est envoyé un signal de 100 kHz environ dont l'amplitude est modulée par la résistance du larynx. Une démodulation de ce signal fournit le signal laryngographique dont les caractéristiques principales sont résumées ci-dessous (cf. Fig. 2).

- Le signal présente un maximum lorsque la glotte est fermée (impédance maximale) et un minimum lorsque la glotte est ouverte (impédance minimale).
- Entre ces deux points caractéristiques, la courbe évolue de façon plus ou moins rapide. La fermeture s'effectue très rapidement (effet Bernouilli) et la courbe présente lors de cette transition, une pente très abrupte (Fig.2, zone a). Par contre, lors de l'ouverture de la glotte (Fig.2, zone b), ainsi que dans l'intervalle où celle-ci reste ouverte (Fig.2, zone c), la courbe est relativement plate et la pente lors de l'ouverture reste modérée.

Afin de réaliser la détection de la fréquence fondamentale sur ce type de signal, il convient de se fixer une référence sur une période laryngée et de localiser celle-ci sur tous les cycles de vibration. La présence d'une transition rapide du signal lors de la fermeture de la glotte incite à prendre comme référence sur cette courbe, le point d'inflexion (point où la dérivée du signal présente un maximum), communément appelé "instant de fermeture glottique". La figure 2 donne un exemple de signal microphonique, de signal laryngographique et de sa dérivée correspondante.

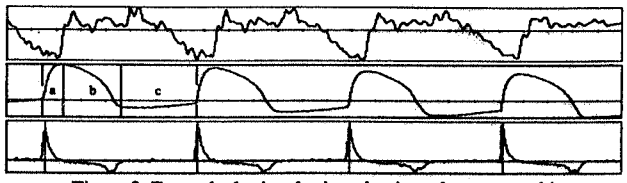


Figure 2: Exemple de signal microphonique, laryngographique et laryngographique dérivé.

Nous avons déterminé, avec la précision donnée par la fréquence d'échantillonnage du signal laryngographique, la position de l'instant de fermeture glottique. Si, par la suite, une précision plus grande dans la mesure de la référence est nécessaire, l'extrapolation suggérée par Hess et al. [5] pourra sans complication être ajoutée à l'algorithme. L'algorithme à tester ne fournissant pas de précison supérieure à celle imposée par la fréquence d'échantillonnage (mesure de la période de la fréquence fondamentale en nombre d'échantillons), nous avons choisi la même précision de mesure pour les deux algorithmes.

On remarque sur les signaux issus d'un laryngographe, la présence d'une composante basse fréquence (20 Hz et moins) superposée au signal utile. Cette composante est dûe aux déplacements verticaux du larynx lors de la phonation. Ce mouvement étant relativement lent par rapport à la fermeture de la glotte, il a une influence négligeable sur la détermination de l'instant de fermeture glottique. Par conséquent, nous avons décidé de ne pas filtrer le signal à l'aide d'un filtre passe-haut, afin de conserver le maximum de précision sur la position de l'instant de fermeture glottique.

Cependant, il s'avère parfois que le signal dérivé présente, dans l'intervalle de fermeture de la glotte, deux ou plusieurs pics distincts, imposant à l'algorithme d'effectuer une décision. Sans nous préoccuper de l'origine de cette irrégularité (pathologie, imprécision du laryngographe, mouvement articulatoire parasite ou autre phénomène), nous avons décidé de considérer les pics les plus représentatifs (pics d'amplitudes les plus élevées) et de déterminer la position du pic par le barycentre des différents pics candidats, avec pour coefficients respectifs, leurs amplitudes. Un exemple de ce type de configuration est donné à la figure 3.

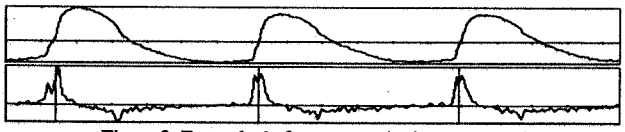


Figure 3: Exemple de fermetures glottiques anormales.

L'algorithme effectue donc une détection de pics sur la dérivé du signal laryngographique. Afin de rendre la détection complètement automatique, nous avons procédé comme suit.

- Le signal présentant un bruit permanent, d'amplitude relativement faible par rapport au signal utile, mais pouvant introduire par moments des pics parasites lors de la dérivation, le signal est lissé par un lissage médian avant la dérivation. La référence se trouvant dans une zone où le signal est croissant, le lissage médian n'a aucun effet sur la position du point de référence.
- Lorsqu'un pic est détecté, il est stocké et un seuil proportionnel à son amplitude est déterminé, seuil au dessus duquel l'amplitude du pic suivant doit se trouver pour pouvoir lui même être considéré.
- Une analyse locale est effectuée sur les différents pics détectés. Elle regroupe les plus proches sous forme de paquets et, dans le cas où ceux-ci contiennent au moins 2 pics, leurs barycentre sont calculés.

Lorsque le signal dérivé est de bonne qualité, les instants de fermeture glottique se manifestent par des pics uniques et d'amplitudes élevées par rapport au bruit du laryngographe. Dans ce cas, un simple seuil est suffisant, mais il s'avère qu'une telle détection est utopique. En effet, il est fréquent de rencontrer des cycles laryngés pour lesquels les fermetures glottiques sont mal effectuées. Ce phénomène se traduit par un pic d'amplitude très faible, inférieure au seuil fixé par le pic précédent (Fig. 4). Il est alors nécessaire d'avoir dans l'algorithme de mesure, une logique de détection particulière pour pallier à ce problème.

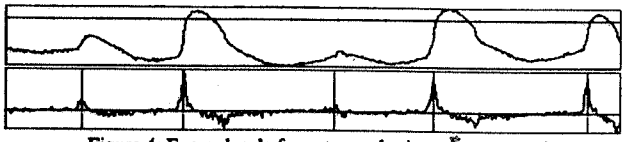


Figure 4: Exemples de fermetures glottiques peu marquées échappant à la première détection.

L'algorithme de mesure procède alors de la façon suivante.

- Dès qu'un pic est détecté, l'algorithme se met en phase de détection de début de voisement. Pour cela, il attend d'avoir au moins trois périodes (soit quatre pics au total) à partir desquelles il détermine la période la plus probable. C'est en général la moyenne des deux périodes les plus proches, sauf si elles sont trop dispersées, auquel cas, c'est la période minimale qui est choisie. Cette valeur devient alors la période de référence.
- Une fois cette période de référence connue, l'algorithme considère chaque nouvelle période et compare sa valeur avec la période de référence. Si la période comparée est trop élevée, l'algorithme cherche si il n'y a pas eu omission de pics dans cette période, auquel cas, la position et l'amplitude de ces pics sont déterminés. Si par contre, la période comparée est voisine de la période de référence, elle devient période de référence et la période suivante est analysée.
- Dès que le signal dérivé ne présente plus de pics, c'est que la zone de voisement est terminée et l'algorithme se met en attente.

Cet algorithme permet de s'affranchir des cas de détections particulières tels que:

- les cycles laryngés peu marqués (Fig. 5.a)
- l'établissement du voisement réalisé de façon irrégulière (Fig. 5.b)
- les zones de vocal fry (Fig. 5.c)
- les mouvements du larynx introduisant un ou plusieurs pics irréguliers (Fig. 5.d).

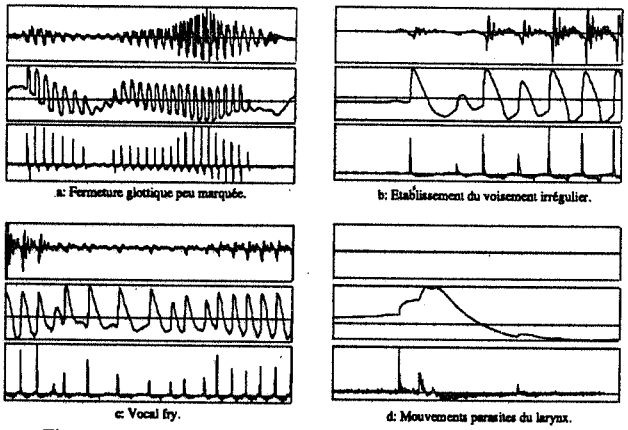


Figure 5: Cas de vibrations anormales pouvant introduire des erreurs.

Cet algorithme permet ainsi de fournir une mesure de référence pour l'évaluation d'un détecteur de fréquence fondamentale. La précision de la mesure peut être accrue par adjonction à l'algorithme d'un filtre interpolateur correspondant, l'interpolation s'effectuant au voisinage des instants de fermeture glottique donnés par cet algorithme. Sa conception permet une implémantation pour un traitement en temps réél et dans le cas d'analyse de fichiers de signaux issus de bases de données ou d'enregistrement personnels, son utilisation est d'une simplicité extrême vu qu'aucun réglage ou paramètre ne sont à fournir.

3. Détection de la fréquence fondamentale à partir du signal issu du microphone.

L'algorithme que nous présentons maintenant est celui de la mesure de la fréquence fondamentale qui effectue la détection à partir du signal microphonique. Cet algorithme, tout comme le précédent, a été conçu en vue d'une implémentation en temps réel, et une attention particulière a été

apportée à l'automatisation des calculs des divers seuils qui lui sont nécessaires. Par conséquent, l'utilisation de cet algorithme ne nécessite aucune connaissance préalable sur le signal à analyser (niveau d'enregistrement, tessiture du locuteur etc.), si ce n'est la fréquence d'échantillonnage.

La gamme de détection de la fréquence à été fixée à [50 Hz, 500 Hz] et les résultats de la mesure de la fréquence fondamentale sont donnés toutes les 10 ms.

L'algorithme réalisé est basé autour de la fonction d'Amdf. La définition de cette fonction pour un signal s(k), pour un retard d donné et pour une fenêtre de NpS points, est:

$$Amdf(d) = \sum_{k=1}^{k=NpS} |s(k) - s(k+d)|$$

Sur cette fonction, la périodicité d'un signal se manifeste par la présence d'un minimum pour le retard correspondant à la période T0 du signal. Cette fonction à l'avantage , par rapport à l'autocorrélation d'être moins sensible au bruit. Par contre, elle est plus sensible aux variations de gain.

L'algorithme de détection effectue donc la recherche du minimum de la fonction d'Amdf correspondant à la période T0 du signal, lorsque celui-ci est voisé. L'inconvénient majeur, source des erreurs classiques de mesure de la fréquence fondamentale pour les fonctions définies dans le domaine temporel, est la présence d'un minimum pour les multiples entiers de T0, ce qui peut revenir à choisir les sous multiples de la fréquence fondamentale (F0/2, F0/3 etc.). Pour cela, la décision doit s'effectuer sur une durée de signal suffisament longue, afin d'éviter les détections érronées classiques [1,2,9,11]. L'algorithme nécessite pour cela trois ou quatre périodes au minimum pour donner une valeur correcte de la fréquence fondamentale. Si la fréquence minimale mesurable est fixée à 50 Hz, correspondant à une période temporelle de 20ms, il faut donc entre 60 et 80 ms de signal. Nous avons donc choisi une valeur de 70 ms, imposant une détection sur 7 trames. Il est bien évident que si la fréquence fondamentale vaut 200 Hz, l'analyse s'effectue sur 14 périodes, mais l'algorithme n'ayant aucune indication préalable sur la valeur moyenne de la fréquence fondamentale à mesurer, on est obligé de considérer le cas le plus défavorable. Par conséquent, les résultats seront fournis avec un retard de 80 ms par rapport au signal. Cet intervalle peut être réduit, mais dégrade les résultats.

Nous allons présenter maintenant les diverses étapes du traitement effectué sur le signal, en précisant les divers paramètres évalués pour mener à bien cette détection, l'organigramme de l'algorithme étant donné à la figure 6.

Afin d'atténuer l'effet dû aux harmoniques supérieures qui peuvent fausser la détermination de la fréquence fondamentale, le signal est filtré par un filtre passe-bas, dont la fréquence de coupure est fixée à 500 Hz.

Une fois le signal filtré, deux énergies sont évaluées, à savoir, l'énergie du signal et l'énergie du signal filtré.

Trois paramètres, dont le but est de déterminer les zones de voisement, sont ensuite évalués. Ces paramètres sont illustrés par la figure 7 où la phrase est "Je vois ces enfants. Ils jouents avec un balai." (Fig.7,a) est prononcée par un locuteur masculin.

Le premier paramètre permet de déterminer les débuts de voisement. Pour une trame n donnée, on détermine le rapport des énergies du signal filtré de la trame n+1 et de la trame n-1. Ce paramètre présente un maximum très prononcé lorsqu'on a un brusque saut d'énergie, soit en début de période voisée, soit au début d'une plosive. Dans ce dernier cas, ce maximum est suivi d'un deuxième maximum, correspondant au début effectif du voisement. Une analyse simple sur quelques trames successives permet de detecter une position fiable du début de voisement (Fig.7,c).

Le second paramètre permet de positionner la fin de voisement. Ce paramètre est plus difficile à mettre en oeuvre que le précédent car la fin de voisement peut présenter une variation plus ou moins lente de l'énergie. Ce paramètre est évalué en prenant en compte la décroissance de l'énergie au cours du temps, le seuil correspondant étant lui uniquement proportionnel à l'énergie de la trame analysée et doit placer la fin de voisement, que celle-ci

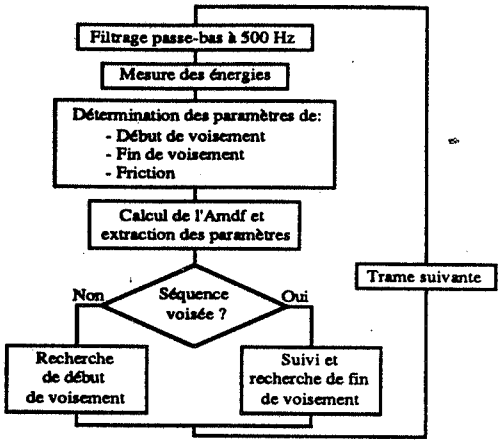


Figure 6: Organigramme du détecteur de fréquence fondamentale à partir du signal microphonique.

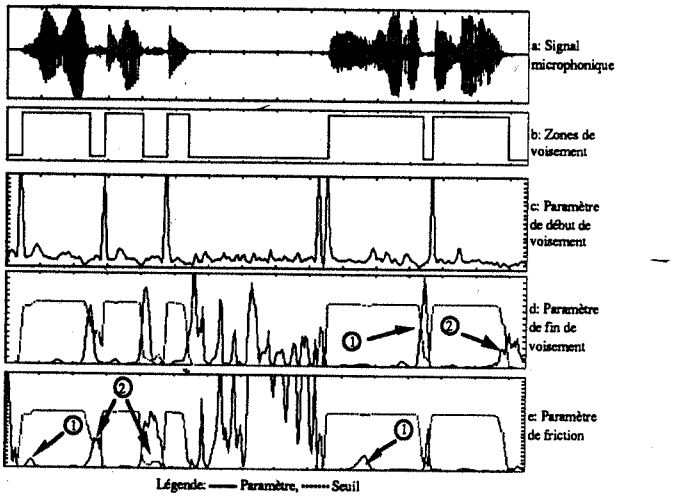


Figure 7: Paramètres mesurés à partir des énergies et détermination des zones de voisement.

soit brusque (Exemple: fin de voisement de "avec" (Fig.7,d, fin N°1)) ou lente (Exemple: fin de voisement de "balai" (Fig.7,d, fin N°2)).

Le troisième paramètre délimite des zones du signal où la friction est importante. Ce paramètre est évalué en fonction de la différence entre les énergies du signal filtré, définissant l'énergie de friction. Le point crucial de ce paramètre est de bien faire la discrimination entre les fricatves voisées (Fig.7,e, zones N°1) et les fricatives non voisées (Fig.7,e, zones N°2). Partant du principe que dans une fricative voisée, l'énergie du signal filtré est élevée (présence du voisement) par rapport à la fricative non voisée correspondante, pour qu'une trame analysée soit déclarée non voisée, il faut qu'elle ait une énergie faible et une énergie de friction élevée. Le seuil de friction calculé est proportionnel à l'énergie du signal filtré.

La combinaison de ces trois paramètres délimite des zones de voisement (Fig.7,b) dans lesquelles le programme analyse les fonctions d'Amdf. Pour chaque trame, l'analyse consiste à repérer les positions des minima ainsi que leurs amplitudes (après normalisation), chaque minima définissant un candidat, la période T0 se manifestant sur la courbe d'Amdf par un creux profond. La figure 8 illustre cette caractéristique et on remarque sur celle-ci la présence des minima dûs aux harmoniques du signal ainsi que ceux dûs aux sous-harmoniques (2T0, 3T0 etc.).

Deux cas interviennent alors, suivant que la trame analysée se trouve en début de voisement, ou au milieu d'une zone de voisement pour laquelle la valeur de la fréquence fondamentale a précédement été estimée.

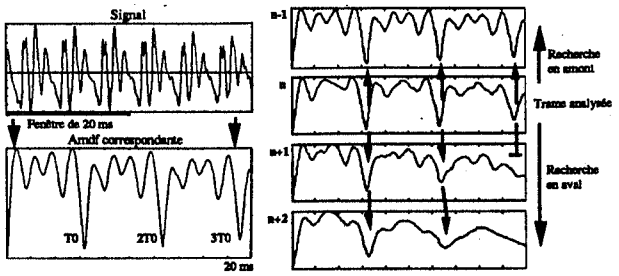


Figure 8: Exemple d'Amdf obtenue pour le son /a/ et exemple de chemins entre Amdfs.

Initialisation

- En début de voisement, l'algorithme considère le nombre de trames nécessaires (7 dans le cas général, moins si la zone voisée l'est sur moins de 7 trames), élimine de ces trames les candidats les moins probables (amplitudes trop élevées) et détermine des chemins de passage pour chacun des différents candidats sélectionnés, entre une trame considérée et ses trames voisines (Fig. 8). Enfin, par programmation dynamique, il choisit le chemin le plus probable en fonction des amplitudes des candidats se trouvant sur le chemin, de la longueur de ce chemin et des positions relatives des chemins entre eux en tenant compte de la présence de chemins parallèles localisés autour de 2T0, 3T0 etc. (Fig. 9).

Suivi

- Lorsque la période de la fréquence fondamentale a déja été estimée, l'algorithme recherche sur la fonction d'Amdf de la trame analysée, le minimum au voisinage de la position du minimum de la trame précédente, afin d'éviter les sauts d'octaves fréquents dans les détecteurs de fréquence fondamentale. La recherche de ce minimum sur l'Amdf cesse dès qu'une trame hors de la zone de voisement est rencontrée (Fig. 9).

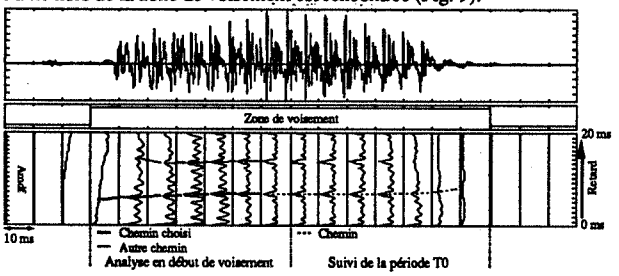


Figure 9: Détection sur une zone voisée /ka/.

Pour accélérer le traitement, les Amdfs sont évaluées en sous-échantillonant le signal à 3 kHz et les points des Amdfs sont calculés avec un pas de décalage de 0.4 ms (Amdf grossière). Avec un signal échantillonné à 16 kHz, le temps de calcul est divisé par un facteur 30. Cependant, afin de donner une valeur de la fréquence fondamentale avec la précision donnée par la vraie fréquence d'échantillonnage, on récalcule l'Amdf au voisinage de la période détectée (Amdf fine). La figure 10 donne le résultat de la mesure obtenu avec les sous-échantillonnages, ainsi que le résultat final obtenu après relocalisation locale de la période T0.

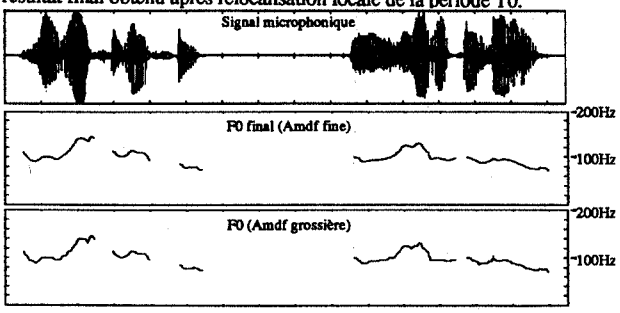


Figure 10: Allure des 2 courbes de F0 mesurées par l'algorithme.

L'algorithme à évaluer a donc été conçu pour mesurer en temps réel la valeur de la fréquence fondamentale. Sa conception est relativement simple et son avantage principal est qu'aucun paramètre n'est à fournir à celui-ci pour effectuer la mesure. Nous nous proposons maintenant d'en évaluer ses performances par comparaison avec les courbes mélodiques de référence données par l'analyse du signal laryngographique.

4 Evaluation

Nous présentons ici les mesures obtenues sur un corpus de test, composé de phrases extraites de la base de donnée EUROM et de 2 locuteurs enregistrés dans notre laboratoire. Nous avons sélectionné des enregistrements comportant des signaux microphoniques et laryngographiques de bonne qualité dont les caractéristiques sont:

- Locuteur N°1: Anglais, masculin, F0 moyen: 117Hz, F0 compris entre 73Hz et 198Hz, 74 secondes d'enregistrement (EUROM).
- Locuteur N°2: Danois, masculin, F0 moyen: 104Hz, F0 compris entre 49Hz (vocal fry) et 165Hz, 61 secondes d'enregistrement (EUROM).
- Locuteur N°3: Français, féminin, F0 moyen: 209Hz, F0 compris entre 80Hz et 300Hz, 32 secondes d'enregistrement (EUROM).
- Locuteur N°4: Français, masculin, F0 moyen: 126Hz, F0 compris entre 40Hz (vocal fry) et 230Hz, 65 secondes d'enregistrement.
- Locuteur N°5: Français, masculin, F0 moyen: 128Hz, F0 compris entre 32Hz (vocal fry) et 250Hz, 58 secondes d'enregistrement.

Afin d'effectuer la comparaison entre les deux courbes précédement présentées, nous avons adopté la convention qui suit.

Les deux mesures à comparer n'étant pas déterminées selon les mêmes critères, il a fallu les normaliser. En effet, l'algorithme à évaluer fournit des résultats à une fréquence fixe puisque les valeurs de la fréquence fondamentale par analyse de l'Amdf sont estimées toutes les 10 ms. L'algorithme de référence, quant à lui, isole toutes les périodes, si bien que dans le cas où l'intervalle de 10ms ne contient pas une seule période, il faut définir une valeur de référence unique afin de pouvoir effectuer la comparaison. Nous avons choisi de définir comme valeur de référence, pour chaque intervalle de 10 ms imposé par l'algorithme à évaluer, la valeur moyenne de la fréquence de base déterminée à partir du signal laryngographique. Dans le cas particulier des débuts ou des fins de voisement, cette valeur est considérée comme nulle si le voisement a lieu sur moins de la moitié de l'intervalle (Fig. 11).

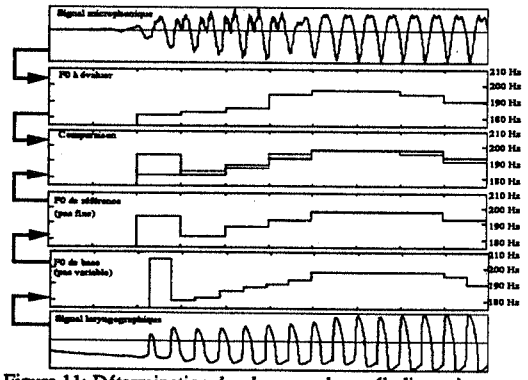


Figure 11: Détermination des deux courbes mélodiques à comparer.

En ce qui concerne les mesures, nous avons considéré les critères classiques d'évaluation [1,8] calculés sur l'erreur de mesure définie par:

$$Err(n) = \frac{Fref(n) - Fev(n)}{Fref(n)}$$

où Fref(n) est la fréquence donnée par l'algorithme de référence et Fev(n), celle donnée par l'algorithme à évaluer.

- Le nombre d'erreurs de mesure grossières correspondant à des erreurs de mesure supérieures (en valeur absolue) à 10%.

- La moyenne et l'écart-type des erreurs fines correspondant à des erreurs de mesure inférieures (en valeur absolue) à 10%.
- Le nombre d'erreurs de voisement (Fref(n)=0 et Fev(n)>0, notée NV=>V) et de dévoisement (Fref(n)>0 et Fev(n)=0, notée V=>NV).

Les valeurs numériques obtenues pour les divers paramètres sont les suivantes:

Locuteur	Moyenne	Ecart-type	Er. grossière	NV=>V	V=>NV
1	1,47%	1,78%	5,19%	9,65%	0,49%
2	1,43%	1,85%	7,16%	10,64%	0,91%
3	1,33%	1,72%	5,34%	7,43%	0,19%
4	1,26%	1,59%	3,36%	11,51%	0,25%
5	1,53%	1,66%	4,73%	11,18%	0,57%

Nous avons fait figurer sur la figure 12 les résultats obtenus pour la meilleure détection (débit lent, peu de variations prosodiques) et la plus mauvaise (débit rapide, variations prosodiques importantes).

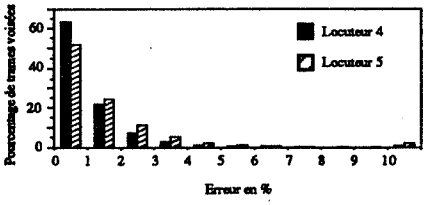


Figure 12: Erreurs de mesure pour deux des cinq locuteurs.

On constate que, pour les cinq locuteurs analysés, les erreurs moyennes sont proches de 1,5% et que les écarts-type restent faibles. Par contre, on remarque que le pourcentage d'erreurs grossières, tout en étant modéré, est élevé. Cependant, en localisant ces trames sur les courbes analysées (Fig.13), on s'apperçoit qu'elles se situent pour 50% en début ou en fin de voisement, et pour 50% au niveau de transitions de type consonne sourde, voyelle sonore ou dans les zones de vocal fry. En supprimant systématiquement les trames de début et de fin de voisement de l'analyse statistique, les pourcentages d'erreurs grossières sont donc diminués de moitié.

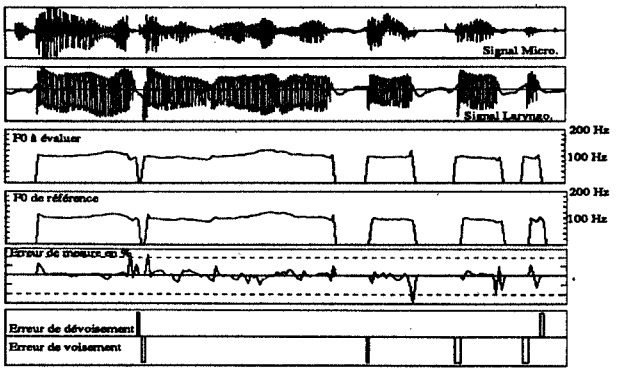


Figure 14: Comparaison et localisation des erreurs.

En ce qui concerne le taux de dévoisement (rapport du nombre de trames déclarées voisées sur le nombre total de trames voisées), celui-ci est très faible (inférieur à 1%), alors que le taux de voisement est relativement élevé. Ici aussi, l'analyse de la localisation de ces trames montre que 30% de ces trames se situent en début de voisement (plosives déclarés principalement), 70% en fin de voisement (propagation de l'onde acoustique entrainant une atténuation progressive du signal par rapport à la fin brusque de la vibration laryngée) et 0% en milieu de voisement (Fig. 14).

5. Conclusion.

L'algorithme de détection de la fréquence fondamentale à partir du signal microphonique, présenté dans cet article, se révèle être un bon algorithme. Cependant, les résultats obtenus peuvent encore être améliorés. Pour cela, nous réalisons un algorithme permettant de positionner des marqueurs de période (ou "Pitch mark") sur le signal microphonique à l'instant de fermeture glottique. L'indication sur la valeur de la période moyenne, nécessaire pour positionner ces marqueurs, sera identique à celle utilisée par cet algorithme (Analyse de l'Amdf par programmation dynamique sur 7 trames), le suivi étant fait par une méthode de réduction de données [7]. Ces marqueurs seront par la suite utilisés dans une synthèse "pitch-synchrone" paramètrique (extraction de formants (par analyse LPC par covariance) et de l'exitation glottique par filtrage inverse).

A titre indicatif, nous avons représenté à la figure 15 la mesure de référence obtenue sur le signal laryngographique, la détection de pitchmarks et le filtrage inverse pour un son /a/ et les performances comparées des deux algorithmes.

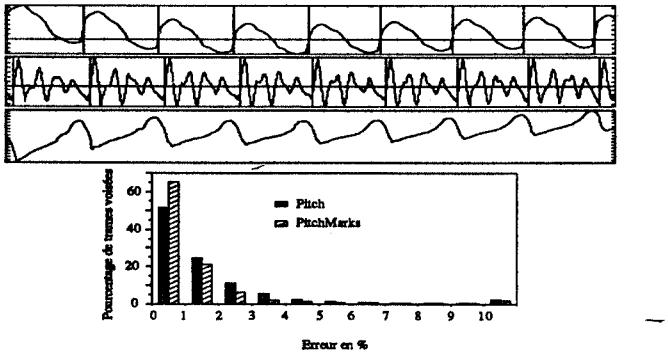


Figure 15: Pitchmarks, filtrage inverse et évaluation.

6. Bibliographie.

[1] Bailly G., "Détection du fondamental par prétraitement Amdf et programmation dynamique", SFA, 15ème Journées d'Etudes sur la Parole, Aix en provence, pp.285-288, Mai 1986.

[2] Chong K.U. & Shih-Chien Y.,"A pitch extraction algorithm based on LPC inverse filtering and Amdf",IEEE Trans. on Acous., Speech and Sig. Proc., 25, pp.565-572, Dec. 1977.

[3] Dubnowski J.J. Schaffer R.W. & Rabiner L.R.,"Real-time digital hardware pitch detector", IEEE Trans. on Acous., Speech and Sig. Proc., 24, pp.2-8, Feb. 1976.

[4] Hess W.J.,"Pitch determination of speech signals - Algorithms and devices", Springer, Berlin, 1983.

[5] Hess W.J. & Indefrey H.,"Accurate time-domain pitch determination of speech signals by means of a laryngograph", Speech communication 6, 1987.

[6] McGonegal C.A. Rabiner L.R. & Rosenberg A.E.,"A Semi Automatic Pitch Detector (SAPD)", IEEE Trans. on Acous., Speech and Sig. Proc., 23, pp.570-574, Dec. 1975.

[7] Miller N.J.,"Pitch Detection by Data Reduction", IEEE Trans. on Acous., Speech and Sig. Proc., 23, pp.72-79, Feb. 1975.

[8] Rabiner L.R. Chang M.J. Rosenberg A.E. & McConegal C.A.,"A comparative performance study of several pitch detection algorithms", IEEE Trans. on Acous., Speech and Sig. Proc., 24, pp.399-418, Oct. 1976.

[9] Rabiner L.R.,"On the use of autocorrelation analysis for pitch detection", IEEE Trans. on Acous., Speech and Sig. Proc., 25, pp.24-33, Feb. 1977.

[10] Rabiner L.R. Sambur M.R. & Schmidt C.E.,"Application of a nonlinear smoothing algorithm to speech processing", IEEE Trans. on Acous., Speech and Sig. Proc., 23, pp.552-557, Dec. 1975.

[11] Ross M.J. Shaffer H.L. Cohen A. Feudberg R. & Manley H.J.,"Average magnitude difference function pitch extractor", IEEE Trans. on Acous., Speech and Sig. Proc., 22, pp.353-362, Oct. 1974.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

REPRESENTATION TEMPS-ECHELLE ET DETECTION DE LA FREQUENCE FONDAMENTALE DU SIGNAL DE PAROLE

S.Montrésor, M.Baudry

Laboratoires d'Informatique et d'Acoustique de l'Université du Maine BP 535 - 72017 - LE MANS Cédex - France

Abstract

Pitch determination algorithm based on a time-scale representation of the speech signal is proposed. Phase spectrum calculated from the coefficients of the decomposition allows us to estimate the instantaneous frequency of the harmonics components of the signal. After presenting the caracteristics of this representation, we briefly described the method employed to extract the instantaneous frequencies from the phase spectrum. Several criterias are shown in order to estimate the fundamental frequency .Finally, we present the first results for which the algorithm has been applied to some test signals.
We discuss the validity of the method in some difficult cases, in particular when the fundamental frequency has weak energy in the power spectrum of the signal.

I - INTRODUCTION

Les performances des systèmes de traitement de la parole dépendent étroitement, très souvent, de la détection du fondamental. Plusieurs dizaines de solutions algorithmiques ont déjà été proposées jusqu'à maintenant. On peut les classer globalement en trois catégories: les méthodes temporelles, les méthodes fréquentielles et les méthodes à maximum de vraisemblance [1]. Chaque algorithme présente bien entendu son lot d'avantages et d'inconvénients suivant les conditions d'analyse (taux de bruit important ,ou distorsions non négligeables du spectre d'amplitudes). Nous présentons ici un algorithme n'entrant dans aucune des catégories évoquées plus haut, et qui repose sur une analyse en temps et en échelle du signal. Suscitant un intérêt grandissant dans la communauté du traitement du signal ,l'analyse temps-échelle semble receler des propriétés intéressantes pour l'étude du signal de parole, tant du point de vue de l'analyse pure [2], que de la représentation du signal de parole [3] ou encore de l'analyse-synthèse [4]. Nous tentons ici d'exploiter l'information de phase obtenue sur les coefficients complexes de l'analyse pour accéder à la structure harmonique du signal, et donc en particulier à la détermination de la fréquence fondamentale. Nous comparons notamment les diagrammes de phases obtenus à partir d'une analyse temps-échelle et obtenus par une analyse de Fourier à court-terme, dans plusieurs cas de situations difficiles: bruit important et distorsions du spectre d'amplitudes.

II - REPRESENTATION TEMPS-ECHELLE

Les premières applications d'analyse de signaux à l'aide de représentations temps-échelles ont été développées par le géophysicien J.Morlet [5] dans le cadre de la prospection pétrolière. Par la suite, une collaboration comprenant plusieurs mathématiciens et physiciens a contribué à construire un cadre mathématique rigoureux permettant d'interpréter les nombreuses simulations effectuées par J.Morlet [6][7], et qui constitue ce que l'on appelle aujourd'hui la transformée en ondelettes.
Le principe de l'analyse par ondelettes est de décomposer un signal sur une base de fonctions localisées en temps et en échelle [8]. Les différentes fonctions sont obtenues d'une part par translation et d'autre part par dilatation ou compression d'une fonction mère appelée ondelette analysante (1) :

$$\Phi_{a,b}(t) = \frac{1}{\sqrt{a}} \Phi\left(\frac{t-b}{a}\right) \qquad (1)$$

b est le paramètre de translation en temps et a est le facteur d'échelle. La fonction $\Phi(t)$ est soumise à des contraintes d'admissibilités pour être ondelette analysante, elle doit en particulier être de somme nulle. Nous nous limiterons dans la suite à des ondelettes analytiques de la forme (2) :

$$\Phi(t) = A(t)\ e^{-iw_o t} \qquad (2)$$

A(t) est l'enveloppe bornée de la fonction.

En posant :

$$f = \frac{w_o}{2\pi a} \qquad (3)$$

on peut interpréter la projection du signal sur les fonctions
$\Phi_{a,b}(t)$ en terme de représentation temps-fréquence à résolution relative constante. Les coefficients d'ondelettes sont obtenus en intégrant le produit du signal par les fonctions (4) :

$$C_{a,b} = \int s(t)\ \frac{1}{\sqrt{a}}\ \Phi\left(\frac{t-b}{a}\right)\ dt \qquad (4)$$

on obtient ainsi le module et la phase des coefficients complexes. Le choix pour l'ondelette Φ d'une fonction analytique permet d'interpréter, pour tout a fixé, la dérivée de la phase des coefficients, par rapport à t, en terme de fréquence instantanée. Prenons pour exemple le cas d'un signal monochromatique $s(t) = A\cos(wt+\phi_0)$,
sa transformée en ondelettes est donnée par (5) :

$$C(t,a) = Ae^{i\phi_0}F(aw)\ e^{iwt} \qquad (5)$$

ou F est la transformée de fourier de Φ.
On a :

$$\frac{d}{dt}\ Arg(\ C(t,a)\ ) = w \qquad (6)$$

Les figures 1 montre des diagrammes de phases obtenus sur des signaux de parole voisée. On distingue très nettement la structure harmonique du signal.

III - EXTRACTION DU FONDAMENTAL

La méthode proposée consiste à explorer la partie basse du spectre du signal qui correspond a la bande de base du signal de parole (entre 50 et 1000 hz). On se donne pour cela un jeu d'ondelettes dont les fréquences centrales se répartissent dans cette bande. On estime ensuite la fréquence instantanée associée à chaque ondelette en calculant au minimum deux coefficients autour du point ou l'on désire estimer le fondamental. L'accroissement de la phase mesuré pour ces deux coefficients donne alors la fréquence instantanée f_i :

$$\phi_{t_2} - \phi_{t_1} = 2\pi f_i \qquad (7)$$

Lorsque la fréquence centrale de l'ondelette balaye le spectre du signal, on retrouve une agglomération des f_i autour des harmoniques du signal. La donnée des fréquences de ces harmoniques permet alors le calcul du fondamental.

Une méthode similaire à été proposée par F.J. Charpentier [9] mais sur une implantation de type T.F. à court-terme.
Nous avons comparé des diagrammes de phase issus d'une analyse par ondelettes et d'une analyse de Fourier. Nous avons, en particulier, constaté une meilleure robustesse de l'analyse par ondelettes en l'absence du fondamental ou en présence de bruit.

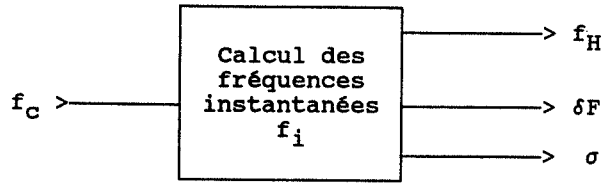
Plusieurs précautions s'imposent, en particulier pour le calcul des f_i. En effet lorsque la fréquence centrale du spectre de l'ondelette tombe sur une transition entre deux harmoniques, le spectre de phase présente quelques distorsions. Les rampes font alors place, en certains endroits, à des segments non linéaires. Les f_i mesurés n'ont alors plus la même signification (figure 2). Une technique simple pour estimer ce type de situation consiste à effectuer plusieurs mesures de f_i avec la donnée de plus de deux coefficients d'ondelette. L'écart à la moyenne de l'ensemble des f_i estimés est un indicateur de la stabilité de la courbe de phase obtenue pour une fréquence donnée. Ce f_i résultant est, bien entendu, la moyenne des f_i estimés. On dispose ainsi de trois paramètres, définis comme suit :

$$f_H = \frac{\sum_{i=1,n} f_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{n} (f_i - f_H)^2}{n}}$$

$$\delta f = f_H - f_c$$

où f_c est la fréquence centrale de l'ondelette.



Ce module constitue la brique de base de l'algorithme de détection. Un processus itératif réinitialise l'entrée du module en fonction de f_H, et assure ainsi une convergence assez rapide de l'ensemble. On s'arrête quand la quantité δf est suffisamment petite (de l'ordre du Hertz). On se situe alors sur une des harmoniques du signal.
La recherche du fondamental s'effectue en testant les différentes possibilités d'affecter un numéro de rang aux harmoniques trouvées. Le choix du bon rang utilise les critères définis précédemment.

IV - RESULTATS

Nous avons d'abord testé l'algorithme de détection sur différents types de signaux synthétiques pour s'assurer de son fonctionnement général correct, puis sur des extraits de signaux voisés normaux ou.soumis à diverses dégradations : atténuation du fondamental ou présence de bruit.
La figure 3 illustre le comportement de la méthode de recherche de fondamental dans ces divers cas. On observe une bonne tenue du détecteur.

L'évaluation a ensuite utilisé un corpus de phrases prononcées par huit locuteurs (quatres hommes et quatres femmes). Chaque locuteur a prononcé vingt phrases. Il y a 80 phrases différentes, phonétiquement équilibrées, extraites du corpus de P.Combescure [10].

Un algorithme complet de détection du fondamental se compose généralement de trois principaux modules :
- un module de prétraitement qui effectue une détection de frontière d'évènement, suivie d'une détection voisée-non voisée (quand elle n'est pas fournie par l'analyse elle-même),
- un module spécifique d'extraction de la fréquence fondamentale fournissant des données brutes,
- enfin un module de post-traitement qui se charge de supprimer les valeurs aberrantes éventuelles, et qui lisse la courbe initiale du fondamental.

Dans le cas présent, le module de prétraitement effectue uniquement une détection son-silence. La détection voisée-non voisée n'est effectuée à aucun moment dans l'algorithme actuel. Le module de post-traitement ne réalise pas de lissage de courbe du fondamental, mais élimine les valeurs trouvées lorsqu'il existe un écart trop important entre deux valeurs consécutives. Lorsqu'il n'y pas de rejet, la fréquence d'initialisation du détecteur est ajustée sur la valeur courante trouvée.
La figure 4 montre des exemples de courbes du fondamental obtenues sur le corpus utilisé. Elles révèlent un fonctionnement correct de l'algorithme dans toutes les zones de voisement du signal.

Remarquons également que les courbes de phases pourraient permettre l'accès à une analyse synchrone au fondamental (figures 2 et 3).

V - CONCLUSION

Nous avons présenté un algorithme de détection du fondamental reposant sur une analyse du signal en coefficients d'ondelette, associée à une méthode de phase stationnaire.

Les tests effectués sur un corpus de phrases phonétiquement équilibrées, prononcées par huit locuteurs dans des conditions normales d'enregistrement, ont montré un fonctionnement général correct de l'algorithme. Ils ont montré également sa robustesse lorsque le fondamental est filtré, ou en présence de bruit.

L'ensemble de ces résultats nous encourage donc à poursuivre dans cette voie. Pour le moment la simplicité des étages de pré et post-traitement n'a pas permis de tester l'algorithme dans des conditions difficiles:

D'autre part, cette méthode est assez peu coûteuse en temps calcul, et permet d'envisager une implantation temps réel.

BIBLIOGRAPHIE

[1] W. Hesse, "Pitch determination of speech signals", Springer Verlag, 1983.

[2] P. Goupillaud, A. Grossmann, J. Morlet, "Cycle-octave and related transforms in seismic analysis", Geo exploration 23 (1984/1985), Elsevier Sciences Publishers, BV Amsterdam, 85-102.

[3] R. Kronland-Martinet, J. Morlet, A. Grossmann, "Analysis of sound patterns through wavelet transforms", Int. J. Pattern Recog. Artif. Intell. pp 273-302, 1987.

[4] C. d'Alessandro, J.S. Lienard, "Decomposition of the speech signal into short time waveforms using spectral segmentation", ICASSP, New-York, 1988.

[5] P. Dutillaux, A. Grossmann, R. Kronland-Martinet, "Application of the wavelet transform to the analysis, transformation and synthesis of musical sounds", Preprint of the 88th AES convention, Los Angeles, 1988.

[6] Y. Meyer, "Ondelette, fonction spline et analyses graduées", Preprint, Université Paris-Dauphine, Paris.

[7] J.M. Combes, A. Grossmann, Ph. Tchamitchian, "Wavelet, time frequency methods and phase space", Springer Verlag, Berlin, 1989.

[8] R. Kronland-Martinet, A. Grossmann, "Time and scale representation obtained through continuous wavelet transforms", in Signal Processing IV, theories and applications (J.L. Lacoume et al, eds), pp 243-252, North Holland, 1988.

[9] F.J. Charpentier, "Pitch detection using the short term phase spectrum", pp 113-116, ICASSP, 1986.

[10] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées", Revue d'acoustique, n°56, pp 34-37, 1981.

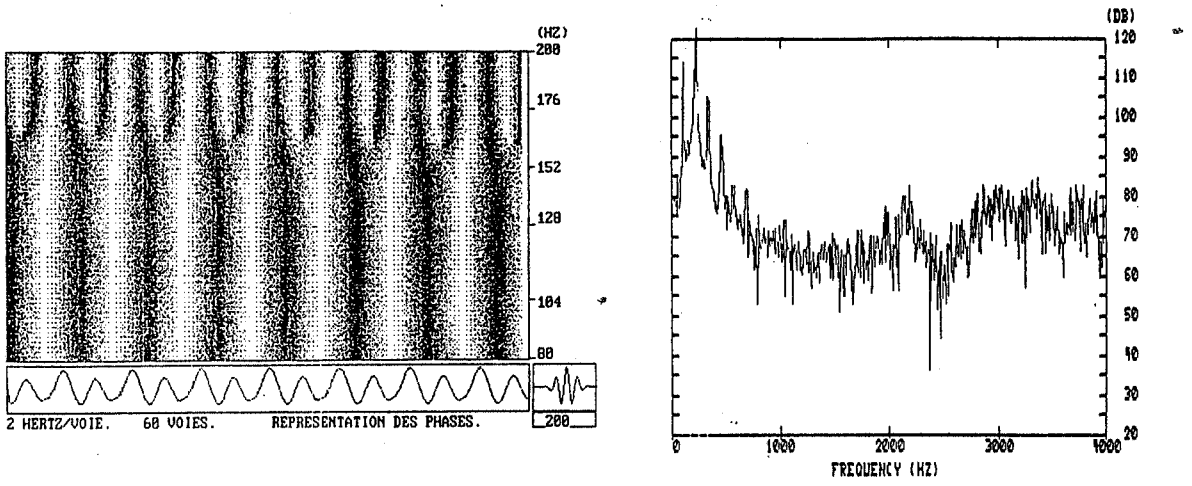


Figure 1a : A gauche, spectre de phases de la transformée en ondelettes d'un son /i/, de durée 72 ms. A droite, spectre FFT du signal analysé.

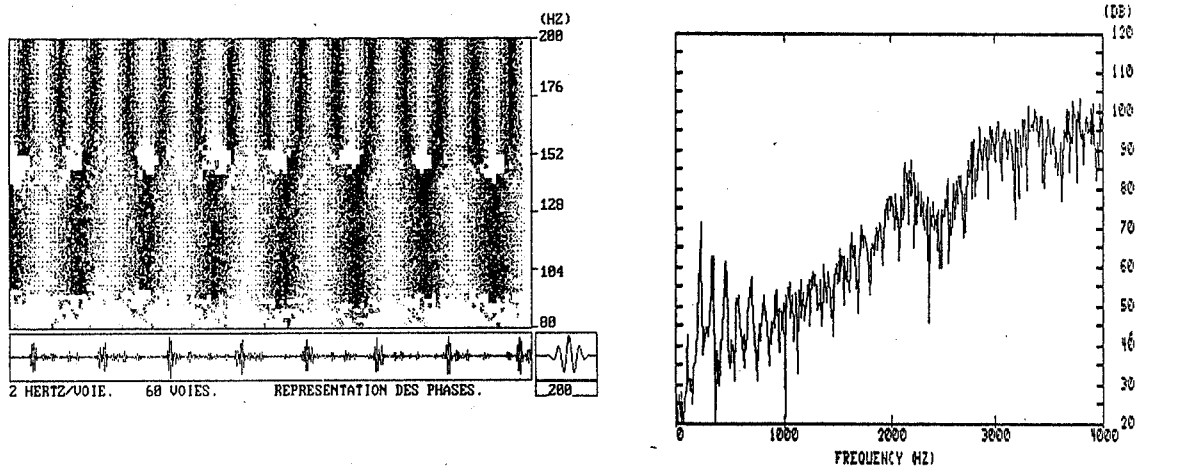


Figure 1b : A gauche, spectre de phases de la transformée en ondelettes d'un son /i/, filtré passe-haut, de durée 72 ms. A droite, spectre FFT du signal analysé.

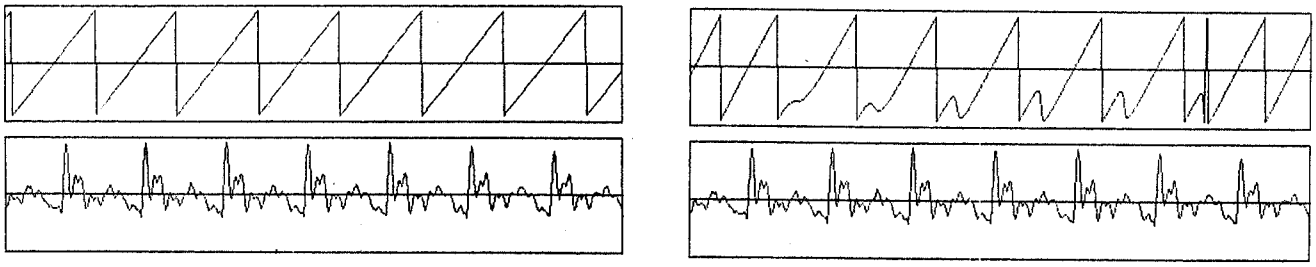


Figure 2 : A gauche, évolution de la phase d'une ondelette de fréquence centrale $f_c = 117$ Hz, analysant un son /a/ de $F_0 = 117$ Hz.
A droite, même signal analysé avec une ondelette de $f_c = 170$ Hz.

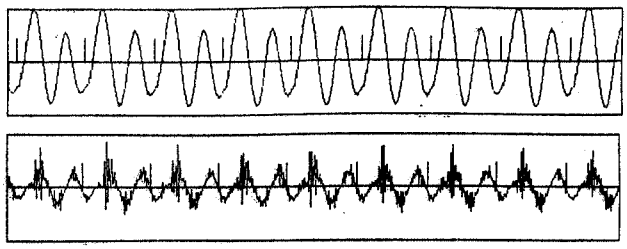


Figure 3a : son /i/ normal et filtré passe-haut.

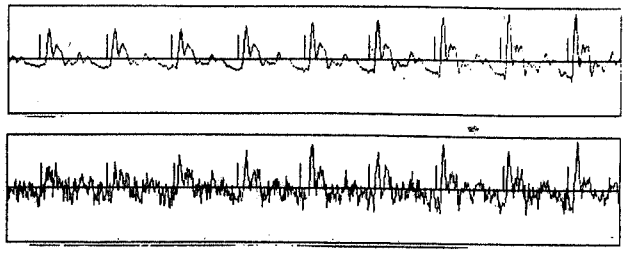


Figure 3b : son /a/ normal et bruité avec un RSB de 0 dB.

Figure 3 : Comportement de l'algorithme de détection sur des exemples de signaux voisés. Les zéros de phase du fondamental sont représentés par un trait vertical.

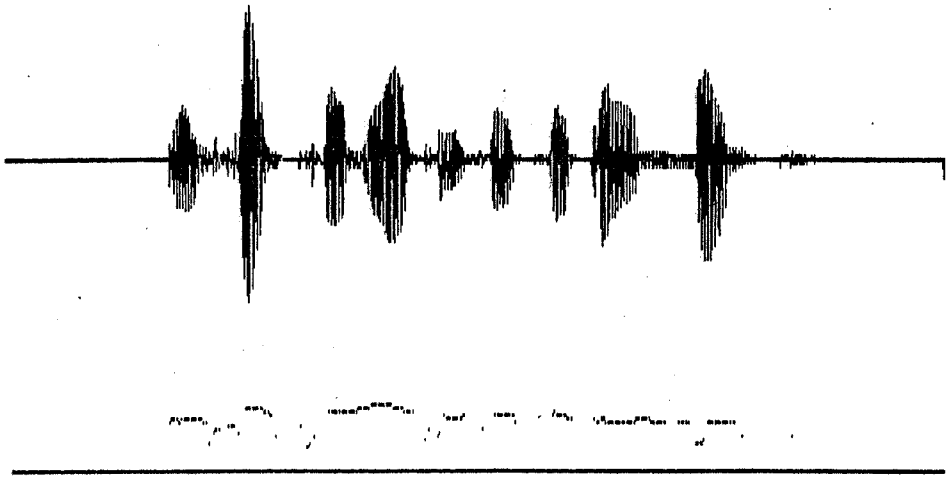


Figure 4a : Courbe de mélodie de la phrase
"un fort crédit est consenti par une banque"
prononcée par un locuteur masculin de F_0 moyen à 160 Hz.

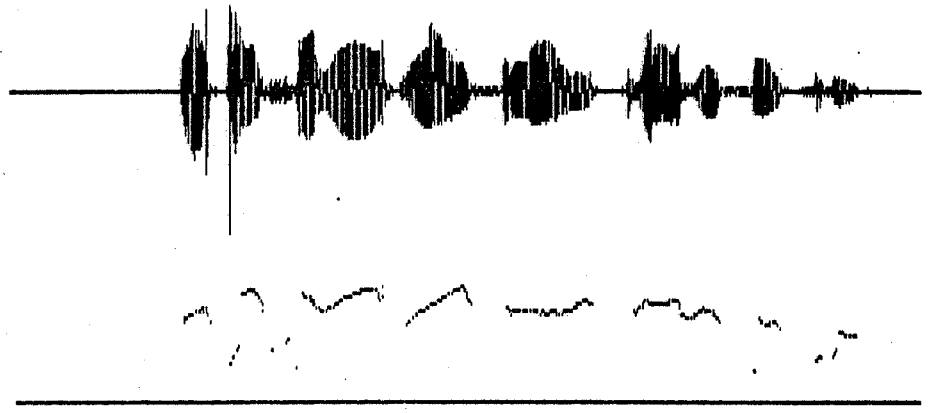


Figure 4b : Courbe de mélodie de la phrase
"Le passereau lance une roucoulade et s'enfuit"
prononcée par un locuteur féminin de F_0 moyen à 250 Hz.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Développement d'un poste de traitement de signal basé sur la programmation synchrone: application au traitement de la parole

C. LE MAIRE, R. ANDRE-OBRECHT, P. LE GUERNIC

IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France

Résumé

Cet article décrit une nouvelle approche pour le traitement de la parole: la programmation synchrone. Le langage SIGNAL est utilisé comme outil pour spécifier et programmer les algorithmes, pour décrire l'architecture des programmes et pour mettre au point les programmes en temps-réel.

1 Introduction

Les insuffisances de la programmation asynchrone traditionnelle pour la programmation de systèmes temps-réel ont conduit à l'élaboration d'un nouveau type de programmation: la programmation synchrone. Celle-ci a notamment été à la base du développement de trois principaux langages: SIGNAL [6] , ESTEREL [3] et LUSTRE [2] .
Dans l'étude que nous présentons, nous appliquons les concepts de la programmation synchrone par le biais du langage SIGNAL et développons des outils afin de programmer les différents étages d'un décodeur acoustico-phonétique, puis construire ce décodeur et enfin développer un environnement autour de ce décodeur. Pour des raisons de clarté, nous nous bornons au module complexe de segmentation [7] et présentons les principaux mécanismes utilisés pour le construire ainsi que son environnement.

2 SIGNAL: un langage synchrone à flots de données pour la programmation de systèmes à temps-réel

Les caractéristiques temps-réel (temps de réponse bornés) et synchrone (déterminisme) permettent de décrire facilement le comportement temporel d'un algorithme. Dans un langage temps-réel synchrone, on fait l'hypothèse que toutes les actions ont une durée d'exécution nulle. On peut donc parler de vue *chronologique* du temps (temps logique) par opposition à la vue *chronométrique* des systèmes asynchrones (temps physique).
La caractéristique flots de données permet de décrire facilement le parallélisme. L'ensemble des calculs à effectuer est représenté par un graphe orienté, dans lequel les arcs sont des chemins de données et les nœuds des opérations. Lors de l'exécution, un nœud de calcul est activé lorsque toutes ses entrées sont prêtes (règle flots de données). La notion de variable est remplacée par celle de flot (suite de valeurs).

Les objets de base sont appelés signaux. Un signal est une paire (flot, horloge) où le flot est une suite ordonnée de données typées de longueur non spécifiée et l'horloge associée à ce flot spécifie les instants auxquels les données sont disponibles. Aussi la notion de temps dans un système est définie par les relations entre les horloges des différents signaux du système, sans faire référence à un temps universel.
Un processus SIGNAL décrit à la fois les relations fonctionnelles et temporelles entre les signaux. Un processus est une boîte noire pouvant communiquer avec le monde externe ou d'autres processus par l'intermédiaire de signaux appelés ports d'entrée et ports de sortie.
Un programme SIGNAL est représenté par un réseau statique de processus interconnectés, ces derniers pouvant être eux-mêmes composés de sous-processus.

Le noyau du langage SIGNAL comprend cinq instructions de base:

- un opérateur *composition* permet la construction de réseaux et exprime le parallélisme entre les processus (on note P|Q). D'autres opérateurs dérivés existent (les compositions séquentielle, parallèle...)
- trois processus statiques — les valeurs des ports de sortie dépendent des valeurs des ports d'entrée aux instants où celles-ci sont disponibles, sans tenir compte des valeurs passées — :
  - les *fonctions* sont des transformations instantanées élémentaires sur les données (arithmétiques, logiques, etc). Par exemple,
    "$x := y + z$" est équivalent à, quelquesoit t, $x_t := y_t + z_t$
  - le *filtrage* "x:=y when c" est un sous-échantillonnage conditionnel. La sortie x prend la valeur de y lorsque le signal booléen est présent et possède la valeur vraie
  - le *mélange* "x:=u default v" donne la priorité au signal u lorsque u est présent, sinon x prend la valeur de v si v est présent
- un processus dynamique, le *retard* "x:=y $ 1" exprime la connaissance du passé, et est équivalent à, quelquesoit t supérieur à 1, $x_t := y_{t-1}$.

La construction de réseaux se fait par interconnexion de processus en identifiant les noms de ports. Cela impose l'utilisation de mécanismes de *renommage* entre ports d'entrée et ports de sortie, de *masquage* de ports, de *rebouclage* pour connecter une sortie et une entrée d'un même port.

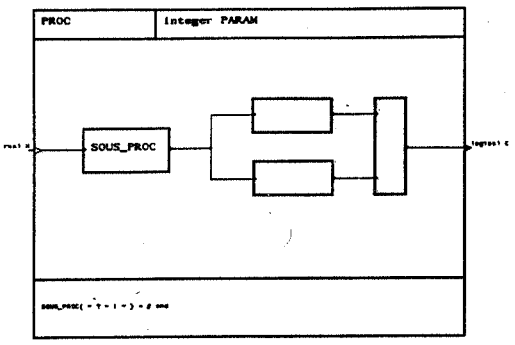


Figure 1 : Vue graphique d'un processus de nom PROC formé de plusieurs sous-processus. Les boîtes représentent les processus, les fils représentent les liens entre les ports d'entrée/sortie.

Les outils développés autour du noyau
Afin de faciliter l'utilisation de SIGNAL, quelques outils standards ont été développés.
Exemples d'outils statiques:

- l'accès à l'horloge d'un signal "hx:= event x"
- la mémorisation d'un signal "x:= y cell c" mémorise y à l'horloge de (c=vrai)
- la composition conditionnelle "if c then P else Q"
- la synchronisation explicite "synchro $x_1, ..., x_n$" contraint les signaux à être synchrone (c.a.d. possédant la même horloge)
- les tableaux de processus "array i to n of P with <signaux d'entrée> end" permettent de disposer des structures de répétition

Exemples d'outils dynamiques:

- les compteurs: "x:=#c" compte le nombre d'occurences du signal booléen c à vrai
- les fenêtres: "wx:=x window n" définit une fenêtre glissante sur x de dimension n
- l'extension du retard: "x:= y $ m" définit le signal x retardé de m par rapport au signal y

3 Un outil pour la spécification et la programmation des algorithmes

Une méthode facilitant l'écriture d'un programme SIGNAL est présentée dans [5] . Les algorithmes étudiés se composent de deux parties distinctes:

- une partie spécifiant les formules utilisées, c.a.d. des systèmes d'équations
- une partie spécifiant les synchronisations, c.a.d. par exemple le traitement du signal échantillon par échantillon ou par blocs glissants, le traitement des périodes d'initialisation...

Les systèmes d'équations sont immédiatement transposables en SIGNAL. En effet l'opérateur de composition est commutatif et associatif et assimilable à l'union d'équations ou de systèmes d'équations. Par exemple

$$\left|\begin{array}{l} x_t = y_t + z_t \\ y_t = (y_{t-1} + z_t) - 10 \\ z_t = (2 * n) + 1 \end{array}\right. \quad \text{se traduit par} \quad \begin{array}{l} (| \, x := y + z \\ | \, lasty := y\$1 \\ | \, y := (lasty + z) - 10 \\ | \, z := (2 * n) + 1 \\ |) \end{array}$$

Les parties synchronisations sont en général des opérations temporelles non adaptées à des langages classiques tels que FORTRAN ou PASCAL, et réalisées en SIGNAL à l'aide d'opérateurs spécifiques:

- le traitement du signal échantillon par échantillon est trivial. Le signal de parole est considéré comme objet signal, les échantillons étant les valeurs successives du flot
- le traitement du signal par blocs glissants est transparent au programmeur qui utilise l'outil "window"
- les périodes d'initialisation sont traitées à l'aide de compteurs. Par exemple, si un test n'est actif qu'après une période de longueur n, on écrira

```
(| ny := #(event y)      % compteur d'evenements y
 | actif := ny ≥ n       % signal booleen
 | testactif := true when actif
 |) !! testactif         % testactif est seul visible de l'exterieur
```

La figure 2 montre l'entrelacement des 4 signaux. Le signal actif est présent à chaque évènement y tandis que le signal testactif n'est présent qu'aux instants où actif est égal à vrai.

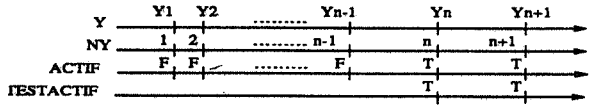


Figure 2 : les horloges des signaux y, ny, actif et testactif.

- la resynchronisation de signaux à l'aide des opérateurs mémorisation et filtrage. Considérons l'exemple où l'on cherche à détecter les maxima significatifs d'une courbe, à savoir qu'un maximum est validé a posteriori après franchissement d'un seuil par la différence entre les valeurs suivantes de la courbe et ce maximum. Soit en SIGNAL:

```
(| nouveaumax := u ≥ zmax
 | zmax := max $ 1
 | max := (u when nouveaumax) default zmax
    % calcul du max
 | detection := true when (max − u ≥ seuil)
    % seuil franchi
 |)
```

Pour obtenir la distance entre deux détections (la frontière est située au maximum validé), il faut compter le nombre d'occurences de Y lors d'un nouveau max, puis le mémoriser jusqu'à l'instant de détection. Soit :

```
(| ny := #(event y) after raz
    % compteur de y depuis la derniere frontiere
 | lseg := ny when nouveaumax
 | lsegment := (lseg cell detection) when detection
    % resynchronisation de lseg a detection = vrai
 |) !! lsegment
```

4 Un outil pour l'architecture des programmes

SIGNAL permet de composer les modules pour construire de façon hiérarchique des systèmes tel qu'un système de reconnaissance de parole. Les modules ne sont pas obligatoirement écrits en SIGNAL mais peuvent être par exemple des modules FORTRAN.
Afin de faciliter les constructions de ces modules et réseaux, un environnement graphique a été développé [4] , ce qui permet d'avoir une vue du programme soit textuelle soit graphique, ce dernier cas laissant apparaître une vue modulaire et hiérarchique du programme.

4.1 Un exemple: la méthode de divergence forward-backward [1]

Cette méthode est basée sur un test statistique, le test de divergence Hinkley.

Le test de divergence Hinkley dérive d'une mesure de distance entre deux modèles autorégressifs d'ordre M, auquel un biais positif est ajouté. Une frontière est détectée après franchissement d'un seuil de la différence entre la valeur courante du test u et le maximum courant max (figure 3).
Dans notre cas, une phase d'initialisation de longueur L est nécessaire pour l'identification des modèles autoregressifs. Un des modèles est identifié séquentiellement sur une fenêtre croissante par la méthode de Burg, l'autre est identifié sur une fenêtre glissante par la méthode d'autocorrélation. Lors d'une détection à l'instant n, le test est réinitialisé et redémarre à l'instant de rupture r'.

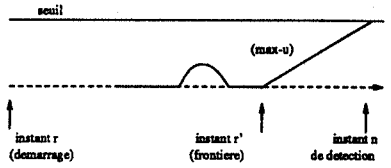


Figure 3 : Le test de divergence Hinkley.

Le principe du retour-arrière: lorsqu'un segment [r,r'] est détecté, et jugé trop long, le test de Hinkley lui est appliqué dans le sens rétrograde. Si une nouvelle frontière r" est détectée, la segmentation reprend dans le sens direct à partir de r". Sinon elle repart en r'.

4.2 Le traitement en SIGNAL

La construction du module de Hinkley
Nous disposons de quatre modules. Deux modules FORTRAN, AUTO et BURG, identifiant les modèles autoregressifs, un module SIGNAL calculant la statistique u (à l'aide des systèmes d'équations) et un module SIGNAL calculant le maximum max et la longueur du segment (r'-r).
Le module SEGMENTATION_HINKLEY est construit en assemblant ces 4 modules avec un module gérant la période d'initialisation pendant laquelle la statistique n'est pas calculée (figure 4). De l'extérieur, le module est vu comme une boîte noire, seuls ses ports d'entrée et sortie étant visibles. Il peut ainsi être inséré à un niveau supérieur.

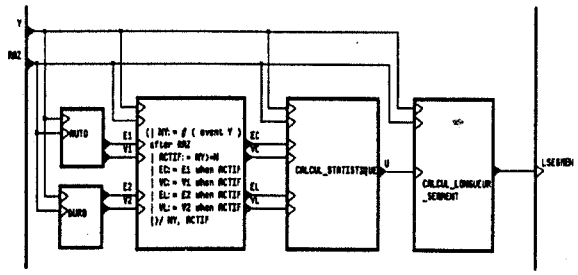


Figure 4 : Le module SEGMENTATION_HINKLEY.

La construction du module de Hinkley avec le retour-arrière
Nous appelons reprise, la période pendant laquelle le signal est retraité. C'est le cas lors d'un retour-arrière (réinjection rétrograde du signal de l'instant r' à l'instant r) ou lors d'un redémarrage (réinjection directe de l'instant r" (ou r') à l'instant n).
Le problème de la gestion des reprises est résolu en SIGNAL en suréchantillonnant le signal d'entrée comme le montre la figure 5 d'entrelacement des signaux.

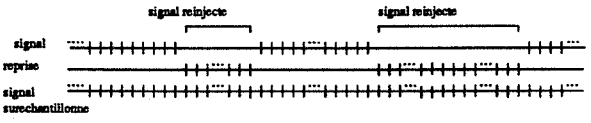


Figure 5 : Les horloges des signaux reprise, signal et du signal suréchantillonné

Le signal d'entrée , dont une partie est stockée dans une fenêtre glissante, est synchronisé de façon explicite avec un signal indiquant que l'on n'est pas en reprise. Cette gestion des reprises s'écrit en SIGNAL:

```
(| wsignal := signal window lfenetre
 | wy := (wsignal cell reprise) when reprise
    % resynchronisation de la fenetre pendant la reprise
 | y := signal default wy[irep]
    % irep est l'indice courant de reprise dans la fenetre
 | onprendsignal := true when (not reprise)
 | synchro signal, onprendsignal
    % synchronisation explicite de signal
 |) !! y
```

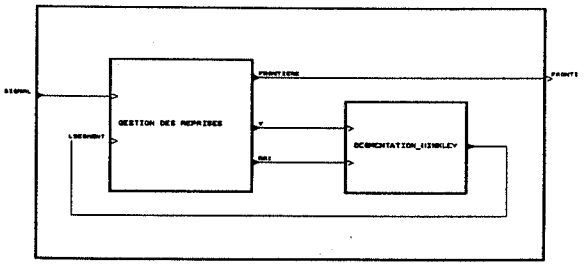


Figure 6 : Le processus SEG_FORWARD_BACKWARD construit à partir du processus SEGMENTATION_HINKLEY.

5 Un outil pour la mise au point de programmes en temps-réel

L'approche synchrone permet la lecture des échantillons du signal au fur et à mesure de l'exécution du programme. De la même façon, elle permet de produire des résultats au cours de cette exécution. Pour tirer avantage de cette approche, nous développons en SIGNAL une boîte à outils permettant l'exploitation immédiate de ces résultats. Nos développements ont pour but l'intervention de l'opérateur pendant l'exécution ou la visualisation des résultats sans modifier le programme source. Pour cela, nous développons un programme SIGNAL d'affichage associé au programme source.

Nous avons développé deux sortes de processus autour du programme SIGNAL source:

- un processus "sonde" permettant d'observer sans modifier l'exécution du programme. Ce processus est associé à un port du programme. La figure 7 montre un processus sonde associé au signal de parole.

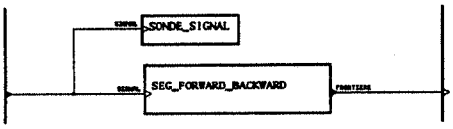


Figure 7 : Un processus sonde est associé au signal de parole.

  Un tel processus n'a pas de sortie dans le programme source et est déclaré en externe afin d'être analysé par le programme d'affichage

- un processus "debug" permettant d'intervenir sur le déroulement de l'exécution est associé à un lien entre deux ports. Par exemple, nous pouvons interdire manuellement le retour-arrière (figure 8).

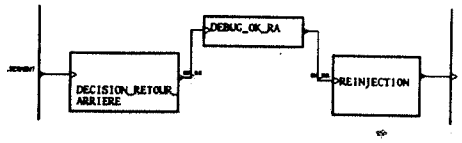


Figure 8 : Un processus debug est associé au signal de retour-arrière

  Le processus debug_ok_ra est un processus SIGNAL, lequel synchronise un appel externe lié au clavier ou à la souris, avec l'horloge du signal debug_ok_ra. Ces processus debug nous permettent pendant l'exécution, de régler les seuils, d'autoriser ou refuser le passage de signaux, d'autoriser ou refuser l'utilisation de modules, etc.

Le programme SIGNAL d'affichage prend en entrée les signaux liés aux processus sonde. Le programme est construit à partir de ces entrées et permet le tracé de courbes, la gestion d'écran, etc. Prenons l'exemple sonde_signal. Nous souhaitons gérer l'écran de façon circulaire, c'est à dire gérer une bande d'effacement de largeur L qui précède ce que nous traçons à l'écran. Le programme SIGNAL est le suivant (nous traçons pixel par pixel):

$(|\ v := (0\ when\ zv + 1 = longueurecran)\ default\ zv + 1$
$|\ zv := v\ \$\ 1$ *% init largeur L*
$|\ effacer\ ?x1 : v, y1 : 0, x2 : v, y2 : hautecran$
% proc. externe qui efface un trait du pt$(x1, y1)$ *au pt* $(x2, y2)$
$|\ synchro\ v, signal$ *% on efface a chaque event signal*
$|)\ /\ v, zv$ *% masquage*

Application à la segmentation automatique en reconnaissance de la parole continue. Nous avons développé sous SunView sur SUN un environnement autour d'un module complexe de segmentation acoustique. Trois modules opèrent en parallèle, deux modules traitant le signal échantillon par échantillon – une segmentation Hinkley avec retour-arrière et une segmentation Hinkley sur le signal filtré – et un module traitant le signal par blocs glissants – une segmentation voisé-non voisé –. Sur la figure 9 nous pouvons voir les différentes fonctions offertes à l'utilisateur (réglage de l'ordre des modèles, interruption d'un ou plusieurs modules, exécution pas à pas...)

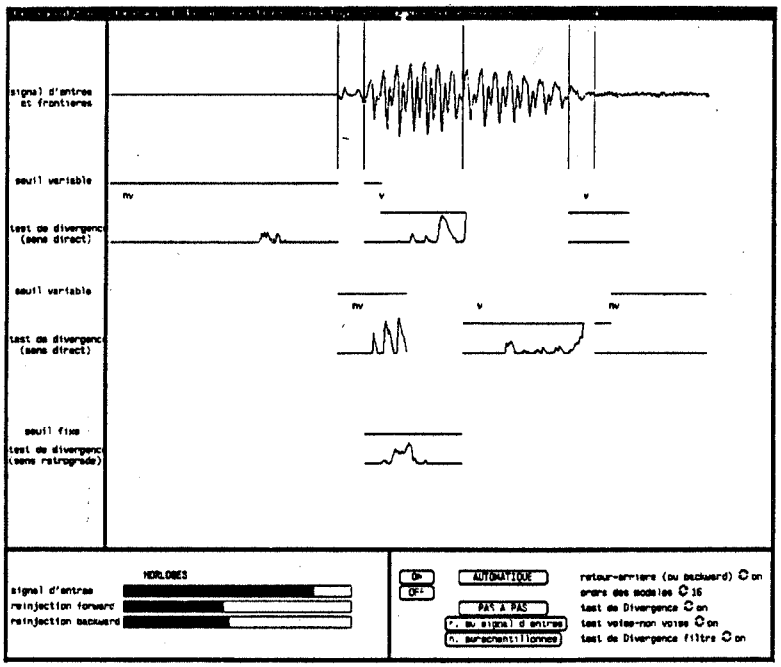


Figure V.10 : Environnement synchrone autour d'un module de segmentation.

6 Conclusion

Le langage SIGNAL permet une description interne du parallélisme et une approche synchrone adaptée à la classe d'algorithmes étudiés. Il possède un outil graphique permettant d'avoir une vue modulaire et hiérarchique du programme. On peut ajouter une vérification complète du comportement temporel d'un programme (absence de blocage...) grâce à un mécanisme de calcul d'horloges qui transforme un programme en un système d'équations.
Actuellement le compilateur génère du code FORTRAN simulant ainsi l'approche synchrone. Un translateur de programmes SIGNAL en OCCAM permettant une future mise en œuvre de ces programmes sur un réseau de transputers est en cours de réalisation. De plus des études sont en cours pour répartir le graphe obtenu à la compilation (caractéristique flots de données) sur une architecture multiprocesseurs.
Pour toutes ces raisons, SIGNAL se présente comme un langage spécifique au traitement du signal et bien adapté pour spécifier et mettre en œuvre des applications en parole.

Bibliographie

[1] R. ANDRE-OBRECHT : *A New Approach for the Automatic Segmentation of Continuous Speech Signals* ; IEEE Trans. on ASSP, ASSP-36 No 1, pp. 29-40, January 1988.

[2] J.L. BERGERAND, P. CASPI, N. HALBWACHS, D. PILAUD, E. PILAUD : *Outline of a real-time Data-Flow Language* ; in Real-Time Systems Symposium, San Diego, December 1985.

[3] G. BERRY, L. COSSERAT : *The ESTEREL Programming Language and its Mathematical Semantics* ; Research report 327, INRIA FRANCE, 1984.

[4] P. BOURNAI, V. KERSCAVEN, P. LE GUERNIC : *Un environnement graphique pour la conception d'applications temps-réel* ; IN2-INRIA-LRI, Cargese, Corse, France, Mai 1989.

[5] F. DECHELLE, Y. SOREL : *Utilisation du langage SIGNAL pour la spécification et la mise en œuvre d'algorithmes de traitement du signal* ; GRETSI, NICE, pp. 657-660, Juin 1987.

[6] P. LE GUERNIC, A. BENVENISTE, P. BOURNAI, T. GAUTIER : *SIGNAL—A Data Flow-Oriented Language for Signal Processing* ; IEEE Trans. on ASSP, ASSP-34 No 2, pp. 362-374, April 1986.

[7] C. LE MAIRE : *Le langage SIGNAL: un Exemple en Segmentation Automatique de la Parole Continue* ; Rapport de recherche, INRIA FRANCE, mars 1990, à paraître.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

UNE STATION DE TRAVAIL D'ANALYSE DE LA PRODUCTION DE LA PAROLE

Bernard TESTON et Benoît GALINDO

INSTITUT DE PHONETIQUE
URA 261 du CNRS
Université de Provence
13621 AIX en PROVENCE

ABSTRACT

We describe a work station for speech production analysis which enables the recording, measurement, marking, segmentation, and processing of sixteen physiological parameters (kinesiography, aerophonometry, electromyography, and electropalatography) in perfect synchronisation with the acoustic signal (spectral analysis) and videofilms of the vocal tract. Articulatory acoustic correlations can now be adequately studied. Given the massive amount of data to be analysed, particular attention was paid to the ergonomical aspects of the work station.

1-INTRODUCTION :

Une des difficultés majeures que l'on rencontre dans la mise en oeuvre des systèmes de reconnaissance automatique de la parole réside dans la grande variabilité de la relation entre les stuctures phonologiques d'une phrase et le niveau de sa représentation acoustique (1). La coarticulation semble en être la raison principale. La possibilité d'utiliser l'information apportée par les phénomènes de coarticulation a été jusqu'a maintenant presque totalement négligée (2). De récents travaux, aussi bien théoriques qu'experimentaux (3), montrent clairement la necessité d'une recherche fondamentale d'envergure sur les mécanismes de production de la parole. Une étude comparée sur plusieurs langues doit contribuer à élargir nos connaissances, non seulement sur leur spécificité mais également, sur les aspects universels de la production de la parole. Pour cela, une action ayant pour objectif, d'identifier et d'interpreter les phénomènes de coarticulation a été mise en oeuvre dans le cadre d'un contrat ESPRIT II BRA (Basic Research Action) ACCOR (Articulatory-Acoustic Correlation in Coar-ticulatory processes), à l'initiative de A.MARCHAL (4). Elle a un double objectif. D'une part, d'identifier et d'interprèter les phénomènes de coarticulation. D'autre part, de les intégrer dans un modèle qui doit permettre de déduire les représentations articulatoires à partir de l'analyse acoustique de la parole.
Cette nouvelle démarche doit conduire à l'amélioration des technologies vocales, et contribuer en particulier à la mise au point de systèmes de reconnaissance automatique plus fiables (5), de synthèse articulatoire de bonne qualité à partir d'un taux d'information réduit, et de synthèse par règles multilingues (6).

Pour mener à bien l'étude des phénomènes de coarticulation il est nécessaire de disposer du maximum d'informations sur les mouvements des organes articulateurs, ou de leurs manifestations induites, en synchronisation avec le signal de parole. Il est également nécessaire de la réaliser sur un corpus important, répété plusieurs fois, par un grand nombre de sujets et, avec la même méthodologie expérimentale chez tous les partenaires du contrat. Pour cela, un des premiers travaux nécessaire à cette action a été d'étudier et de réaliser une station de travail d'analyse de la production de la parole destinée à acquérir, éditer, traiter, et interpréter les paramètres articulatoires et acoustiques à partir des différentes techniques de capture de leurs manifestations.

2-LES DIFFERENTS SYSTEMES D'ANALYSE DES MOUVEMENTS ARTICULATOIRES :

Il est possible d'analyser les mouvements des organes articulateurs à trois niveaux différents. Tout d'abord, au niveau des commandes neuromotrices des muscles mis en jeu dans la parole, ensuite au niveau direct de leurs manifestations, enfin au niveau des phénomènes qu'ils induisent.
Le premier niveau est représenté essentiellement par l'électromyographie (7). Le second englobe l'analyse du mouvement, soit au moyen des techniques d'imagerie cinématographique (films du visage de face et de profil ,cinéradiographie du conduit vocal dans le plan sagittal), soit au moyen de capteurs de déplacement (kinésiographie représentée par les dispositifs MOUVETRACK (8) et ARTICULOGRAPH (9)) auxquels on peut rattacher l'électropalatographie (EPG) qui permet de rapporter sur un plan la topographie des contacts lingaux palatins (10). Le troisième niveau est représenté par les phénomènes aérodynamiques et acoustiques qui évoluent tout au long du conduit vocal (aérophonométrie) et, dont l'ultime réalisation représente le signal de parole porteur d'information. Il est possible d'obtenir de cette méthode des renseignements trés importants sur les variations du conduit vocal (11). Elle est surtout utilisée pour l'étude des occlusions buccales et vélaires (12 et 13).
Dans le cadre de l'action ACCOR, la kinésiographie l'électropalatographie et l'aérophonomètrie doivent être utilisées en priorité pour leur facilité de mise en oeuvre et leur absence d'egressivité pour les sujets. Cependant, il est prévu sur la station de travail, de pouvoir traiter des signaux électromyographiques et de synchroniser des vidéofilms.

3-DEFINITION DE LA STATION DE TRAVAIL :

A-Caractéristiques d'acquisition:

Notre objectif est d'analyser le plus grand nombre possible de paramètres sur le fonctionnement du conduit vocal. Le point particulier réside dans la simultanéité d'acquisition des différentes données. La fréquence d'acquisition est fonction du taux d'information nécéssaire à la description convenable des différents paramètres.
Les signaux issus des capteurs kinésiographiques, aérophonomètriques et électromyographiques fluctuent au rythme des mouvements des organes articulateurs, une largeur de bande de 1 kHz est suffisante pour décrire leurs variations avec une bonne précision. Elle permet également une bonne synchronisation avec les vidéofilms au standart européen de 50 trames par seconde. L'EPG représente un cas particulier par le fait qu'il est nécessaire de multiplexer les 64 contacts palatins dans un cycle d'observation dont la durée maximale ne doit pas dépasser 10 ms. La largeur de bande du signal de parole est choisie à la valeur de 8 kHz en conformité avec le programme européen SAM. La dynamique des signaux physiologiques et de la parole naturelle étant de l'ordre de 60 dB, une résolution de conversion de 12 bits a été jugée suffisante.

B-Choix du système de base:

Le standart PC AT et MS DOS nous paraissent être les mieux adaptés à nos besoins pour les raisons suivantes : puissance de calcul suffisante avec le couple 80286-87, grand choix de cartes additionnelles d'acquisition et de traitement, bonne résolution graphique en VGA (640-480) suffisante pour l'édition des signaux, enfin, cout modéré des materiels au standart PC. Notre choix s'est porté sur un PC HEWLETT-PACKARD Vectra ES12 équipé d'un disque dur de 80 Mo et d'une sauvegarde de 60 Mo (Figure 1).
Le système d'acquisition est centré sur une carte BURR-BROWN PCI 20041C-2 équipée de 40 entrées analogiques multiplexées et de 2 sorties. Sa résolution est de 12 bits et sa fréquence d'acquisition maximale de 89 kHz.La configuration est complétée par un moniteur monochrome alphanumérique, un moniteur couleur VGA et une souris.

4-ACQUISITION DES PARAMETRES ARTICU-LATOIRES:

A-Principe général :

Bien que ses caractéristiques soient plus ambitieuses, nous nous sommes inspirés, pour la réalisation de notre système d'acquisition, de la philosophie de la carte développée par IBM pour supporter l'EPG et quelques autres paramètres (14). Nous disposons de 40 entrées analogiques de 1 kHz de largeur de bande, réparties en 5 groupes de 8: 1 groupe pour l'entrée acoustique N°1 un autre pour l'entrée acoustique N°2, 2 groupes pour les 16 entrées physiologiques et 1 groupe pour l'EPG. Grace a la possibilité de programmation aléatoire du multiplexeur, on peut faire varier la fréquence d'échantillonnage de 16 kHz (pour un groupe de 8 entrées) à 80 kHz (pour les 5 groupes). Le système d'acquisition est ainsi trés souple, il peut s'adapter à chaque expérience sans encombrer le système de traitement.

B-L'interface d'acquisition :

Il est constitué par les circuits de conditionnement des signaux physiologiques et acoustiques, des circuits d'adaptation de l'EPG et de génération de l'horloge d'échantillonnage. Les 16 entrées physiologiques sont équipées de filtres antirepliement dont la fréquence de coupure est de 1 kHz. Le niveau d'entrée est de + ou - 10 Volts. Il est possible de supprimer le filtre sur une entrée, elle est alors réservée aux signaux de synchronisation des vidéofilms.
Les deux entrées acoustiques sont équipées de filtres identiques avec une fréquence de coupure de 8 kHz (également 16 kHz pour le canal 1), de préamplificateurs, de vu-mètres et, de détecteurs de seuil pour les acquisitions automatiques.
Les deux sorties acoustiques sont équipées de filtres identiques aux entrées et d'amplificateurs de puissance pour l'écoute au casque ou sur haut-parleur. L'EPG choisi pour l'action ACCOR, a été développé par l'Université de READING. Il est commercialisé par la firme MILLGRANTS Ltd. Nous avons réalisé, pour sa connection, un dispositif de liaison qui permet l'acquisition des données de l'EPG au moyen de 8 signaux analogiques codés sur 8 bits. Ainsi les données de l'EPG sont synchrones avec les autres paramètres qu'elle que soit la configuration du multiplexeur avec une erreur maximale de 0, 5 ms.

C-Le programme d'acquisition :

C'est le programme qui gère le déroulement des acquisitions. Avant toute manipulation, nous utilisons un programme baptisé CORPUS qui permet de créer les corpus des phrases à prononcer et, de les présenter au sujet lors des séquences d'acquisition.
Le programme d'acquisition s'appelle ACQUERIR. Sa première partie est réservée à la configuration de l'acquisition qui permet à l'utilisateur d'indiquer au système le nombre d'entrées utilisées, et le type des signaux (acoustique, physiologique etc..).
En général, l'acquisition est déclenchée par le sujet en validant l'horloge d'échantillonnage. La phrase apparait simultanément sur l'écran monochrome (face au sujet) et sur l'écran couleur VGA face au manipulateur. Ce dernier peut ainsi contrôler les niveaux d'acquisition au moyen de "bargraph" equipés d'indicateurs de saturation. La phrase prononcée, le sujet stoppe l'acquisition en arrêtant l'horloge. Les informations sont envoyées directement dans le disque dur sous la forme d'un fichier brut contenant les différents paramètres entrelacés.Il n'y a pas, en théorie de limite à la durée d'acquisition, elle est fonction de la place libre sur le disque. En pratique, elle est effectuée phrase par phrase pour faciliter la manipulation des fichiers. Les fichiers d'acquisition peuvent être ensuite sauvegardés sur la bande megnétique de 60 Mo qui peut contenir 6 minutes de corpus effectif à la fréquence maximale d'acquisition du dispositif (160 Ko / s). Chaque fichier est conservé avec toutes ses caractéristiques d'acquisition et le corpus dont il est issu (figure 2).

5-LE PROGRAMME D'EDITION DES PARAMETRES ARTICULATOIRES :

Ce programme appelé EDITSIGN est le coeur de la station de travail. Il permet d'éditer tous les paramètres enregistrés et de les segmenter, les marquer, les mesurer les traiter, avec une grande simplicité d'utilisation. Cette dernière est obtenue grâce à un programme de gestion de fenêtres spécifique d'un rapport performance/encombrement mémoire élevé, à l'utilisation de menus déroulants, d'icones et de nombreux programmes utilitaires. L'ergonomie est encore améliorée par l'utilisation d'un écran monochrome pour les informations alphanumériques(fig 2).
La première opération d'édition consiste à éclater les fichiers bruts d'acquisition qui sont décomposés en fichiers

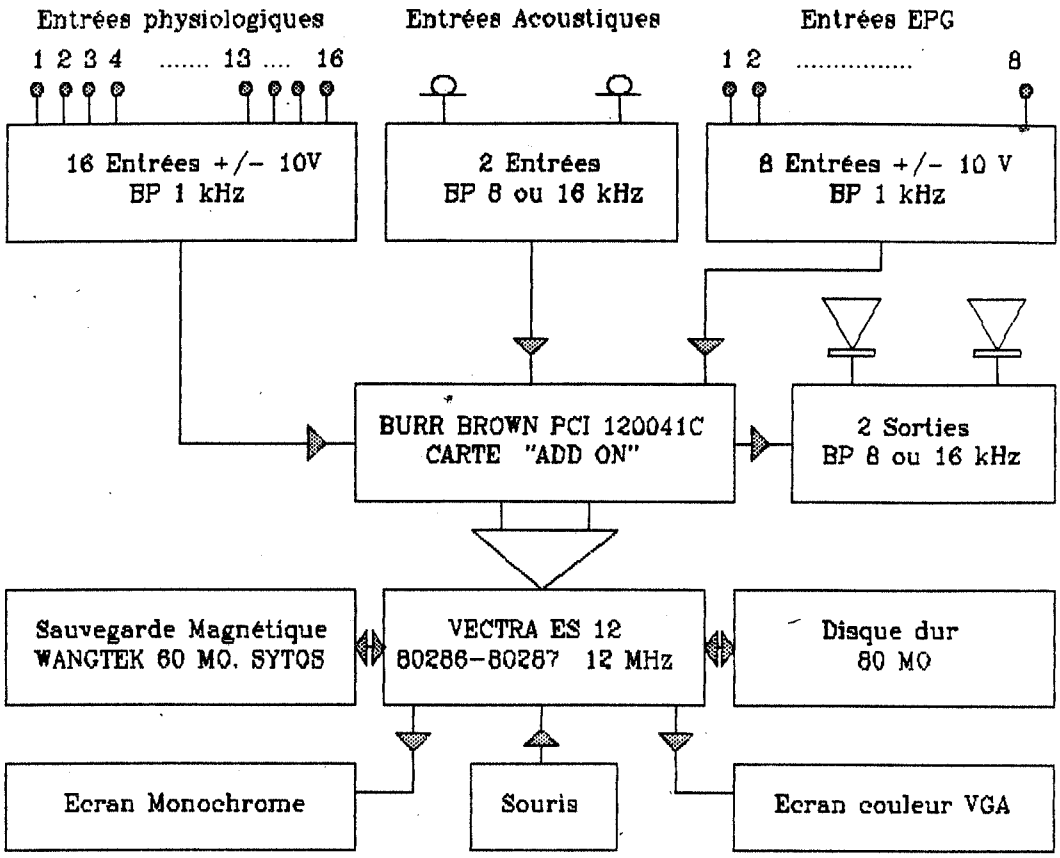


FIGURE 1
Configuration de la station de travail

paramètres. On peut ensuite les visualiser sur l'écran VGA (10 paramètres max). L'affichage est réalisé sur une phrase, c'est à dire entre le départ et l'arret de l'acquisition.Il est possible de déplacer les courbes vers la gauche ou la droite (scroll) et de faire des "Zoom" entre dux curseurs en changeant le pas d'affichage, dont le minimum est de un échantillon par pixel. Toutes les opérations sur les paramètres (calibration, segmentation, mesure, découpage, écoute, positionnement de marqueurs, etc. . .)sont réalisées à partir de curseurs. On peut en positionner 30 par écran. Chaque courbe affichée dispose d'une ligne pour positionner des marqueurs. Lorsque l'on travaille sur un paramètre, ses caractéristiques d'acquisition et d'affichage apparaissent sur l'écran monochrome ainsi que toutes les mesures réalisées. On peut choisir dans une fenètre mathématique les traitements effectués sur des curseurs tels que: mesures d'amplitudes linéaires ou logarithmiques, visualisation des configurations palatines, calculs des spectres (FFT); ou entre des curseurs tels que: mesure de durées, calculs de sommes , de différences, d'intégrations, de moyennes, d'écarts-type. Une fenètre particulière permet de faire appel à des fonctions de traitement externes au programme EDITSIGN, sous forme de programmes exécutable sous MS DOS, au moyen d'un interprétteur de commandes. Les fonctions externes s'appliquent sur les courbes. Nous disposons des fonctions externes suivantes: Somme de 2 courbes, différence entre 2 courbes, valeur absolue, intégration par paliers de durées variables, élévation au carré, racine carrée, logarithme, exponentielle, retards variables, splines cubiques et quadratiques, quantifications en niveaux.Toutes les mesures sont conservées sous la forme de fichiers résultats en code ASCII, interfacés avec les programmes de traitement au moyen d'un utilitaire de tri. Nous utilisons pour les traitements statistiques le logiciel SYSTAT.
Lorsque l'EPG est utilisé, les configurations palatines apparaissent dans une fenètre particulière, celle qui correspond au curseur de travail apparait en surbrillance. L'analyse spectrale FFT apparait à la demande dans une fenètre spéciale, elle correspond également à la position du curseur de travail (figure 3 A et B).

6-CONCLUSION :

Il est impossible de décrire succintement un logiciel aussi complexe que EDITSIGN et de mettre en évidence toutes ses possibilités. Il est décrit en détail par ailleurs (18).
Jamais auparavent, il n'a été possible, d'obtenir un tel ensemble d'informations simultannées sur le fonctionnement articulatoire. Par exemple; avoir la connaissance à un instant donné des débits d'air (buccal et nasal), de la pression intraorale, de l'aperture et de la protrusion labiale, de l'articulation de la langue et du spectre du signal acoustique résultant, représente pour le chercheur un avantage évident.
L'analyse de grands corpus répétés plusieurs fois, avec de nombreux sujets, sur de multiples paramètres, dans différentes langues, fait que le travail d'acquisition et de

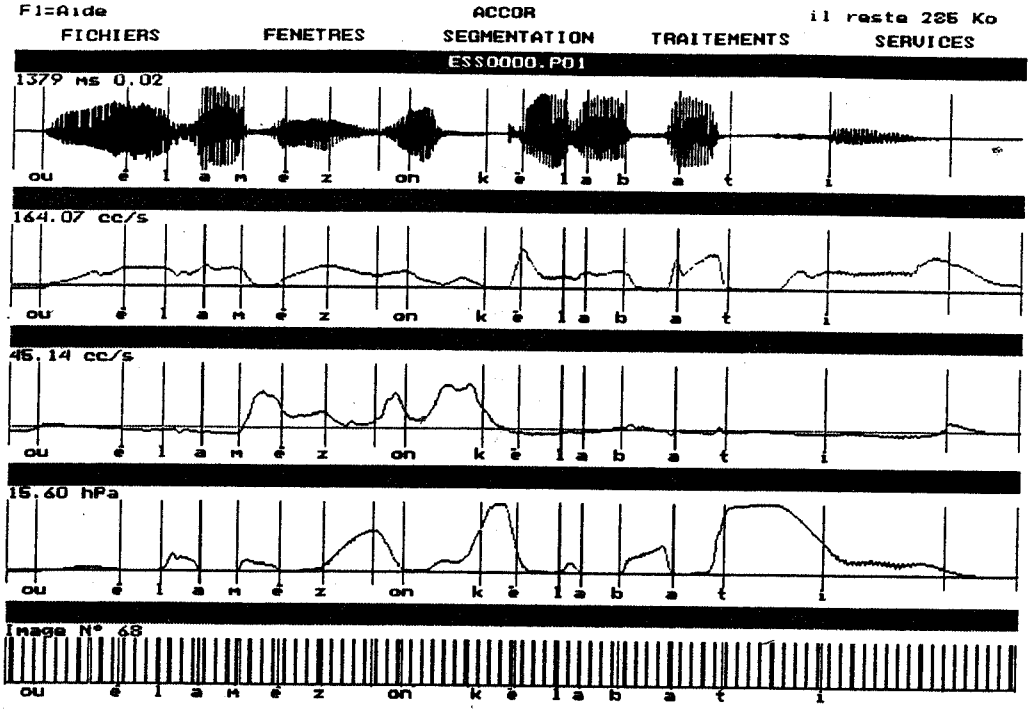


FIGURE 3 A

Phrase "où est la maison qu'elle a bâtie?" avec segmentation phonétique large.

P01: Phonogramme buccal. P02: Débit d'air buccal (en CC/S). P04: Débit d'air nasal (en CC/S). P08: Pression intra orale (en hPa). P10: Images vidéo.
Les valeurs qui se trouvent en haut à droite des fenêtres des paramètres correspondent au curseur de travail situé entre les marques "z" et "on".

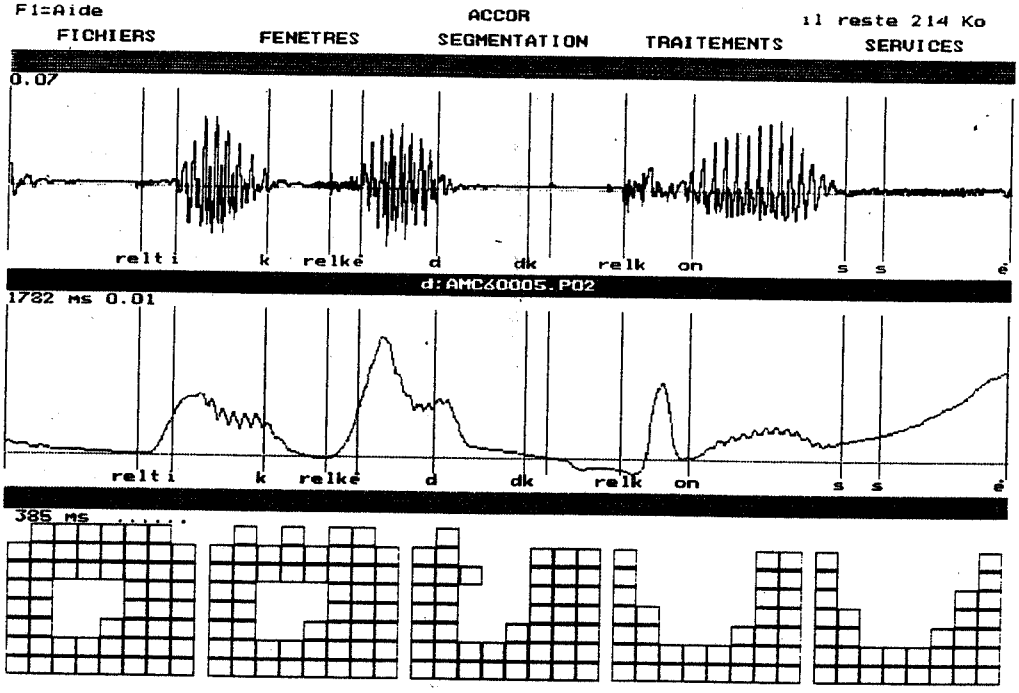


FIGURE 3 B

Phrase "ticket de concert", avec segmentation infra phonèmique.

La représentation palatine centrale correspond au curseur non marqué entre "d" et "k".

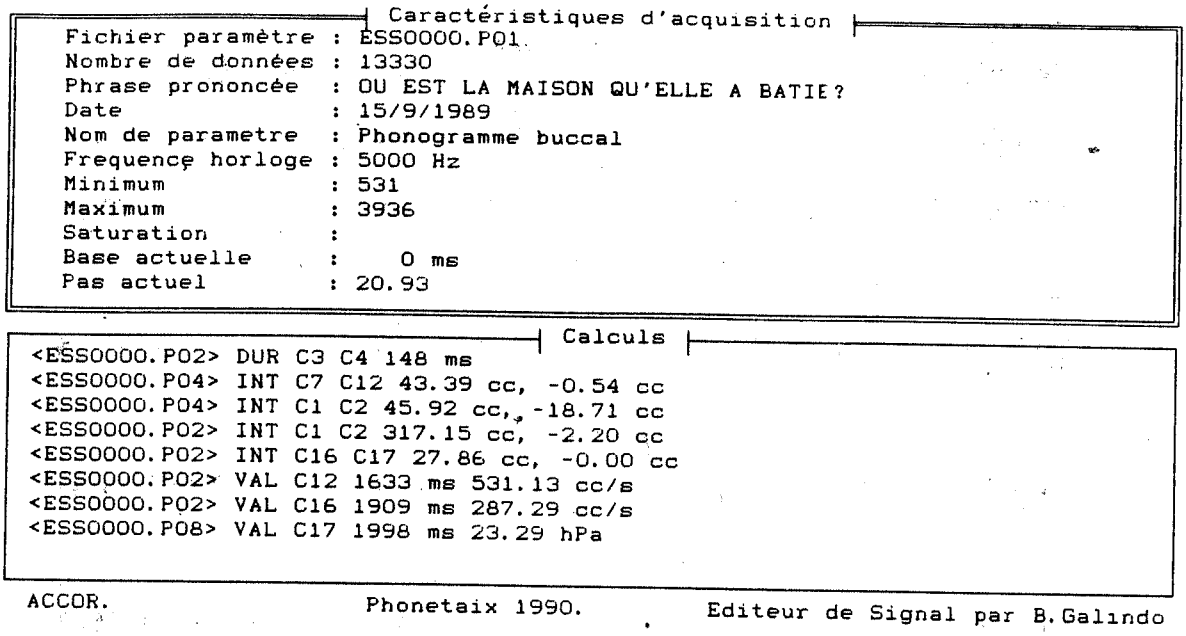


FIGURE 2
Informations d'acquisition et résultats de mesure sur l'écran monochrome correspondant à la figure 3 A.

dépouillement prévu dans le cadre de l'action ACCOR est considérable. Il vient de commencer, et la station de travail répond bien à ce que l'on attend d'elle. Compte tenu de sa spécificité nous l'avons baptisée PHYSIOLOGIA. Nous espérons que son utilisation pourra contribuer efficacement à l'amélioration des connaissances sur les processus articulatoires. Il est à noter que tous les programmes sont indépendants et peuvent fonctionner avec d'autre systèmes d'acquisition moins complexes. Nous développons actuellement des modules de traitements acoustiques (sonagramme,banc de filtres) qui augmenterons encore les possibilités de la station de travail.

BIBLIOGRAPHIE

1-LINDBLOM, B. (1987)
Adaptative variability and absolute constancy in speech signal : Two themes in the quest for phonetic invariance.
Proceedings 11th ICPhS, Vol 3, 9-18.
2-ZUE, V. and SCHWARTZ, R. (1980)
Acoustic processing and phonetic analysis.
In **TRENDS IN SPEECH RECOGNITION**, LEA, W. A. Editor, Prentice Hall, Englewood Cliffs, 101-124.
3-LINELL, P. (1982)
The concept of phonological form and the activity of speech production and speech perception.
JOURNAL OF PHONETIC, Vol 10, 37-72.
4-Articulatory-acoustic correlation in coarticulatory processes : A cross language investigation (ACCOR).
Work program, ESPRIT 2 (BRA), action 3279, 1989.
5-CHOW, Y. L. and al. (1986)
The role of word-dependent coarticulatory effects in a phoneme based speech reconigtion system.
ICASSP 86, paper 30-9, 1593-1596.
6-HOLMES, J. N. (1988)
SPEECH SYNTHESIS AND RECOGNITION.
Chapter 2-6, Van Nostrand Reinhold, Wakingham.
7-GENTIL, M. and GAY, T. (1984)
Temporal organization of interarticulator muscle activity in american english monosyllables.
JASA, 76(S1) S15(A).
8-ALME, A. M. and Mc ALLISTER, R. (1987)
Labial coarticulation in stutterers and normal speakers.
SPEECH MOTOR DYNAMICS IN STUTTERING, PETERS and HULSTIJN ed, Spinger Verlag,
9-HONG, G. SHONLE, P. W. and CONRAD, B. (1987)
Electromegnetic articulography.
Proceedings 11th ICPhS, Vol 1, 27-30.
10-HARDCASTLE, W. J. and al. (1989)
New development in electropalatography : A state-of-the-art report.
CLINICAL LINGUISTICS AND PHONETICS, Vol 3, N°1, 1-38.
11-SCULLY, C. (1989)
Articulatory synthesis.
Proceedings Speech Production and Speech Modelling NATO Seminar, 17-29 July, Bones France, 36 p. in press.
12-FARNETANI, E. (1985)
Profilo aerodinamico di alcune consonanti italiane.
PHONIATRICA ITALIANA, Vol 7, 33-45.
13-AUTESSERRE, D. and al. (1989).
Movements of the lips and velum in speech : Variations in aerodynamic parameters.
EUROSPEECH 89, Vol 2, 437-440.
14-CALDER, G. (1988)
A multichannel data acquisition unit for IBM PC.
Technical report, IBM UKSC, 60 p.
15-GALINDO, B. et TESTON, B. (1989)
Physiologia : Un logiciel d'analyse des paramètres physiologiques de la parole.
TIPA, Vol 14, 40p, sous presse.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Amélioration de la Parole Bruitée par un Filtrage Sélectif

Douglas O'Shaughnessy et Hélène Valbret

INRS-Télécommunications, Université du Québec
3 Place du Commerce, Ile des Soeurs, Québec, Canada H3E 1H6

1. Introduction

Les objectifs de rehaussement de la parole dégradée sont d'améliorer la qualité générale d'un signal de parole, son intelligibilité, et de diminuer la fatigue des personnes qui l'entendent. Ces recherches ont été motivées par ces problèmes fréquemment rencontrés:

- la mauvaise qualité des systèmes de communication, qui introduisent souvent dans le signal des distorsions (échos ou bruit) qui se révèlent désagréables à l'écoute et qui provoquent une diminution de l'intelligibilité du signal transmis.
- les difficultés d'audition chez certains malades;
- la chute rapide, avec la dégradation du signal d'entrée, des performances des systèmes de compression (pour les chaînes de transmission numérique) et des systèmes de reconnaissance automatique de la parole.

Lorsqu'un signal de parole est transmis acoustiquement dans un environnement bruité, transmis électroniquement ou enregistré sur ruban, il est sujet à des dégradations qui limitent sa qualité et son intelligibilité. Dans le cas où les signaux sont destinés à être traités par des algorithmes automatiques de codage ou de reconnaissance, les dégradations limitent le fonctionnement de ces algorithmes. Les dégradations les plus communes sont dues à l'ajout de bruit (soit par l'environnement, soit par la transmission, soit au niveau de l'enregistrement), à un mauvais positionnement du microphone, et à la présence d'autres sources sonores.

Le rehaussement de la parole par des techniques de traitement des signaux numériques peut rendre la parole plus intelligible et peut contribuer à améliorer la performance des algorithmes qui l'utilisent comme signal d'entrée [1]. Quelques exemples d'application sont les suivants: les systèmes de radio mobile, les dispositifs d'aide aux malentendants, les conférences audio, la téléphonie, et la surveillance policière.

Dans certains cas, lors d'enregistrements audio, un microphone est situé dans une salle, assez éloigné de la source sonore que l'on désire capter. Le signal est parfois transmis au magnétophone par radio ou par une ligne téléphonique. De plus, les conditions d'enregistrement sont souvent loin d'être idéales: 1) les locuteurs peuvent parler doucement mais aussi se déplacer sans prendre le soin de demeurer près du microphone, 2) des sons indésirables provenant d'une source située dans la salle, par ex., autres locuteurs, musique, télévision, ou encore provenant des salles voisines ou de la rue, peuvent être captés simultanément, 4) la position et la qualité du microphone peut affecter le signal audio, 5) le médium de transmission peut restreindre la gamme de fréquence, peut avoir des caractéristiques non linéaires, et être source de bruit.

2. Systèmes de rehaussement

2.1 Suppression d'interférence avec deux microphones

Il arrive souvent qu'il soit possible d'utiliser deux microphones pour capter un signal de parole. En plaçant alors un microphone près de la source (qui produit le signal primaire) et l'autre près de la source d'interférence (qui produit le signal secondaire), on peut, en soustrayant cette dernière version de la version primaire, réduire les effets de distorsion dans le signal primaire [2, 3]. Avant d'effectuer la soustraction, il faut traiter le signal secondaire pour simuler les effets de réverbération et d'addition de bruit qui altèrent la partie du signal primaire issue de la source d'interférence. Quand un microphone est loin d'une source sonore, la qualité du signal enregistré est réduite par l'ajout de bruit provenant de l'environnement et par la réverbération causée par les objets présents dans la salle. Ce problème nécessite un filtrage adaptatif, qui se sert d'égalisateurs afin de modéliser les effets de l'environnement.

2.2 Suppression d'interférence avec un microphone

Lorsqu'un seul signal de la parole bruitée est disponible, la performance des systèmes de rehaussement à date a eu peu de succès à augmenter l'intelligibilité de la parole, malgré son succès fréquent à diminuer le bruit. Ces

approches rendent la parole plus « propre » et moins fatiguante pour l'écoute, mais c'est rare qu'une personne peut comprendre mieux la parole après le rehaussement qu'avant.

Les principales approches de rehaussement de la parole dégradée, qui se servent d'un seul signal, sont exposées ci-dessous. Trois types de techniques ont fait l'objet de nombreuses recherches [1]:

- la soustraction spectrale des signaux d'interférence,
- la suppression des fréquences non harmoniques,
- la resynthèse utilisant un vocodeur.

2.2.1 La soustraction spectrale des signaux d'interférence

Lorsque le signal d'interférence est un signal de bruit blanc additif, le signal dégradé peut se décomposer de la façon suivante,

$$x(n) = s(n) + b(n),$$

où x est le signal dégradé, s le signal exempt de bruit et b le signal d'interférence. Dans le domaine fréquentiel, on obtient alors

$$X(e^{j\omega}) = S(e^{j\omega}) + B(e^{j\omega})$$

où X, S, B sont les transformées de Fourier de x, s, b respectivement.

On cherche à recouvrer le signal s, connaissant le signal x. Dans la plupart des cas, on ne dispose pas d'un enregistrement séparé du signal d'interférence. Or le bruit est un signal aléatoire dans le temps. On ne peut pas donc en connaître les caractéristiques temporelles et de ce fait, effectuer directement la soustraction temporelle des signaux x et b. Par contre, il est possible d'évaluer le spectre du signal de bruit, comme nous le verrons dans le paragraphe suivant. En soustrayant au spectre X l'estimation du spectre de bruit $\tilde{B}$, on obtient une valeur approchée $\tilde{S}$ du signal sans bruit, qu'il suffit alors de transformer en signal temporel pour avoir le signal rehaussé.

Pour évaluer le spectre de bruit, on supposera qu'il demeure localement stationnaire, c'est à dire que son amplitude spectrale, mesurée juste avant que le locuteur se mette à discourir, est identique sur toute la période parlée. Ces hypothèses étant admises, toutes les informations spectrales nécessaires à l'obtention d'une fonction approchée du spectre de bruit seront obtenues en analysant les portions de l'enregistrement exempt de parole. On mesure ainsi, lorsque le locuteur est estimé être silencieux, la valeur moyenne du bruit $\beta(e^{j\omega})$, et la phase du signal dégradé $\theta_x(e^{j\omega})$. L'estimation du spectre de bruit est alors obtenue en remplaçant l'amplitude du spectre de bruit $|B(e^{j\omega})|$ par $\beta(e^{j\omega})$, et sa phase $\theta_b(e^{j\omega})$ par $\theta_x(e^{j\omega})$ [4]. Donc,

$$\tilde{B}(e^{j\omega}) = \beta(e^{j\omega})\theta_x(e^{j\omega})$$

et

$$\tilde{S}(e^{j\omega}) = X(e^{j\omega}) - \tilde{B}(e^{j\omega}) = X(e^{j\omega}) - \beta(e^{j\omega})\theta_x(e^{j\omega}).$$

Le rehaussement par soustraction spectrale n'apporte aucune amélioration en ce qui concerne l'intelligibilité mais accroît la qualité [4]. Cependant, si le système de soustraction spectrale est utilisé en tant que préprocesseur d'un vocodeur à prédiction linéaire, on note, à la sortie de ce dernier, une nette amélioration de l'intelligibilité et de la qualité du signal [4].

2.2.2 La suppression des fréquences non harmoniques

Les sons voisés sont quasi-périodiques. Cette périodicité se traduit dans le domaine fréquentiel par des harmoniques, multiples de la fréquence fondamentale correspondant à la période temporelle du signal. La méthode que l'on considère dans cette section, est basée sur cette observation. Elle consiste à filtrer le signal de parole dégradé, en ne laissant passer que les harmoniques de la parole. Un tel filtre est appelé un « peigne ». Il permet ainsi de diminuer le bruit dont l'énergie est répartie sur tout le domaine fréquentiel, en particulier entre les harmoniques de la parole voisée.

Cette approche suppose la périodicité du signal dégradé, ce qui restreint cette méthode aux seuls sons voisés. En réalité, cette restriction est mineure. En effet, non seulement les sons voisés composent la majeure partie des signaux de parole mais ils fournissent aussi les informations concernant les articulations de deux consonnes adjacentes. D'autre part, l'opération de filtrage dépend fortement de la détection de la fréquence fondamentale. Or Frazier [5] s'est rendu compte que, malgré une connaissance précise de la fréquence fondamentale, cette méthode déforme le signal de façon sensible, ces distorsions étant dues à la variabilité temporelle des signaux. Afin de réduire ces dernières, Frazier a alors proposé un filtre adaptatif, qui s'ajuste automatiquement aux variations de la fréquence fondamentale.

Dans les expériences renouvelées pour différents filtres (longueurs de filtre de 3, 7, 13 périodes du signal) et pour divers rapports signal à bruit blanc, Lim [6] a trouvé que l'intelligibilité décroît au fur et à mesure que le bruit augmente. Pour un rapport signal à bruit donné, l'intelligibilité diminue lorsque la longueur du filtre croît. D'autre part, ces tests indiquent que, malgré des informations précises sur la fréquence fondamentale, le peigne adaptatif n'apporte pas de hausse significative de l'intelligibilité, lorsque le signal d'interférence est un bruit blanc additif. L'utilisation des filtres les plus longs conduit même à une baisse de l'intelligibilité, à cause d'une diffusion temporelle de l'information spectrale.

2.2.3 La resynthèse utilisant un vocodeur

La production de la parole peut être considérée comme une opération de filtrage pendant laquelle une source de bruit excite un filtre représentant les configurations du conduit vocal. Lorsqu'il s'agit de sons voisés, la source est périodique; par contre, dans le cas d'un son sourd, un bruit blanc excite le filtre. Le conduit vocal renforçait certaines zones de fréquences du signal excitateur (les formants); si nous supposons que les configurations du

conduit vocal évoluent lentement, on peut considérer que celui-ci est stable sur un court laps de temps (typiquement 20 ms). Le conduit vocal peut alors être modélisé par un filtre linéaire « tous pôles » dont la fonction de transfert est de la forme,

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}},$$

où les coefficients a_k sont les coefficients LPC. Un tel filtre s'appelle filtre à prédiction linéaire.

Si les paramètres nécessaires à la modélisation de la parole sont évalués avec précision à partir du signal dégradé, on pourra alors affirmer avoir extrait la parole du bruit. Le problème de rehaussement de la parole devient alors un problème d'estimation de paramètres.

Lim [7] a évalué les performances de cette méthode. Des phrases ont été générées par filtrage d'une source de bruit aléatoire ou d'un train d'impulsions à travers un filtre « tous pôles » dont les coefficients sont connus. Les données obtenues sont dégradées par l'ajout d'un bruit blanc aléatoire. Lim a calculé alors les coefficients à prédiction linéaire à partir du signal dégradé. Ces coefficients sont ensuite comparés aux coefficients du filtre qui a servi à la synthèse des phrases de référence.

Il apparaît que cette technique améliore de façon sensible la qualité de la parole pour divers rapports signal à bruit (10–20 dB). Cependant, pour des rapports signal à bruit très faibles (lorsque le bruit est suffisamment important pour que quelques pics spectraux du signal dégradé disparaissent sous le bruit ou que ceux-ci soient très différents de ceux des données de référence), la qualité du signal de parole resynthétisé se détériore; les dégradations se traduisent par un manque de naturel et la génération de « tons musicaux ».

3. Filtrage Sélectif Spectral

Chacune des méthodes de rehaussement traditionnelles mentionnées ci-dessus a ses difficultés face à un signal verbal soumis au bruit. Si possible, il est important de ne pas déformer la parole pendant qu'on essaie d'enlever le bruit, parce que l'objectif est de hausser la qualité de la parole. Nous avons exploré une nouvelle méthode de supprimer le bruit dans les gammes de fréquences dont l'énergie de la parole est faible. D'autres méthodes suppriment le bruit à travers le spectre, souvent déformant la parole dans les fréquences importantes où se trouvent des résonances (formants). Donc, face à la tâche de supprimer des distorsions sans déformer la parole, nous avons développé une méthode pour identifier des gammes de fréquences qui ne contiennent qu'une faible énergie de la parole, et pour supprimer complètement l'énergie dans ces gammes.

Cette approche s'applique aux signaux dont une large gamme de RSB, approximativement 0–30 dB. Pour des sons voisés de tels signaux, quelques formants de basse fréquence sont nettement plus élevés que le niveau du bruit, et pour des sons non voisés, l'énergie importante de haute fréquence est souvent plus élevée que celle du bruit. Donc, de simples mesures de l'énergie et de la fréquence suffisent pour identifier les gammes de fréquence à supprimer.

Quant aux signaux dont le RSB est plus élevé que 15 dB, le problème se simplifie. Au lieu de calculer une transformée Fourier, nous avons utilisé la mesure intitulée « taux de croisement à l'amplitude nulle » (TCAN). La valeur du TCAN est le nombre de fois (par seconde) que le signal change sa valeur algébrique. Un son voisé est dominé par de l'énergie de basse fréquence; donc, son TCAN est faible. Un son non voisé est dominé par de l'énergie de haute fréquence; donc, son TCAN est haut. Normalement, le bruit de fond contient de l'énergie de toutes les fréquences, et le TCAN d'un tel bruit (sans parole) a une valeur moyenne. Donc, un simple ensemble de règles suffit pour distinguer quatre catégories de parole bruitée: 1) voyelle (dont l'intensité est forte et le TCAN est faible), 2) fricatif non voisé (dont l'intensité est moyenne et le TCAN est haut), 3) fricatif voisé (dont l'intensité est moyenne et le TCAN est haut), 4) bruit seulement (dont l'intensité est faible et le TCAN est moyen).

Chez les voyelles, le signal dégradé est passé par un filtre passe-bas, afin de supprimer le bruit en hautes fréquences (la gamme sans l'énergie de parole souhaitée). Dans les fricatifs non voisés, le signal dégradé est passé par un filtre passe-haut, afin de supprimer le bruit en basses fréquences. Pour les parties du signal identifiées comme contenant seulement du bruit, la sortie est mise à zéro. Chaque transition entre sections associées aux différentes catégories est sujette à un lissage pendant 60 millisecondes, pour éviter un brusque changement dans la qualité de la parole quand le filtrage change.

Ce processus a été appliqué avec succès à un enregistrement audio préparé pour diffusion à la télévision. Après avoir enregistré plusieurs conversations (qui duraient 30 minutes au total), on a découvert que le signal audio contenait un bruit partout, dont le RSB était de 20-30 dB. Ce bruit n'avait pas un spectre plat, mais une prépondérance en basse fréquence avec une légère pente décroissant en fréquence. Quoique ce bruit ne se conformait pas au bruit blanc, l'approche mentionnée ci-haut s'applique. Les personnes écoutant la parole sortante ont trouvé que la parole était préservée sans déformation, tandis que le bruit était nettement supprimé.

Pour les voyelles, la largeur de bande préservée était réduite à 0–4 kHz au lieu de 0–7 kHz dans l'original souhaité. Ceci est à cause du bruit qui dominait, dans les voyelles, la gamme de fréquences supérieure à 4 kHz. Ce n'était pas possible de reconstituer la parole voisée dans ces hautes fréquences. Donc, la qualité et l'intelligibilité de cette parole restent hautes, mais les très hautes fréquences, qui ont contenu un bruit fort, étaient coupées.

De plus, le filtrage sélectif a été évalué, par des tests auditifs sans formalisme, avec la parole bruitée aux RSBs jusqu'à 0 dB. Quand le niveau de bruit augmente, le bruit dans la parole traitée ne monte pas directement, mais les gammes de fréquence préservées diminuent, ce qui réduit son intelligibilité graduellement. Même à un RSB de 0 dB.

la parole traitée reste plus intelligible que la parole bruitée originale.

Notre approche diffère de celle de McAulay [8] en plusieurs détails: 1) notre décision de supprimer l'énergie dans une bande de fréquence est plus simple, basée seulement sur le niveau d'énergie vis-à-vis le niveau du bruit, 2) leur réalisation a nécessité l'usage d'un vocodeur de 19 canaux dans le contexte de transmission par téléphone (donc une largeur de bande de seulement 180–3720 Hz), 3) nos gammes de fréquence ne sont pas fixes, mais correspondent aux largeurs de bande des formants, 4) nous nous servons d'un critère de continuité (nous ne permettons pas l'énergie d'une bande de fréquence à aller et à venir d'une trame à la prochaine), 5) nous exploitons directement l'existence des formants (nous présumons un formant à chaque 1000 Hz).

Les expériences de Lim et al [6] ont démontré que le filtrage de peigne simple n'aide pas à hausser l'intelligibilité de la parole bruitée. Cependant, ils ont présumé que les caractéristiques de la parole ne varient pas rapidement. Sauf pour quelques périodes pendant des voyelles longues, cette présomption n'est pas valide. Donc, quand la parole est lissée avec un filtre de peigne d'une longue durée, la parole changeante détériore, surtout à une frontière voisée-non voisée. En plus de notre filtrage sélectif, nous avons utilisé un filtrage de peigne adaptatif. Le nombre de périodes du filtre varie avec la stabilité de la parole. Si l'énergie dans les différentes bandes de fréquence reste stable et la fréquence fondamentale estimée ne change que lentement, nous utilisons un filtre de peigne de 3–4 périodes; lors d'un changement moyen, nous réduisons la durée du filtre à 2 périodes. Lors d'un changement rapide, nous n'utilisons pas de filtrage de peigne.

Afin d'éviter les difficultés d'un filtre fixe, nous avons changé les traits du filtre en fonction d'une analyse de la parole. Pour des RSBs assez sévères (par ex., vers 0 dB), les deuxième et troisième formants, qui sont très importants pour la perception, tendent à disparaître sous le bruit. Pour certains sons, cependant, ces formants sont suffisament forts pendant certaines périodes de temps (par ex., le deuxième formant des voyelles articulées en arrière de la bouche, et le troisième formant des hautes voyelles articulées en avant de la bouche) à apparaître au dessus du bruit. Quand un formant disparaît sous le niveau de bruit à cause d'un changement de la forme du conduit vocal, qui cause des changements dans les fréquences et dans les amplitudes des formants, nous interpolons ce formant dans la direction dont ce formant se déplace au moment de la disparition.

4. La qualité de la parole rehaussée

Le signal obtenu après rehaussement par le filtrage sélectif souffre d'une réduction de la largeur de bande fréquentielle, comparé à la parole originale. Cependant, les gammes de fréquence éliminées ont été auparavant complètement déformées par le bruit. Donc, l'élimination de ces fréquences n'empêche pas son intelligibilité, et même augmente son naturel, et réduit la fatique des personnes qui l'écoutent.

Le rehaussement d'un signal de parole, par notre méthode de filtrage sélectif, suppose une détection fiable des gammes de fréquence à préserver. Pour les sons voisés, de telles gammes correspondent aux formants de basse fréquence, qui sont très importants pour la réception et la compréhensibilité du signal. Pour un RSB faible (0 dB et moins), cette détection est très difficile à réaliser. Donc, dans la pratique, notre méthode ne fonctionne que pour les cas où le RSB est plus élevé que 0 dB.

Lorsqu'un RSB est de 0 dB ou moins, pour la plupart des sons voisés, seul le premier formant (au plus) est visible dans un spectrogramme. Quant aux sons sourds, on n'a plus aucune possibilité d'en déterminer la répartition énergétique (sauf peut-être un fort fricatif palatal). En effet, ces derniers sont des signaux de faible énergie. L'addition d'un bruit blanc obscurcit uniformément le spectre du son dégradé et empêche d'en déduire les caractéristiques spectrales. Par contre, dans le cas des voyelles et d'un RSB plus élevé que 0 dB, quelques formants sont très marqués. Ils apparaissent nettement au dessus du bruit sur un spectrogramme. Notre méthode de rehaussement est alors efficace et le signal reste intelligible même pour des RSBs aussi bas que 0 dB.

5. Conclusion

La technique de rehaussement basée sur un filtrage sélectif présente un intérêt certain. Dans le cadre de cette recherche, nous avons cherché à améliorer la qualité de la parole pour des signaux dégradés par du bruit blanc. Lorsqu'un bruit blanc se superpose au signal, nous avons proposé d'éliminer complètement l'énergie aux fréquences où le bruit est supérieur à la parole. Le filtrage de peigne peut aider le rehaussement si la longueur de ce filtre varie avec la stabilité de la parole. Finalement, la prolongation des formants relativement faibles en amplitude par une interpolation de direction aide surtout aux frontières voisées-non voisées; ces périodes de transition spectrale sont souvent très importantes pour la perception.

References

1. J. Lim, *Speech Enhancement* (Englewood Cliffs, NJ: Prentice-Hall), 1983.
2. S. Boll & D. Pulsipher, « Suppression of acoustic noise in speech using two microphone adaptive noise cancellation », IEEE Trans. ASSP, **ASSP–28**. no. 6, pp. 752–753, 1980.
3. G. Kang & L. Fransen, « Experimentation with an adaptive noise-cancellation filter », IEEE Trans. Circuits and Systems, **CAS–34**, no. 7, pp. 753–757. 1987.
4. S.F.Boll, « Suppression of Acoustic Noise in Speech Using Spectral Subtraction », IEEE Trans. ASSP. **ASSP-27**, No.2, 113-120, avril 1979
5. R.H.Frazier, S.Samsam, L.D.Braida, A.V.Oppenheim, « Enhancement of Speech by

Adaptive Filtering », IEEE Int.Conf. ASSP, 251-253, 1976.

6. J.S.Lim, A.V.Oppenheim, L.D.Braida, « Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition », IEEE Trans. ASSP,**ASSP-26**, No.4, 354-358, août 1978.
7. J.S.Lim, A.V.Oppenheim, « All-Pole Modeling of Degraded Speech », IEEE Trans. ASSP, **ASSP-26**, No.3, 197-210, juin 1978
8. R. McAulay et M. Malpass, « Speech enhancement using a soft-decision noise suppression filter », IEEE Trans. ASSP-28, no. 2, 137–145, Avril 1980.

7 RECONNAISSANCE ET DIALOGUE ORAL

Président: J.P. TUBACH
ENST-Paris,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de parole continue

Dominique François et Dominique Fohr

CRIN/INRIA-Lorraine, BP 239, 54506 Vandœuvre, France

Résumé

In order to increase the accuracy of continuous speech acoustic-phonetic decoding, the APHODEX project started some yearsago. The project was ended by the realisation of an expert system that implement the knowledge of an phonetician expert who is able to read speech spectrograms.
The system is now operational and we can obtain a phone-lattice, each phone is level-headed with a coefficient.
Because our aim is to increase the accuracy of this sytem, we have undertook the Aphodex evaluation. Results concerning a multi-speaker corpus of continuous speech are presented and discussed after a summary of the expert system works.

C'est dans le but d'améliorer le décodage acoustico-phonétique en parole continue que le projet APHODEX a vu le jour il y quelques années. Son aboutissement a été la réalisation d'un système expert basé sur la connaissance d'un phonéticien capable de lire des spectrogrammes de parole.
Le système est maintenant opérationnel et permet d'obtenir, à partir du signal numérisé et de façon entièrement automatique, un treillis de phonèmes pondérés par des coefficients de vraisemblance.
Dans le but d'améliorer les performances de ce système, nous avons entrepris l'évaluation d'APHODEX. Nous présenterons et discuterons les premiers résultats de cette évaluation faite sur un corpus multilocuteur de parole continue après avoir résumé le fonctionnement du système expert.

1 Description du système APHODEX

1.1 Introduction

Aphodex est un système expert pour le décodage acoustico-phonétique multilocuteur de la parole continue. Pour améliorer les algorithmes existants de décodage acoustico-phonétique,nous avons entrepris l'analyse et la modélisation du savoir-faire d'un expert en lecture de spectrogrammes: François Lonchamp de l'institut de phonétique de Nancy II. Après une première phase d'acquisition de l'expertise (règles et stratégies), nous avons formalisé cette expertise sous la forme d'un système expert à règles de production.

1.2 Principales caractéristiques

L'expertise présentant des particularités spécifiques, nous avons développé notre propre moteur d'inférence. Ses principales caractéristiques sont:

- remise en cause de la segmentation possible à tout moment,
- déroulement en parallèle de analyse sur plusieurs segmentations,
- prise en compte des phénomènes contextuels,
- gestion de l'incertitude en ce qui concerne l'interprétation des mesures (raisonnement incertain),
- fabrication d'un véritable treillis phonétique,
- échange d'informations avec les niveaux supérieurs (utilisation d'Aphodex en vérification d'hypothèses lexicales).

1.3 Architecture

Aphodex se compose de quatre parties: un module de prétraitement, un moteur d'inférence, une base de règle et un ensemble de procédures d'extraction d'indices du signal de parole. Le module de prétraitement (procédural) a pour but de segmenter grossièrement la phrase prononcée et de déterminer la classe phonétique à laquelle appartient chaque segment. Dans une deuxième phase, le moteur d'inférence, à l'aide des règles, rafine la segmentation et étiquète chaque segment. Pour chaque segment généré par le prétraitement on étudie le paquet de règles correspondant à la classe phonétique du segment. Une règle se présente suivant cette syntaxe :

```
R <numero>
CONTEXTE_DROIT [liste de phonèmes]
CONTEXTE_GAUCHE [liste de phonèmes]
DEJA_SUGGERE [liste de phonèmes]
 SI
   Prémisses
 ALORS
   Conclusions.
```

Les listes de phonèmes décrivant les contextes droit et gauche du segment limitent l'application de la règle à une situation contextuelle particulière. La règle sera appliquée si les prémisses sont vérifiées. Dans le cas où un indice figurant dans celles-ci n'est pas connu, on active une procédure d'extraction. S'il s'agit d'une règle de déduction, on affecte une pondération à la liste de phonèmes résultat. S'il s'agit d'une règle d' action, on active une procédure qui, par exemple, fusionne ou découpe des segments pour corriger d'éventuelles insertions ou omissions. Le résultat se présente sous la forme d'un treillis de phonèmes pondérés.
APHODEX obtient ce treilli de manière entièrement automatique a partir du signal de parole, ceci avec ou sans visualisation graphique du déroulement de la lecture du spectrogramme.

1.4 Treillis

Le treillis est composé de noeuds qui contiennent les informations suivantes:

- le debut et la fin en milliseconde du segment auquel correspond ce noeud,
- les règles ayant été activées,
- les indices utilisés,
- une liste de phonèmes pondéres
- le contexte gauche supposé (les règles sont contextuelles)
- le contexte droit supposé

L'activation de règles contextuelles génèrent donc des chemins possibles dans le treillis. Un exemple de décodage complet est donné à la figure 1.

2 Evaluation

En quoi consiste l'évaluation d'un système de décodage acoustico-phonétique ? Après le décodage d'une phrase nous disposons d'un treillis de phonèmes tenant compte de l'analyse contextuelle et pour la même phrase nous connaissons la transcription phonétique manuelle.

Nous voulons connaître le taux de reconnaissance du système expert. Le taux de reconnaissance est le pourcentage de phonèmes correctement reconnus par rapport au nombre de phonèmes de la transcription exacte de ce qui a été décodé. Il serait intéressant de calculer ce taux pour tous les phonèmes

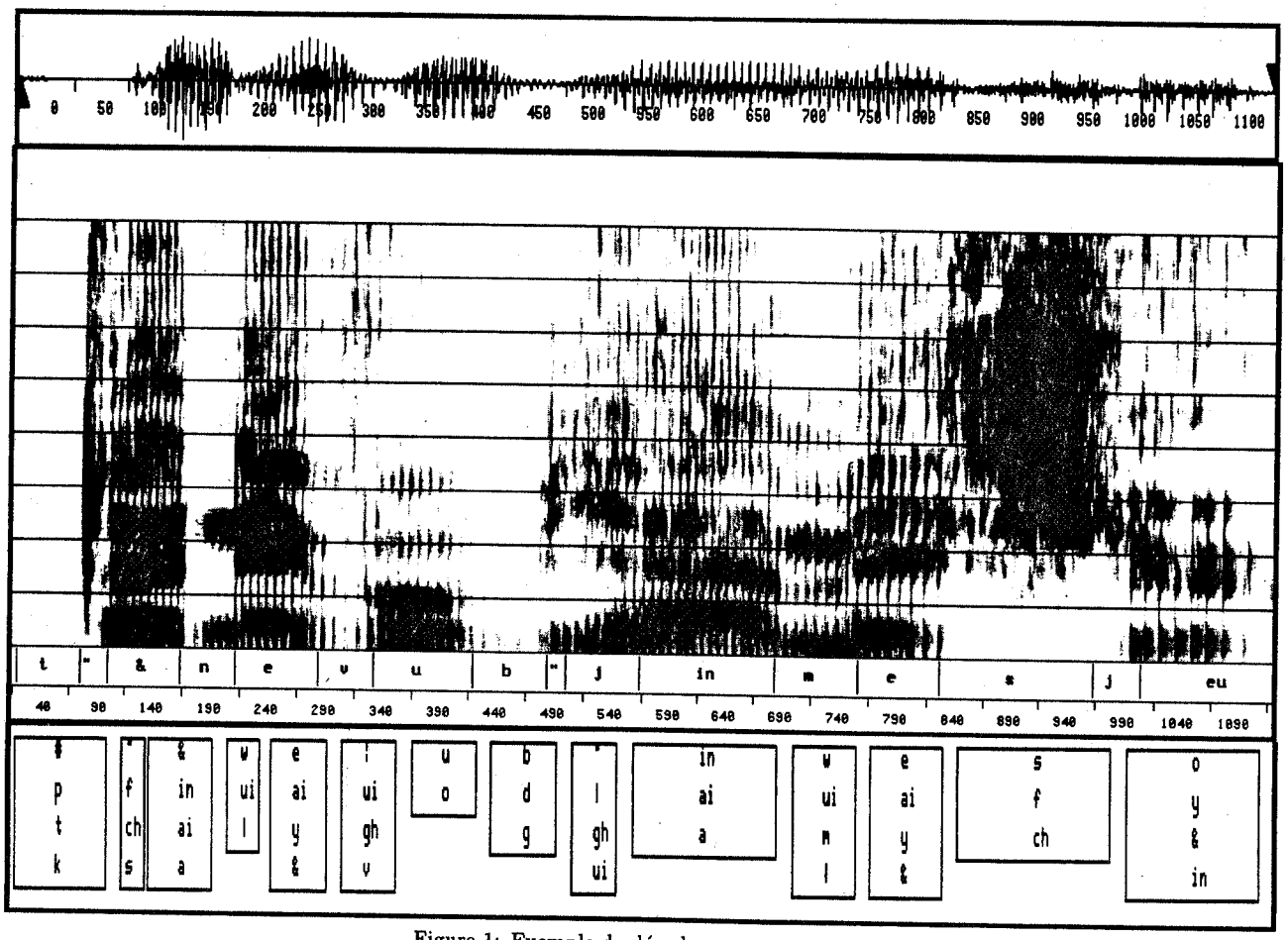


Figure 1: Exemple de décodage.

rencontrés et, si l'on veut être encore plus précis, de donner les phonèmes confondus et leur nombre.

L'évaluation d'un treillis est chose difficile. Plusieurs méthodes sont envisageables. Si l'on veut prendre en compte le fait que l'analyse d'un phonème s'est déroulée en hypothèsant les contextes gauche et droit, l'évaluation du décodage devra se faire à travers les nombreux chemins du treillis. On sera donc amené à chercher les N meilleurs chemins du treillis, avec plusieurs possibilités pour juger la qualité d'un chemin : déterminer un score par la somme des maxima des coefficients de vraissemblance de chaque noeud, considérer la pertinence des résultats de chaque noeud au vu de la transcription phonétique exacte. Un projet ESPRIT II a pour but de comparer des systèmes de reconnaissance à l'aide de plusieurs méthodes de ce type. En ce qui concerne Aphodex nous nous somme limités à une étude qui puisse être rapide, nous avions besoin de pouvoir mesurer le progrès après chaque modification du système, dans le but d'accroître ses performances.

Nous avons recencé trois voies pour effectuer des modifications, ameliorer le prétraitement du signal de parole, affiner les connaissances en modifiant la base de règles et améliorer la fiabilité des mesures et extractions d'indices sur le spectrogramme. Pour chaque classe phonétique, après chacune de ces modifications nous garderons après évaluation une image des performances.

Le résultat d'une telle étude est mis sous forme de matrice de confusion, pour laquelle chaque ligne détaille les performances de reconnaissance pour un phonème. La comptabilisation des erreurs ne peut être faite qu'après une mise en correspondance des phonèmes du treillis et des phonèmes de la transcription exacte.

Celle-ci a été réalisée manuellement pour plusieurs corpora de parole. Le corpus multilocuteur "La bise et le soleil" a été transcrit par notre expert phonéticien. Un autre corpus a été étiqueté à l'aide du logiciel d'étude de la parole SNORRI, avec possibilité d'écoute. Seule la réalité acoustique a été prise en compte, une prononciation avec élision telle que [m ɛ d s ɛ̃] pour le mot médecin sera étiqueté par la suite phonétique précédente. C'est avec ce corpus que l'évaluation d'Aphodex a été réalisée. Il s'agit de 17 phrases francaises lues, chacune des phrases a ete prononcée quatre fois par dix locuteurs masculins. Une étiquette de la transcription phonétique est constituée du début et de la fin du phone et du codage du phonéme.

La segmentation effectuée par APHODEX étant la plupart du temps différente de la segmentation manuelle, il va tout d'abord falloir effectuer un recalage temporel des deux séquences.

2.1 Le recalage temporel.

Il s'agit de mettre en correspondance l'étiquetage manuel et le treillis phonétique, pour cela nous avons à notre disposition deux éléments :

- La position des segments sur l'échelle temporelle.
- La proximité des phonèmes du treillis par rapport au phonème étiqueté, au sens où par exemple [i] et [u] sont deux phonèmes proches.

Le premier élément permet un critère éliminatoire dans la correspondance : si le décalage entre deux segments est trop important on ne pourra pas les mettre en correspondance. Cette condition sélectionne sur un critère simple et réaliste, en effet le décalage de la segmentation automatique existe, mais ne peut dépasser une certaine limite. Il ne serait pas raisonnable de mettre en correspondance deux segments distants de 160 ms (deux fois la largeur moyenne d' un segment).

Pour chaque étiquette de la phrase nous établirons une liste de segments susceptibles d' être celui qui décode le morceau de parole désigné par l'étiquette. Chaque segment sera évalué sur les deux éléments cités précédemment pour pouvoir faire partie de la liste de candidats.

- On calcule le décalage comme la somme de la distance entre les marques de début et de la distance entre les marques de fin. Ce décalage est calculé en milliseconde.

- On calcule une note en évaluant la proximité du segment par rapport à l'étiquette.

Une pondération tenant compte du lien de parenté entre le phonème prononcé et le phonème reconnu est appliquée sur chacun des coefficients de ces trois meilleurs phonèmes et ce produit constituera sa note.

Ces pondérations sur la parenté des phonèmes deux à deux sont issus d'une prématrice de confusion.

Pour un segment existe un ou plusieurs noeuds dans le treillis suivant les différentes situations contextuelles envisagées par APHODEX. Dans un premier temps, afin d'accélérer le traitement, nous ne travaillerons pas directement sur le treillis, à partir de celui-ci nous construirons un noeud particulier par segment contenant les N meilleurs phonèmes au sens des coefficients de vraissemblance accordés par le système expert.

De cette façon, si les segments sont décalés temporellement, un segment ayant comme liste de phonèmes résultat [t 10 ; p 8] pourra être placé plus correctement en face de l'étiquette [p] de la suite [a p o] que le segment voisin de liste résultat [a 8 ; a 10 ; o 5]. On sélectionne ainsi pour le segment le phonème de meilleure note.

Chaque candidat appartenant à la liste d'une étiquette donnée se voit donc attribuer deux mesures : un décalage et une note. Ensuite on parcours une première fois toutes les étiquettes en regardant dans la liste de chacune si un segment est à retenir d'une façon évidente et sûre. On a ainsi des repères qui permettent d'éliminer des candidats dans la liste des étiquettes voisines.

La suite de l'algorithme se déroule en plusieurs passages jusqu'au recalage complet, c'est à dire jusqu'à l'épuisement des correspondances possibles. S'il reste alors des étiquettes ou des segments libres, il s'agit d'omissions ou d'insertions.

2.2 Les résultats.

Après le décodage de chaque phrase le recalage est effectué puis la matrice de confusion est mise à jour.

Pour que les pourcentages d'une matrice de confusion soient significatifs il faut qu'un grand nombre de phonèmes soient recencés. Les phrases du corpus décodées par APHODEX comptent 16848 phonèmes. Toutefois certains phonèmes y sont en trop petit nombre pour permettre une interprétation statistique. Ainsi les phonèmes [v, ch, gh, s, j, w, o] comptent moins de 200 occurences dans le corpus.

La matrice de confusion est présentée en figure 2.

On remarquera tout d'abord un taux d'omissions non négligeable: 11%.Le phonème le plus touché est le [l] qui est omis à 40%. la pluspart du temps en debut de phrase. Les omissions concernant le [R] sont de l'ordre de 20%. Ces deux phonèmes représentent à eux seuls presque la moitié des omissions. Les insersions sont peu nombreuses :6%, il s'agit principalement des phonèmes [ui, j, w, R]. Ces deux types d'erreur sont dues à la première phase du système lors de la segmentation.

Les confusions se distinguent sur la matrice en concentration dans les grandes classes, les confusions interclasses sont présentées par une matrice de confusion des grandes classes en figure 3.
Le nombre de phonèmes de chaque classes est indiqué en première colonne, puis les confusions avec chacune des autres classes avec en seconde ligne le total des phonèmes bien reconnus. Les classes les mieux reconnues sont les occlusives. les voyelles nasales et les voyelles orales avec un taux de reconnaissance dépassant 85%. Les phonèmes [R] et [l] font considérablement chutter le taux de leur classe qui dépasse à peine 50%. Un gros effort devra donc être porte sur ces deux derniers. Les fricatives souffrent d'un manque de règles pertinentes. les seules règles utilisées sont basées sur une mesure de la limite inférieure du bruit de friction.

	nb	#	p	b	t	d	k	g	f	v	ch	gh	s	z	m	n	nj	j	w	R	l	ui	on	an	un	i	e	ai	a	)	o	u	y	&	ommis	
#	1228	998	136	3	26	1	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	59	81%	#
p	542	1	441	.	.	81	.	2	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	14	81%	p
b	514	.	.	417	28	.	1	.	.	.	.	.	.	.	4	1	.	1	18	.	.	13	.	.	.	.	.	.	.	.	.	3	1	.	27	81%	b
t	966	.	4	223	651	.	.	7	.	.	2	.	1	.	.	.	.	1	3	2	3	1	.	.	.	.	.	.	.	.	.	.	.	.	67	67%	t
d	1079	2	116	58	1	623	1	.	.	.	1	1	.	.	6	7	.	7	55	1	3	6	.	.	1	2	1	.	.	1	.	1	2	.	183	58%	d
k	608	.	5	95	2	12	421	.	.	1	.	1	.	.	1	.	.	1	5	2	.	1	.	.	.	.	.	.	.	.	.	.	.	.	60	69%	k
g	422	.	43	26	10	.	.	196	.	.	.	.	.	.	.	.	.	6	31	1	3	4	.	.	.	4	1	.	.	.	.	.	.	.	97	46%	g
f	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		f
v	185	.	4	30	.	.	.	.	.	12	1	.	.	.	5	.	.	29	17	57	4	3	.	.	.	.	.	.	.	.	.	.	.	.	22	6%	v
ch	64	.	.	.	.	.	.	.	.	.	61	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	95%	ch
gh	82	.	.	.	.	.	.	.	2	.	13	63	.	.	.	.	.	.	2	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	77%	gh
s	352	.	.	.	.	.	.	.	.	2	.	18	305	21	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	87%	s
z	180	1	1	.	.	.	.	.	4	.	.	.	32	128	.	1	.	1	2	2	.	.	.	.	.	2	.	.	.	.	.	.	.	.	6	71%	z
m	332	.	4	16	1	.	.	.	.	.	.	.	.	.	153	18	.	6	55	2	.	1	1	.	3	6	1	.	1	.	.	3	1	.	60	46%	m
n	401	2	.	35	.	3	.	.	.	.	.	.	.	.	41	67	.	25	123	1	6	1	.	.	3	1	.	.	.	1	2	.	.	.	89	17%	n
nj	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		nj
j	137	.	.	.	.	.	.	.	.	2	.	5	.	.	.	.	.	5	22	.	10	17	.	.	.	4	5	.	.	.	.	.	.	.	67	4%	j
w	137	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	92	2	.	1	.	2	.	.	.	.	.	1	1	1	.	.	36	67%	w
R	1533	37	28	38	4	.	1	.	51	10	134	26	88	6	3	.	.	8	109	560	23	7	2	3	10	4	15	5	4	4	8	1	1	.	330	37%	R
l	1063	38	15	28	3	1	.	.	.	1	4	2	1	.	1	5	.	11	13	7	501	1	.	.	.	3	.	.	.	.	2	.	1	.	425	47%	l
ui	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0		ui
on	145	.	.	1	.	1	.	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	84	2	24	7	7	4	.	.	3	2	.	.	6	58%	on
an	369	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	8	261	45	1	13	2	.	1	6	2	2	2	18	71%	an
un	333	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	5	.	.	.	8	288	4	6	6	.	1	.	.	.	.	10	86%	un
i	840	1	3	3	.	2	.	.	1	.	4	17	1	4	.	.	.	7	2	.	1	35	.	.	.	688	22	1	.	.	.	1	2	.	43	82%	i
e	365	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	354	.	.	.	.	.	.	.	7	97%	e
ai	693	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	1	5	.	.	.	.	.	.	6	649	10	2	.	.	.	1	16	94%	ai
a	1503	.	1	1	.	.	.	.	.	.	.	1	.	.	.	.	.	1	1	17	.	1	.	.	2	1	1	621	358	8	.	.	.	7	36	90%	a
)	449	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	2	.	2	.	5	.	28	3	389	5	.	.	9	4	87%	)
o	217	2	1	1	.	.	.	.	.	.	.	.	.	.	.	2	.	8	24	5	.	5	.	.	.	6	3	.	.	5	130	.	1	.	24	60%	o
u	610	1	2	12	1	.	.	.	.	.	.	1	.	.	5	5	.	6	5	5	.	6	.	.	4	43	3	1	.	2	12	393	20	7	69	64%	u
y	247	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	3	.	.	1	12	7	.	.	.	.	.	218	.	5	88%	y
&	1252	2	8	4	1	.	.	.	.	.	.	.	.	.	.	1	.	6	2	26	3	.	.	1	4	6	3	23	15	2	.	.	1052	88	84%	&
ins		10	47	22	5	5	0	0	9	14	12	22	13	4	14	11	0	50	210	133	10	67	25	15	41	71	40	13	19	40	36	25	16	10		69 %	

Figure 2: Matrice de confusion du système.

	nb	occlusives	fricatives	autr.cons.	voy. nas.	voy. orales	omis	
occlusives	5359	4632 (86%)	8 (0%)	189 (4%)	2 (0%)	18 (0%)	507 (9)	occlusives
fricatives	863	36 (4%)	665 (77%)	124 (14%)	.	3 (0%)	34 (4)	fricatives
autr.cons.	3603	255 (7%)	330 (9%)	1897 (53%)	24 (1%)	76 (2%)	1007 (28)	autr.cons.
voy. nas.	847	3 (0%)	.	14 (2%)	720 (85%)	69 (8%)	34 (4)	voy. nas.
voy. orales	6176	46 (1%)	33 (1%)	192 (3%)	14 (0%)	5579 (90%)	292 (5)	voy. orales
ins		115 (2)	74 (9)	514 (14)	81 (10)	270 (4)		69 %

Figure 3: Matrice de confusion des grandes classes.

3 conclusion

Nous avons mené cette étude dans le but de poursuivre nos efforts dans le décodage acoustico-phonétique de la parole continue multilocuteur. Les résultats obtenus pour les voyelles nous laissent à penser qu'il sera difficile de les améliorer sans adaptation au locuteur. Les résultats nous encouragent à persévérer en ce qui concerne les consonnes principalement dans deux directions : l'amélioration de la base de règles et des procédures d'extraction d'indices accoustiques.

4 Références

[1] Carbonell N. et alii "An Expert System For the Automatic Reading of French Spectrograms", IEEE-ICASSP. paper 42-8. 1984.

[2] Lonchamp F. "Reading Spectrograms : The View from the Expert", in J.P. Haton (ed), Fundamentals in Computer Understanding : Speech and Vision. Cambridge University Press. 1987

[3] Fohr D. "APHODEX : un systeme expert en decodage acoustico-phonetique de la parole continue" These d'Universite en informatique, Nancy. 1986

[4] Carbonell N. and Pierrel J.M. "Architecture and Knowledge Sources in a Human-computer Oral Dialogue System". NATO Workshop on Multimodal Dialogues including Voice. Corse 1986

Reconnaissance multilocuteur de voyelles par un réseau connexionniste auto-organisateur

Franck Poirier

Télécom Paris
CNRS URA - 820
46 rue Barrault
75634 Paris Cedex 13

Résumé :

On présente une expérience de reconnaissance des voyelles du français, en multilocuteur, par un réseau connexionniste auto-organisateur. Plusieurs tailles de réseau, différentes règles d'adaptation et différentes entrées ont été testées. Les résultats obtenus sont comparés à ceux des k plus proches voisins.

Mots clés :
Reconnaissance phonétique, réseau connexionniste, carte phonotopique.

1- Introduction

Les performances d'un système de reconnaissance analytique de la parole dépendent en grande partie de la qualité du décodage acoustico-phonétique. Nous présentons une première expérience de reconnaissance des voyelles du français par un réseau connexionniste auto-organisateur [Koh, 88]. Le problème majeur du décodage acoustico-phonétique est la difficulté de définir les modèles statistiques associés aux différents allophones d'une langue. La grande variabilité du signal de parole et la coarticulation entre phonèmes, entraînent un recouvrement des réalisations spectrales des différents phonèmes.

Depuis plusieurs années, en reconnaissance de la parole, les réseaux de neurones formels ont montré leur capacité à prendre en compte la variabilité intra-locuteur et inter-locuteur.

2- Analyse

La base de sons utilisée se compose d'un ensemble de 105 syllabes isolées prononcées par 40 locuteurs de sexe masculin. Cette base comprend 11 voyelles dont 8 orales /a, o, œ, e, ɛ, u, y, i/ et 3 nasales /ã, ɛ̃, ɔ̃/, sur les 16 qui composent le français.

Le signal de parole est échantillonné à 16kHz sur 16 bits. Un fichier de parole, au format GRECO, est créé par locuteur. Une segmentation automatique permet de détecter le "début" de chaque syllabe qui correspond au relâchement de l'occlusion (burst) pour les syllabes commençant par une occlusive sourde ou bien au début du voisement pour les autres syllabes. Le point de relâchement correspondant à l'ouverture du conduit oral est également détecté pour les syllabes commençant par une consonne nasale.

Pour chaque syllabe, l'énergie du signal est calculée, depuis le point T_0 de relâchement ou de début de voisement jusqu'à la fin de la syllabe. Puis 8 vecteurs de 8 coefficients MFCC sont calculés sur des fenêtres d'environ 30ms (504 points), avec recouvrement de 50%. L'analyse débute 3*252 points avant T_0 et se termine 6*252 points après T_0. Ainsi, pour chaque syllabe, la partie stable de la voyelle est représentée par 8 vecteurs de $\Re^8$.

L'apprentissage a été effectué sur 15 locuteurs, soit environ 1500 occurrences de voyelles, les 10 autres locuteurs (1000 occurences) ont servi pour tester le réseau.

3- Réseau auto-organisateur

3-1 Fonction de transfert du neurone formel

Cette partie a uniquement pour but de donner un aperçu des réseaux auto-organisateurs. Les bases neuro-physiologiques et tout les développements théoriques ne seront pas développés [Koh-2, 88].

Le fonctionnement d'un réseau auto-organisateur s'inspire "grossièrement" du fonctionnement des neurones. Par la suite, par abus de langage, on parlera indifféremment de neurone ou de cellule pour désigner un neurone formel. Chaque neurone reçoit et transmet des signaux de ou vers d'autres cellules par l'intermédiaire de ses synapses. La relation entrée-sortie du neurone peut être décrite par l'équation : $s_i = \sigma(\sum_{j=1..n} \mu_{ij}\, e_{ij})$

- s_i : sortie du $i^{ème}$ neurone
- σ : fonction de transfert non linéaire
- e_{ij} : activité de la $j^{ème}$ entrée du $i^{ème}$ neurone
- μ_{ij} : coefficient d'efficacité du couplage synaptique

En général, la fonction σ est remplacée par l'identité, ce qui donne l'équation :

$$s_i = \sum_{j=1..n} \mu_{ij}\, e_{ij} \qquad (1)$$

L'intérêt du neurone est sa faculté d'apprentissage, elle est décrite par la règle de Hebb (Hebb, 49) qui modifie les couplages synaptiques de façon proportionnelle au produit de l'entrée et de la sortie.

La modification des poids est décrite par l'équation :

$$d\,\mu_{ij} / dt = \alpha\, s_i\, e_{ij} - \beta(s_i)\,\mu_{ij} \qquad (2)$$

avec α : coefficient d'apprentissage
$\beta(s_i)$: fonction d'oubli

Le premier terme (terme de Hebb) correspond au couplage de la sortie s_i aux signaux d'entrées. Il rend le neurone sélectif à certaines entrées. Le deuxième terme correspond à un effet de perte non-linéaire (saturation).

Dans un réseau auto-organisateur, on distingue deux types de connexions. Les neurones formels sont disposés sur un maillage régulier de dimension 1 ou 2 et reçoivent les mêmes entrées $[e_1\ e_2\ ...\ e_n]$ par l'intermédiaire d'un ensemble de connexions externes. Un ensemble de connexions internes (ou latérales) réalise la fonction d'inhibition latérale [Bur, 88]. Pour ces dernières connexions, l'efficacité du couplage synaptique ω_{ik} entre les cellules i et k ne dépend que de la distance entre ces deux cellules. La fonction ω_{ik} a la forme d'un chapeau mexicain. Au cours de l'apprentissage, chaque cellule devient sélective à certaines entrées (ou active pour ces entrées). La fonction ω_{ik} a pour rôle de renforcer la réponse des cellules voisines et de former ainsi des îlots d'activité (ou bubbles d'activité). A des fins d'efficacité de calcul, cette fonction est remplacée par la notion de voisinage sur les cellules.

En tenant compte des connexions latérales, l'équation d'entrée-sortie devient :

$$s_i = \Sigma_{j=1..n}\ \mu_{ij}\, e_{ij} + \Sigma_{k\neq i}\ \omega_{ki}\, s_k$$

Equation qu'on peut simplifier en posant :

$m_i = [\mu_{i1}\ \mu_{i2}\ ...\ \mu_{in}]$
$x = [e_1\ e_2\ .....\ e_n]$

d'où $s_i = m_i^T . x + \Sigma_{k\neq i}\ \omega_{ki}\, s_k \qquad (3)$

Quand on laisse le réseau évoluer suivant les équations (2) et (3), après une initialisation aléatoire des poids synaptiques, l'expérience montre qu'il se forme, au bout d'un certain temps, des îlots d'activité. Cela signifie qu'à partir d'un certain apprentissage, pour une entrée donnée x, un groupe de neurones voisins a une activité maximum alors que le reste des neurones est à l'état de repos (propriété d'auto-organisation).

3-2 Algorithme d'apprentissage

Des hypothèses simplificatrices permettent de simplifier et d'accélérer l'apprentissage tout en conservant la propriété d'auto-organisation du réseau. Par un changement d'échelle sur les coefficients synaptiques μ_{ij} et sur les entées e_{ij}, il est possible de rendre s_i égale à 0 à l'état de repos ou 1 pour l'état de maximum d'activité.

Dans ce cas, on peut imposer que $\beta(s_i)$ ne prenne que deux valeurs :

$\beta(s_i) = 0$ si $s_i = 0$
$\beta(s_i) = \alpha$ si $s_i = 1$

L'équation d'apprentissage (2) devient :

si le neurone i est appartient à l'îlot d'activité
$d\,\mu_{ij} / dt = \alpha\,(e_{ij} - \mu_{ij})$
sinon $d\,\mu_{ii} / dt = 0 \qquad (4)$

L'équation (4) montre clairement que les poids synaptiques tendent à suivre de façon adaptative le signal d'entrée à l'intérieur d'un îlot d'activité.

Lors de l'apprentissage, les échantillons (vecteurs) sont présentés successivement au réseau. Sans attendre la formation d'un îlot d'activité, on fait l'hypothèse qu'il serait centré autour du neurone c qui rend maximum $m_c^T . x$ et on applique l'équation (4) pour modifier les couplages synaptiques μ_{ij}. Comme l'équation (4) fait converger les normes des vecteurs m_i vers une constante, il est alors équivalent de choisir c tel que $\| x - m_c \|$ soit minimum.

En final, cela donne l'algorithme d'apprentissage suivant :

1. Initialisation des poids m_i , t = 0
2. Apprentissage
2-1 Soit x(t) un vecteur d'apprentissage
déterminer c tel que $\|x(t) - m_c\|$ soit minimum
2-2 $m_i = m_i + \alpha(t)\,[x(t) - m_i]$ pour i dans $N_c(t)$
2-3 t = t + 1
2-4 Répéter 2-1 à 2-3 jusqu'à convergence

$N_c(t)$ est le rayon de l'îlot d'activité. L'expérience montre que pour assurer une bonne auto-organisation du réseau, N_c et α doivent être des fonctions décroissantes en fonction du temps. Le réseau converge alors vers une approximation des entrées x. Cela signifie que la densité des vecteurs m_i suit à peu près les mêmes variations que celles des entrées x.

3-3 Auto-organisation et quantification vectorielle

L'algorithme d'apprentissage des réseaux auto-organisateurs effectue une quantification vectorielle de $\Re^n$. Une entrée x est quantifiée par le vecteur d'efficacité synaptique m_i le plus proche. Comme tout processus de quantification vectorielle, le réseau permet de coder x par m_i ou encore de reconnaître x comme un élément de la classe dont m_i est un prototype.

Pour la reconnaissance, un réseau auto-organisateur permet une classifi-cation automatique (ou apprentissage sans professeur) sur un ensemble de vecteurs d'apprentissage. L'espace $\Re^n$ est alors partitionné en régions de Voronoï V_j qui correspondent aux vecteurs m_j du réseau. Une règle doit permettre d'associer à chaque neurone une classe C_i.

Cette phase de l'apprentissage correspond à la calibration du réseau. Le neurone j est associé à la classe C_i qui maximise l'expression $n_{ij} / n_{.j}$

n_{ij} : nombre d'échantillons de la classe i qui activent le neurone j
$n_{.j}$: nombre d'échantillons qui activent le neurone j

La calibration du réseau permet de dresser une table, appelée carte phonotopique, qui associe une étiquette à chaque neurone du réseau.

4- Apprentissage

Il faut d'abord déterminer la structure du réseau, les entrées et certains paramètres de l'apprentissage.

Pour la structure, les neurones sont arrangés sur une grille carrée de dimension 2 qui semble préférable à la dimension 3. Les réseaux de taille 8*8, 12*12, 16*16 et 20*20 ont été étudiés.

L'entrée du réseau est un vecteur de 8 coefficients MFCC. Les 8 positions différentes de la fenêtre d'analyse ont été utilisées pour savoir laquelle donnait les meilleurs résultats et tester si le maximum d'énergie est un bon critère pour localiser les voyelles. Les paramètres suivants sont utilisés pour adapter les coefficients synaptiques :

$\alpha(t)$: coefficient d'apprentissage
$N_c(t)$: rayon de l'îlot d'activité

4-1 Taille du réseau

La taille du réseau est variable afin d'étudier son influence sur les performances et le comportement du réseau en terme de généralisation sur de nouvelles données. Une première remarque s'impose, avec un réseau de taille 20*20 et environ 1500 échantillons d'apprentissage, soit moins de 4 échantillons par neurone, on ne peut pas espérer de bonnes facultés de généralisation. Dans ce cas, au cours de l'apprentissage, le réseau enregistre tous les "cas particuliers". Pour ces raisons, les expériences sur un réseau 20*20 n'apparaissent pas dans les résultats.

En autocohérence, les résultats croissent avec la taille du réseau. Par contre, les facultés de généralisation sont meilleures avec le plus petit réseau de taille 8*8 puisqu'il obtient le meilleur taux de reconnaissance, de l'ordre de 60%, sur les données de test. Les deux autres réseaux de taille 12*12 et 16*16 obtiennent respectivement pour la 6ème fenêtre d'analyse 58,8% et 58,4% de reconnaissance. Dans l'optique d'un apprentissage plus important, il faut préférer des réseaux de taille 12*12 ou 16*16 qui offrent un bon compromis de performances entre les données d'apprentissage et de test.

4-2 Entrée du réseau

On note une forte variation du taux de reconnaissance en fonction de la position de la fenêtre d'analyse, aussi bien sur les données d'apprentissage que de test. Entre la troisième et la sixième fenêtre (distantes de 47ms), la différence atteint près de 10%. Les fenêtres 5, 6, 7 donnent les meilleures résultats.

D'autre part, même sur les meilleures fenêtres, le taux de reconnaissance n'est pas satisfaisant. Pour l'améliorer, il est envisageable de caractériser une voyelle par plusieurs vecteurs de coefficients MFCC. Cette approche est présentée dans la suite de l'article, après la description des expériences avec un seul vecteur en entrée du réseau.

4-3 Paramètres de l'apprentissage

Les paramètres doivent être déterminés de façon à permettre une bonne convergence du réseau et un apprentissage rapide. Deux méthodes ont été récemment proposées [Bra, 89].

La première, appelée méthode exponentielle, utilise un voisinage $N_c(t)$ circulaire dont le rayon $R(t)$ décroît exponentiellement, en conjonction avec un coefficient $\alpha(t)$ à décroissante linéaire. Seul les neurones de $N_c(t)$ voient leur efficacité synaptique modifiée.

La deuxième méthode, appelée méthode centrale, adapte à chaque fois le poids de tous les neurones du réseau. Le coefficient d'apprentissage $\alpha(t)$ décroît toujours en fonction du temps mais aussi en fonction de la distance qui sépare un neurone du neurone activé par l'échantillon présenté au réseau.

Sur des données réelles de parole, la méthode centrale donne de meilleurs résultats. Cette deuxième règle d'adaptation est celle utilisée par la suite.

Il faut également déterminer le nombre d'itérations et l'état du réseau au cours de l'apprentissage. Une itération correspond à la présentation d'un vecteur en entrée du réseau. Pour obtenir la convergence, le nombre d'itérations a été porté à 20 000. A chaque présentation de l'ensemble du corpus d'apprentissage, le nombre de vecteurs reconnus à cet instant est évalué.

5- Résultats

5-1 Représentation de la voyelle par 8 coefficients MFCC

Les résultats obtenus en fonction de "bons" paramètres d'apprentissage sont maintenant présentés plus en détail. Les résultats sont donnés pour un réseau de taille 12*12 et pour la 6ème fenêtre d'analyse.

En fin d'apprentissage, on obtient, comme résultat de la calibration du réseau, une carte phonotopique qui contient, par ligne, les informations suivantes pour chaque neurone :

1ère : voyelle v_1 qui active le plus le neurone
2ème : nombre d'échantillons de v_1 qui ont activé le neurone (n_{i1})
3ème : nombre total d'échantillons activateurs $n_{.1}$
4ème n_{i2}
5ème : deuxième voyelle v_2 qui active le neurone

La figure 3 montre la carte phonotopique obtenue. On constate que le nombre d'échantillons qui activent chaque neurone varie de 1 (ligne 3, colonne 5) à 23 (ligne 3, colonne 3), ce qui indique que les densités de probabilité des entrées varient fortement d'un endroit à l'autre. Il arrive fréquemment que n_{i1} soit inférieur à la moitié de $n_{.1}$ Cela entraîne nécessairement une dégradation du taux de reconnaissance. D'autre part, on remarque que la

répartition des étiquettes est assez homogène. Chaque voyelle étiquette entre 9 (/i/) et 16 (/o/ , /e/) neurones.

Sur les résultats par locuteur et par voyelle, on note des variations du taux de reconnaissance de 50% à 78%. La variation est encore plus forte entre les voyelles orales avec 84% pour /y/, 82% pour /o/ et les voyelles nasales avec 43% pour /ɛ̃/ et 42% pour /ɔ̃/.

La figure 1 correspond à la matrice de confusion du réseau sur les données d'apprentissage. On peut noter que la plupart des confusions se produisent entre voyelles orales et leur correspondante nasale (/a/ et /ã/ ou /ɛ/, /ɛ/ et /ɛ̃/, /o/ et /ɔ̃/) ou entre voyelles qui ne diffèrent que par un seul trait /e/ et /ɛ/ (voyelles ouvertes d'avant), /o/ et /u/ (voyelles fermées d'arrière), /a/ et /ɔ̃/. Ces confusions sont classiques.

Il convient de remarquer que le faible taux de reconnaissance des voyelles nasales tout comme les confusions signalées entre voyelles orales et nasales sont dus au fait que, le plus souvent, seule la partie orale de la voyelle est analysée.

5-2 Représentation de la voyelle par 3 vecteurs de 8 MFCC

Comme les meilleurs résultats sont obtenus avec les fenêtres d'analyse 5, 6 et 7, plusieurs structures de réseaux et plusieurs stratégies d'apprentissage ont été testées avec ces trois fenêtres.

La première stratégie consiste à concaténer les 3 vecteurs de $\Re^8$ pour former un seul vecteur de $\Re^{24}$ qui sert d'entrée au réseau. Chaque neurone j possède alors un vecteur de coefficients synaptiques m_j de longueur 24.

La deuxième stratégie consiste à soumettre au réseau l'un après l'autre, les vecteurs v_1, v_2, v_3. Les vecteurs de coefficients synaptiques reste alors de longueur 8. Tout se passe comme si l'apprentissage se faisait sur 3 fois plus de vecteurs.

La figure 2 montre les résultats obtenus avec les deux stratégies précédentes sur les données d'apprentissage (en abscisse) et de test (en ordonnée). Les noms des réseaux sont de la forme suivante :

t [M] / n[nn] [C]20

- t : taille du coté du réseau, 12 ou 16
- M : indique la 2ème stratégie
- nnn : numéros des fenêtres d'analyse
- C : indique la méthode centrale

Il apparait clairement que la première stratégie n'apporte que 2% à 3% d'amélioration sur un réseau de taille 12*12, avec un taux de reconnaissance de l'ordre de 64% sur les données de test. Par contre, le temps d'apprentissage et de reconnaissance augmente proportionnellement à la longueur des vecteurs. Ces résultats décevants sont certainement dus à la non compacité des nuages dans l'espace $\Re^{24}$.

Pour la deuxième stratégie, on considère toujours que la densité de probabilité est uniforme dans la région de Voronoï V_j correspondant au neurone j et qu'elle peut être approchée par : $P(C_i \mid V_j) = n_{ij} / n_{.j}$

Le problème est que pour identifier une voyelle v inconnue, on dispose de 3 vecteurs v_1, v_2, v_3 qui activent respectivement les neurones j_1, j_2, j_3 et qu'il faut combiner ces informations pour une décision de reconnaissance.

On considère que la voyelle inconnue v se trouve dans la réunion des trois régions de Voronoï $V_{j1} \cup V_{j2} \cup V_{j3}$. Les densités de probabilités conditionnelles peuvent être estimées par :

$$P(C_i \mid V_{j1} \cup V_{j2} \cup V_{j3}) = \frac{n_{ij1} + n_{ij2} + n_{ij3}}{n_{.j1} + n_{.j2} + n_{.j3}}$$

La figure 2 montre que cette stratégie améliore les résultats de 10% par rapport à ceux obtenus avec une seule fenêtre. Le taux de reconnaissance atteint 70% sur les données de test.

5-3 Comparaison avec la méthode des plus proches voisins (k-PPV)

Pour évaluer les résultats précédents, il convient de les comparer à ceux obtenus par une décision classique sur les plus proches voisins. Un vecteur x, de classe inconnue, est reconnu comme un élément de la classe la plus représentée parmi ses k plus proches voisins. La distance utilisée est la suivante : $d^2(x,y) = (x - y)^T \, COV^{-1} \, (x - y)$
où COV est la matrice de covariance de l'ensemble des vecteurs d'apprentissage. Les meilleurs résultats sont obtenus pour k = 9 et un seuil de rejet de l'ordre de 10%. Le taux de reconnaissance est alors de 58,6%.

Il apparaît que le réseau auto-organisateur, sans rejet, est légèrement supérieur aux k-PPV avec 10% de rejet et que de plus la décision de reconnaissance est beaucoup plus rapide.

6- Conclusions et perspectives

Il faut replacer ces résultats dans le cadre de la reconnaissance de voyelles multilocuteur et polycontextuelles. Ce problème est particulièrement difficile en français qui possède un système vocalique très riche. Les résultats sont sensiblement supérieurs à ceux obtenus à partir d'un système de reconnaissance d'indices binaires [Bon,86] [Bon, 87]. D'autre part, ils sont également supérieurs à ceux obtenus pour un seul locuteur masculin en voyelles isolées sur le suédois, qui comporte il est vrai plus de 17 voyelles [Bra, 89].

Les résultats obtenus sont donc très encourageants et montre l'intérêt d'approfondir cette première expérience de reconnaissance de voyelles par des réseaux auto-organisateur. Pour cette application, dans l'avenir, il semble nécessaire d'augmenter le nombre de locuteurs, de mieux localiser la partie stable de la voyelle et de réfléchir à une meilleure représentation de chaque voyelle.

Bibliographie

[Bon, 86] A. Bonneau, G. Mercier, M. Gérard, M. Rossi, Le décodage acoustico-phonétique à l'aide du système expert Serac-Iroise. 15ème JEP, Aix en Provence, mai 1986.

[Bon, 87] A. Bonneau, M. Rossi, Reconnaissance des voyelles et des traits vocaliques en français. 16ème JEP, Hammamet, 1987.

[Bra, 89] P. Brauer, P. Knagenhjelm, Infrastructure in Kohonen maps. IEEE, S12.13,1989.

[Bur, 88] Y. Burnod, An adaptative neural network, the celbral cortex. Masson, 1988.

[Koh, 88] T. Kohonen, Self-Organization and Associative Memory. Springer Verlag, Series in Information Sciences, 1988.

[Koh-2,88] T. Kohonen, The "Neural" Phonetic Typewriter. IEEE Computer, mars 1988.

	a	o	œ	e	ɛ	u	y	i	ã	ɛ̃	ɔ̃
a	101 69				7 5		1 1	1 1	12 8	21 14	3 2
o		119 82	4 3			9 6	2 1		2 1		10 7
œ	1 1	10 8	81 61	5 4	2 2	9 7	14 11	3 2	2 2		5 4
e		1 1	6 5	99 75	5 4	1 1	7 5	11 8		1 1	1 1
ɛ	1 1	2 2	4 3	24 18	85 64	1 1	7 5	2 2		3 2	3 2
u		25 17	4 3	3 2		85 58	8 5	10 7			11 8
y				4 3	12 9	1 1	1 1	111 84	3 2		
i		1 1		13 10		9 7	11 8	96 73			2 2
ã	20 14	3 2	2 1		1 1	1 1	1 1		101 69	7 5	10 7
ɛ̃	26 18		9 6	3 2	23 16	2 1	3 2	2 1	8 5	63 43	7 5
ɔ̃	1 1	19 13	4 3	2 1		18 12	5 3	5 3	29 20	2 1	61 42
T	150 10	180 12	118 8	161 10	124 8	136 9	170 11	133 9	154 10	97 6	113 7

Figure 1 - Matrice de confusion

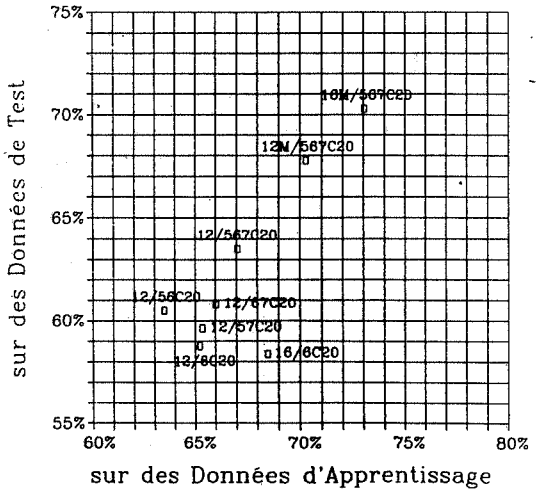


Figure 2 - Taux de reconnaissance

i 17 19 1 e	i 15 19 1 œ	y 5 9 2 i	y 8 17 5 u	i 9 13 2 ɔ	i 6 11 2 y	u 9 11 2 i	o 9 11 2 u	u 8 9 1 o	u 16 17 1 ɔ	u 10 14 2 o	u 4 11 4 ɔ
u 8 14 2 i	u 10 19 3 i	ɔ 7 18 4 u	u 10 21 4 o	y 4 9 2 œ	ɔ 4 6 1 œ	u 8 16 4 ɔ	ɔ 7 11 3 u	o 11 12 1 u	o 2 9 2 œ	o 4 10 3 u	o 10 14 2 ɔ
ã 6 14 5 ɔ	i 8 20 7 u	ɔ 9 23 5 o	u 2 4 1 œ	œ 1 1 0 a	o 6 11 3 u	o 12 16 3 u	o 12 17 2 u	o 6 9 2 u	o 15 17 1 u	o 2 3 1 ɔ	ɔ 3 6 2 ã
ɛ 6 11 3 ɛ̃	y 5 11 3 ɛ	y 7 11 4 œ	y 8 10 1 œ	œ 12 13 1 u	œ 8 12 2 ɛ̃	o 6 10 3 u	ɔ 8 13 2 o	o 8 13 3 ɔ	ã 3 7 3 ɔ	ɔ 9 11 1 ɛ	ɔ 6 11 3 ã
e 7 15 6 i	y 4 11 4 i	y 15 17 2 œ	y 9 16 5 ɔ	œ 14 16 1 y	œ 4 10 4 ɛ̃	œ 10 14 3 o	o 6 9 1 u	ɔ 4 6 1 ɛ	ã 4 9 4 ɔ	ã 11 12 1 ɛ̃	ã 11 12 1 ɔ
i 11 14 2 e	e 7 15 4 i	e 3 7 3 y	e 3 5 1 œ	ɛ 6 8 1 a	œ 7 7 0 a	o 6 13 4 œ	o 4 6 2 ã	ã 4 7 1 œ	ã 7 11 2 a	ã 12 14 2 ɔ	ã 7 11 4 ɔ
i 15 15 0 a	e 8 10 1 ɛ	e 4 8 3 ɛ	e 7 12 2 y	y 4 9 3 ɛ	ɛ 3 4 1 e	ɛ 3 4 1 ɛ̃	ã 2 5 1 œ	ã 7 13 3 ɔ	ã 11 14 3 a	a 6 11 5 ã	a 6 8 1 ã
i 8 10 2 e	e 6 6 0 a	e 5 5 0 a	e 8 11 3 ɛ	e 5 8 2 ɛ	ɛ 4 5 1 ɛ̃	ɛ̃ 7 10 2 a	ɔ 4 8 2 a	ã 5 6 1 ɔ	ã 4 7 3 a	a 16 16 0 o	a 6 12 3 ã
i 7 12 2 e	e 6 8 1 ɛ	e 6 11 5 ɛ	ɛ 7 8 1 e	ɛ 9 11 1 e	ɛ 5 8 3 ɛ̃	ɛ̃ 5 10 3 a	ɛ̃ 4 7 2 ã	a 5 8 3 ɛ̃	a 6 13 6 ã	a 11 11 0 o	ɛ 5 9 2 a
e 7 14 3 i	e 5 7 2 ɛ	e 7 10 2 ɛ	œ 1 3 1 e	ɛ 7 9 2 e	ɛ 6 13 6 ɛ̃	ɛ̃ 4 6 1 a	ɛ̃ 4 5 1 a	a 5 8 3 a	ɛ̃ 6 9 2 ã	a 3 6 3 ɛ̃	a 5 12 5 ɛ̃
y 7 8 1 e	e 5 9 2 ɛ	œ 6 12 2 ɛ	y 9 11 2 œ	ɛ 7 8 1 a	ɛ 4 6 2 ɛ̃	ɛ̃ 4 7 2 a	ɛ̃ 7 12 4 ɛ̃	a 6 10 2 a	ɛ̃ 5 9 3 ɛ̃	a 6 8 1 ã	a 12 14 2 ɛ̃
y 14 16 1 e	y 9 9 0 a	œ 7 16 3 ɔ	œ 11 14 2 y	y 3 6 3 ɛ̃	ɛ 9 13 3 ɛ̃	ɛ̃ 7 9 1 ɛ	ɛ̃ 4 7 2 a	ɛ̃ 6 9 3 a	ã 7 12 4 ɔ	ɛ̃ 5 7 2 a	a 7 10 2 ã

Figure 3 - Carte phonotopique

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

METHODOLOGIES POUR L'EVALUATION PHONETIQUE

C. Bourjot A. Boyer D. Fohr J.P. Haton

CRIN/INRIA Lorraine BP239 54506 VANDOEUVRE Cedex FRANCE

Abstract

This paper deals first with a methodology to perform semi-automatically labelling and its possible use for phonetic assessment which requires huge natural speech multilingual multispeaker corpora. The method is described and the results obtained during the labelling are compared with manual labelling. Then two methodologies to assess an acoustic phonetic decoder based on dynamic programming are proposed. They have been used to test APHODEX and the main results are indicated.

1 INTRODUCTION

Dans [BOURJOT,1988], nous préconisons deux types d'évaluation, une évaluation analytique des systèmes qui porte sur les performances de chaque composant d'un système complexe (décodage acoustico phonétique, lexique, syntaxe,..) et une évaluation globale qui mesure les performances générales et fournit une idée sur les qualités d'un système dans son environnement. Ce travail s'intéresse plus particulièrement à l'évaluation des systèmes de décodage acoustico-phonétique (DAP). Evaluer un système de DAP consiste à déterminer un certain nombre de résultats comme la matrice de confusion, la matrice d'omission, la matrice d'insertion, le taux de reconnaissance de phonèmes, le taux de reconnaissance par phonème ou par classe de phonèmes, le contexte des erreurs,... Ces diverses estimations vont permettre d'obtenir une vision globale du comportement du système pour faire du diagnostic dans le but entre autres de faciliter l'utilisation des sorties du DAP par les autres modules du système de reconnaissance (par exemple le lexique), ou d'améliorer le système de DAP.

La détermination d'un ensemble de tests qui conduit à une évaluation fiable nécessite habituellement l'utilisation de très grandes bases de données de manière à obtenir des résultats statistiquement valides liés au système de reconnaissance et non pas à l'élocution des locuteurs. Pour la conception de ces bases de données, on pourra se reporter à [Zue,1989]. Nous proposons de comparer deux méthodologies pour réaliser l'évaluation du DAP, l'une utilisant une base de données étiquetée, l'autre non. Afin de fournir une aide à l'étiquetage de corpus de parole, nous avons conçu un système d'étiquetage semi-automatique que nous allons décrire dans une première section. Les méthodologies d'évaluation feront l'objet de la section suivante. La troisième section portera sur l'évaluation d'un système de DAP développé au CRIN, APHODEX, à l'aide des deux méthodes. La dernière partie portera sur la comparaison des deux méthodologies et indiquera les principales conclusions que nous avons tirées.

2 ETIQUETAGE SEMI-AUTOMATIQUE

Faire de l'étiquetage semi-automatique est un des objectifs du projet ESPRIT SAM. Il s'agit de concevoir un outil d'aide à l'étiquetage de bases de données parole. Dans ce but, plusieurs voies ont été explorées, toutes sont issues des travaux antérieurs en DAP. Certains utilisent directement un système de DAP. Par exemple, [DALSGAARD,1989] utilise un réseau de neurones. Ce système est dépendant du locuteur. D'autres utilisent un algorithme de mise en correspondance [SWENDSEN,1989], [PERENNOU,1989]. La méthode que nous proposons repose sur un algorithme de programmation dynamique. Elle nécessite un petit corpus étiqueté début-fin (c'est à dire chaque phonème est, repéré par une marque de début et une marque de fin dans le signal), la transcription standard des phrases à étiqueter et une base de règles phonologiques.

2.1 ETIQUETAGE AUTOMATIQUE

Un préprocesseur calcule le taux de passage par zéro, le centre de gravité de l'énergie globale, l'énergie totale, l'énergie dans différentes bandes de fréquence, la transformée de Fourier rapide, le pitch en utilisant un algorithme de Rabiner adapté. Ces différents paramètres permettent de réaliser une classification en grandes classes phonétiques (noyaux vocaliques, fricatives, plosives, voyelles fricatives, sonnantes). Cette suite de grandes classes est ensuite mise en correspondance par un algorithme de programmation dynamique avec les suites de phonèmes obtenues en appliquant des règles phonologiques sur la transcription standard (forme pleine). Pour minimiser les erreurs faites par le processus d'alignement, les distances locales pendant l'étape de mise en correspondance sont calculées à partir d'une matrice de confusion reflétant en première approximation un modèle d'erreur du système. C'est pourquoi un apprentissage est nécessaire pour déterminer la matrice du système en exécutant un alignement entre un petit corpus étiqueté manuellement début-fin et les sorties du DAP. Puisque la méthode repose à la base sur la détermination de grandes classes, la station d'étiquetage semi-automatique est indépendante du locuteur et son adaptation à d'autres langues européennes semble facile.

2.2 RESULTATS

Les tests ont été réalisés dans une salle de terminaux, sans précautions particulières. A l'heure actuelle, cette station est utilisée au CRIN pour étiqueter les bases de données parole et 40 phrases ont été traitées. Il est intéressant de comparer sur 10 phrases les performances de l'étiquetage automatique avec l'étiquetage manuel réalisé par un expert phonéticien. La figure 1 résume les principaux résultats de cette comparaison en terme de décalage des frontières. Signalons que la précision maximale est de 8ms (durée d'une fenêtre d'analyse). Nous pouvons tout de suite conclure que 85 des phonèmes ne nécessiteront pas de correction manuelle.

Le système de correction manuelle qui est intégré à cette station est interactif : il propose une interface graphique en couleurs, avec multifenétrage et menus déroulants. Un exemple d'écran est donné figure 2. On peut comparer sur cette figure les résultats de l'étiquetage automatique et de l'étiquetage manuel. L'essentiel de la correction manuelle consiste à modifier la position des frontières, mais il est également possible d'insérer, de supprimer ou de modifier un phonème. Un phonéticien estime que cette station permet de diviser par 2 le temps d'étiquetage d'une phrase.

3 EVALUATION

3.1 METHODE D'EVALUATION AVEC UN CORPUS ETIQUETE

Cette méthode nécessite un corpus etiqueté début-fin. L'algorithme de programmation dynamique utilise les frontières et c'est pourquoi une matrice de confusion élémentaire, une contrainte locale simple et un terme correcteur suffisent. En effet, les frontières sont utilisées durant le processus de mise en correspondance pour guider l'appariement des phonèmes : un phonème ne peut être apparié à un autre que si l'un est inclus dans l'autre ou si seul un petit décalage les sépare. Nous utilisons la contrainte de base avec les pondérations (1,1,1). La distance locale entre deux phonèmes est calculée à l'aide d'une matrice de confusion dont les coefficients sont :

- 1- (n-1) * epsilon si les deux unités sont identiques (n est le nombre total d'unités, epsilon un coefficient de l'ordre de 0.001),
- epsilon autrement.

Cette matrice ne nécessite aucune connaissance a priori sur les unités et est donc très facile à déterminer.

3.2 METHODE AVEC UN CORPUS NON ETIQUETE

Le corpus n'est plus étiqueté mais on dispose de la transcription pleine de la phrase (tous les phonèmes possibles sont présents). Les principales prononciations de la phrase sont dérivées à partir de cette transcription en utilisant des règles phonologiques. Ce sont essentiellement des règles d'omission puisque l'on part de la forme pleine.

Dans cette méthode, il existe trois sources possibles d'erreurs :

- les erreurs faites par le système de DAP,
- les erreurs de la transcription,
- les erreurs de la mise en correspondance.

Le problème consiste donc à minimiser les erreurs dues à la méthode d'évaluation pour déterminer celles faites par le DAP. Il est donc fondamental de trouver le chemin de mise en correspondance correct et non pas le meilleur au sens d'une quelconque distance. c'est pourquoi nous avons introduit dans le processus de comparaison des connaissances sur le décodeur à évaluer [BOURJOT,1989]. Cette évaluation a donc lieu en deux étapes (figure 3) :

- une étape d'adaptation détermine des résultats intermédiaires qui seront utilisés dans l'etape suivante pour aider la programmation dynamique à trouver le chemin correct. Elle calcule automatiquement une estimation grossière des erreurs les plus courantes du DAP en utilisant un petit corpus étiqueté. Une première estimation de la matrice de confusion du système, le taux d'insertion et le taux d'omission, le nombre maximal d'insertions ou d'omissions consécutives sont évalués.

La méthode utilisée pour cette première étape est celle qui a été décrite au paragraphe précédent. Il est bien sûr évident que le choix du petit corpus est essentiel; il doit contenir un nombre suffisant de locuteurs masculins et féminins pour ne pas refléter les caractéristiques d'un locuteur ou d'un sexe, il doit contenir un nombre suffisant de phonèmes pour que les résultats soient quand même relativement fiables.

- la deuxième étape est l'étape d'évaluation proprement dite. Elle utilise les résultats de l'étape précédente. Dans l'algorithme de programmation dynamique, la contrainte locale est affectée de pondérations calculées à partir du nombre maximal d'insertions ou d'omissions consécutives réalisées par le système. L'algorithme de programmation dynamique tend à maximiser la probabilité de coincidence entre la sortie du DAP et les différentes prononciations de la phrase obtenues en appliquant des règles de phonologie sur la transcription pleine standard. Toutes les prononciations possibles sont en effet comparées à la sortie du DAP et celle qui obtient la probabilité de coincidence maximale est retenue pour la construction des différentes matrices et taux résultats.

Ces deux méthodes ont été utilisées pour évaluer APHODEX, un système de décodage acoustico-phonétique qui a été développé au CRIN [CARBONNEL,1987]. Le corpus qui a servi aux tests consiste en la répétition des phrases phonétiquement équilibrées de Combescure prononcées par deux locuteurs. Cela fait un total d'environ 500 phonèmes. Pour la deuxième méthode, l'adaptation a été réalisée sur un autre corpus. Il a été étiqueté par un expert phonéticien à l'aide de la station d'étiquetage semi-automatique décrite dans le premier paragraphe. Les résultats sont contenus figure 4. Remarquons au préalable que le nombre total de phonèmes est différent pour les deux méthodes bien que le corpus utilisé soit le même. Ceci est dû au fait que pour chaque méthode il est calculé à l'aide de la transcription qui sert effectivement à la détermination des erreurs (étiquetage pour l'une, déduite de la transcription qui obtient le meilleur score pour l'autre). La deuxième méthode trouve moins d'omissions car elle utilise des règles d'omission phonologiques. Par exemple, une règle phonologique prévoit l'omission du e muetdans le mot "porte". Si le décodeur n'a pas trouvé le e muet alors qu'il a été prononcé, aucune erreur ne sera comptabilisée dans la deuxième méthode alors qu'elle sera prise en compte par la première.

4 DISCUSSION

Nous nous proposons de comparer dans ce paragraphe les deux évaluations en terme de corpus, d'algorithme et d'adaptabilité à d'autres unités ou d'autres langues. Dans la première méthode, un changement de corpus impose d'étiqueter le nouveau corpus, ce qui est une tâche longue et fastidieuse même si l'on dispose d'une station d'étiquetage semi-automatique. De plus, il n'est pas improbable que plusieurs phonéticiens interviennent, et se pose alors le problème de l'homogénéité de l'étiquetage. Mais il est bien entendu que dans ce cas l'information utilisée par l'évaluateur est plus proche de la réalité. Par contre, la deuxième méthode permet de changer facilement de corpus puisque on peut facilement obtenir la transcription standard par exemple par une conversion automatique graphème-phonème. La transcription est homogène mais ne reflète pas exactement ce qui a été prononcé.

L'algorithme utilisé dans la première méthode est très simple et ne requiert que peu d'informations sur le système (seuls la nature et le nombre des unités à traiter sont nécessaires). Le deuxième doit compenser le manque d'informations par un algorithme de programmation dynamique plus sophistiqué.

L'adaptabilité de la première méthode aux autres langues est possible, elle sera d'autant plus facile que la station d'étiquetage est multilingue. La seconde demande une base de règles phonologiques de la langue cible. L'adaptabilité des deux méthodes à d'autres unités n'est pas simple car l'étiquetage du corpus doit être refait.

5 CONCLUSION

Nous avons présenté deux méthodes d'évaluation de DAP qui toutes les deux sont basées sur l'utilisation de grandes bases de données de parole. La deuxième méthode qui utilise de la parole non étiquetée fournit de moins bons résultats que la première. Néanmoins, les deux méthodes sont importantes et complémentaires. En effet, étiqueter de la parole est un travail fastidieux et long.

Il est possible d'améliorer la deuxième méthode en introduisant la notion de contexte dans l'algorithme de programmation dynamique. De plus, pour les deux méthodes, il serait intéressant que l'évaluateur s'adapte automatiquement au DAP au fur et à mesure de l'évaluation en actualisant la matrice de confusion. L'ajout de règles phonologiques dans la seconde méthode n'est pas forcément une amélioration : des erreurs du système pourront être considérées comme des variantes de prononciation. Il serait plus intéressant de disposer d'une transcription à l'écoute qui reste toujours moins fastidieuse à obtenir que l'étiquetage demandé par la première. Nous envisageons de séparer les résultats de l' évaluation de la segmentation et de l'évaluation de l'étiquetage. La segmentation sert pour des études phonétiques où l'extraction de prototypes est déterminante (détermination d'invariants, apprentissage, adaptation au locuteur, vérification du locuteur...), l'étiquetage est utilisé dans les systèmes de reconnaissance de la parole naturelle.

6 BIBLIOGRAPHIE

[BOURJOT,1988] : Bourjot C., Boyer A., Mari J.F., Methodology about assessment of large vocabulary systems, 7th FASE Symposium, book 1, pp161-169, EDINBURGH 1988.

[BOURJOT,1989] : Bourjot C., Boyer A., Fohr D., Phonetic decoder assessment, EUROSPEECH 89, Paris, 1989.

[CARBONNEL,1987] : Carbonnel N., Fohr D., Haton J.P., APHODEX, an acoustic phonetic decoding expert system", International Journal of Pattern Recognition and Artificial Intelligence, vol 1 no-2, pp207-222, 1987.

[PERENNOU,1989] : Perennou G., De Calmes M., Vigouroux N., Pecatte J.-M., Phonetic string alignment for an automatic labelling of speech corpora, ESCA SPEECH INPUT/OUTPUT ASSESSMNENT AND SPEECH DATABASES, NOORDWIJKERHOUT, pp541-544, 1989.

[DALSGAARD,1989] : Dalsgaard P., Semi automatic phonemic labelling of speech data using a self organising neural network, EUROSPEECH 89, Paris, 1989.

[SVENDSEN,1989] : Svendsen T., A two pass semi automatic algorithm for segmentation and labelling, 1989.

Nbre phon.	Nbre correct	erreur < 32ms	erreur <56ms	erreur grave
206	151	25	14	19
	73%	12%	7%	9%

Figure 1
résultats de l'étiquetage automatique

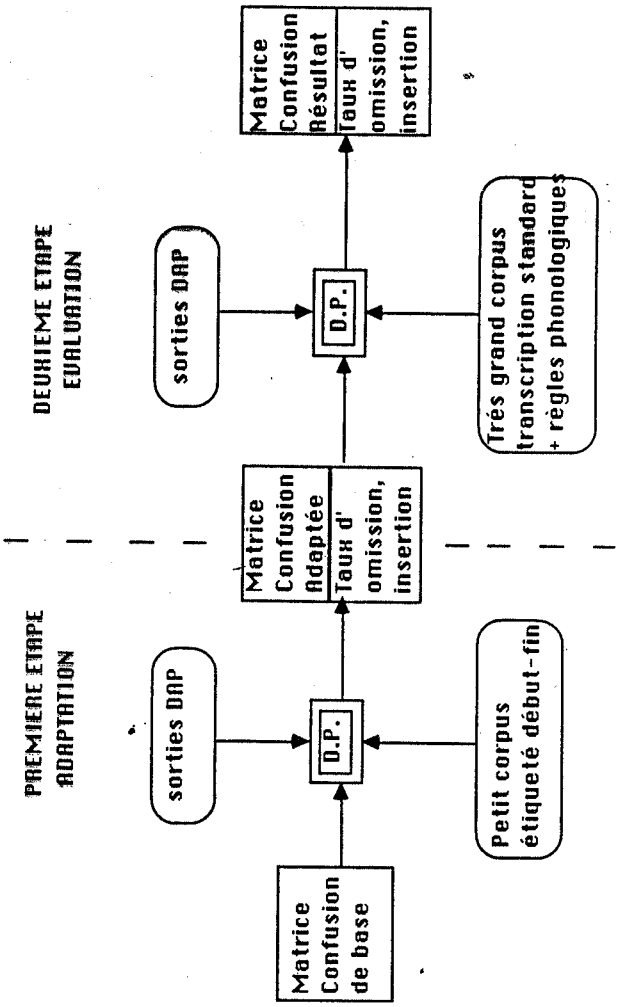
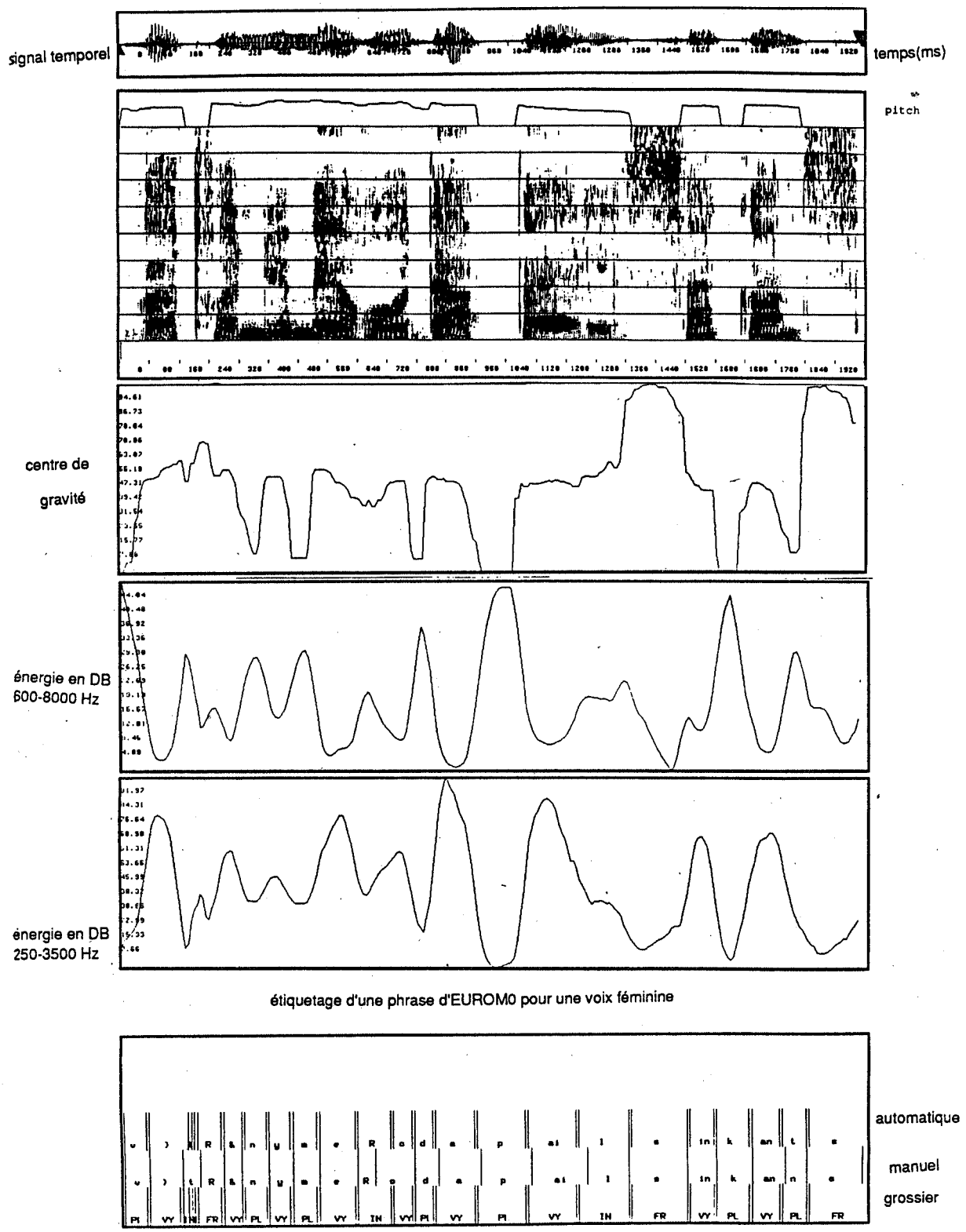


Figure 3 : les deux étapes de l'évaluation



étiquetage d'une phrase d'EUROM0 pour une voix féminine

Figure2 : exemple d'écran, résultats de l'étiquetage manuel et automatique

	nb	PL	VY	FR	IN	FV	PF	omm
p	9	7	0	1	0	0	0	1
b	3	2	0	0	0	0	0	1
t	13	12	0	0	0	0	0	1
d	6	4	0	0	0	0	0	2
k	8	7	0	0	0	1	0	0
g	1	1	0	0	0	0	0	0
f	5	0	0	5	0	0	0	0
v	7	0	0	2	4	0	0	1
ch	2	0	0	2	0	0	0	0
gh	7	0	0	7	0	0	0	0
s	10	0	0	10	0	0	0	0
z	1	0	0	1	0	0	0	0
m	10	0	0	0	7	0	0	3
n	9	0	0	0	6	1	0	2
nj	1	0	0	0	1	0	0	0
j	4	0	1	0	1	0	0	2
w	2	0	0	0	1	0	0	1
R	17	0	0	2	6	0	0	9
l	14	0	0	0	10	0	0	4
ui	1	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0
on	5	0	4	0	0	0	0	1
an	7	0	7	0	0	0	0	0
in	2	0	0	0	1	0	0	1
un	1	0	1	0	0	0	0	0
i	12	0	8	0	0	3	0	1
e	17	0	15	0	0	2	0	0
ai	8	0	7	0	0	0	0	1
a	17	0	16	0	0	0	0	1
)	3	0	2	0	0	0	0	1
o	5	0	5	0	0	0	0	0
u	8	0	7	0	0	0	0	1
y	4	0	3	0	0	0	0	1
eu	2	0	2	0	0	0	0	0
oe	1	0	1	0	0	0	0	0
&	17	0	16	0	0	0	0	1
*	36	1	0	20	4	0	0	11
hh	1	0	1	0	0	0	0	0
#	18	18	0	0	0	0	0	0
??	0	0	0	0	0	0	0	0
!	0	0	0	0	0	0	0	0
!!	0	0	0	0	0	0	0	0
?	0	0	0	0	0	0	0	0

	PL	VY	FR	IN	FV	PF
ins	0	3	1	7	0	0
omm	0	0	0	0	0	0

nombre de plosives presentes : 58
nombre de plosives trouvees : 51 (88 %)
nombre de plosives inserees : 0 (0 %)
nombre de fricatives presentes : 25
nombre de fricatives trouvees : 25 (100 %)
nombre de fricatives inserees : 4 (16 %)
nombre de noyaux presentes : 109
nombre de noyaux trouvees : 99 (91 %)
nombre de noyaux inserees : 5 (5 %)
nombre de inconnu presentes : 58
nombre de inconnu trouvees : 32 (55 %)
nombre de inconnu inserees : 12 (21 %)

méthode 1

	nb	PL	VY	FR	IN	FV	PF	omm
p	10	9	0	0	0	1	0	0
b	2	1	0	0	1	0	0	0
t	13	12	0	0	1	0	0	0
d	8	5	0	0	2	0	0	1
k	8	6	1	0	0	0	0	1
g	1	1	0	0	0	0	0	0
f	5	0	0	4	1	0	0	0
v	7	0	0	1	6	0	0	0
ch	2	0	0	2	0	0	0	0
gh	7	0	1	6	0	0	0	0
s	9	0	0	9	0	0	0	0
z	2	0	0	2	0	0	0	0
m	10	0	0	0	8	0	0	2
n	5	0	0	0	4	0	0	1
nj	1	0	0	0	1	0	0	0
j	3	0	0	1	0	0	0	2
w	2	0	0	0	0	0	0	2
R	17	0	0	7	5	0	0	5
l	14	0	0	0	9	0	0	5
ui	1	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0
on	5	0	5	0	0	0	0	0
an	7	0	7	0	0	0	0	0
in	2	0	1	0	0	1	0	0
un	1	0	1	0	0	0	0	0
i	12	0	8	1	0	2	0	1
e	18	0	13	1	0	3	0	1
ai	7	0	7	0	0	0	0	0
a	17	0	17	0	0	0	0	0
)	3	0	3	0	0	0	0	0
o	5	0	5	0	0	0	0	0
u	8	0	7	0	0	0	0	1
y	4	0	4	0	0	0	0	0
eu	3	0	3	0	0	0	0	0
oe	0	0	0	0	0	0	0	0
&	14	0	14	0	0	0	0	0
*	15	0	0	15	0	0	0	0
hh	0	0	0	0	0	0	0	0
#	18	18	0	0	0	0	0	0
??	0	0	0	0	0	0	0	0
!	0	0	0	0	0	0	0	0
!!	0	0	0	0	0	0	0	0
?	0	0	0	0	0	0	0	0

	PL	VY	FR	IN	FV	PF
ins	0	2	2	10	0	0
omm	0	0	0	0	0	0

nombre de plosives presentes : 60
nombre de plosives trouvees : 52 (87 %)
nombre de plosives inserees : 0 (0 %)
nombre de fricatives presentes : 25
nombre de fricatives trouvees : 23 (92 %)
nombre de fricatives inserees : 12 (48 %)
nombre de noyaux presentes : 106
nombre de noyaux trouvees : 101 (95 %)
nombre de noyaux inserees : 4 (4 %)
nombre de inconnu presentes : 53
nombre de inconnu trouvees : 27 (51 %)
nombre de inconnu inserees : 21 (40 %)

méthode 2

Figure 4 : résultats des deux méthodes

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

L'algorithme VITERBI-BLOC pour la reconnaissance de la parole continue

Abdelaziz KRIOUILE, Jean-François MARI, Jean-Paul HATON

CRIN-INRIA
B. P.. 239 54506 Vandoeuvre les Nancy, France

ABSTRACT

In this paper. we present a new version of the VITERBI algorithm well suited to continuous speech recognition. This version, called VITERBI-BLOCK, allows a segmentation and a labelling of an unknow utterance during its processing when, in the same time, the classical VITERBI algorithm waits until the end of the sentence before backtracking the best path. This algorithm is based on a local optimun comparison of the different probabilities computed by the VITERBI algorithm during the time-warping of a shift-window of fixed length in the signal and the different HMMs.

Different criteria of comparison between the probabilities of different path have been tried and assessed. The results are promising.

1 Introduction

Les modèles markoviens cachés ont largement contribué à l'évolution des systèmes de reconnaissance de la parole. Cependant, l'objectif d'obtenir un taux de reconnaissance phonétique convenable pour la parole continue est loin d'être atteint. Les systèmes de ce type de reconnaissance, basés sur les HMMs. utilisent le plus souvent, pour le décodage, une généralisation de l'algorithme de VITERBI [9.3] à un réseaux décrivant la grammaire des suites d'unités phonétiques possibles. Notre premier système, décrit en [5], utilisait ce type de généralisation. Dans ce papier nous allons décrire une autre version de l'algorithme de VITERBI applée VITERBI-BLOC et qui a donné lieu à un deuxième système pour la reconnaissance phonétique [6]. Les résultats expérimentaux qui vont être présentés porteront sur la comparaison des deux systèmes.

2 L'algorithme VITERBI-BLOC

2.1 Description de l'algorithme

Ce paragraphe fera l'objet d'une nouvelle version de l'algorithme de VITERBI. bien adaptée à la reconnaissance de la parole continue tout en surmontant le problème posé par la durée du modèle HMM. Cette version appelée VITERBI-BLOC permet une segmentation et un étiquetage de la phrase inconnue au fur et à mesure de son exécution. Par contre. l'algorithme de VITERBI attend la fin de la phrase pour "backtraking" le meilleur chemin [1]. La proposition de cet algorithme, utilisant celui de VITERBI par bloc de trames, a été dictée par plusieures raisons, parmi lesquelles nous citons :

- Avant de continuer le décodage de la phrase inconnue, il vaut mieux chercher la meilleure classe candidate correspondante à un meilleur segment courant.
- D'un point de vue probabiliste, il est plus facile de décoder sans erreurs un segment court qu'un segment long. Cet algorithme apporte une solution à ce problème de durée de segment en donnant des chances de décodage équiprobables à des segments de longueurs différentes.
- Reprendre, en quelques sortes, la stratégie utilisée efficacement dans le cas des mots isolés. Cette stratégie est basée sur la comparaison des probabilités calculées par l'algorithme de VITERBI appliqué aux différents modèles HMMs.

La reconnaissance par l'algorithme VITERBI-BLOC est réalisée par un processus itératif. Le principe de ce processus est basé sur une procédure de comparaison des probabilités calculées par l'algorithme de VITERBI. Ces probabilités sont celles correspondantes aux états finaux des différents modèles HMMs dans une fenêtre de trames de largeur fixe et à distance fixe de la trame de départ

de l'algorithme [figure 1]. Cette distance, ainsi, que la largeur de la fenêtre sont définies empiriquement. En effet les occurrences d'apprentissage nous donnent une idée sur les longueurs des segments possibles de chaque phonème.

Ce processus de reconnaissance est schématisé par les étapes suivantes :

1. trame courante ⟵ trame de départ
2. Exécution de l'algorithme de VITERBI pour chaque modèle à partir de la trame courante
3. Mémorisation des probabilités obtenues dans la fenêtre de comparaison pour chaque modèle
4. Comparaison de ces probabilités et choix du segment correspondant au modèle favorable
5. trame courante ⟵ trame juste après le segment choisi
6. Recommencer 2, 3, 4 et 5 jusqu'à la fin de la suite d'observations à décoder.

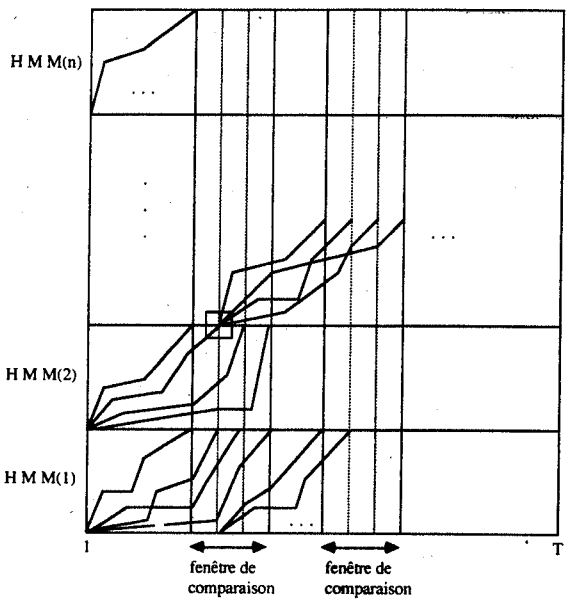


figure 1: L'algorithme VITERBI-BLOC

2.2 Critères de comparaison

La procédure de comparaison des probabilités, calculées par l'algorithme de VITERBI, des différents modèles est le noyau de l'algorithme VITERBI-BLOC. C'est le choix du critère de comparaison qui conditionnera l'efficacité de l'algorithme. Nous allons développer par la suite quelques critères qui ont servi de test pour les performances de l'algorithme.

2.2.1 Critère MAXDIFF

Le premier critère de comparaison, appelé MAXDIFF, n'utilise que des traits numériques tels que la comparaison et la différences des valeurs numériques. Il est basé sur les trois étapes suivantes :

1. Recherche de la probabilité maximale de celles obtenues pour les différents modèles, au niveau de chaque trame de la fenêtre de comparaison.
2. Calcul de la plus petite différence entre cette probabilité maximale et les autres de la même trame.
3. Comparaison de ces différences obtenues aux différentes trames de la fenêtre de comparaison. La plus grande différence permettra de choisir le modèle favorable et la trame finale du segment correspondant.

2.2.2 Critère MAXDUREE

Ce critère un peu plus compliqué que MAXDIFF, nous ramène à modifier légèrement l'algorithme d'apprentissage pour obtenir en plus une matrice C, déterminant des distributions de probabilités (discrètes ou continues) sur les différentes longueures en trames qu'une unité phonétique peut avoir, où $C[i,t]$ est la probabilité d'avoir la longueur t pour le modèle i.

Ce critère MAXDUREE peut être résumé par les étapes suivantes :

1. trame courante ⟵ trame de dédut de la fenêtre de comparaison
   MAXDUREE ⟵ 0
2. Recherche de la probabilité maximale entre celles obtenues pour les différents modèles et MAXDUREE, au niveau de la trame courante
3. MAXDUREE deviendra cette valeur maximale multipliée par un coefficient. Ce coefficient est une puissance d'ordre α de la probabilité $C[i,t]$ où i est le modèle correspondant à cette valeur maximale. L'exposant α est une constante déterminée empiriquement.
4. trame courante ⟵ trame suivante
5. Recommencer 2, 3, et 4 jusqu'à la trame de la fin de la fenêtre de comparaison
6. Le modèle et la trame correspondants à l'origine de la valeur gardée dans MAXDUREE, seront, respectivement, le modèle favorable choisi et la trame finale du segment qui lui correspond.

3 Expérimentation

3.1 Conditions expérimentales

Pour mesurer et comparer les performances des deux sys-

tèmes, nous avons utilisé le corpus, acquis au CRIN, de 57 phrases phonétiquement équilibrées de Combescure [2] prononcées à un rythme naturel d'élocution par cinq locuteurs masculins non professionnels. Ces phrases ont été numérisées à 12 kHz sur 10 bits et segmentées manuellement par un expert. Le signal issu de chaque phrase du corpus est paramétrisé par la méthode MFCC : Sur une fenêtre de Hamming de 20 ms de durée, on calcule 12 coefficients cepstraux en tenant compte d'une échelle MEL. Cette fenêtre est déplacée de 10 ms afin de permettre un recouvrement des intervalles d'analyse. L'algorithme de classification utilisé est celui initialement proposé par BUZO et GRAY [4,8]. Les expériences ont été faites avec deux dictionnaires de 64 prototypes chacun. Les deux dictionnaires, DIC5 et DIC12, sont respectivement obtenus après classification de 5 et 12 phrases. Nous quantifions ainsi les 57 phrases paramétrées. Chaque phrase est représentée par une suite de prototypes éléments de DIC5 ou de DIC12 suivant que l'on utilise l'un ou l'autre.

Nos tests portaient sur la segmentation de la parole continue. Nous avons construit, par l'algorithme de BAUM-WELCH, un modèle markovien caché pour chaque classe phonétique et un dernier modèle pour le silence. Les classes phonétiques utilisées dans nos expériences sont :

- PV : Les plosives voisées.
- PS : Les plosives sourdes.
- FV : Les fricatives voisées.
- FS : Les fricatives sourdes.
- VY : Les voyelles.
- RE : Le reste des phonèmes.

Pour estimer un HMM par classe phonétique (7 modèles), nous avons utilisé 40 phrases quantifiées du corpus et leur segmentation manuelle pour l'apprentissage. Les 17 autres ont servi pour tester les deux versions de l'algorithme de reconnaissance proposées. Chaque phrase reconnue est comparée à sa segmentation manuelle.

3.2 Résultats expérimentaux

Plusieurs expérimentations ont été faites dans le but de déterminer le plus convenablement possible :

- Le type de modèle initial (nombre d'états,transitions...),
- Le nombre des itérations de l'algorithme d'apprentissage sur les occurences de chaque classe issues des 40 phrases,
- La constante α (l'exposant utilisé dans MAXDUREE).

En moyenne, les meilleurs résultats sont ceux fournis par :

- le modèle initial illustré sur la figure [figure 2],
- 3 itérations,
- α égal à 2.

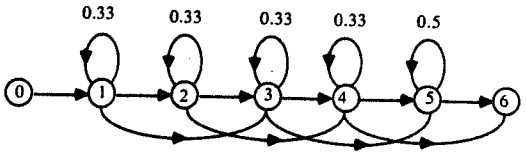


figure 2: Modèle initial choisi

Ce sont ces choix que nous avons utilisés pour les résultats exposés par la suite.

3.2.1 Matrice durée

Pour le critère de comparaison MAXDUREE utilisé par l'algorithme VITERBI-BLOC, nous avons determiné pour nos expériences une matrice durée C formée de distributions de probabilités discrètes. Ces distributions sont calculées à partir des échantillons de longueurs des classes phonétiques. Ces échantillons sont obtenus à partir des occurences de chaque classe se trouvant dans les 40 phrases d'apprentissage. Pour donner une possibilité d'avoir des longueurs de classe non existantes dans le corpus d'apprentissage, nous leur avons attribué une probabilité de petite valeur (10^{-4} dans nos expériences).

3.2.2 Résultats du corpus test

Les taux de reconnaissance du tableau (1) illustrent les résultats de la reconnaissance des phrases test par les différentes versions, de l'algorithme de VITERBI, proposées :

- VITERBI-réseau : L'algorithme de Viterbi généralisé à une grammaire décrivant les suites de classes phonétiques possibles [5].
- VITERBI-réseau-3meilleurs : C'est l'algorithme VITERBI-réseau modifié pour avoir les trois meilleurs alignements. En combinant ces trois chemins, nous obtenons un treillis de classes phonétiques qui ne fera qu'augmenter le taux de reconnaissance.
- VITERBI-BLOC-MAXDIFF et VITERBI-BLOC-MAXDUREE : Ce sont les versions de l'algorithme VITERBI-BLOC utilisants, respectivement, les critères MAXDIFF et MAXDUREE.
- VITERBI-frontières : Cette version suppose que les frontières des segments de la phrase à décoder sont connues. Il ne reste plus qu'à étiqueter ces segments.

Version de l'algorithme de Viterbi	DIC 5 (%)	DIC 12 (%)
VITERBI-réseau	59	60
VITERBI-réseau-3meilleurs	60	60
VITERBI-BLOC-MAXDIFF	62	62
VITERBI-BLOC-MAXDUREE	63	64
VITERBI-Frontières	67	68

tableau (1): Taux de reconnaissance du test

L'algorithme VITERBI-frontières est un bon test de la suffisance ou non du corpus d'apprentissage. Son taux de reconnaissance (68%) montre que notre corpus de 40 phrases ne permet pas un très bon apprentissge. Cependant, notre apprentissage est accéptable et la comparaison des taux de reconnaissance est significative. Les algorithmes VITERBI-BLOC-MAXDIFF et VITERBI-BLOC-MAXDUREE permettent un gain de 3% à 4% par rapport au taux de reconnaissance fourni par VITERBI-réseau. Ce qui est une amélioration très intéressante, surtout que l'amélioration idéale donné par VITERBI-frontières, est de 8%.

Sans passer par l'étape de segmentation, une application directe de ces systèmes pour reconnaître les phonèmes est tout à fait envisageable.

4 Conclusion

L'algorithme VITERBI-BLOC représente une nouvelle contribution aux algorithmes de la reconnaissance de la parole continue , basés sur les HMMs [7]. Les résultats illustrent son importance par rapport aux versions de l'algorithme de Viterbi généralisé à une grammaire pour donner le meilleur chemin. Ce meilleur chemin global obtenu, n'entraîne forcément ni la meilleure segmentation, ni la meilleure concaténation des unités phonétiques de reconnaissance.

L'algorithme VITERBI-BLOC peut être amélioré par des bons choix de critères de comparaison. Nous pensons aussi qu'une matrice durée C utilisant des distributions de probabilités continues au lieu de discrètes améliorera les résultats de l'algorithme VITERBI-BLOC-MAXDUREE.

VITERBI-BLOC est un algorithme ouvert à l'intervention d'autres processeurs pour améliorer l'identification des segemnts pendant le décodage. En effet, par le biais des choix des critères de comparaison, il permettra une forte intéraction avec d'autres processeurs pour confirmer ou infirmer un segment décodé. Une intéraction avec des techniques à bases de connaissances, est à l'étude.

References

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. PAMI.*, PAMI-5(2):179–190, 1983.

[2] P. Combescure. Vingt listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 14(56). 1981.

[3] G. D. Forny. The Viterbi algorithm. In *Proc. IEEE*, pages 268–278, Mar 1973.

[4] A. H. Gray and J. D. Markel. Distance measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:380–391, 1976.

[5] J. P. Haton, N. Carbonell, D. Fohr, J. F. Mari, and A. Kriouile. Interaction between stochastic modeling and knowledge-based techniques in acoustic-phonetic decoding of speech. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing 1987*, pages 868–871, DALLAS,TEXAS, 1987.

[6] A. Kriouile. Les modèles markoviens cachés et leur application à la reconnaissance automatique de la parole. *Thèse de Doct. Univ. de NANCY 1, à paraître*, 1990.

[7] A. Kriouile, J. F. Mari, and J. P. Haton. Some improvements in speech recognition algorithms based on hmm. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Albuquerque, New Mexico, 1990.

[8] Y. Linde, A. Buzo, and R. M. Gary. An algorithm for the vector quantizer design. *IEEE. Trans. on Communication*, com. 28(1), Jan. 1980.

[9] A. J. Viterbi. Error bounds for convolutional codes and asymtotically optimum decoding algorithm. *IEEE Trans on Information Theory*, IT-13:260–269, April 1967.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés *

Régine André-Obrecht

IRISA/CNRS, Campus de Beaulieu, 35042 Rennes Cedex

Abstract

This paper describes the introduction of an automatic segmentation algorithm as a pre-processing for a speaker independent recognition system based on hidden Markov models.

An acoustic analysis is performed over each segment and the Markov modelling is based on a hierarchical description of the application, the recognition of numbers from 0 to 999. The whole acoustical network is derived from the a priori syntactical, lexical and phonetical knowledges and uses as basic units, the pseudo-diphones.

The recognition performances show that the segmental analysis doesn't loose essential information with respect to the sliding-block analysis. The comparison between articulatory interpretations of the segmentation and sequences of pseudo-diphones found during the recognition, validates our qualitative analysis of the acoustical segments, and this approach seems very promising for future recognition systems.

1 Introduction

Les systèmes actuels de reconnaissance automatique de parole continue sont en majorité, et pour les meilleurs d'entre eux, fondés sur des méthodes statistiques, de type "source de Markov cachée". L'avantage de cette approche est la capacité d'apprentissage sur de grandes bases de données de parole. Cependant, pour être efficaces, ces systèmes requièrent une description exhaustive de l'application et intégrent finalement très peu d'informations au niveau de la paramétrisation acoustique.

En effet, afin d'atteindre des applications pour lesquelles le vocabulaire devient important ($\geq$ 1000 mots) et la complexité syntaxique devient croissante, la modélisation acoustique globale de chaque mot s'est avérée insuffisante [4,6,9,10] : bien qu'automatique, un apprentissage correct nécessiterait alors un ensemble de données trop volumineux. Une alternative naturelle est donc de chercher à modéliser chaque mot à l'aide d'un niveau intermédiaire faisant intervenir un nombre restreint d'entités. Une description "phonétique" permet de décrire chaque mot en unités phonétiques de base et chacune d'elles fait l'objet d'une modélisation acoustique. Il reste qu'un système de reconnaissance ne peut se contenter de cette représentation simpliste ; de nombreuses règles de coarticulation entre les unités sont ajoutées, et pour tenir compte de la variabilité contextuelle il faut introduire des modèles acoustiques dépendant du contexte (modélisation d'allophones [7]) ou intégrer le contexte en regroupant certaines unités en macro-unités et créer les modèles acoustiques appropriés (modélisation de triphones [9]). Un nouveau compromis entre données d'apprentissage et fiabilité apparaît. Ces différentes approches ne semblent pas satisfaisantes car le problème de la variabilité du signal de parole ne semble que contourné.

L'étude que nous amorçons a pour but de proposer pour le décodage phonétique de nouvelles idées quant à la modélisation et la paramétrisation en prenant plus en compte la variabilité du signal. Elle a eu comme point de départ, les résultats qualitatifs obtenus par segmentation automatique du signal ; cette segmentation met en évidence des zones à l'intérieur desquelles il est pertinent de vouloir extraire une information contextuelle (indication d'évolution, cible acoustique, etc.) [5]. Nous introduisons, dans un premier temps, cette segmentation comme pré-traitement dans des systèmes de reconnaissance classiques, à base de modèles de Markov. Les réseaux stochastiques ainsi obtenus se trouvent énormément simplifiés, et malgré la perte d'information (facteur 4), les performances sont très correctes et approchent celles des meilleurs systèmes.

Ces expériences valident l'hypothèse que les segments sont une réalité acoustique ; il reste à trouver comment extraire et intégrer proprement l'information qu'ils véhiculent.

2 Module acoustique

Le prétraitement acoustique se décompose en deux temps : un algorithme de segmentation automatique traite le signal de parole et décide d'instants de rupture ; chaque segment est ensuite analysé pour fournir un vecteur de paramètres. Le but de cette segmentation n'est pas de poser des marques de discontinuités, mais de localiser des zones qui correspondent aux différentes réalisations des sons ; l'analyse acoustique de chaque segment pourra être plus adaptée à la nature de l'information recherchée, suivant qu'elle se veut statique ou dynamique.

L'algorithme de segmentation est basé sur un test statistique : deux modèles auto-régressifs (L.P.C.) sont estimés sur des fenêtres différentes et comparés à l'aide de la divergence de Kullback [2]. Cette méthode diffère des traitements classiques dans la mesure où un diagnostic de rupture est donné à chaque échantillon de signal et non toutes les 10ms ; la localisation de certains événements demande cette précision.

*Cette étude fait l'objet d'une convention d'études entre le CNET et l'IRISA

Qualitativement, les frontières obtenues correspondent à deux types de changements (figure 1a) :

- les discontinuités majeures ; elles correspondent à des événements phonétiques, interprétables au niveau articulatoire, et structurant le signal en "coordinations" [1].
- les changements à l'intérieur de chaque coordination ; ils révèlent des zones transitoires et des zones stables, là où une cible acoustique est atteinte ou quasiment atteinte.

Il est malheureusement difficile de parvenir automatiquement à cette classification des frontières ; seules celles correspondant à des événements phonétiques robustes tels que les débuts et fins de voisement et de friction peuvent être fiablement étiquetées [5]. Une étude [3] a montré qu'il était sans espoir de vouloir distinguer statistiquement les zones dites transitoires des zones homogènes : il sera toujours préférable de représenter une quelconque portion de signal, aussi stable soit elle, par un modèle évolutif ; ce qui ne fait que confirmer qu'en parole continue, une cible acoustique est rarement atteinte.

C'est pourquoi, l'analyse acoustique est faite sans connaissance a priori sur le segment ; les paramètres extraits pour chaque segment seront identiques qu'il s'agisse d'une zone consonantique, vocalique, transitoire ou non... Ils seront obtenus à partir d'une analyse cepstrale ou une analyse LPC.

Dans une première étape, une fenêtre de 10ms ou 20ms suivant la longueur du segment est centrée sur celui-ci. Après préaccentuation et transformée de Fourier, 6 coefficients cepstraux (MFCC), l'énergie et la durée du segment définissent une observation vectorielle. Les segments de longueur inférieure à 10ms ne sont pas paramétrés ; il s'en suit que le nombre d'observations est légèrement inférieur au nombre de segments.

3 Le module de reconnaissance

L'application, à savoir la reconnaissance des nombres de 0 à 999, est décrite de manière exhaustive par un réseau stochastique global, intégrant de nombreuses connaissances a priori.

Chaque niveau de connaissances est modélisé par une source de Markov cachée :

- les descriptions syntaxique et lexicale de l'application donnent naissance à une première source où chaque observation correspond à un mot du vocabulaire,
- le vocabulaire est transcrit phonétiquement en unités de base pour donner un réseau phonétique par mot du lexique ; dans notre étude, les unités de base sont les pseudo-diphones
- chaque unité de base est décrite par un réseau acoustique élémentaire, où les observations sont de même nature que les sorties de l'analyseur acoustique.

Quelques règles phonologiques sont adjointes à l'ensemble pour obtenir le réseau final.

3.1 Les pseudo-diphones et leur modélisation acoustique

Chaque mot du vocabulaire est donc décrit comme une succession d'unités, une pour chaque son et une pour chaque transition entre sons. Cette entité, le pseudo-diphone, est choisie, dans cette approche, comme unité élémentaire car elle correspond à la description segmentale obtenue au niveau acoustique et conduit par conséquent à des modèles de Markov cachés extrêmement simples et précis.

Les résultats de segmentation conduisent à classer les pseudo diphones en 8 catégories ; telles que les voyelles orales, nasales, les consonnes voisées,...

Pour chacune de ces catégories, nous définissons un type de modèle acoustique en utilisant des connaissances a priori : la durée moyenne en nombre de segments du son considéré et sa stabilité spectrale permettent de définir le nombre d'états et le nombre de lois d'observations, distributions acoustiques associées à chaque transition. A titre d'exemple, une voyelle nasale est formée de deux zones spectrales différentes ; l'une est appelée classiquement partie orale et l'autre partie nasale. Le tableau 1 donne la liste des modèles acoustiques utilisés.

Des règles phonologiques sont introduites pour définir les coarticulations entre mots, différencier certains pseudo-diphones suivant leur contexte (allophones) et traiter certains sons délicats. Des difficultés surgissent lorsque certains sons sont, dans des contextes précis, trop fortement coarticulés. Nous sommes

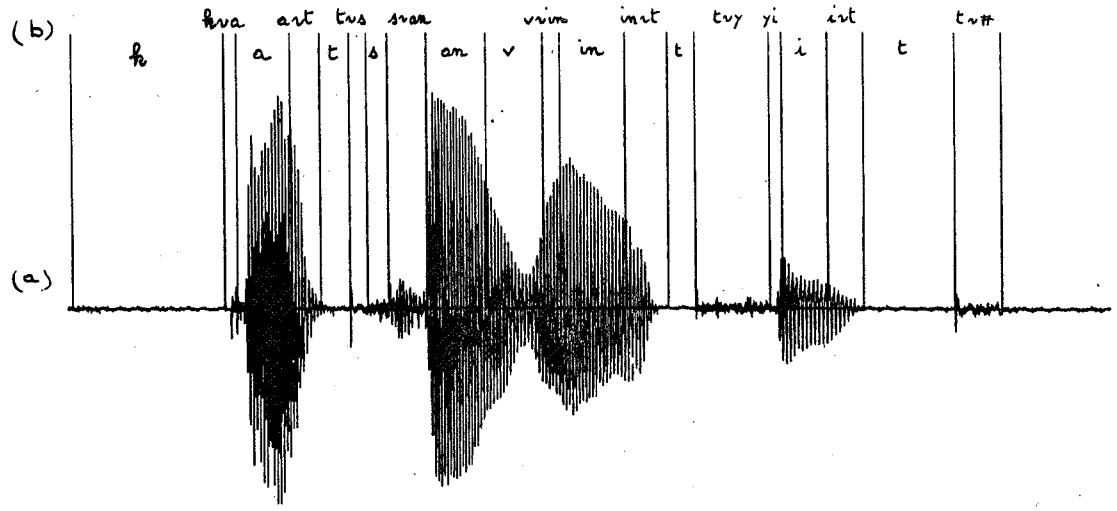


Figure 1 : a. Signal segmenté
b. Chemin optimal à travers le réseau PSD-SI représenté par les transitions non vides.

alors amenés à définir de nouvelles entités et parfois de nouvelles structures acoustiques. Pour cette application précise, seule l'unité *tr* est créée pour traiter le phonème /r/ en contexte /t/, et un modèle acoustique approprié est spécifié.

Le réseau global est obtenu en utilisant le compilateur de réseaux développé par le CNET [7] (figure 2) ; il sera référencé comme étant le modèle PSD.SI dans la suite de ce papier.

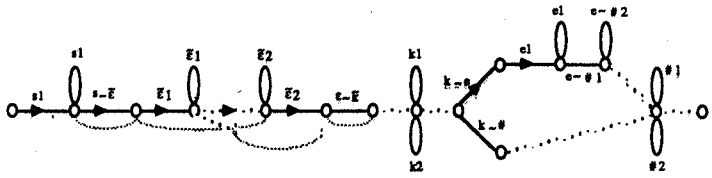


Figure 2 : Extrait du réseau global représentant le mot "5" en fin de phrases

3.2 Mise en oeuvre du module de reconnaissance

(Ce module est mis en oeuvre à l'aide des logiciels développés par le CNET).

Les distributions acoustiques associées aux transitions non vides de la chaîne de Markov sont des gaussiennes de matrices de covariance diagonales. L'apprentissage des paramètres du modèle (lois de transition et lois d'observations) est réalisé à l'aide d'une adaptation de l'algorithme de Baum-Welch ; ne sont utilisées dans les formules de réestimation que les transitions appartenant à des chemins proches des chemins optimaux [8].

4 Tests de Reconnaissance. Comparaison avec les modèles ALL-V2 et PSD-A3

La base de données est formée d'environ 5 300 prononciations, à savoir 1 000 chiffres, 3 000 nombres à 2 chiffres et 1 300 nombres à 3 chiffres ; ce corpus a été enregistré avec 70 locuteurs dans une pièce standard, sans isolation particulière. Cette base est découpée en deux parties égales ; l'une constitue l'ensemble d'apprentissage et correspond aux enregistrements de 34 locuteurs, l'autre constitue l'ensemble-test. Aucun locuteur n'est commun aux deux ensembles, on mesure donc des performances de reconnaissance en mode indépendant du locuteur. Le signal est échantillonné à 12.8 KHz.

Les tests de reconnaissance sont effectués de manière à mettre en évidence l'influence des paramètres intervenant soit au niveau du modèle de Markov, soit au niveau de l'analyse acoustique des segments. La structure du réseau acoustique global est celle du modèle PSD-SI définie précédemment et sera inchangée au cours des expériences ; en particulier les fautes de modélisation au niveau coarticulation ne sont certainement pas toutes corrigées, et sont fatales dans une approche segmentale.

Il est à noter également que les distributions gaussiennes sont toutes doublées afin de mieux prendre en compte la variabilité inter-locuteur ; le meilleur chemin à travers ce graphe correspond alors à un maximum de gaussiennes pondérées.

Les résultats sont présentés sous forme de tableaux ; les tests de reconnaissance sont réalisés sur l'ensemble d'apprentissage lui-même et sur l'ensemble test.

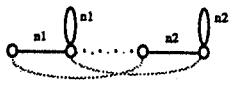
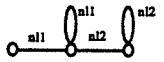
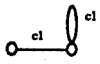
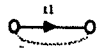
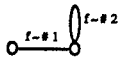
Sons	Modèles
voyelle	
voyelle nasale	
voyelle nasale longue	
consonne	
explosion de plosives non voisées	
silence	
transitoire	
fin de phrases	

Tableau 1 : Structure des modèles acoustiques en fonction du type de son ; les transitions vides correspondent aux lignes pointillées et les transitions non vides portent des lois d'observations.

Ces deux ensembles sont eux-mêmes divisés en 3 sous-ensembles, l'ensemble des nombres à 1 chiffre (NB1), nombres à 2 chiffres (NB2) et à 3 chiffres (NB3).

Les résultats de reconnaissance sont donc donnés sur ces 6 sous-ensembles, sur l'ensemble d'apprentissage (apprentissage total) et l'ensemble test (test total).

4.1 Influence du choix des coefficients acoustiques

Dans la version de base, chaque segment est paramétré par 6 coefficients MFCC, son énergie et sa durée. Cette paramétrisation nécessite l'emploi d'une transformée de Fourier rapide, donc des fenêtres de 128 ou 256 échantillons dans notre cas. Dans la mesure où les segments ne sont pas de longueur égale à une puissance de 2, il est intéressant de comparer cette analyse cepstrale à une analyse de type LPC plus souple ; elle est réalisée sur la totalité du segment lorsque sa durée est inférieure à 20ms (256 échantillons) et sinon sur 20ms.

Trois tests de reconnaissance ont permis de comparer trois ensembles de coefficients acoustiques :

- 8 coefficients : 6 MFCC, énergie, durée

- 10 coefficients : énergie résiduelle du modèle LPC à l'ordre 8, 8 LPCC, durée
- 14 coefficients : énergie résiduelle du modèle LPC à l'ordre 12, 12 LPCC, durée.

Les performances (tableau 2) ne montrent aucun avantage à utiliser les modèles LPC, puisque pour atteindre les mêmes taux il faut considérer un modèle d'ordre 12.

4.2 Influence de la prise en compte des durées

Dans les systèmes de reconnaissance classiques, où le traitement acoustique est réalisé sur des blocs d'analyse glissants de taille fixe, il est courant d'introduire une modélisation des temps de bouclage sur chaque état. Etant donné la nature de la segmentation, cette prise en compte n'a aucune raison d'être ; par contre la durée d'un segment est une information, semble-t-il, pertinente : une zone transitoire est toujours plus courte qu'une zone bruitée correspondant par exemple à une fricative. Cette hypothèse est confirmée en comparant les taux de reconnaissance lorsque la durée du segment est considérée comme un paramètre d'observation ou non (tableau 2). Le gain de 2% est significatif pour l'intervalle de confiance correspondant de 1%.

Nature des coefficients acoustiques	Apprentissage				Test			
	NB1	NB2	NB3	Total	NB1	NB2	NB3	Total
6 MFCC + E	96,8	93,4	86,3	92,3	94,9	89,2	90,0	90,4
12 LPCC + σ	98,2	94,8	90,2	94,3	93,2	91,3	86,1	90,4
8 LPCC + σ + T	97,0	94,4	90,8	94,0	90,6	88,8	88,0	88,9
8 LPCC + σ + T	97,4	95,5	91,9	95,0	93,4	92,6	89,7	92,1
6 MFCC + E + T	96,2	94,3	88,9	93,4	95,7	92,3	93,4	93,2

Tableau 2 : Taux de reconnaissance en fonction des différentes analyses acoustiques.
NB1 désigne l'ensemble des chiffres,
NB2 l'ensemble des nombres à 2 chiffres,
NB2 l'ensemble des nombres à 3 chiffres
L'intervalle de confiance sur l'ensemble test pour un taux de 90 % est de ± 1%

4.3 Influence de paramètres dynamiques

Il est à craindre que lorsque l'on ramène la représentation d'un segment à sa paramétrisation statique sur une fenêtre de 20ms, quantité d'informations soit perdue. En particulier, l'information dynamique interne disparaît. L'influence des variations temporelles de coefficients est testée en ajoutant aux 6 coefficients MFCC leur régression au cours du segment ; que l'on ajoute 3 ou 6 dérivées, les taux de reconnaissance n'augmentent pas. Il en résulte que les variations temporelles ne doivent pas être intégrées sous cette forme.

4.4 Influence des paramètres d'apprentissage

De nombreux tests ont été effectués lors de l'étude de ce système avec une analyse du signal par blocs glissants [8] ; il n'a pas semblé utile de remettre en question les paramètres tels que le seuil sur les écart-types des gaussiennes, le seuil de convergence des algorihtmes.

Une seule expérience est réalisée pour tester l'importance de l'initialisation des lois gaussiennes avant l'apprentissage. Un sous-ensemble de l'ensemble des données d'apprentissage est étiqueté manuellement et permet cette initialisation. La différence entre les taux de reconnaissance de l'ordre de 3,5% ne permet aucun doute sur l'utilité de l'initialisation. Il est à craindre que l'étiquetage ayant été fait indépendemment des résultats de segmentation, l'initialisation ne soit pas totalement optimale !.

4.5 Comparaison avec l'approche "bloc glissant"

Comme il est dit précédemment, le compilateur de réseaux et les différents algorithmes de reconnaissance et d'apprentissage ont été développés au CNET, dans le but de créer des systèmes de reconnaissance des nombres de 0 à 999 indépendants du locuteur. Ces systèmes ont pour différence essentielle avec celui proposé présentement d'analyser le signal acoustique par blocs de 20ms toutes les 10ms. Les systèmes les plus performants proposés par le CNET ont pour unité de base :

- soit l'allophone, phonème en contexte (modèle appelé ALL V2),
- soit le pseudo-diphone (modèle appelé PSD-A3).

Les paramètres acoustiques sont les 6 coefficients MFCC, l'énergie et la différence d'énergie entre trames adjacentes. Le tableau 3 permet de comparer la taille des réseaux et les performances obtenues par ces deux modèles à celles obtenues par le modèle PSD-SI, proposé ci-dessus où les paramètres acoustiques sont les 6 coefficients MFCC, l'énergie et la durée du segment. Le taux de reconnaissance est certes dégradé, que ce soit sur l'ensemble d'apprentissage ou l'ensemble test (-2%). Par contre le réseau global est simplifié, le nombre de transitions et de gaussiennes décroît notablement, l'apprentissage comme la reconnaissance en sont accélérés.

	Nombre d'états	Nombre de transitions	Nombre de gaussiennes	Taux d'erreur	
				apprentissage	test
ALL-V2	797	2149	320	5,28	4,28
PSD-A3	872	1898	544	4,86	5,60
PSD-SI	335	1260	319	6,70	6,80

Tableau 3 : Comparaison entre les 3 systèmes de reconnaissances ALL-V2, PSD-A3, PSD-SI
Le nombre de coefficients acoustiques par observation est identique (= 8)

5 Conclusion : Interprétation des chemins optimaux

Etant donné les connaissances a priori que nous cherchons à introduire dans les différents modèles acoustiques, il est intéressant de se demander si, après apprentissage, le calage des segments sur les suites de pseudo-diphones correspondantes s'est effectué correctement, à savoir si un segment transitoire (resp-stable) correspond effectivement à un pseudo-diphone transitoire

(resp-stable). Pour effectuer cette évaluation qui ne peut être que qualitative, nous mettons en correspondance le signal segmenté et le chemin optimal trouvé par l'algorithme de Viterbi à travers le réseau acoustique global ; le chemin est matérialisé par la suite des transitions non vides empruntées et marquées du nom du pseudo-diphone.

Cette démarche a permis lors de l'apprentissage de corriger de nombreuses erreurs de coarticulation dûes à une mauvaise connaissance phonologique de notre part. L'examen minutieux des résultats confirme l'interprétation a priori que nous faisions des segments (figure 1), ils représentent une réelle unité acoustique, indépendante du locuteur et robuste à l'analyse.

Nous avons donc introduit une analyse acoustique basée sur une segmentation automatique du signal dans un système de reconnaissance indépendant du locuteur, basé sur une modélisation markovienne. Le nombre de transitions et le nombre de lois d'observations du réseau global ont par conséquent diminué. La phase d'apprentissage comme celle de reconnaissance sont donc plus rapides. Cette qualité doit être bénéfique surtout pour des applications plus "grosses". Le nombre de paramètres de chaque loi d'observations n'ayant pas augmenté, les performances d'un tel système en mode indépendant du locuteur restent très correctes (93% ± 1% de reconnaissance). Il faut espérer qu'une meilleure modélisation statistique et une meilleure extraction de l'information permettraient par ce biais d'atteindre des applications plus complexes.

Références

[1] C. ABRY, C. BENOIT, L.J. BOË, R. SOCK, Un choix d'événements pour l'organisation temporelle du signal de parole. *14ème JEP*, Paris 1985.

[2] R. ANDRÉ-OBRECHT, A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. on ASSP*, vol.36, no1, January 1988.

[3] R. ANDRÉ-OBRECHT, B. DELYON, Etude des segments transitoires à l'aide de modèles AR évolutifs et du critère d'Akaïke. *GRETSI 12ème Colloque*, Juin 1989.

[4] L.R. BAHL, Large vocabulary natural language continuous speech recognition. *ICASSP* 1989.

[5] B. DELYON, R. ANDRÉ-OBRECHT, H.Y. SU, Expériences en vue du décodage phonétique. *Journal d'Acoustique*, vol.1, Septembre 1988.

[6] L. FISSORE, A word hypothesizer for a large vocabulary continuous speech understanding system. *ICASSP*, 1989.

[7] D. JOUVET, J. MONNÉ, D. DUBOIS, A new network based speaker independent connected word recognition system. *ICASSP*, Tokyo 1986.

[8] D. JOUVET, *Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques.* Thèse de Docteur Ingénieur ENST, Juin 1988.

[9] K.F. LEE, The SPHINX Speech Recognition System. *ICASSP*, Glasgow, 1989.

[10] D.B. PAUL, The LINCOLN robust continuous speech recognizer. *ICASSP*, 1989.

MARS : Un Système de Reconnaissance de l'Arabe Moderne

M. Djoudi D. Fohr J. P. Haton

CRIN — INRIA Lorraine
Campus Scientifique
B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX
France

Résumé

Nous proposons dans cet article une architecture d'un système de reconnaissance automatique de l'Arabe moderne. Le système a pour but la compréhension de phrases dans un contexte multilocuteur. Sa mise en œuvre tient compte bien sûr des particularités de la langue. Nous allons décrire dans ce qui suit le module de décodage phonétique qui n'est autre que le système SAPHA développé pour le décodage acoustico-phonétique de l'Arabe [7]. Ensuite, nous aborderons le décodeur linguistique en évoquant les aspects de la morphologie de la langue, de la transcription phonétique-orthographique et de la structure syntaxico-sémantique des phrases. Enfin, nous parlerons de l'importance de la prosodie dans un système de compréhension de l'Arabe parlé et les points qu'il faut résoudre.

1 Introduction

L'arabe moderne, dite aussi littéraire ou standard a fait l'objet de plusieurs travaux anciens ou récents, ayant trait soit à l'aspect phonétique [1] [2] [10] [5] soit à la composante linguistique de la langue [15] [12] [14] [9]. Toutefois, le problème de reconnaissance automatique n'a été que très peu abordé jusqu'à présent [6]. Nous allons décrire dans cet article le système MARS (Modern Arabic Recognition System) qui se veut un système de reconnaissance de phrases en parole continue et dans un contexte multilocuteur. Le système reçoit en entrée une phrase et rend en sortie une (ou éventuellement plusieurs) interprétation sémantique de la phrase prononcée en utilisant des sources de connaissances diverses: phonétiques, phonologiques, morphologiques, syntaxiques, sémantiques et prosodiques.

2 Architecture du système

La structure générale du système fait apparaitre deux grands sous systèmes. D'une part, un système de décodage acoustico-phonétique, en l'occurrence SAPHA et d'autre part, le décodeur linguistique SALAM. Chaque sous système utilise diverses sources de connaissances. Le décodeur phonétique fournit au système linguistique, en mode proposition, un treillis phonétique. Il peut être réactivé, en mode vérification, par le décodeur linguistique pour savoir si un phonème donné est l'étiquette d'un morceau de signal donné.

3 Le décodeur phonétique : SAPHA

Le système SAPHA [7] est composé d'un ensemble de modules et englobe les trois étapes : acoustique, phonétique et phonologique que comporte le niveau inférieur d'un système de reconnaissance automatique de la parole [11]. Ces module sont :

3.1 Le module acquisition

C'est la première étape dans tout processus de reconnaissance de la parole. A ce niveau, l'acquisition de la parole est faite à partir d'un microphone ou d'une cassette. De même, on peut écouter un morceau de parole et jouer sur la valeur de la fréquence d'échantillonnage. Les fonctions de lecture, de sauvegarde, de coupure et d'affichage du signal sur une console graphique sont aussi prévues à ce niveau [13].

3.2 Le module acoustique

Ce module se charge d'extraire les paramètres acoustiques à partir du signal temporel, il s'agit en particulier de :

- L'énergie du signal.
- La densité de passages par zéro par seconde.
- Les pics à partir des coefficients LPC.
- La fréquence fondamentale.
- Les caractéristiques fréquencielles en utilisant un algorithme de transformée de Fourier rapide (FFT) directement du signal ou bien à partir des coefficients LPC.

3.3 Le module de segmentation

Il consiste à segmenter le signal de parole en grandes classes phonétiques [8] [3] en utilisant des algorithmes procéduraux non contextuels et reposant sur des critères simples, le but essentiel de la segmentation est de :

- réduire l'explosion combinatoire lors de la reconnaissance,
- offrir un cadrage pour le module d'étiquetage .

Les classes concernées sont les voyelles, les plosives et les fricatives. Les autres phonèmes seront considérés comme faisant partie de la classe des inconnues.

3.4 Le calcul d'indices

L'extraction des indices phonétiques pertinents est une étape très importante dans le processus de décodage phonétique. Les valeurs des indices seront utilisées lors de la phase d'étiquetage. Ces indices sont: la durée du segment, le degré de voisement, la position de la barre d'explosion [4]. les caractéristiques de cette barre, le suivi de formant, les transitions formantiques, le centre de gravité et la limite inférieure du bruit de friction.

3.5 Le module d'étiquetage

C'est un système expert à base de règles de production et c'est à ce niveau que se fait le décodage proprement dit. En partant des segments fournis par le module de segmentation, le module tente de trouver les bon phonèmes prononcés en utilisant les indices extraits lors de l'étape précédente et les connaissances se trouvant dans la base de règles.

3.6 Le module phonologique

La sortie du module d'étiquetage est un ensemble de phonèmes par segment. Le module phonologique essaye d'éliminer certaines solutions au regard du contexte et des règles phonologiques qui régissent la langue arabe. Dans le système, le module phonologique est intégré dans le module d'étiquetage, et les règles phonologiques sont contenues dans la base de connaissances du système expert.

4 Le décodeur linguistique : SALAM

SALAM comme Système Approprié pour le décodage Linguistique de l'Arabe Moderne est conçu pour recevoir en entrée une suite de phonèmes sous forme de treillis phonétique pour donner en résultat l'interprétation sémantique de la phrase prononcée. Le décodeur linguistique SALAM comporte plusieurs modules :

4.1 Le module morphologique

L'Arabe possède un système morphologique régulier qui n'a pas beaucoup changé depuis plusieurs siècles. En effet, il existe trois types de racines : triradicale, quadriradicale et 5-radicale. Les verbes, les noms et autres adjectives sont dérivés à partir de ces racines selon un modèle régulier, généralement, parmi les modèles suivants :

- Insertion de voyelles spécifiques.
- Préfixage.
- Suffixage.
- Ajout d'une consonne médiane.

Toutefois, les verbes sont dérivés uniquement des racines tri et quadriradicales. Cette régularité devrait être exploitée par le module morphologique. Les quelques irrégularités peuvent être prises comme des exceptions que le module doit stocker dans une table à part.

4.2 Le transcripteur phonético-orthographique

L'algorithme de passage de la transcription phonétique à la forme orthographique du mot doit prendre en compte certaines caractéristiques de la langue en particulier :

- La quantité de la voyelle (brève ou longue).
- La gémination et le tanwin.
- L'assimilation ou non du /l/ avec les consonnes lunaires et solaires.
- Les lettres non prononcées.
- La prononciation du /t/ "fermé" à la fin des noms et des adjectives.
- Les mots irréguliers.

L'absence d'un clavier bilingue, nous contraint de

negliger cette phase et de travailler sur la structure phonétique interne du mot.

4.3 Le module syntaxico-sémantique

La structure grammaticale de l'Arabe standard est rigide, et obéit à des règles qu'on peut formaliser. Cette rigidité n'est pas évidente lorsqu'il s'agit du langage parlé. Comme restriction, nous prenons, dans un premier temps, une grammaire artificielle qui engendre la plus grande partie des phrases. Une phrase correcte de l'Arabe peut être soit nominale (commence par un nom) ou verbiale (commence par un verbe). Dans les deux cas, on peut bien introduire une "adat" pou exprimer par exemple une négation ou une intérrogation. Le verbe s'accorde en genre et en nombre avec son sujet seulement lorsque le sujet le précede. Dans le cas contraire, il n'y a accord que dans le genre .
Le module syntaxico-sémantique se charge de traduire le treillis de mot fournis par le module phonologique en des phrases ayant un sens en utilisant les règles syntaxiques et sémantiques. Il travaille sur les racines des mots d'une part et d'autre part, sur le résultat du système phonétique sur les voyelles. Les consonnes ont plus de rôles à jouer sur ce plan. Les voyelles servent à règler certains problèmes d'accord.
La vocalisation d'un message, qui consiste à mettre des voyelles sur les consonnes n'est nécessaire en Arabe que pour enlever certaines ambiguités dans le sens de la phrase.

4.4 Le système prosodique

Le rôle principal de la prosodie est de déterminer la nature de la phrase. Les indices prosodiques à extraire sont :

4.4.1 La durée

La durée relative d'un phonème dépend de son environnement, de la vitesse d'élocution et d'autres facteurs et elle est très significative dans la langue arabe. Le calcul de la durée vocalique moyenne permet d'avoir une idée sur la vitesse d'élocution.

4.4.2 L'accent

Les syllabes ne sont pas produites avec la même intensité, elles peuvent être accentuées ou non. Nous distinguons en Arabe deux types d'accents :

- Un accent inhérent ou primaire qui se situe au niveau du mot et dépend de la structure syllabique de ce dernier.
- Un accent global ou secondaire qui apparait au niveau de la phrase.

Les règles qui régissent les syllabes accentuées d'un mot de l'Arabe seront formalisées et utilisées par le module prosodique.

4.4.3 L'intonation

La fréquence fondamentale ou pitch est fonction du voisement des phonèmes. Le modèle du pich dépend de la phrase. En Arabe, il existe cinq types de phrases :

- Les phrases déclaratives.
- Les commandes.
- Les questions.
- Les appels.
- Les exclamations.

L'extraction du modèle du pitch permet d'avoir une idée sur la nature de la phrase prononcée.

5 Conclusion

Nous avons présenté dans cet article une architecture d'un système de reconnaissance de l'Arabe moderne. Un système de reconnaissance de la parole continue dans un contexte multilocuteur. La partie décodage phonétique est en grande partie opérationnelle; pour trois locuteurs masculins, le système SAPHA arrive à faire la reconnaissance analytique des phonèmes avec un taux global de l'ordre de 80 pourcent. Il reste à introduire de nouvelles connaissances pour améliorer le pourcentage de reconnaissance phonétique et développer les algorithmes du décodeur linguistique.

Références

[1] Salman H. AL. Ani. *ARABIC PHONOLOGY An Acoustical and Physiological Investigation.* Mouton & Co N.V., The Hague, 1970.

[2] J. F. Bonnot. *Recherche expérimentale sur la nature des consonnes emphatiques de l'Arabe classique.* Technical Report 9, Travaux de l'institut de phonétique de Strasbourg, 1977.

[3] A. Callec, S. Monne, M. Querre, O. Travarain, and G Mercier. Automatic segmentation of phonetic units and training in the keal speech recognition system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2000–2003, Paris, 1982.

[4] M. Djoudi. *Détection et localisation de la barre d'explosion en parole continue et dans un contexte multilocuteur.* Technical Report, Centre de Recherche en Informatique de Nancy, 1986.

[5] M. Djoudi. *Etude phonétique de l'Arabe standard.* Technical Report 89-R-057. Centre de Recherche en Informatique de Nancy, 1989.

[6] M. Djoudi, D. Fohr, and J. P. Haton. Phonetic Study for Automatic Recognition of Arabic. In *Proceedings of European Conference on Speech and Technology*, pages 268–271, Paris, September, 1989.

[7] M. Djoudi, D. Fohr, and J. P. Haton. SAPHA : un système expert pour le décodage acoustico-phonétique de l'arabe standard. In *Première Conférence Maghrébine sur le Génie logiciel et l'Intelligence Artificielle*, Constantine, September, 1989.

[8] D. Fohr. APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue. *Thèse de Doct. Univ. de NANCY 1*, 1986.

[9] S. Ghazali. Elements of Arabic Phonetics. In *Applied Arabic Linguistics and Signal & Information Processing*, pages 51–58, 1987.

[10] A. Giannini and M. Pettorino. *The Emphatic Consonants in Arabic.* Giardini editori e stampatori, 1982.

[11] Gillet and et al. SERAC : Un système expert en reconnaissance acoustico-phonétique. *Actes du 4ème congrès Reconnaissance des Formes et Intelligence Artificielle*, 1984.

[12] Ibn Jinni. *Sirr SinaaEat Al IEraab.* Mustapha Al Halabi, 1954.

[13] Y. Laprie. *Notice d'utilisation de Snorri.* Technical Report, Centre de Recherche en Informatique de Nancy, 1988.

[14] S. De Sacy. *Grammaire arbe.* De Sacy, 1810.

[15] Sibawayh. *EL KITAB, traité de grammaire arabe.* H. Derembourg, 1889.

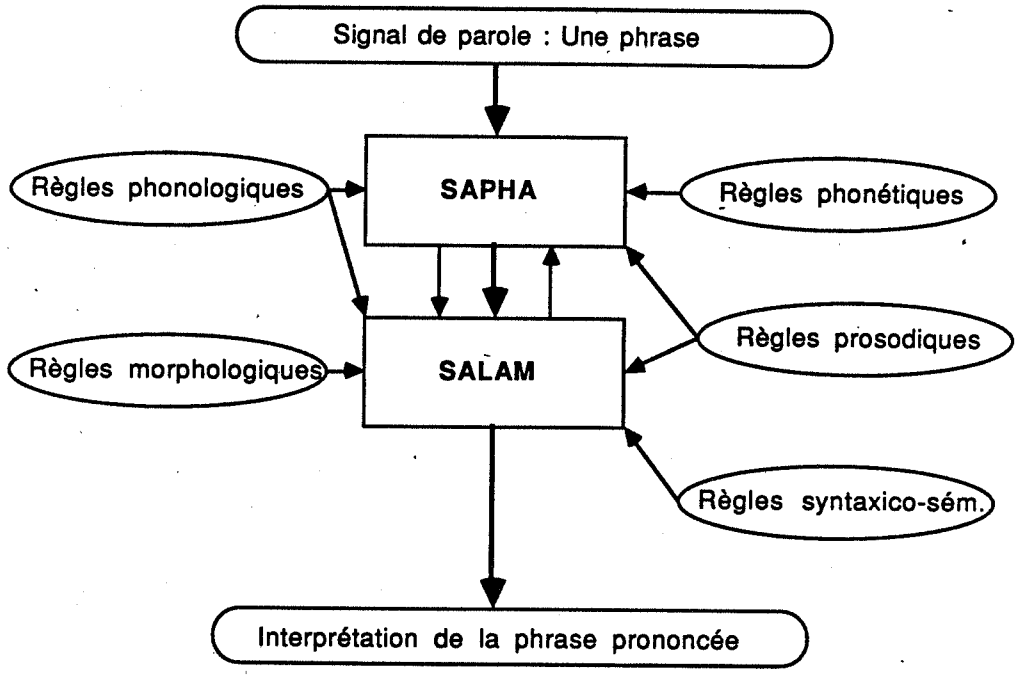


Figure 1 : Architecture de MARS

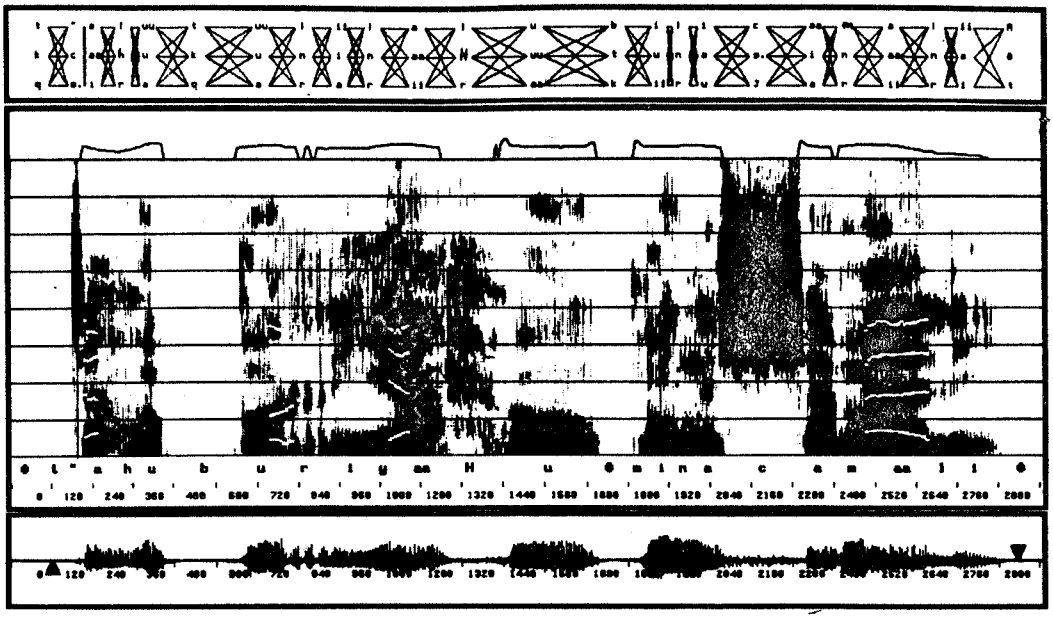


Figure 2 : Etiquetage phonétique d'une phrase

Traduction : "Les vents soufflent du nord"

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Calcul Dynamique de Pondération Sémantique dans un Algorithme DTW

S. Bornerand, F. Néel, G. Sabah

LIMSI-CNRS BP 133 91403 ORSAY CEDEX

Abstract

This paper describes a method to compute semantic weights and to adapt a DTW algorithm to use evolving weights during the recognition process. The weights are obtained from word co-occurrence frequencies which were calculated from a training corpus. Tests were made using an air-traffic controller application where the language is described by a word-pair grammar. We evaluated the system, both with and without weights, on a speech database containing almost 500 utterances from 6 speakers.

1. Introduction

Depuis la première application de la programmation dynamique à la parole [VINTSJUK,1968], la technique a évolué en fonction des spécificités du langage et de la parole. Les évolutions propres à la parole concernent le passage du mode isolé de diction de mots à un mode de prononciation en mots enchaînés [SAKOE, 1971]. Parallèlement à ce progrès, la linguistique a apporté sa contribution par l'intermédiaire des contraintes syntaxiques qui ont été introduites dans les systèmes de reconnaissance [SAKOE, 1979][GAUVAIN, 1982] sous la forme de grammaire régulière. Mais, ces enrichissements demandent de plus en plus de puissance de calcul. Et, malgré le développement de microprocesseur dédié à la programmation dynamique [QUENOT, 1986], les besoins d'un algorithme de comparaison dynamique (DTW ou Viterbi) en reconnaissance de mots enchaînés avec contraintes syntaxiques sont tels que le nombre de règles syntaxiques et le nombre de mots du vocabulaire sont limités.

Or, on constate que pour répondre aux exigences des applications qui nous sont proposées aujourd'hui en parole, il est impératif de pouvoir étendre le vocabulaire et la syntaxe. Dans cette perspective, la syntaxe ne peut plus être formalisée par une grammaire régulière; elle devient trop volumineuse pour être acceptée par les systèmes. Une alternative opérationnelle consiste à utiliser des syntaxes binaires -règles de succession de mots- entraînant une baisse sensible des performances de ces systèmes. L'objectif de ce travail est de pallier ces pertes par l'utilisation de contraintes sémantiques qui agissent en accord avec les contraintes syntaxiques pour réduire au maximum le nombre de comparaison des images acoustiques de mot. La définition des contraintes sémantiques s'appuie sur le résultat d'études en psycho-linguistique qui ont montré que des relations sémantiques pouvaient être déduites d'observations portant sur les associations de mots [EVENS, 1983]. En partant de ce point de vue réducteur de la sémantique linguistique, nous utilisons les fréquences d'association de mots pour calculer les pondérations sémantiques introduites dans l'algorithme DTW. Contrairement à l'approche présentée dans [MATROUF, 1990] où les pondérations sont mises à jour entre deux messages, nos pondérations évoluent au cours de la reconnaissance d'une phrase en fonction des solutions partiellement reconnues.

2. La Fréquence de Cooccurrence

Définition: **La relation de cooccurrence , notée C,** est une relation réciproque entre deux mots qui s'établit lorsqu'ils appartiennent à un même contexte. Dans ces travaux, **le contexte** est la phrase.

Pour un mot appartenant à un contexte, il existe un ensemble de mots qui possèdent une relation de cooccurrence C avec ce mot. La réunion de ces ensembles regroupant la totalité des relations de cooccurence qu'un mot possède dans une langue, s'appelle **le voisinage** d'un mot. La relation de cooccurrence entre un mot m_i et un mot m_j peut être affectée d'une grandeur mesurable que l'on appelle **la fréquence de cooccurrence,** notée $F(m_i, m_j)$. Cette valeur est représentative du nombre de cooccurrence de m_i avec m_j, dans le langage observé.

Exemple:

voisinage(*prenez*) = { *vitesse, noeuds, niveau, zéro* }

F(*prenez, niveau*) = 0.18 F(*prenez, noeuds*) = 0.26

F(*prenez, vitesse*) = 0.26 F(*prenez, zéro*) = 0.27

L'obtention des fréquences de cooccurrence s'appuie sur deux opérations: **le comptage et la normalisation.** Nous allons présenter deux méthodes de calcul des fréquences de cooccurrence qui diffèrent par leur normalisation.

Pendant la phase de comptage, l'utilisation d'un corpus contenant un ensemble de N phrases du langage permet de calculer les valeurs des variables suivantes:
- $\mathbf{m(m_i,m_j)}$ représente le nombre de fois où le mot m_j appartient aux contextes d'apparition du mot m_i;
- $\mathbf{n(m_i)}$ représente le nombre d'occurrences du mot m_i.

L'expérimentation montre qu'un voisinage contenant trop d'éléments tend à disperser les relations de cooccurrence qui perdent alors toute valeur de

représentativité de la cohérence sémantique entre les mots. Pour éviter cette dispersion, il faut restreindre le voisinage en utilisant **un seuil de rejet**. De plus, ce phénomène peut se répéter systématiquement pour un même mot qui n'apparaîtrait plus dans aucun voisinage. Ces termes sont appelés **des mots vides**. Ils sont définis par la propriété suivante:

$\forall i$, m_i est un mot vide $<=>$ voisinage(m_i) = 0 (1)

Pour simplifier le processus de construction des voisinages, certains mots peuvent être catégorisés à priori comme des mots vides afin de mieux contrôler la dispersion.

2.1. Construction des Voisinages

Dans un premier temps, on ajoute au voisinage de m_i toutes les cooccurrences détectées dans le corpus sans condition. Puis en fin d'analyse, un filtrage minimum supprime tous les éléments placés dans la catégorie des mots vides. A partir de ces voisinages, un second filtrage optionnel élimine les éléments parasites m_j caractérisés par des fréquences de cooccurrence $F(m_i,m_j)$ qui sont inférieures au seuil de rejet défini par le logarithme du nombre d'occurrence $n(m_j)$. Le résultat, appelé **voisinage canonique**, regroupe tous les contextes canoniques d'occurrence d'un mot.

2.2. Normalisation Absolue

Cette normalisation donne une fréquence de cooccurrence, notée **Fa**, indépendante du voisinage.

$\forall i \forall j$ / $m_j \in$ voisinage(m_i), $Fa(m_i,m_j) = m(m_i,m_j) / n(m_i)$ (2)

2.3. Normalisation Relative

Cette normalisation fournit une fréquence de cooccurrence, notée **Fr**, relative au voisinage.

$\forall i \forall j \forall k$ / $m_j \in$ voisinage(m_i), $m_k \notin$ voisinage(m_i),
$Fr(m_i,m_j) = m(m_i,m_j) / \Sigma_k m(m_i,m_k)$ (3)

3. La Force de Cohésion

Sachant que chaque mot possède un voisinage qui définit des relations de cooccurrence, on introduit alors **une force de collocation** pour chaque mot m_i d'une séquence S, définie par une fonction notée $f(m_i/S)$, telle que

$\forall i \forall j$ / $m_i \in S$, $m_j \notin$ voisinage(m_i)$\cap$S,
$f(m_i/S) = \Sigma_j F(m_i,m_j)$ (4)

Un mot apparaissant dans un contexte qui lui est habituel, donne une force de collocation élevée. Inversement, un contexte accidentel fournit une faible force de collocation. On constate que la force de collocation donne l'estimation de la concordance d'un mot avec le contexte S de son occurrence. Nous définissons alors la **force de cohésion** d'une séquence S de mots par une combinaison linéaire de ses forces de collocation tout en imposant une plage de variation comprise entre 0 et 1.

3.1. La Force de Cohésion Ha

La force de cohésion **Ha** s'obtient en calculant la moyenne des forces de collocation appliquées aux fréquences de cooccurrence Fa et divisées par le nombre de mots, sur l'ensemble des mots de la séquence. Ce qui donne:

$\forall i \forall j$ / $m_i \in S$, $m_j \notin$ voisinage(m_i)$\cap$S,
$Ha(S) = \Sigma_i \Sigma_j Fa(m_i,m_j) / card(S)\,(card(S)-1)$ avec $0<Ha(S)<1$ (5)

3.2. La Force de Cohésion Hr

On obtient la force de cohésion **Hr** en calculant directement la moyenne des forces de collocation appliquées aux fréquences Fr sur l'ensemble des mots de la séquence. Ce qui donne:

$\forall i \forall j$ / $m_i \in S$, $m_j \notin$ voisinage(m_i)$\cap$S,
$Hr(S) = \Sigma_i \Sigma_j Fr(m_i,m_j) / card(S)$ avec $0<Hr(S)<1$ (6)

Lorsque les forces de collocation sont élevées, la force de cohésion tend vers son maximun. Inversement, la force de cohésion s'annule avec les forces de collocation. **Ainsi, la force de cohésion peut être considérée comme une pondération sémantique qui reflète la cohérence sémantique d'un énoncé.**

4. Intégration DTW

L'objet de cette démarche est de nuancer les résultats des algorithmes DTW par l'utilisation de pondération sémantique. Les algorithmes DTW déterministes retournent une solution optimale correspondant à la séquence de mots reconnus délivrant **la distance cumulée minimale**. Le critère d'optimisation peut être exprimé sous la forme simplifiée suivante:

MIN($\Sigma_i d_i$) où d_i représente la distance entre deux structures élémentaires de comparaison à l'instant i (7)

L'introduction de pondération sémantique permet de calculer un score qui doit être optimal lorsque la distance cumulée est minimale et la pondération maximale. Il suffit alors de remplacer la pondération par **une pénalité** pour obtenir le nouveau **critère d'optimisation de l'algorithme DTW modifié**:

MIN ($\Sigma_i d_i$ - alpha log PONDERATION) (8)

Cette modification du critère ne préserve pas le principe d'optimalité. L'algorithme ne donne plus la solution optimale mais délivre seulement la meilleure solution pour le critère exposé ci-dessus.

Cependant, la pondération ne doit pas être appliquée en fin de reconnaissance pour trouver la meilleure séquence parmi un ensemble de solutions. L'apport de celle-ci doit être pris en compte de manière continue. A chaque étape de calcul, la recherche du minimum doit dépendre de la distance cumulée et de la pondération associée.

4.1. Principe général

Le principe de l'algorithme DTW est fondé sur l'obtention de la solution optimale à partir d'un **calcul récurrent** de la **distance cumulée g**. On définit alors une **pondération cumulée w** par les équations locales suivantes:

intra-unité
$$\begin{cases} g_{i+1} = MIN_j\,(g_{i,j} + d_{i+1} + p^n{}_{i,j}) \\ p^n{}_{i+1} = p^n{}_{i,j} \\ w^n{}_{i+1} = w^n{}_{i,j} \end{cases}$$

inter-unité
$$\begin{cases} g_{i+1} = g_i + d_{i+1} + p^{n+1}{}_{i+1} \\ p^{n+1}{}_{i+1} = pen(w^n{}_i) \\ w^{n+1}{}_{i+1} = eval\,(w^n{}_i, m) \end{cases}$$

avec $w^n{}_i$: pondération cumulée d'une séquence de n mots à l'instant i;
d_i: distance locale à l'instant i;
g_i: distance cumulée à l'instant i;
$p^n{}_i$: pénalité calculée à partir de la pondération $w^n{}_i$;
eval(), pen(): définies dans le chapitre suivant;
m : mot qui accroît la séquence.
(9)

La première, appelée intra-unité, définit la progression du calcul à l'intérieur des unités acoustiques, présentement des mots, dans laquelle j indexe des chemins autorisés. La seconde, appelée inter-unité, définit les contraintes de calcul lors du passage d'un mot à un autre.

Nos développements sont directement issus du système de reconnaissance AMADEUS [QUENOT, 1986] [TUBACH, 1989] réalisé au LIMSI par J.L. GAUVAIN [GAUVAIN, 1990]. Les modifications apportées permettent de prendre en compte la pondération cumulée et les pénalités dans les équations ci-dessus.

4.2. Calcul des Pondérations

4.2.1. Modèle Probabiliste

On adopte ici une démarche similaire à l'approche probabiliste dans le sens où **la pondération** d'une séquence de mots s'obtient par le produit **des probabilités conditionnelles** p_i des mots qui la composent. En remplaçant dans l'expression (8) la pondération par le produit des probabilités conditionnelles, on obtient un critère d'optimisation qui nous ramène au système d'équations défini dans le paragraphe §4.1.

Dans [DEROUAULT, 1985], l'auteur définit une distribution qui prédit un mot m dans un contexte ctx comme suit:

$\forall i \in ctx,\ p(m/ctx) = p(m) * \sum_i Fa(m,m_i)$

avec p(m) : fréquence lexicale du mot m trouvée dans un dictionnaire;
Fa : fréquence de cooccurrence définie en (2);
m_i : mot en place i dans le contexte ctx;
ctx: nombre de mots qui seront retenus à droite et à gauche de m.

(10)

Avec un algorithme DTW, seul le contexte gauche d'un mot est accessible pour calculer dynamiquement des probabilités. En posant p(m) = 1 pour s'abstraire des fréquences lexicales, nous obtenons alors **la probabilité conditionnelle dynamique** suivante:

$\forall j / m_j \in S_n,\ p(m/S_n) = \sum_j Fa(m, m_j)$

avec S_n : séquence des n premiers mots d'une phrase;
m_j : mot en place j dans S_n.

(11)

Un premier système d'équations est obtenu en définissant les fonctions eval() et pen() de la manière suivante:

SYSTEME PA

$\{ eval(w^n, m) = \sum_i Fa(m, m_i)$

$\{ pen(w^n) = - alpha \sum_j \log Fa(m, m_j)$ avec j = 1 à n

4.2.2. Modèle avec Estimation

A chaque instant et pour chaque séquence, **une fonction d'évaluation** délivre **une force de cohésion** considérée comme **la pondération** utilisée dans le cas de la solution optimale. Mais, cette pondération optimale étant inconnue pour des séquences intermédiaires, on la remplace par **une estimation ~P** qui est réévaluée après chaque modification de séquence de mots.

$$\text{MIN}\left(\sum_i d_i * \text{PONDERATION}\right) <-> \text{MIN}\sum_i\sum_j (d_{i,j} * {\sim}P) \quad (11)$$

De la même manière que précédemment, il s'agit de trouver un optimum qui minimise la distance acoustique et maximise la force de cohésion. Là encore, il suffit de transformer la pondération ~P en pénalité afin de nous ramener au système d'équations défini au chapitre §4.1 dans lequel la fonction eval() correspond désormais à la fonction d'évaluation.

4.2.2.1. Estimation avec Fa

La décomposition de la force de cohésion Ha (5) en un calcul récurrent a permis d'identifier les fonctions pen() et eval() du système d'équations qui suit:

SYSTEME EA

$\forall j / m_j \in$ voisinage(m)$\cap S_n$,

$\{ eval(w^n, m) = w^n + \sum_j (Fa(m,m_j) + Fa(m_j,m))$

$\{ pen(w^{n+1}) = - alpha \log (w^{n+1} / (n\,(n+1)))$

Sachant que l'algorithme DTW fonctionne avec un **mécanisme d'élagage (pruning)**, il est souhaitable d'introduire les pénalités les plus faibles afin d'éviter une élimination prématurée d'un chemin S_{n+1}. Au lieu d'injecter directement la pénalité p^{n+1}, on retranche au préalable la pénalité minimale à tous les chemins S_{n+1}.

4.2.2.2. Estimation avec Fr

A partir de la force de cohésion Hr exprimée en (6), on peut aussi définir les fonctions eval() et pen() par le système d'équations suivant:

SYSTEME ER

$\forall j / m_j \in$ voisinage(m)$\cap S_n$,

$\{ eval(w^n, m) = w^n + \sum_j (Fr(m,m_j) + Fr(m_j,m))$

$\{ pen(w^{n+1}) = - alpha \log (w^{n+1} / (n+1))$

On peut aussi appliquer ici le raisonnement tenue pour EA en ce qui concerne l'injection des pénalités.

4.2.3. Probabilité Vs Estimation

Les probabilités apprises statistiquement sur un grand nombre de réalisations sont considérées comme des prédictions. Le calcul d'une probabilité permet de prédire l'occurrence d'un mot sachant son contexte. Les estimations obtenues dynamiquement à l'aide d'une fonction d'évaluation sont considérées comme des vérifications. Le calcul d'une estimation évalue le coût de l'occurrence d'un mot dans un contexte.

5. Les Résultats

5.1. L'Application

Le souci d'obtenir un terrain de comparaison avec des systèmes existants nous a poussé à choisir une application de contrôle aérien qui est actuellement en cours de développement au LIMSI [MATROUF, 1987]. Le langage de l'application comprend 212 mots. Les chiffres sont inclus dans ce vocabulaire mais les nombres sont exclus du langage. La description de la syntaxe contient 928 règles binaires.

Un défaut des langages applicatifs tels que la phraséologie utilisée dans les dialogues pilote-contrôleur, réside dans la présence de nombreuses occurrences des chiffres, considérés comme des mots vides. En effet, ils apparaissent dans l'indicatif qui se trouve en début de chaque message et en place réservée aux valeurs numériques. Les indicatifs composés d'un nom de compagnie et d'un numéro de vol permettent de désigner le destinataire du message. Cependant, les chiffres des indicatifs possèdent une forte correllation avec le nom de la compagnie aérienne puique un même numéro de vol est toujours associé à une même compagnie. Pour réintégrer les numéros de vol dans le vocabulaire des mots pleins, nous avons distingué les chiffres de l'indicatif en les codant sous forme alphabétique alors que les chiffres donnant les valeurs numériques sont les chiffres arabes.

Exemple: "air-france zéro sept sept quebec prenez le cap 2 0 0"

5.2. L'Apprentissage des Fréquences de Cooccurrence

N'ayant pas à notre disposition un véritable corpus sur la phraséologie des dialogues pilote-contrôleur, une génération automatique a permis de collecter un ensemble de 10.000 phrases à partir d'une description de la syntaxe

sous forme de grammaire régulière respectant les contraintes sémantiques du langage. Ces 10.000 phrases comprennent un nombre de mots allant de 8 à 20 avec une moyenne proche de 12. A partir de ce corpus, nous avons construit **quatre matrices** correspondant respectivement aux voisinages simples et aux voisinages canoniques pour les deux types de fréquences de cooccurrence. Pour les voisinages simples, le nombre d'éléments variait de 10 à 150 mots avec une moyenne de 27 à 66 mots par voisinage en fonction du type de fréquences de cooccurrence, et le nombre de mots vides était de 26. En effet, certains mots n'apparaissant pas dans le corpus sont automatiquement classés comme mot vide. Ils correspondent aux chiffres et aux indicatifs non compris dans le langage. Pour les voisinages canoniques, le nombre d'éléments de chaque voisinage a considérablement diminué, variant de 5 à 50 mots avec une moyenne de 14 à 16 mots par voisinage, tandis que le nombre de mots vides restait voisin du précédent.

5.3. Une pré-sélection

Un ensemble de tests préliminaires a permis de faire la sélection parmi les différents systèmes d'équations (PA, EA, ER), les deux types de fréquence de cooccurrence (Fa, Fr) et les deux voisinages, ainsi que de régler les paramètres des programmes. Le corpus était composé de 262 phrases prononcées par un locuteur avec une longueur moyenne de 10 mots. Les différentes configurations testées {système + matrice} ont donné des résultats dont on a pu tirer un certain nombre de conclusions valides pour l'application. Le corpus utilisé étant artificiel, il ne donne aucune information sur la probabilité des phrases. Les fréquences de cooccurrence n'ont aucune valeur dans l'absolu. Elles ont un sens uniquement dans un voisinage donné. Dans ces conditions, on constate que les fréquences Fr donnent de meilleurs résultats que les fréquences Fa dans ces tests. Il apparaît alors que **la meilleure combinaison rassemble les fréquences relatives de cooccurrence, les voisinages canoniques et le système d'équations sans soustraction du minimum.** Celle-ci sera donc utilisée pour évaluer le système.

5.4. L'évaluation

Les tests ci-dessous ont permis d'évaluer le système avec et sans pondération afin de tirer les avantages, les inconvénients et les limites de l'emploi des pondérations dynamiques pour l'application du contrôle aérien. La base de données contient 473 phrases provenant d'enregistrements effectués auprès de 5 locuteurs (4 masculins, 1 féminin). Chacun ayant prononcé une centaine de phrases contenant en moyenne 10 mots; le nombre variant de 6 à 14 mots. L'apprentissage des références a été fait en contexte sur un ensemble de 150 phrases. La taille du dictionnaire des références acoustiques varie de 330 à 486 mots en fonction du locuteur.

473 phrases	CME1 sans pondération	CME2 avec pondération	VCME avec pondération
taux de reco. de phrase	71 %	75 %	76 %

Le tableau ci-dessus indique le taux de reconnaissance des phrases calculé sur les résultats de toute la base. La colonne CME2 correspond au système avec pondération utilisant les équations ER avec un voisinage canonique. La colonne VCME reprend la même configuration que CME2 mais introduit un facteur "alpha" variable en fonction de l'émergence de la meilleure solution. Lorsque l'émergence est forte, les pondérations sont presque ignorées. Par contre, une faible émergence met en avant les pondérations afin d'introduire une différence et de modifier l'émergence.

Le système sans pondération CME1 obtient une performance de 71 % de reconnaissance sur l'ensemble de la base. L'utilisation de pondérations calculées dynamiquement permet d'obtenir un taux de reconnaissance de 75 %. Cependant, l'analyse des résultats a mis en évidence un phénomème de forte dégradation des phrases erronées. On constate que l'utilisation des pondérations introduit une baisse sensible au niveau du taux de reconnaissance des mots par rapport au comportement sans pondération. Pour éviter cet écueil, la prise en compte variable des pondérations faite dans VCME a permis de diminuer la dégradation et d'améliorer les performances en atteignant 76% de reconnaissance de phrases. En prenant ce dernier système, on peut dresser la liste suivante des erreurs corrigées ou introduites par VCME au regard des résultats obtenus sans pondération.

Erreurs rattrapées par VCME	
Insertions	34
Omissions	0
Substitutions	15
Multiples	1

Erreurs introduites par VCME	
Insertions	4
Omissions	10
Substitutions	15
Multiples	0

On constate que la majorité des erreurs rattrapées sont des insertions. En effet, VCME corrige 34 insertions faites par CME1 alors qu'il n'engendre que 4 insertions. Pour les omissions, on obtient l'effet contraire. VCME introduit 10 omissions alors que CME1 n'en avait faite aucune. **Il semble alors que l'utilisation de pondération tend à forcer le meilleur choix le plus tôt possible.** Par contre, il faut constater que VCME introduit autant de faute de substitution qu'il n'en rattrape. Nous avons 15 substitutions dans chaque tableau. Cela signifierait que VCME n'est guère plus discriminant pour choisir un terme plutôt qu'un autre, que ne l'est CME1. Mais, les caractéristiques du langage de l'application n'y sont pas étrangères. En effet, la plupart des erreurs engendrées par VCME sont de la forme:

Exemple:
message: alpha fox-trot zéro sept sept à droite le cap 2 5 0.
reconnaissance: alpha fox-trot zéro sept sept un droite le cap 2 5 0.

Or, cette erreur ne peut pas être rattrapée car le mot "un" donne une meilleure pondération que le mot "à" lorsque l'on ne considère que la partie gauche de la phrase par rapport à la position tenue respectivement par l'un des deux mots. Ce type d'erreur est essentiellement dû à la forme syntaxique des messages. Les deux principaux constituants syntaxiques (<indicatif> suivi de

<commande>) sont sémantiquement indépendant. Lorsque VCME commence l'analyse de la partie <commande> d'un message, il préfère, lors d'une faible émergence, continuer la partie <indicatif>.

Malgré ce problème dépendant du langage, l'utilisation de pondération sémantique apporte un **gain de 5%** par rapport au taux de reconnaissance du système sans pondération. Cependant, nous avons comparé un système utilisant des pondérations avec un système sans pondération. Or, l'utilisation de pondération réduit implicitement la perplexité du langage avantageant ainsi le système avec pondération. Pour une évaluation plus fine, il faudrait maintenant étudier le comportement du système avec des pondérations statiques et dynamiques apprises sur un corpus réel.

6. Discussion

Dans tous nos systèmes d'équations, la pondération sémantique est obtenue à partir de la force de collocation appliquée à chaque mot m_i d'une séquence S. On s'aperçoit que l'utilisation d'un contexte général -constitué par S- indépendant du mot rencontré peut provoquer une atténuation de la force de collocation. L'introduction des voisinages canoniques permet d'adapter le contexte à chaque mot m_i. En ce sens, le calcul s'effectue uniquement sur les prédictions des relations de cooccurrence lexicale qui sont décrites dans le voisinage canonique. Cependant, quel que soit le contexte du mot, le même voisinage canonique sera utilisé. Dans un langage homogène, les mots sont toujours utilisés dans des contextes qui sont proches les uns des autres. Les contextes renvoient à un nombre très réduit de structures syntaxiques; toutes faisant références au même thème. Ces conditions se satisfont alors d'un voisinage statique tel qu'il a été mis en oeuvre dans ce travail. Mais, dès que l'on aborde un langage hétérogène, qui traite de sujets divers avec un grand vocabulaire, les mots apparaissent dans des contextes variés qui ne peuvent plus être caractérisés uniquement par des associations lexicales. Le voisinage est donc calculé dynamiquement en fonction des catégories syntaxiques et des types de voisinage du mot m_i. Ces derniers sont alors appelés catégories sémantiques ou conceptuelles. Le principe des relations de cooccurrence peut être généralisé aux catégories sémantiques dans lequel les voisinages sont alors obtenus par des descriptions formelles [INOUE, 1989].

7. Conclusions et Perspectives

Les deux principaux volets de ce travail ont été la définition des pondérations à partir de fréquence de cooccurrence lexicale et leur intégration dans un algorithme DTW. Une évaluation a permis de valider l'hypothèse selon laquelle l'utilisation de pondération sémantique améliore la reconnaissance en parole continue pour une application donnée, tout en se situant dans le cadre d'un apprentissage automatique: la syntaxe (règles de succession) et la sémantique (fréquences de cooccurrence) peuvent être extraites d'un corpus.

L'extension de la méthode à un langage quelconque demande une généralisation du principe des relations de cooccurrence. A ce titre, nous poursuivons cette étude par le développement d'une représentation sémantique qui supporte la définition de voisinage conceptuel et un mode d'exploitation permettant d'obtenir les pondérations sémantiques.

Références

A.M. Derouault, 1985, "Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques", p.25-28, p.79-85, Th. d'état ès Sciences, Paris VII.

M.W. Evens, B.E. Litowitz, J.A. Markowitz, R.N. Smith, O. Werner, 1983, "Lexical-Semantic Relations: A Comparative Survey", p.34, p.182-186, Linguistic Research Inc., P.O. Box 5677 - Station 'L',Edmonton, Alberta, Canada T6C 4G1.

J.L. Gauvain, 1982, "Reconnaissance de mots enchaînés et détection de mots dans la parole continue", Thèse de 3 cycle, Paris XI.

N. Inoue, T. Morimoto, K. Ogura, 1989, "A Linguistic Knowledge Base for Applying Semantic Information to a Speech Understanding System", p.194-197, EUROSPEECH'89.

J.L. Gauvain, 1990, "AMADEUS: Principe et structure", Notes et documents LIMSI.

A.K. Matrouf, F. Néel et J.Mariani, 1987, "Système de dialogue orienté par la tâche: une application en avionique", JEP'87.

A.K. Matrouf, F. Néel, J.L. Gauvain and J. Mariani, 1990, "Adapting Probability-Transitions in DP-Matching Process for a Task-Oriented Dialogue", ICASSP'90, .

G. Quenot, J.L. Gauvain, J.J. Gangolf and J. Mariani, 1986, "A dynamic time warp VLSI processor for continuous speech recognition", p.1549-1552, ICASSP'86, Tokio.

H. Sakoe, S. Chiba, 1971, "A Dynamic Programming Approach to Continuous Speech recognition", paper 20C-13, ICASSP'71, Budapest, Hungary.

H. Sakoe, 1979, "Two-Level DP-matching - A Dynamic Programming-based pattern matching algorithm for connected word recognition", p.588, IEEE Trans. Acoust. Speech, Signal Processing, vol.ASSP-27, december 1979.

J.P Tubach, C. Gagnoulet, J.L. Gauvain, 1989, "Advances in Speech—Recognition Products from France", Proceedings of Speech Tech'89, 2-4 may, New-York.

Vintsjuk, 1968, "Reconnaissance de mots par programmation dynamique", p.81-88, Kybernetika 1, publication en Russe, 1968.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Reconnaissance monolocuteur des phonèmes du français au moyen de réseaux à masques temporels

Laurence DEVILLERS

L.I.M.S.I/C.N.R.S B.P 133, 91403 ORSAY Cedex, France.

ABSTRACT

A particular Multi-Layer Perceptron (MLP) so-called Time Delay Neural Network (TDNN) is applied to classify French phonemes using a backpropagation algorithm for a speaker-dependent recognition task. Its main feature consists of time-replicated units connected to a sequence of narrow, overlapping receptive fields. Some properties of TDNN are investigated in the first part of this paper, especially the time-shift and time distortion invariance. It is shown that a TDNN is more robust to both coarticulation and time distortions than a MLP. However, this "variability" is not dynamically managed by the system and must be present in the learning data. A modular architecture of TDNN is investigated in the second part and we present recognition results on 30 French phonemes.

I. INTRODUCTION

La complexité de la reconnaissance de la parole est due en grande partie à l'importante variabilité fréquentielle et temporelle du signal de parole. Chaque son élémentaire est déformé par les sons qui l'entourent par un effet de coarticulation. Le débit du locuteur, son état émotionnel et physique entraînent des modifications du signal de parole. Un système efficace de reconnaissance de la parole doit pouvoir classifier ces sons malgré les distorsions temporelles.

Un système connexionniste de type réseaux à masques temporels (TDNN) [1] utilisant un algorithme d'apprentissage de rétro-propagation du gradient est un modèle de mémoire associative fondé sur des critères statistiques. A la différence d'un Perceptron Multi-Couches (MLP) [4], les connexions entre les couches du réseau sont partielles, les matrices de poids sont dupliquées le long de l'axe temporel afin d'apprendre les caractéristiques acoustiques et leurs relations temporelles indépendamment de leur position dans le temps. Un tel système doit permettre en principe une invariance par translation et par distorsion temporelle.

Dans la partie II de cet article, nous exposons la stratégie d'apprentissage. Nous présentons, dans la partie III, des expériences sur différents paramètres des réseaux à masques temporels et nos conclusions sur les propriétés d'invariance en translation et distorsion des TDNN.
Des résultats sur des réseaux dédiés à la séparation de classes de phonèmes : occlusives sourdes, sonores, fricatives sourdes, sonores, liquides, nasales et voyelles sont donnés dans cette même partie.

Dans le but d'apprendre à un réseau une tâche plus complexe, par exemple la classification de 30 phonèmes du français, nous avons construit une architecture modulaire de TDNN [2] en réutilisant les représentations internes de réseaux pré-entraînés à reconnaître des sous-classes de phonèmes. Cette expérience est décrite dans la partie IV

II. STRATEGIE D'APPRENTISSAGE

Le but de l'apprentissage est de trouver la meilleure configuration de poids pour séparer les classes d'échantillons d'étiquette différente. La méthode d'apprentissage et l'architecture du réseau permettent de déterminer automatiquement une partition de l'espace de représentations, ce qui revient à trouver les poids des connexions qui minimisent une fonction de coût : l'erreur quadratique entre les sorties désirées et les sorties obtenues.

Au cours d'un apprentissage utilisant le gradient stochastique, les poids sont modifiés après chaque présentation d'un segment.

Dans la méthode globale [3] qui est la plus exhaustive mais aussi la plus coûteuse en temps, on modifie les poids du réseau après avoir calculé l'erreur sur tout le corpus d'apprentissage.

Dans cette étude, on a choisi de modifier les poids du réseau après un cycle d'apprentissage, un cycle équivaut à une présentation d'un segment de chacun des phonèmes à classifier. L'ordre de présentation des phonèmes dans le réseau n'influe pas sur l'apprentissage et chaquephonème a la même fréquence d'apparition. L'application de ce gradient "quasi-stochastique" nécessite d'utiliser avec circonspection les paramètres de modifications des poids pour ne pas tomber dans un minimum local.

1. Initialisation des poids

Les valeurs initiales des poids sont un facteur important dans le choix des paramètres de ce modèle. Les poids ont été initialisés à de petites valeurs aléatoires afin que l'écart type des sommes pondérées corresponde aux points de courbure maximum de la sigmoïde [5], c'est à dire que les sommes pondérées se situent dans la partie linéaire de la sigmoïde en début d'apprentissage.

2. La fonction de transfert d'activité

La sigmoïde a été choisie entre -1 et 1 avec une saturation à une activité de 2000.
$S(a) = (e^{ka} - 1)/(e^{ka} + 1)$
a: activité incidente sur la cellule
S(a): activité de sortie de la cellule
k: paramètre positif permettant de régler la sigmoïde.

La dérivée de la sigmoïde n'est jamais nulle; soit S' la dérivée de S, si S(a) = 1 alors S'(a) = S'(2000). Un apprentissage peut être trop rapide; l'activité de sortie des cellules est à 1 ou à -1 alors que l'apprentissage n'a été fait que sur une partie du corpus. Il semble intéressant dans ces conditions de pondérer la matrice de poids entre la couche cachée et la dernière couche d'un coefficient Z pour donner la possibilité au réseau de continuer l'apprentissage.

3. Ajustement des coefficients de modification des poids

La recherche du minimum global n'est pas un problème résolu. Pour une tâche de classification réduite comme la classification des occlusives voisées, le problème est d'un ordre raisonnable.

Equation de modification des poids
Wij(t) = Wij(t-1) - EPS * GRAD(Wij)
GRAD(Wij): dérivée de l'erreur par rapport au poids.
EPS : coefficient de modification du gradient est pondéré par la racine du nombre total de connexions arrivant sur une cellule.

Pour estimer les performances du réseau et superviser l'apprentissage, on utilise plusieurs critères : le taux de reconnaissance sur le corpus d'apprentissage, l'évolution de l'erreur quadratique et le taux de généralisation. L'erreur quadratique est calculée sur tout le corpus d'apprentissage, elle correspond à la moyenne des différences au carré entre les sorties désirées et les sorties obtenues. Le taux de généralisation représente le taux de reconnaissance sur le corpus de test.
On arrête l'apprentissage lorsque le taux de généralisation diminue. Pendant l'apprentissage, il existe un seuil à partir duquel l'erreur quadratique continue de décroître, les performances du réseau sur l'ensemble d'apprentissage continue d' augmenter et pourtant la faculté de généralisation du réseau diminue. Le réseau devient trop spécialisé aux particularités des données du corpus d'apprentissage.

Il est important d'utiliser des coefficients de modification EPS de taille réduite pour avoir un meilleur contrôle de l'apprentissage et une meilleur performance de généralisation. Dès que l'erreur quadratique converge pendant l'apprentissage, il faut diminuer EPS jusqu'à arriver en fin d'apprentissage à une erreur minimale et à une bonne généralisation des données.

Différents tests ont été effectués sur les variations de EPS en fonction du nombre de cycles d'apprentissage. Un bon apprentissage a été obtenu en faisant varier à chaque passe d'apprentissage de façon inversement proportionnel EPS et le nombre de cycles utilisés. De façon plus générale, on obtient de bons résultats avec le critère [5] :la série (EPS(t)) diverge tandis que la série $(EPS(t))^2$ converge; t = nombre de cycles.

III EXPERIENCES

1. Données

Pour un locuteur, nous disposons d'un corpus de logatomes alternant consonnes et voyelles dans différents contextes de type /apataka/, d' un corpus de 50 phrases de parole continue à débit naturel (environ 11 phonèmes par secondes) et de ces mêmes 50 phrases à débit lent (environ 8 phonèmes par secondes) [6]. Ces phrases contiennent l'inventaire exhaustif des diphones CV (consonnes - voyelles) et VC (voyelles - consonnes) du français.

Les données proviennent d'un signal échantillonné à 10KHz, analysé par une FFT toutes les 12.8ms, puis filtré sur 16 canaux répartis suivant une échelle de Bark (100Hz - 5000Hz). Chacune des seize valeurs du vecteur aussi appelé trame est codée suivant une échelle logarithmique sur 8 bits. Puis, l'énergie moyenne est retranchée à chaque coefficient et rajoutée comme 17ième coefficient du vecteur. Les valeurs de ces vecteurs sont ramenées entre -1 et 1 afin d'être homogènes avec les bornes choisies pour les valeurs d' activité des cellules du réseau.

Le corpus a été étiqueté et segmenté manuellement, l'unité phonétique choisie est le phonème. L'ensemble de ces phonèmes comprend 13 voyelles, 17 consonnes et une unité correspondant au silence.

Ce corpus recouvre une grande diversité phonémique en ce sens que les phonèmes sont issus de corpus très différents; naturel (parole continue) et artificiel (logatomes et parole à débit lent) mais il n'y a souvent qu'une occurrence pour chaque contexte phonétique. La base de données est de taille réduite; le corpus d'apprentissage correspond à environ 60% du corpus mixte (2/3 des logatomes, la première moitié du corpus des phrases à débit normal, la deuxième moitié du corpus de phrases à débit lent) soit 2173 segments et les 40% restant servent de corpus de test soit 1964 segments.

La construction de ce corpus mixte a été motivée par plusieurs raisons; le corpus des logatomes ne recouvre pas tous les contextes existants dans le corpus de phrases et la différence de débit de parole entraîne des distorsions temporelles et fréquentielles. Un corpus mixte permet d'avoir une représentation plus robuste des phonèmes.

2. Variation des paramètres du réseau

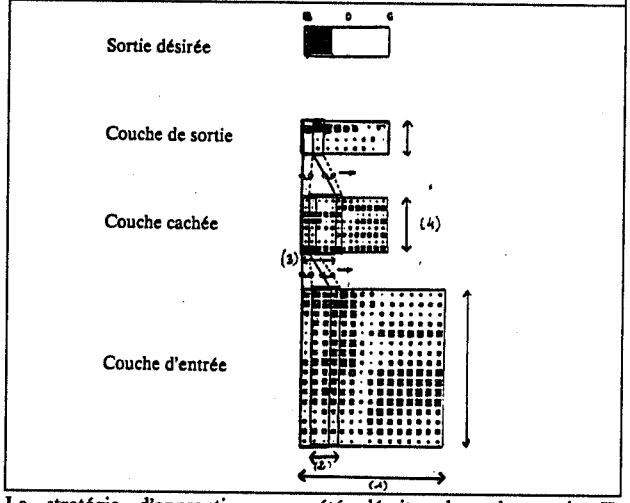
Nous avons essayé d'évaluer l'importance de la taille de la fenêtre d'analyse, des couches hiérarchisées, du contexte pris dans chacune de ces couches, du nombre de cellules dans la couche cachée afin de connaître la robustesse du modèle aux distorsions temporelles.

Une architecture de TDNN à une couche cachée à été implémentée. Une telle architecture semble suffisante pour séparer des nuages de données complexes tels que des classes de phonèmes. La dernière couche du réseau comporte une suite de vecteurs que l'on associe à des étiquettes phonétiques. On appelle ces étiquettes phonétiques, des "quasi-phonèmes". Une chaine de quasi-phonèmes représente un segment phonétique.

Les différentes expériences de cette partie ont été menées sur la séparation de trois phonèmes acoustiquement proches: **b d g**. La figure 1 représente un shéma de TDNN, les numéros entre parenthèses sont des repères donnés dans cette figure.

(1): La taille de la fenêtre d'analyse,
(2): Le contexte de la couche d'entrée,
(3): Le contexte de la couche cachée,
(4): Le nombre d'unités dans la couche cachée.

TDNN permettant de classifier les occlusives b d g: (Fig.1) [1]



La stratégie d'apprentissage a été décrite dans la partie II. L'architecture du réseau utilisé par défaut est la suivante: la taille du contexte utilisé dans la couche d'entrée est de 3, la taille du contexte dans la couche cachée est de 5, le nombre d'unités dans la couche cachée est de 8, La taille de la fenêtre d'analyse utilisée dans la couche d'entrée est une séquence de 10 trames, chaque trame correspondant à 12.8ms de parole. Dans ces conditions, on obtient une décision phonétique, c'est à dire un quasi-phonème (QPh) pour 7 trames d'entrée du signal de parole (7 x 12.8ms ~= 90ms de parole).

Dans certaines expériences, nous avons précisé le taux de reconnaissance sur les segments; pour un segment de 10 trames en entrée, le pas de translation étant de 1 sur chacune des couches, on obtient en sortie du TDNN 4 vecteurs auquels on associe 4 quasi-phonèmes, on reconnait un segment si au moins la moitié des quasi-phonèmes ont la bonne étiquette phonétique.

Tous les tests de cette étude ont été fait sur un seul locuteur, sur une partie de la base de données comportant des segments d'occlusives sourdes.

Taille de la fenêtre d'analyse

L'entrée du réseau est une fenêtre fixe de trames de parole centrée sur la partie stable d'un segment phonétique; pour les occlusives, on a choisit d'aligner les segments de façon à avoir un nombre identique de trames avant le début de l'explosion.

Chaque segment phonétique porte ses caractéristiques propres mais aussi les effets de coarticulation de son contexte qu'il est important de prendre en compte pour classifier les phonèmes. La taille moyenne des segments du corpus utilisé est de 6 trames. 7 trames issues de la fenêtre d'entrée permettent d'obtenir un vecteur de sortie dans l'architecture par défaut que nous utilisons. Une fenêtre de taille supérieure à 7 trames permettra de coder plusieurs positions du même segment dans un contexte différent. 7 trames trop excentrées pour un phonème ne gènent pas ce système à apprentissage statistique qui rejette automatiquement les exemples marginaux. L'apprentissage offre donc une généralisation robuste.

Nous présentons des résultats sur des variations de tailles de fenêtre d'analyse. On peut en déduire que plus la taille de la fenêtre d'analyse est grande, plus on apprend des séquences différentes pour une même étiquette, le réseau est donc plus robuste à un mauvais alignement des segments. Mais inversement la taille de la fenêtre d'analyse ne doit pas être choisie trop large au risque d'apprendre des positions discriminantes pour une autre étiquette phonétique.

Nous avons choisi d'utiliser une fenêtre de taille fixe identique pour l'apprentissage de tous les segments phonétiques. Cette décision paraissait dangereuse car les segments ont des tailles très variables. Le réseau dans cette situation code un large contexte pour des petits segments et pas de contexte pour d'autres. Les résultats obtenus sont intéressants et prouvent la bonne résistance du réseau à une segmentation très grossière.

Le tableau suivant donne des pourcentages de reconnaissance sur le corpus d'apprentissage (noté "AUTO" pour auto-cohérence) et sur le corpus de test. La notation "QPh." correspond à un quasi-phonème c'est à dire à un vecteur sortie du TDNN.

Variation de la taille de la fenêtre d'analyse		
fenêtre d'analyse	AUTO QPh.	TEST QPh.
12	96.8	88.3
10	96.4	91
8	95.8	90.8
7	94.7	90

Taille du contexte dans les couches

Pour montrer l'importance du contexte dans les différentes couches, nous avons effectués plusieurs tests que nous décrivons dans ce paragraphe.

En réduisant le contexte à 1 dans toutes les couches, les résultats montrent une baisse de taux de reconnaissance d'environ 20% sur les quasi-phonèmes (un peu moins sur les segments) par rapport aux scores du TDNN obtenus avec la configuration décrite par défaut. L'utilisation d' un contexte de plusieurs trames dans les couches permet donc d'améliorer notablement les résultats.

Contexte dans la 1ère couche : 1 trame Contexte dans la 2ème couche : 1 trame		
	QPh.	Segment
Auto	68.7	85.3
Test	64.9	81.1

La taille du contexte de la couche d'entrée doit permettre de discriminer les transitions rapides. Les variations plus lentes sont prises en compte par le contexte de la couche cachée qui doit être suffisamment grand pour discriminer un phonème

Les tests suivants ont été effectués en faisant varier les tailles de contexte dans les couches.

Variation de la taille de la fenêtre d'entrée, la taille de la fenêtre de la couche cachée est de 5 trames.		
nombre de trames	AUTO QPh.	TEST QPh.
1	92.8	84.1
2	95.4	87.6
3	96.4	91
4	97.4	90.5

Variation de la taille de la fenêtre de la couche cachée, la taille de la fenêtre d'entrée est de 3 trames.		
nombre de trames	AUTO QPh.	TEST QPh.
3	94.1	84.3
4	94.4	85.5
5	96.4	91
6	97.9	89.2

Les résultats montrent que lorsque les tailles de contexte dépassent un certain seuil, 3 pour la couche d'entrée et 5 pour la couche cachée, le réseau devient "sur-entrainé". il n'apprend plus seulement ce qui est utile pour classifier des phonèmes, il apprend des spécificités propres au corpus d'apprentissage. La conséquence est de faire baisser les taux de reconnaissance du corpus de test. Le réseau offre une moins bonne généralisation des données.

La taille des contextes définit la taille des matrices de poids des connexions du réseau. Dans le cas de contextes importants, un corpus de données très grand sera nécessaire pour obtenir un bon taux de généralisation. La solution MLP avec un contexte minimal apporte des résultats beaucoup moins bons. Le réseau fait une généralisation trop grossière et n'intègre pas l'aspect dynamique de la parole. En conclusion, il faut choisir des contextes suffisants pour coder les variations temporelles tout en étant conscient des limitations données par la taille du corpus d'apprentissage.

Tests sur le nombre d'unités cachées.

Les unités cachées permettent de définir des hyperplans, c'est à dire des frontières entre les classes de phonèmes. Plus le nombre d'unités cachées est grand, plus on pourra délimiter finement le nuage de phonèmes par des axes discriminants.

Les résultats obtenus sur la classification de **b d g** en faisant varier le nombre de cellules de la couche cachée montrent une baisse des taux

de reconnaissance lorsque la couche contient plus de 8 cellules. Le choix d'un nombre de cellules plus important que 8 dans notre étude nécessite plus de données d'apprentissage.

Classification de b d g : Comparaison après 6600 cycles d'apprentissage.					
NOMBRE D'UNITES DANS LA COUCHE CACHEE alignement des occlusives à 5 trames avant l'explosion					
nombre d'unités	nombre de connexions	TCPU VAX750	erreur quadratique	AUTO QPh.	TEST QPh.
5	2392	1H00	0.115	94.2	86.5
6	2868	1H00	0.113	94.9	87.8
8	3820	1H30	0.109	95.4	89
9	4296	2H00	0.119	95.3	87.2
11	5248	2H00	0.125	94.9	87.3
12	5724	2H30	0.139	94.9	86.4

3. invariance en translation et distorsion

Le TDNN intègre les fenêtres temporelles de la même couche en utilisant les mêmes poids de connexions ce qui permet de reconnaître les événements dans le temps. Le pas de translation des fenêtres des deux couches est de 1. Chaque trame apparaît dans toutes les positions possibles dans la fenêtre d'analyse du signal. Sur les bords de la fenêtre d'analyse, on ne code pas toutes les informations de contexte, pour cette raison on ne peut pas parler réellement d'invariance en translation.

Dans un réseau où 7 trames de la couche d'entrée donnent un vecteur de sortie, une fenêtre d'analyse de 10 trames en entrée permet de coder 4 positions du segment phonétique. L'invariance en translation du système correspond à pouvoir identifier le phonème appris à quatre positions différentes dans le temps. Nous préférons alors employer le terme de "quasi-invariance" en translation.

La robustesse à la distorsion temporelle de ce modèle est due uniquement au fait que l'on apprend plusieurs positions du phonèmes et plusieurs occurences du phonème dans la fenêtre d'analyse.

Nous décrivons ci-dessous un test très simple pour montrer une situation type que le réseau ne peut pas gérer dynamiquement. On effectue un apprentissage sur une fenêtre de 7 trames, on ne code donc qu'une position du phonème. L'alignement choisi est de 3 trames avant l'explosion de l'occlusive. les tests ont été menés sur des segments d'occlusives voisées qui ne contiennent plus qu'une trame avant l'explosion, les résultats montrent que le réseau est incapable de gérer une distorsion temporelle non apprise. Ce type de problème serait bien traité par la programmation dynamique ou un algorithme de Viterbi.

Taille de la fenêtre d'analyse : 7 trames		
Réseau	AUTO-COHERENCE alignement des segments avec 3 trames avant le début de l'explosion QPh.	AUTO-COHERENCE alignement des segments avec 1 trame avant le début de l'explosion QPh.
b d g	94.2	61.1

4. Résultats sur différentes classes

Nous avons entraîné plusieurs réseaux à classifier des phonèmes acoustiquement proches comme les occlusives sourdes p t k et le silence #, les liquides ... Ces classes de phonèmes recouvrent une totalité de 30 phonèmes (silence non compté). Le choix des paramètres choisis pour les tailles de contexte, d'analyse... sont le résultat des expériences menées précédemment. Ces choix correspondent pratiquement à ceux décrits par A. Waibel [1]. Les résultats sont cependant moins bons . Nous avons plusieurs explications simples à donner, le français est une langue acoustiquement plus ambigue que le japonais, la base de données utilisée n'est pas suffisamment grande et l'apprentissage n'est pas optimal. Des tests futurs sur une plus grande base de données apporteront plus de comparaisons possibles.

Réseaux	AUTO-COHERENCE QPh./segment		TEST QPh./segment	
p t k #	93.2	98.3	84.1	89.9
b d g	95.4	98.4	85.3	90.5
f s ʃ	100	100	99.4	99.2
v z ʒ	100	100	97.8	99.2
m n ɲ	92.9	94.2	83.9	86
l r	98.3	99.4	97.3	98.1
a e ɛ	95.2	96.3	89.8	92.5
ɑ̃ œ̃ ɔ̃	99.1	99.6	96.9	98.1
œ ə o ɔ	86.1	89.3	80.8	84.7
u y i	98.7	100	97.3	98.4

IV. ARCHITECTURE MODULAIRE DE TDNN

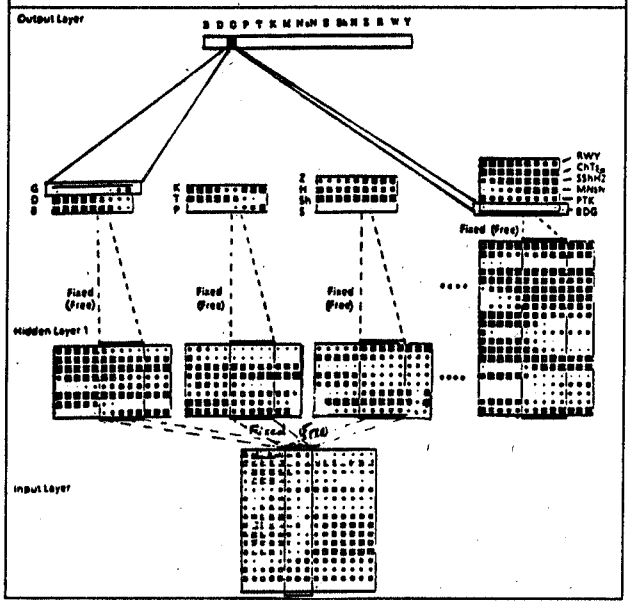
Nous avons présenté dans la partie III des résultats de classification phonétique avec des TDNN sur des classes de phonèmes acoustiquement proches. Une caractéristique intéressante de ces réseaux est l'élaboration de représentations internes dans les couches cachées. Ces informations peuvent être réutilisées pour construire une architecture globale [2] afin de résoudre une tâche plus difficile : la classification de 30 phonèmes. Les matrices de connexions de ces réseaux contiennent des poids permettant une séparation spatiale de phonèmes particuliers, il suffit d'apprendre à ces "sous-réseaux" spécialisés à rejeter tous les phonèmes qu'ils n'ont pas appris.

Ces "sous-réseaux" participent à un ré-apprentissage, ils apprennent des vecteurs de sorties désirées toujours à -1 pour les phonèmes non appris et renforcent les poids des connexions pour les phonèmes qu'ils connaissent. On ajoute à cette architecture un réseau entraîné à coder les caractéristiques phonétiques entre les différentes classes apprises par les sous-réseaux. Ce réseau de "macro-classes" permet de renforcer la discrimination des phonèmes; on se sert d'une information fine qui vient des sous-réseaux et d'une information plus grossière venant du réseau de "macro classes" qui renforce l'emergence d'un sous-réseau.

La figure 2 représente un schéma d'architecture modulaire permettant de classifier les consonnes, cette figure est issus d'un article d' A. Waibel. L'architecture que nous avons construit est du même type mais permet de classifier 30 phonèmes. Elle contient 11 sous-réseaux (en comptant le réseau de Macro-classes) Les connexions entre les couches sont "libres" ou "fixées"; seules les connexions entre la couche cachée et

la dernière couche participent au premier ré-apprentissage, une opération d'"affinement" de tous les poids du réseau est effectuée à la fin de ce premier ré-apprentissage, les connexions sont alors toutes libres et participent à un nouvel apprentissage. Les matrices de connexions entre la couche cachée et la dernière couche sont pondérées d'un coefficient Z qui permet de "désaturer" l'activité transmise et de relancer l'apprentissage.

Architecture modulaire de TDNN pour classifier les consonnes d'après A.Waibel(Fig.2) [2]



Pour résoudre une tâche complexe telle que la séparation de 30 phonèmes, cette architecture modulaire permet une souplesse d'utilisation intéressante et scinde le problème en sous-problèmes beaucoup plus simples à résoudre où le minimum global (ou un minimum satisfaisant) est plus facile à trouver.

Au même titre que certaines méthodes telles que le recuit simulé ou d'autres méthodes de rétro-propagation qui permettent d'éviter d'être pris dans un minimum local au cours de l' apprentissage, l'utilisation d'une architecture modulaire de réseaux de neurones est un moyen de guider l'apprentissage en utilisant nos connaissances des données à traiter, par exemple le fait que certains groupes de phonèmes sont plus difficiles à séparer.

Nous présentons dans le tableau suivant des résultats obtenus avec un "ré-apprentissage" de 600 cycles sur notre corpus de donnée. Cette étude doit être poursuivie dans le futur avec plus de cycles d'apprentissage. Nous envisageons aussi d'utiliser d'autres séparations de classes phonétiques.

Architecture modulaire de TDNN pour classifier 30 phonèmes du français				
Architecture globale	**AUTO QPh./segment**		**TEST QPh./segment**	
sans ré-apprentissage	70.2	77.3	68.9	76.5
avec ré-apprentissage	86.9	92	81.4	86.7

V. CONCLUSION ET PERSPECTIVES

Les modèles neuronaux de type TDNN tenant compte de l'aspect dynamique de la parole sont plus robustes aux phénomènes de coarticulation et de distorsion temporelle que les Perceptrons Multi-Couches classiques. Cependant cette "variabilité" n'est pas gérée dynamiquement, elle doit être présente dans les données d'apprentissage.

La construction d'une architecture modulaire à partir de TDNN déjà entraînés est l'un des côtés attractifs de ce modèle; elle scinde le problème en sous-problèmes plus simples à résoudre et permet de guider l'apprentissage en utilisant des connaissances sur les données à traiter. Les premiers résultats obtenus sont encourageants.

Cependant, les TDNN n'assurent pas une invariance parfaite aux distorsions temporelles auquelles les réseaux devraient résister et ils n'apportent aucun mécanisme qui permettrait de combiner les entités "quasi-phonèmes" en entités de plus haut niveau.

Pour améliorer ce système, on peut envisager d' appliquer un algorithme de programmation dynamique afin de prendre en compte les distorsions temporelles et de combiner les quasi-phonèmes ou phonèmes en entités de plus haut niveau. Il existe à l'heure actuelle de nombreux tests sur des systèmes hybrides HMM-NN[7] qui n'ont pas encore prouvé tout leur intérêt.

Nos perspectives sont d'appliquer une architecture modulaire de TDNN à la reconnaissance de la parole continue sur un corpus (*DARPA Ressource Management*) qui permette des comparaisons avec des modèles tels que les HMM.

REFERENCES BIBLIOGRAPHIQUES

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang: Phoneme recognition using Time Delay Neural Networks, Technical Report ATR Interpreting Telephony Research Laboratories, Oct. 1987.

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang: Modulatity and Scaling in Large Phonemic Neural Networks. Technical Report TR-1-0034, ATR interpreting Telephony Research Laboratories, Août 1988.

[3] P. Haffner, A. Waibel, H. Sawai, K. Shikano: Fast Back-Propagation Learning Methods for Neural Networks in Speech, ATR Interpreting Telephony Research Laboratories, 1988.

[4] P. Blanchet: Utilisation du perceptron multi-couche pour la compression d'information, Notes et documents LIMSI : 88-14, Sept. 1988

[5] L. Bottou: Reconnaissance de la parole par réseaux multi-couches, Neuro-Nimes, Nov. 1988.

[6] M. Adda-Decker: Evaluation d'unités de décision pour la reconnaissance de la parole continue, Thèse de doctorat de l'université Paris XI, Dec. 1988.

[7] H. Bourlard, N. Morgan: Merging Multilayer Perceptrons and hidden Markov Models: some experiments in continuous speech recognition, TR-89-033, July 1989.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Un compilateur d'ATN pour le traitement de la Parole

E. Reynier et J. Caelen

ICP, URA CNRS 368, INPG/ENSERG - Université Stendhal
46, Av. Félix Viallet, F-38031 Grenoble Cedex

ABSTRACT

This paper describes some aspects of the linguistic expert system in continuous speech recognition, which makes part of the DIRA (Integrated Dialogue and Automatic Recognition) project, being developed in our laboratory. This system consists in a rule compiler and an analyzer.

The compiler accepts context free or context sensitive rules and transform rules. It produces a transition network (ATN) which contains the nodes corresponding to the syntactic, lexical or phonetic categories and the transition arcs, whose traversal is conditioned by rules and predicates. These conditions arise both from semantic attributes for syntactic semantic analysis, and from phonologic rules describing the phonetic lexical network.

In fact, it is possible to produce one multi-layer network in which interactions between syntactic-semantic level and lexical level are strong.

INTRODUCTION

La parole est un signal temporel qui, après segmentation en unités acoustiques, se présente comme une suite d'éléments discrets. Cette suite d'éléments peut toujours --du moins formellement-- être considérée comme une séquence de symboles ordonnés respectant une certaine organisation syntaxique. Dès lors, la parole offre une double structuration: (a) hiérarchique --sur l'axe paradigmatique-- des unités acoustiques aux phrases en passant par les phonèmes, les syllabes, les mots, les syntagmes, etc. et (b) séquentielle --sur l'axe syntagmatique-- à chaque niveau de la hiérarchie: il existe une syntaxe de phrase formée de mots décrite par des règles de grammaire, comme il existe une syntaxe des mots formés de phonèmes décrite par les règles phonologiques. Cette propriété a été largement utilisée en reconnaissance de la parole [Lowerre, 76], [Jelinek, 75] pour généraliser la notion de réseaux à tous les étages de l'analyse, du décodage acoustico-phonétique à l'analyse sémantique. Cette idée a été également reprise dans le système DIRA (Dialogue oral Intégré et Reconnaissance Automatique) [Caelen, 88].

Dans ces conditions il est intéressant de disposer d'un compilateur de réseaux qui à partir de la description formelle de ces unités, quel que soit le niveau envisagé, et de la grammaire correspondante, puisse produire ces réseaux en vue d'une analyse ultérieure. On sait [Woods, 70] que les ATN (Augmented Transition Networks) --ou leurs équivalents les RP (Réseaux Procéduraux)-- sont des outils parmi les plus puissants pour résoudre les problèmes posés par la représentation et l'analyse des connaissances syntactico-sémantiques. Le système DIRA utilise déjà une technique d'ATN pour le niveau acoustique [Tattegrain et Caelen, 89]. Il est donc primordial de tenter d'uniformiser toutes ces approches "réseaux" avec un même formalisme pour donner une cohérence maximale au système de reconnaissance. De plus il faut prévoir une implantation qui permette un fonctionnement en prédiction, qui consiste à fournir une liste de symboles candidats après une séquence analysée correcte, en prenant en compte les informations disponibles à tous les niveaux de description afin de réduire la combinatoire du niveau acoustique.

Cet article propose donc de décrire les éléments d'entrée du compilateur, unités, grammaire, le compilateur lui-même et enfin les réseaux compilés (Fig. 1). Des exemples seront volontairement pris indifféremment dans les domaines de la phonologie, lexique, syntaxe ou sémantique afin de bien montrer la généralité de l'approche mais aussi le souci d'intégrer les particularités de chacun de ces domaines.

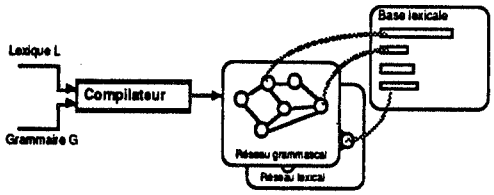


Fig. 1: Le compilateur de réseau ATN.

1. DESCRIPTION DES DONNEES D'ENTREE DU COMPILATEUR

Le compilateur reçoit en entrée la grammaire G et le lexique L de l'application. Il accepte les grammaires de type le plus général possible afin d'offrir un panorama aussi large que possible à l'expérimentateur. Ainsi il est prévu de pouvoir manipuler des grammaires contextuelles ou transformationnelles ainsi que des grammaires lexicales fonctionnelles. Une grammaire G est le 5-uplet:

G=(A, Vn, Vt, Va, R)

A: ensemble des axiomes,
Vn: vocabulaire non terminal,
Vt: vocabulaire terminal,
Va: vocabulaire auxiliaire,
R: ensemble des règles décrivant la grammaire G.

(par convention: V=Vn U Vt U Va, V*=monoïde libre engendré par V, et V+=V*-{ø})

Ces règles sont assorties de deux champs supplémentaires: un champ "contexte" et un champ "actions". Leur forme générale est décrite par le langage de description $\mu(G)$. Le lexique est écrit dans le langage de description $\mu(L)$.

1.1. Langage de description de la grammaire $\mu(G)$

Ce langage permet de décrire la grammaire de l'application dans un formalisme communément adopté par les linguistes et avec des notations proches de celles de Bachus-Naur. Le descriptif de la grammaire comporte deux parties: la déclaration des vocabulaires $\mu Dv(G)$, éventuellement vide, et l'ensemble des règles $\mu R(G)$.

1.1.1. La hiérarchie des vocabulaires

Il est possible de déclarer les éléments des vocabulaires de la grammaire G en une hiérarchie de classes. Ceci permet de simplifier

l'écriture des règles en se référant à tous les éléments d'une classe par un seul nom et de définir les différents niveaux de description de la grammaire. Cette hiérarchie doit tenir compte de la relation d'ordre entre les constituants tels que unité acoustique, phonème, mot, etc. Cette hiérarchie --appelée μDv(G)-- est une structure parenthésée définie comme suit:

{Nc_{11}: C_{11} {Nc_{21}: C_{21} { ... } ... {Nc_{1j}: C_{1j} ... } ... }}

Nc_{1j}: nom de la jème classe au niveau d'imbrication n°1

C_{1j}: liste des constituants de la classe Nc_{1j} définis en extension

Exemple 1:

{PHRASE : P_affirmative, P_interrogative
{SYNTAGME : SV, SN, SP
{CAT_LEXICAL : Verbe, Nom, Dét, Prép, Adj, Adv, ...
{PHONEME : #, a, i, u, p, t, k, f, s, ʃ, l, r, ...
{SEP_MORPHOLOGIQUE : |, +}}}}}

Avant toute analyse, il sera nécessaire de préciser la classe de niveau de description le plus élevé qui correspond à l'ensemble des axiomes et la classe de niveau le plus bas qui désigne le vocabulaire terminal de l'analyse courante. Ceci permet à l'utilisateur de demander des analyses partielles pour mettre au point une grammaire en considérant chaque niveau de description indépendamment des autres. L'analyseur choisit dynamiquement le niveau de profondeur de l'analyse en fonction de l'état pris par les deux variables de commande suivantes:

- $AXIOME: désigne la classe dont les éléments constituent l'ensemble A des axiomes ou par défaut l'élément de la partie gauche de la première règle non transformationnelle,
- $VOC_TERMINAL: désigne la classe dont les éléments constituent le vocabulaire Vt.

Dès lors, Va contient tous les symboles appartenant aux classes de niveau inférieur à celui de $VOC_TERMINAL et Vn = C(A U Vt U Va) où C désigne le "complémentaire".

1.1.2. Les variables (ou registres):

Les variables permettent de manipuler tous les objets de la grammaire et du lexique dans l'ATN [Woods, 70]. Leur type n'est pas imposé à priori, il est fixé en cours d'utilisation. Cette notion de variable est très large et peut englober tour à tour des concepts de valeur, de pointeur, de prédicat et même de procédure.

Une variable peut être locale à une règle ou globale à toute la grammaire. Elle est référencée par son préfixe, L$ pour les variables locales et R$ pour les variables globales, suivi de son nom. Il n'est pas nécessaire de déclarer les variables utilisées dans la grammaire car elles sont identifiables à leur préfixe.

1.1.3. Les règles

Chaque règle est écrite en respectant la syntaxe μR(G) suivante:

E: X op Y << a ^ b >> c / Ac;

E est l'identificateur (facultatif) de la règle,
X est le membre gauche de la règle tel que X ∈ Vn, ou X ∈ V+,
op={->, =>,==>} avec les conventions suivantes:
-> pour les règles non transformationnelles
=> pour les règles transformationnelles facultatives
==> pour les règles transformationnelles obligatoires
Y est le membre droit tel que Y ∈ V*,
a est le contexte gauche éventuellement vide, tel que a ∈V*,
b est le contexte dominant éventuellement vide, tel que b ∈ V*,
c est le contexte droit éventuellement vide, tel que c ∈ V*,
Ac est une liste d'actions, éventuellement vide qui contient autant de champs --séparés par une virgule-- qu'il y a d'éléments dans le membre droit plus un,
{: -> => ==> << ^ >> /, ;} sont des séparateurs attachés à chaque champ.

Ces règles permettent de décrire la syntaxe de l'application et les règles phonologiques pour construire le réseau lexical à partir des transcriptions phonétiques du lexique. Les règles transformationnelles sont celles de la théorie de Chomsky. Toutes les formes de récursivité sont admises sans restrictions. Nous reviendrons sur les parties contexte et action.

Remarques:

1- l'identificateur associé à une règle permet de se référencer à celle-ci plus facilement.
2- la notation X -> Y <<a >>b est équivalente à la notation aXb -> aYb, elle est cependant plus commode pour le compilateur parce que plus proche de la notation des règles hors-contexte X-> Y.
3- le compilateur accepte les règles dont la partie droite est de la forme de Backus-Naur (BNF).

1.2. Langage de description du lexique μ(L)

Le langage de description du lexique suit un modèle de schéma ("frame") c'est-à-dire que les éléments lexicaux sont des instances de classes hiérarchisées qui héritent des propriétés de leurs ascendants. L'ensemble des classes est doublement hiérarchisé: (a) selon une relation syntaxique et (b) selon une relation sémantique. Les instances du lexique sont rattachées à ces deux structures: ceci permet de séparer les deux entrées syntaxique et sémantique du lexique afin d'obtenir des méthodes d'accès plus claires et plus performantes selon le niveau d'analyse cherché. Ce modèle suit d'assez près le formalisme de Description Fonctionnelle [Kay, 79]. Toutes ces informations lexicales sont compilées pour être stockées directement dans l'ATN. En analyse, tout objet lexical auquel on accède est recopié afin de pouvoir être modifié dynamiquement sans modifier la base lexicale. Sur un élément du vocabulaire terminal de la grammaire, on accède aux instances dont la classe syntaxique mère est de même nom. On peut aussi accéder à une classe syntaxique associer à un élément du vocabulaire non terminal, pour construire une structure fonctionnelle de la phrase.

Une classe syntaxique ou sémantique se présente sous la forme:

{ <nom-de-classe> SORTE_DE <nom-classe-mère>;
<nom-d'attribut> : valeurs possibles
... }

et une instance lexicale sous la forme:

{ <nom-d'instance>
attributs généraux (transcription phonétique, orthographique ...)
*** *une classe d'héritage syntaxique obligatoire* ***
EST_UN <nom-classe-syntaxique>;
attributs hérités par la classe syntaxique ...
***et de 0 à n classes d'héritage sémantiques* ***
EST_UN <nom-classe-sémantique>;
attributs hérités par la classe sémantique ... }

Dans ce langage de frame, les attributs n'ont pas, à proprement parler, de facettes mais plutôt des valeurs typées qui ont implicitement des facettes. Les valeurs possibles d'un attributs, dans la définition des classes et des instances, sont de trois sortes:

- une liste d'identificateurs (scalaires) ou d'entiers, un intervalle,
- une remontée de valeurs d'attribut,
- une liste de masques d'unification.

Tout objet lexical hérite des attributs de sa classe mère. Une nouvelle spécification d'un domaine de valeurs d'un attribut hérité ne peut être qu'un sous-ensemble de celui de la classe mère. Par contre, la définition d'un nouvel attribut ne subit aucune restriction.

Remontée de valeurs d'attribut:

Cette opération permet d'assigner l'attribut d'un objet lexical à partir d'objets contenus par d'autres attributs.

{ SN SORTE_DE C$; (*racine de l'arbre syntaxique)*
dominant: < CNom, CPronom >
genre: A$genre(A$dominant)
... }

par cette écriture, après l'unification d'un Nom ou d'un Pronom avec l'attribut dominant d'un SN, l'attribut genre du SN sera affecté par le genre du Nom ou du Pronom.

Masque d'unification :

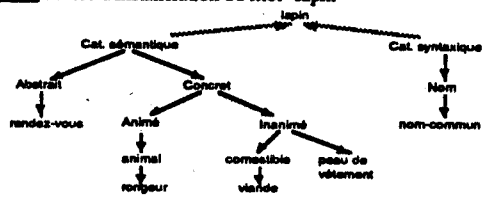
C'est une structure composée de classes et d'instances contraintes par des valeurs d'attributs. Elle inclue toutes les instances appartenant aux classes ou désignées explicitement et dont les valeurs d'attributs concordent avec les contraintes exprimées.

Si on veut définir le sujet du verbe "marcher" comme étant un groupe nominal dont le dominant est de la classe sémantique des objets animés et

dont le moyen de locomotion est les jambes, ou comme étant un pronom, on ecrira:

```
{ marcher
      PHON: "marʃE"
  EST_UN C$Verbe;
      sujet: <C$GN (A$dominant = <S$animés (A$locomotion =
            jambes)>)>, <C$Pronom> ... }
```

Exemple 2: l'arbre d'instanciation du mot "lapin"




```
LEXIQUE:
{ Lapin

      ORTH: "lapin"
      PHON: "lape~"
  EST_UN C$Ncommun;
      genre: masculin /M$fémin1
      nombre: singulier
  EST_UN S$rongeur;
      habitat: sauvage, domestique
      complément: <S$espèce, S$couleur>
  EST_UN S$viande;
      préparation: gril, four, sauce
  EST_UN S$rendez_vous;
      type: manqué
      attaché_à: <P$poser( A$COD = <C$SN( A$dominant = I$)>) >
      genre: masculin
      nombre: singulier }
.....
{ pose
      ORTH: "pose"
      PHON: "pozE"
      PROTO: P$poser
  EST_UN C$Verbe;
      COD:  <C$SN> }
.....
{ SN  SORTE_DE  C$;
      dominant: <C$Nom>
      quantifieur: <C$Dét>
      complément: <C$SP( A$prép=<C$Prép_compl_nom>,
                  A$dominant=A$complément( A$dominant( I$)))>
      genre: A$genre( A$dominant( I$))
      nombre: A$nombre( A$dominant( I$)) }
.....
{ rongeur SORTE_DE S$mammifère
  taille: <S$mesure (A$intervalle = 5 .. 50, A$unité = cm)>
  poids: <S$mesure (A$intervalle = 0.2 .. 5, A$unité = kg)>}
{ mammifère SORTE_DE S$animal
  etc...}
{ mesure SORTE_DE S$abstrait
  intervalle: ENTIER..ENTIER
  unité: CHAINE }
.....
REGLES MORPHO-PHONOLOGIQUES:
{ fémin1 SORTE_DE M$;     /* règle 1 de construction du féminin */
  genre: féminin
  ORTH:  -> e;
  PHON: e~ -> inE }
.....
```

Cet exemple montre les principaux aspects de la description lexicale. A chaque entrée est associée une représentation orthographique et une représentation phonétique. L'attribut PROTO permet de regrouper plusieurs instances (formes conjuguées de verbes par exemple) sous une même appellation, sans avoir pour autant à définir une classe spécifique. L'attachement procédural (ici des règles de dérivation morphologique et phonologique) permet de représenter les formes fléchies des entrées lexicales. Le cadre rectionnel des verbes peut être précisé comme dans les LFG-grammaires [Bresnan/Kaplan, 82].

2. FONCTIONNEMENT DES REGLES ET LIEN AVEC LE LEXIQUE

Les règles décrites formellement ci-dessus possèdent deux champs particuliers, "contextes" et "actions" dont le rôle est explicité ci-après.

2.1. Les contextes

Un contexte est composé d'éléments de V, de marqueurs grammaticaux et d'opérateurs binaires '.' (stricte) et '+' (large) ou unaire '¬' (négation), toutes ces composantes pouvant être associées par l'intermédiaire des opérateurs de liste {} et d'option [] comme dans la forme de Backus-Naur.

Il existe deux marqueurs grammaticaux, DEBUT_DE(X) et FIN_DE(X), avec $X \in Vn$ qui permettent de se référer à des positions syntagmatiques absolues dans les contextes. Les opérateurs binaires représentent respectivement la relation séquentielle stricte ou large entre des éléments de V, la relation large étant prise par défaut. Une relation stricte entre deux éléments de V signifie qu'il ne doit pas exister de marqueurs grammaticaux entre ces deux éléments, sinon il s'agit d'une relation large. L'opérateur négation peut être appliqué à un contexte tout entier pour interdire une règle dans ce contexte-ci. Il existe trois types de contextes: gauche, droit ou dominant. Le contexte dominant permet de conditionner une règle par les syntagmes qui la contienne. Les marqueurs grammaticaux n'ont alors aucun sens. Le contexte gauche (resp. droit et dominant) est conventionnellement précédé du signe << (resp. >> et ^).

Ce formalisme permet de décrire des contextes très précis en prenant en compte la structure profonde et la structure de surface.

Exemple 3: cas de la place du pronom personnel dans la phrase énonciative:

```
P_énonc -> [SP] SN Verbe [SN];
SN -> Pro_pers   ^ P_énonc >>SV;
```

"Je mange du lapin." est une phrase valide en ce qui concerne la place du pronom personnel.

Exemple 4: cas de l'élision du "e muet" en fin d'un mot suivi par un mot commençant par une voyelle ou une semi-voyelle:

```
C' ∂ =>  C'   >> FIN_DE(CAT_LEXICALE) V';
```

(∂: e muet, C': consonne+semi-voyelle, V': voyelle+semi-voyelle)

2.2. Les actions

La notion d'action est plus classique dans les ATN que celle de contexte vue ci-dessus. Ces actions sont déclenchées sur les transitions du réseau lorsque des prémisses spécifiées dans la règle et évaluées au moment du parcours de la transition sont vérifiées. Une prémisse est une ELBF (expression logique bien formée) qui prend les valeurs 'vrai' ou 'faux'. Les termes sont des prédicats P ou des prédicats évaluables c'est-à-dire des procédures Pe qui une fois exécutées retournent la valeur L(Pe)='vrai' ou 'faux'. Par souci de concision il est possible d'écrire assez simplement des structures de contrôle d'actions --inspirées du langage C-- par exemple de la manière suivante:

$$ELBF(P,Pe) = (Pe_1+(P_2.Pe_3)+P_4)$$

qui est équivalente à l'algorithme suivant:

si L(Pe1) alors 'vrai' sinon si P2 et L(P3) alors 'vrai' sinon si P4 alors 'vrai' sinon 'faux'.

Cette manière d'exprimer les tests et les procédures permet une concision intéressante de l'écriture des actions attachées aux transitions. Chaque transition peut-être gouvernée par une ELBF éventuellement vide, c'est-à-dire toujours vraie.

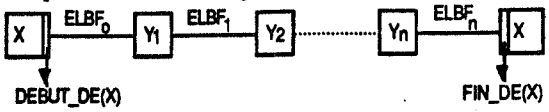
Ainsi la liste d'actions Ac s'écrit:

$$Ac = ELBF_0, ELBF_1, ELBF_2, ..., ELBF_n$$

pour une règle telle que:

$$X \rightarrow Y_1\ Y_2\ Y_3 ..... Y_n\ /\ ELBF_0, ELBF_1, ELBF_2, ..., ELBF_n;$$

qui est représentée dans le réseau grammatical par:



La liste d'actions des règles peut être utilisée pour contraindre la syntaxe de base, par exemple en comptant le nombre de réalisations d'une règle récursive, mais elle est avant tout faite pour construire la description fonctionnelle de la phrase analysée. Les arguments possibles d'un prédicat sont les registres, les scalaires, et les éléments des règles de grammaire qui correspondent en fait aux objets de même nom définis dans le lexique. Les données associées à ces arguments ont une structure parallèle pour permettre le stockage de plusieurs hypothèses en même temps.

Parmi les procédures et les prédicats prédéfinis et connus du compilateur on peut citer:

- les prédicats:
  TEQ (Xa, Xb): 'vrai' si la variable Xa est égale à Xb,
  MATCH(Att, X): 'vrai' si la variable X appartient au domaine défini par l'attribut Att, ...
- les procédures générales:
  ASG(Xa, Xb): assignation de la valeur de Xb à Xa, ...
- les procédures de gestion de listes:
  INTER(Xa, Xb): retourne l'intersection des listes contenues dans Xa et Xb: 'vrai' si non vide, ...
- les procédures arithmétiques:
  ADD(Xa, Xb): retourne la somme des valeurs de Xa et Xb
  INC(Xa): incrémentation de Xa ... ces procédures retournent la valeur 'vrai' après exécution.
- la procédure d'unification:
  UNIF(Att, X): 'vrai' si l'unification de l'attribut Att avec la variable X réussit.

Exemple 5:

GN -> Nom Adj / @,@,INTER(A$genre(D$Nom), A$genre(D$Adj))
. UNIF(A$qualificatif(D$Nom), D$Adj) . RET(D$Nom);

Cette règle s'applique lorsque le genre du nom et de l'adjectif sont compatibles. On effectue alors l'unification de l'attribut qualificatif de Nom avec Adj et on retourne au GN le Nom ainsi complété. On notera que la vérification du genre se fait sur la transition de sortie de la règle.

2.3. Contraintes sur les accès lexicaux

L'analyse syntactico-sémantique est toujours liée au lexique selon les deux perspectives suivantes:

- la vérification --où il s'agit de confirmer ou d'infirmer qu'une suite de mots est syntactiquement correcte-- exige un accès au lexique pour rechercher les attributs syntactico-sémantiques des mots à vérifier, ces mots étant connus par leur transcription phonétique,

- la prédiction (très importante pour le traitement de la parole) -- où il s'agit de fournir une liste de mots candidats possibles après une séquence correcte-- exige aussi un accès au lexique à partir des attributs syntactico-sémantiques prédits par l'analyseur syntaxique mais en examinant cette fois tous les chemins possibles dans l'ATN.

Dans les deux cas il est évident que la relation syntaxe-sémantique-lexique est très forte et doit être prévue au moment de la compilation de l'ATN afin d'accélérer le temps de la recherche (accès "statique" précompilé). Cette relation est prise en compte sous forme de contraintes d'accès portant sur les catégories syntaxiques, catégories sémantiques, attributs, etc. Il est évident que des "actions" pourraient résoudre le même problème, mais le temps d'exécution serait plus long puisque les accès seraient calculés à chaque fois. Ces contraintes d'accès sont indiquées directement dans les règles immédiatement après chaque terme du membre droit des règles qu'il soit du vocabulaire terminal ou non-terminal.

Contraintes d'accès sur le vocabulaire terminal Vt

Il existe 2 types de contraintes lexicales sur les éléments du vocabulaire terminal:

- soit du type restriction à une sous-classe syntactico-sémantique (masque d'unification).
  ex: Nom<C$propre & S$animé(A$type=personne)>,
- soit du type restriction par des valeurs d'attributs.
  ex: Nom(A$genre = masculin).

Ces contraintes permettent de limiter a priori la recherche lexicale lorsqu'on est en mode de prédiction et de restreindre le champ d'application d'une règle en mode vérification.

Exemple 6: Accès à des sous-classes fixes ou à des instances lexicales données

SP -> Prep<Ide, Ià> Nom<C$nom-propre & S$ville>;

restraint la règle aux cas où la préposition est 'de' ou 'à' et le nom est un nom propre de ville. A la compilation on crée un lien direct entre le premier terme droit "Prep" et les instances lexicales "de" et "à" de même pour le 2ème vis-à-vis des catégories syntaxique "nom-propre" et sémantique "ville". Cette règle aurait pu s'écrire aussi à l'aide d'actions de la manière suivante:

SP -> Prep Nom / @, MATCH (<Ide, Ià>, D$Prep),
MATCH(<C$nom-propre & S$ville>, D$Nom);

avec les problèmes de temps d'accès soulevés ci-dessus puisqu'il n'y a plus dans ce cas de compilation des liens.

Contraintes d'accès sur le vocabulaire non-terminal Vn

Ce cas est un peu plus général dans la mesure où il traite de métarègles en permettant de préciser derrière un terme des contraintes portant sur les règles elles-mêmes. On peut faire l'analogie avec le vocabulaire terminal en considérant qu'une règle est une instance de la classe représentée par l'élément de Vn en partie gauche, dont les attributs sont les parties droites. On retrouve donc 2 types de contraintes d'accès:

- soit en modifiant l'objet du lexique associé à ce non-terminal.
  ex: SP(A$domin = <C$nom-propre>),
- soit en précisant les règles applicables, leurs éléments en partie droite pouvant eux-mêmes être soumis à des contraintes.
  ex: SP<E$règle1, E$règle2(D$Prép=<I$de>, D$SN=<C$SN (A$dominant = <C$nom-propre>)>)>.

3. IMPLANTATION DU RESEAU

Le réseau grammatical est un graphe dans lequel les noeuds représentent les éléments du vocabulaire V, et les arcs les transitions qui portent les conditions et les actions relatives au parcours. Chaque élément non terminal x ∈ Vn de la grammaire est représenté par une paire de noeuds: DEBUT_DE(x) noeud d'entrée et FIN_DE(x) noeud de sortie. Chaque élément terminal y ∈ Vt n'est représenté que par un seul noeud. Après compilation un seul réseau est produit, il n'y a donc pas d'appel à des sous-réseaux: c'est grâce à cette implantation que les règles contextuelles et transformationnelles sont réalisables car tous les chemins possibles à travers le réseau sont explicites.

L'ATN est construit de façon à permettre à l'analyseur de le parcourir aussi bien de gauche à droite que de droite à gauche: pour cela il y a un double chaînage entre les noeuds. Toutefois pour exécuter les actions écrites pour un raisonnement de gauche à droite, l'analyse, en exploration de droite à gauche à partir d'un point d'ancrage est donc conçue en deux étapes: (a) recherche d'un ou plusieurs noeuds à gauche puis (b) analyse gauche-droite (Fig. 2).

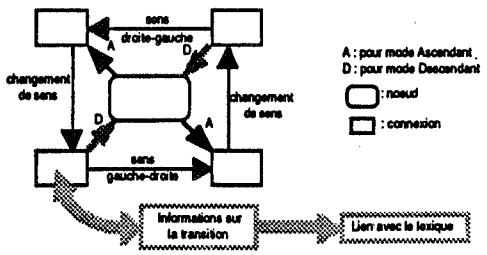


Fig. 2: cellule de base pour l'implantation d'un noeud.

Si un élément apparaît plusieurs fois dans la grammaire, il n'est représenté qu'une fois. Ainsi un noeud se trouve dans un environnement

où tous ses contextes de réalisation apparaissent. Cette factorisation permet à l'analyseur d'avoir un fonctionnement en mode ascendant performant.

4. PARCOURS DE L'ATN EN ANALYSE

Une analyse va consister à parcourir le réseau selon le type de fonctionnement fixé par le superviseur [Caelen, 88]. Cette analyse est effectuée par le contrôleur de réseau appelé plus simplement analyseur. Deux modes d'analyse sont prévus: le mode ascendant et le mode descendant. Pour chaque mode et à tout moment de l'analyse, deux fonctionnements sont possibles: (a) un fonctionnement en vérification et (b) un fonctionnement en prédiction. Pour vérifier une chaîne d'entrée, l'analyseur cherche un chemin dans le réseau à partir du noeud courant. En prédiction, l'analyseur propose tous les noeuds possibles, successeurs à la distance k, du noeud courant.

L'analyseur construit toutes les solutions syntactiquement et sémantiquement correctes en parallèle. Chaque solution est construite par les actions associées aux règles utilisées à partir des informations lexicales. En fin d'analyse, il fournit un arbre de solutions syntactiquement et sémantiquement correctes (la structure des constituants), ainsi qu'une liste de solutions fonctionnelles (la structure fonctionnelle au sens de Bresnan). Le nombre de règles ne nuit pas à la rapidité d'analyse mais le facteur de branchement de chaque noeud y est déterminant.

Exemple 7: L'analyse de la phrase "Le petit lapin de garenne ne l'a pas vu.", avec une grammaire et un lexique adéquates, donne la structure fonctionnelle suivante:

```
(PForme = Négative
  Mode = Indicatif
  Temps = passé_composé
  Aux        (I$a)
  Verbe (I$vu)
  Sujet (SN  dominant (Nom    (I$lapin)
                              qualificatif (Adj   (I$petit)))
             quantifieur (Dét (I$le))
             complément (SP   dominant (Nom   (I$garenne))
                              modulateur (Dét (I$de)))
             genre = masculin
             nombre = singulier)
  COD (SN    dominant (Pronom (I$l'))
             genre = masculin, féminin
             nombre = singulier))
```

5. APPORT EN TRAITEMENT DE LA PAROLE

-du point de vue de l'integration:

Le module linguistique ainsi construit peut s'insérer dans un système de reconnaissance de la parole entre le niveau bas comme le décodage acoustique ou la reconnaissance de mots et le niveau du dialogue. L'interface avec le niveau acoustique est assurée par un réseau lexical construit à partir de la transcription phonétique de chaque mot contenue dans le lexique et de règles phonologiques décrites avec le langage μ(G). Avec un reconnaisseur de mots, il est nécessaire de stoquer dans le lexique l'index de chaque mot dans la table du reconnaisseur. L'interface avec le module de dialogue se fait par la structure fonctionnelle construite pendant l'analyse et dont la forme est entièrement définie dans le lexique. La grande paramétrisation de ce système permet de le transporter facilement d'une application à l'autre.

-du point de vue de l'analyse:

Dans un fonctionnement en prédiction tous les niveaux de description sont mis à contribution; ceci permet de prédire un ensemble de séquences toutes exactes au regard des informations lexicales, syntaxiques et sémantiques disponibles pour guider les niveaux bas de reconnaissance. Grâce au langage de description du lexique, la sémantique d'une application peut être décrite de façon trés précise et ainsi être utilisée au mieux pour la prédiction.

-du point de vue des performances:

Tous les liens entre la grammaire et le lexique sont traités au moment de la compilation, pour accélérer les accès lexicaux. Le temps de parcours du réseau compilé est proportionnel au facteur de branchement moyen de la grammaire, ce qui fait de ce compilateur un outil tout à fait utilisable pour des applications réelles en reconnaissance de la parole.

5. CONCLUSION

La grammaire compilée offre la même puissance de description qu'un ATN mais la description de la structure syntaxique est à la fois plus simple et plus performante grâce aux règles de réécritures, contextuelles et transformationnelles. De plus le langage à objets de la base lexicale offre un formalisme puissant pour décrire des informations lexicales statiques et dynamiques (qui vont modifier l'analyse syntaxique).

La grande paramétrisation dans la description de la grammaire et du lexique permet de développer des grammaires dans divers formalismes. Cela permet au syntacticien une plus grande souplesse dans l'écriture de grammaires tournées vers des applications ciblées sans se restreindre à un formalisme issu d'une seule école de pensée. La variété de fonctionnements possibles de l'analyseur autorise un système superviseur à le commander pour différentes stratégies d'analyse, comme cela est le cas dans le système DIRA..

Ce compilateur est écrit en C, les structures objet sont en C++. Actuellement, toutes les fonctionnalités décrites ici, exceptées les règles transformationnelles et l'analyse ascendante, sont opérationnelles et sont déjà utilisées dans un système de reconnaissance de mots connectés.

6. REFERENCES

[Bresnan/Kaplan, 82] J. Bresnan and R.M. Kaplan
Introduction: Grammar as mental representations of language. *The mental representations of grammatical relations.* in Bresnan ed., Cambridge Mass. & London, MIT Press, 1982

[Caelen, 88] J. Caelen
Meta-stratégie en reconnaissance dans le système DIRA-RAP. 17ième J.E.P., Nancy, 1988, pp.173-179

[Chomsky, 80] N. Chomsky
Rules and representations. New York: Columbia University Press, 1980

[Fillmore, 71] C. Fillmore
Types of lexical information, in *Semantics: an interdisciplinary reader.* Cambridge University Press, Steinberg and Jakobovits, 1971, pp. 370-392

[Jelinek, 75] F. Jelinek and al.
Design of a linguistic Statistical Decoder for the recognition of Continuous Speech. I.E.E.E. Trans. of Inf. Theory, vol. II-21, n°3, 1975, pp. 250-256

[Kay, 79] M. Kay
Functional grammars, Actes 5th. annual meeting of the Berkeley linguistic society, 1979, pp. 142-158

[Lowerre, 76] B.T. Lowerre
The Harpy Speech Recognition System. Ph. D. Dissertation, C.M.U.-C.S.D., Carnegie Mellon University, 1976

[Tattegrain et Caelen, 89] H. Tattegrain et J. Caelen
Phonetic unit localization in a multi-expert recognition system, Actes d'EUROSpeech-89, Paris, 1989

[Winograd, 83] T. Winograd
Language as a cognitive process. Vol. 1: Syntax - Reading, Mass., etc., Addison-Wesley ed., 1983

[Woods, 70] W.A. Woods
Transition Network Grammars for Natural Language Analysis. C.A.C.M., Vol. 13, n°10, 1970, pp. 591-602.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

OPTIMISATION DE MODELES DE LANGAGE BASES SUR DES SCHEMAS

* FISET, Jean-Yves,** DESCOUT, Raymond,* ROBERT, Jean-Marc

*Ecole Polytechnique de Montréal, Département de Génie Industriel
C.P. 6079, Succursale A, Montréal, Québec, H3C 3A7
**Centre Canadien de Recherches sur l'Informatisation du Travail
1575, Boul. Chomedey, Laval, Québec, H7V 2X2

SOMMAIRE

L'utilisation d'un reconnaisseur de parole dans une application donnée requiert l'emploi d'un modèle convenable du langage utilisé. Dans certains cas, un modèle basé sur des schémas est adopté. Ce texte présente les résultats d'une expérience où un reconnaisseur de parole simulé envoyait un message partiellement entaché d'erreurs à un modèle de langage basé sur des schémas. L'habileté du modèle à classifier le message dans le bon schéma a été évaluée pour divers paramètres: le nombre de schémas dans le modèle, le recouvrement entre les schémas et la longueur des messages à classifier dans le schéma approprié. Les résultats obtenus indiquent que les facteurs les plus importants sont le recouvrement entre les schémas et la longueur des messages.

INTRODUCTION

La disponibilité de systèmes de reconnaissance de la parole de plus en plus performants a stimulé l'intérêt envers cette technologie dans les interfaces usagers-machines modernes. L'utilisation d'un reconnaisseur de parole implique l'utilisation d'un modèle du langage décrivant la syntaxe et le vocabulaire permis. Pour certains langages orientés vers la tâche, il semble que l'on puisse utiliser des schémas comme modèles de langage [1]. Un schéma est une structure de données représentant une situation stéréotypée [2], par exemple une catégorie donnée de messages. Un modèle de ce genre contient donc un certain nombre de schémas, représentant un nombre équivalent de catégories de messages. Ce genre de modèle a été utilisé notamment dans le cas du contrôle aérien [3]. L'exploitation d'un modèle de langage basé sur des schémas se fait en deux étapes:

. Un message reçu doit d'abord être classifié dans le schéma auquel il appartient.
. Le message est alors instantié par rapport au contenu du schéma retenu.

Cette recherche veut aider à préciser les conditions favorables à l'utilisation de modèles de langage basés sur des schémas. Pour y parvenir, on examine l'impact de 3 facteurs sur la capacité d'un tel modèle à classifier des messages. Le modèle utilisé ici ne correspond pas à une application en particulier: plutôt, il s'agit d'un outil permettant d'étudier le comportement d'une représentation par schémas. Les facteurs étudiés sont le nombre de schémas dans le modèle, le recouvrement (i.e. le nombre d'éléments communs aux divers schémas) et la longueur des messages à classifier.

METHODE

On a réalisé une simulation informatique d'un système comprenant un reconnaisseur de parole et un modèle de langage basé sur des schémas. La simulation a été écrite en XLISP, une version orientée-objet de LISP. Le système simulé comprend un reconnaisseur de parole qui reconnait imparfaitement des messages et les transmet au modèle de langage où ils sont classifiés. La figure 1 montre un schéma-bloc du système:

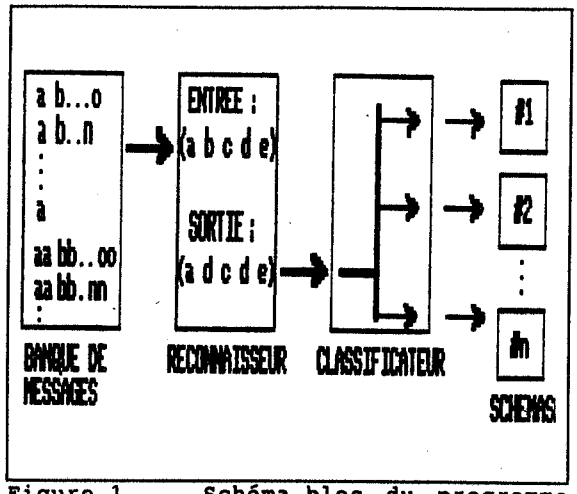


Figure 1 Schéma-bloc du programme de simulation

-Le module "banque de messages" contient une liste de messages qui seront envoyés au module "reconnaisseur". Pour chaque schéma, il y a 15 messages comptant chacun de 1 à 15 éléments. Par exemple, dans le cas d'un modèle de langage comptant 4 schémas, on aura 4 groupes de messages, soit au total 60 messages (i.e. 4 groupes X 15 messages/schéma). Dans le contexte, on considère qu'un élément est un mot.

- Le module "reconnaisseur" représente fonctionnellement un reconnaisseur de parole. Ce dispositif est susceptible de faire deux erreurs de base: substitution d'un élément du langage avec un autre et refus de reconnaitre un élément valide du langage (omission). Le module "reconnaisseur" reçoit un message du module "banque de messages", l'entache d'erreurs jusqu'à un certain point et l'envoie ensuite au module "classificateur". Le nombre d'erreurs introduites par le module "reconnaisseur de parole" dépend du taux de reconnaissance du reconnaisseur simulé. Par exemple, un reconnaisseur ayant un taux de reconnaissance de 80% introduirait en moyenne 2 erreurs sur un message de 10 éléments. Le processus d'introduction d'erreurs est simple: on échange le nombre voulu d'éléments d'un message avec d'autres éléments, choisis au hasard, parmi tous les éléments de la banque de messages. Conceptuellement, le nouvel élément pourrait être un mot du vocabulaire permis (confusion) ou un silence (omission). Dans cette simulation, un taux de reconnaissance de 80 % est utilisé, ce qui peut être raisonnable pour un reconnaisseur de parole opérant sans syntaxe sur un vocabulaire restreint.

- Le module "classificateur" reçoit un message partiellement entaché d'erreurs et tente de le classifier dans le schéma correspondant. La stratégie utilisée est la suivante: un message est réputé appartenir au schéma avec lequel il a le plus d'éléments en commun. Si un message a un même nombre d'éléments en commun avec plus d'un schéma, sa classification est indéterminée.

Le modèle du langage est constitué par l'ensemble des schémas qui représentent le langage. Dans cette simulation, chaque schéma contient quinze éléments. Dans le cas où le recouvrement est nul, les éléments sont exclusifs à chaque schéma. Si par contre il y a recouvrement, un certain nombre d'éléments est commun à tous les schémas. Par exemple, avec 20 % de recouvrement, il y aura 3 (i.e. 20 % * 15 éléments maxi.) éléments au maximum qui seront communs à tous les schémas (et à tous les messages). Lors d'une simulation, chaque message du module "banque de messages" est généré, entaché d'erreurs et classé 100 fois pour augmenter la fiabilité statistique des résultats. Dans le cas du modèle le plus simple (2 schémas), on a généré, entaché d'erreurs et classé 3000 messages (2 schémas * 15 messages/schéma * 100 fois/message). Pour le modèle le plus gros (10 schémas), on a traité 15000 messages. L'expérimentation complète a nécessité le traitement de 45000 messages.

RESULTATS

Les figures 2 à 5 montrent le pourcentage de messages qui ont été correctement classifiés lors de la simulation. Chaque figure correspond à un niveau de recouvrement maximum (ou nominal) distinct entre les différents schémas. Comme les messages comptent un nombre discrets d'éléments, le recouvrement que l'on retrouve dans chaque figure n'est pas constant. Par exemple, pour un message de 5 éléments, un recouvrement de 20 % signifie que 1 élément sur 5 a été substitué avec un autre. Par ailleurs, pour un message de 6 éléments et avec le même recouvrement nominal, le recouvrement n'est plus que de 16.7 % (i.e. 1 élément sur 6). Quand le recouvrement effectif diffère du recouvrement nominal, le recouvrement effectif selon la longueur du message apparait dans une table après la figure correspondante.

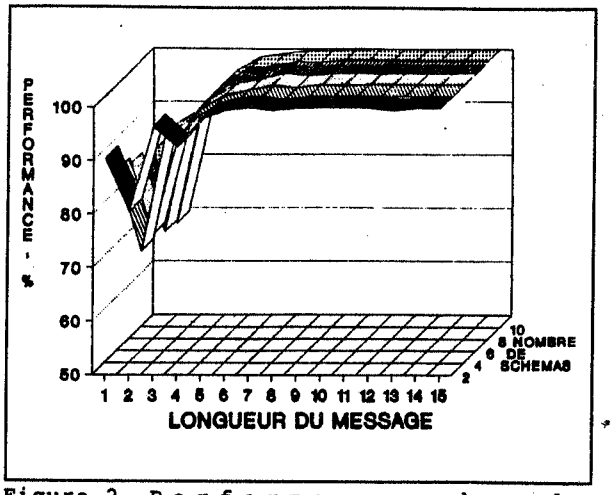


Figure 2 Performance de la classification avec un recouvrement nominal de 0 %

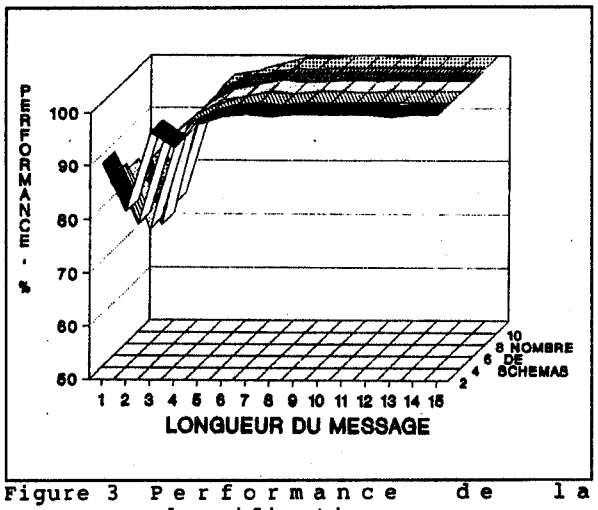


Figure 3 Performance de la classification avec un recouvrement nominal de 10 %

TABLE I Recouvrement effectif, avec un taux nominal de recouvrement de 10 %

	LONGUEUR DU MESSAGE														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# d'éléments communs	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
% recouvrement effectif	0	0	0	0	0	0	0	0	0	10	9	8	8	7	7

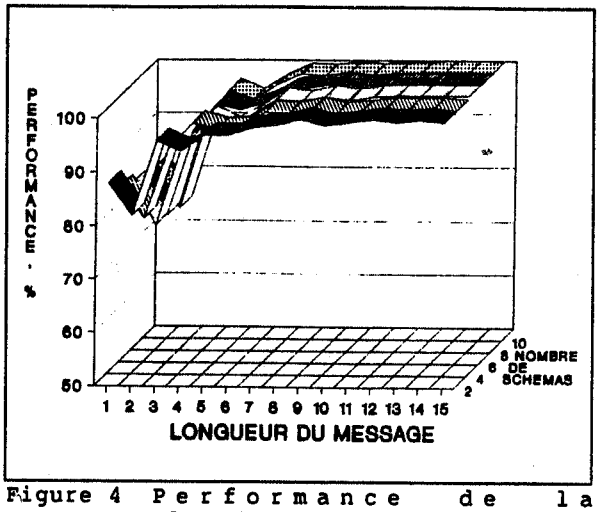


Figure 4 Performance de la classification avec un recouvrement nominal de 20 %

TABLE II Recouvrement effectif, avec un taux nominal de recouvrement de 20 %

	LONGUEUR DU MESSAGE														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# d'éléments communs	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3
% recouvrement effectif	0	0	0	0	20	17	14	13	11	20	18	17	15	14	20

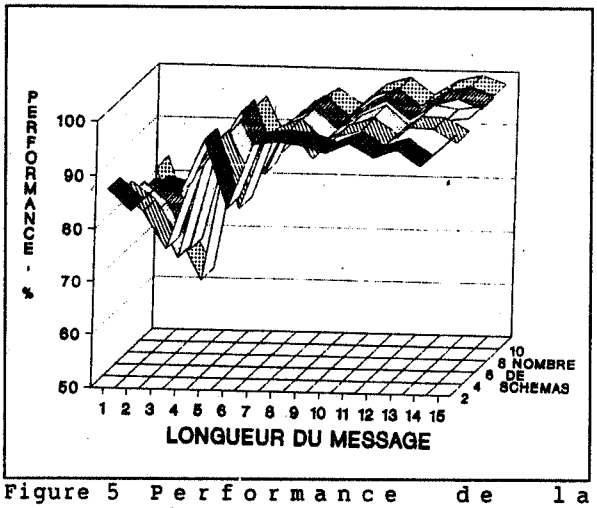


Figure 5 Performance de la classification avec un recouvrement nominal de 33 %

TABLE III Recouvrement effectif, avec un taux nominal de recouvrement de 33 %

	LONGUEUR DU MESSAGE														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# d'éléments communs	0	0	1	1	1	2	2	2	3	3	3	4	4	4	5
% recouvrement effectif	0	0	33	25	20	33	29	25	33	30	27	33	31	29	33

L'influence de chacune des variables indépendantes (nombre de schémas, recouvrement et longueur des messages) apparait dans les figures suivantes:

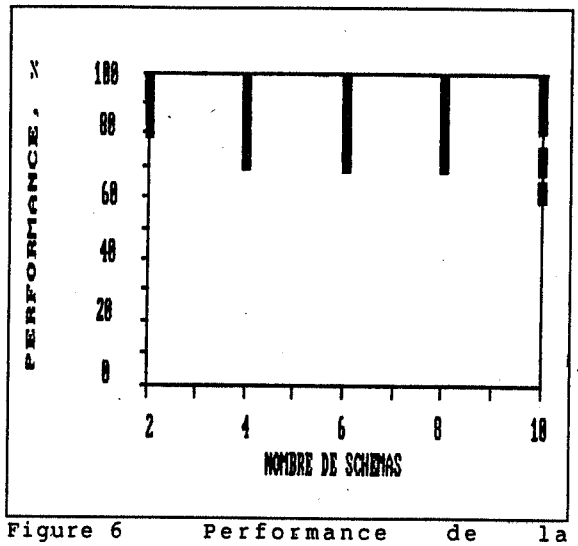


Figure 6 Performance de la classification par rapport au nombre de schémas

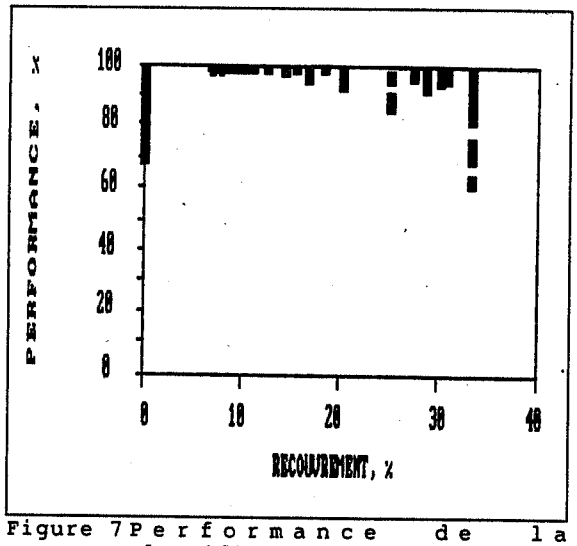


Figure 7 Performance de la classification par rapport au recouvrement

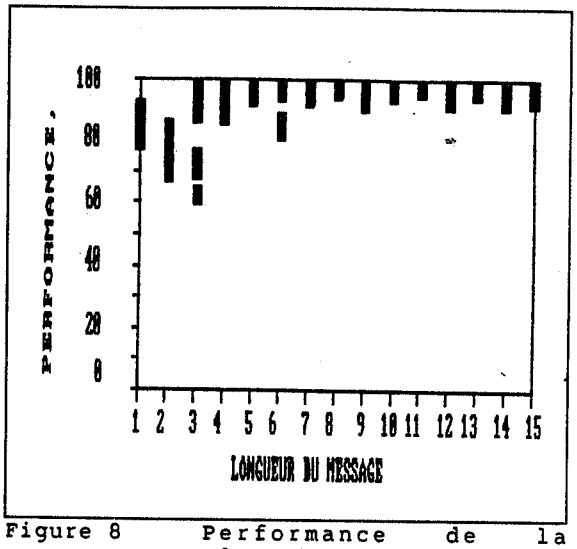


Figure 8 Performance de la classification par rapport à la longueur du message à classifier

DISCUSSION

Comme on s'y attendait, la meilleure performance en termes de classification est obtenue pour des messages longs lorsqu'il n'y a pas de recouvrement entre les schémas, comme le montrent les figures 2 à 5.

L'influence du nombre de schémas (figure 6) dans le modèle ne semble pas aussi déterminante qu'on aurait pu le penser. Dans les cas où il y a très peu (i.e. 2) ou beaucoup (i.e. 10) de schémas, la performance est la meilleure ou la pire. Pour un nombre de schémas compris entre ces extrêmes, la performance semble se situer sur un plateau.

Par contre, l'effet du recouvrement semble marqué (figure 7). Avec un recouvrement de moins de 10 %, la dégradation de la performance est modérée (si on excepte le cas où le recouvrement est de 0 %). Si le recouvrement excède 10 %, la dégradation augmente rapidement. La singularité que l'on observe à 0% de recouvrement est intriguante, mais on peut probablement l'expliquer si on tient compte de l'influence de la longueur des messages. En effet, il y a beaucoup de messages courts où le recouvrement effectif est 0 %, comparativement au nombre de messages longs où le recouvrement effectif est plus élevé. Or, comme le montre la figure 8, il semble que la longueur du message joue un rôle assez important, d'où peut-être la dispersion observée. Pour ce qui est de l'effet de la longueur des messages, on observe que la performance augmente de façon régulière jusqu'à ce que le message atteigne environ 7 éléments. Par la suite, il n'y a pas d'amélioration sensible de la performance.

Enfin, le fait que les messages de 1 élément soient mieux classifiés que les messages de 2 éléments appelle quelques commentaires. En effet, il est toujours possible de classifier un message de 1 élément (même à tort), alors qu'il peut ne pas être possible de classifier un message de 2 éléments (comment classifier le message [a aa]?).

CONCLUSION

Cette recherche a permis d'explorer par simulation la capacité d'un système composé d'un modèle de langage basé sur des schémas et d'un reconnaisseur de parole à classifier des messages. Les résultats montrent que dans le cas d'un système utlisant un reconnaisseur de parole ayant un taux de reconnaissance de 80 %, on obtiendra la meilleure performance en classification si on rencontre les conditions suivantes:

- La longueur du message est égale ou supérieure à 7 éléments.
- Le recouvrement entre les schémas est inférieur à 10 %.
- La classification s'opère entre 2 schémas. On aura une performance légèrement inférieure, mais uniforme, pour une classification entre 4, 6 ou 8 schémas.

Les résultats suggèrent que l'on peut prédire, dans des conditions données de nombre de schémas, de recouvrement, de longueur de message et de taux de reconnaissance, la performance en classification d'un modèle de langage basé sur des schémas. Par exemple, la figure 5 suggère qu'une classification entre 2 schémas, avec un recouvrement de 33 %, d'un message de 12 mots sera moins bonne que celle obtenue pour une classification entre 4 ou même 8 schémas avec un recouvrement de 20 %. On pourrait ainsi établir, à priori, si le langage à modéliser se prête à une représentation basée sur des schémas.

Cette recherche s'est limitée au processus de classification, sans considération pour le processus d'instanciation. Il est concevable qu'un mécanisme d'instanciation puissant corrige une partie d'un message entaché d'erreur, permettant éventuellement une reclassification du message. Il serait aussi intéressant d'examiner l'influence du taux de reconnaissance sur le processus de classification, pour les différentes valeurs des divers facteurs étudiés ici.

Enfin, bien que cette étude ait été menée dans une perspective de modèle de langage au niveau des mots, il est vraisemblable que les résultats soient applicables à un modèle utilisant des unités autres que des mots ou encore à des domaines connexes utilisant le formalisme des schémas.

References:

[1] Falzon, P., Ergonomie Cognitive du Dialogue, Presses Universitaires de Grenoble, 1989, 175 p.

[2] Minsky, M., A Framework for Representing Knowledge, Readings in Knowledge Representation, Morgan Kaufmann Publishers, Inc., 1985, Los Altos, California, p. 246-262

[3] Néel, F., Matrouf, K., Gauvain, J.L., Mariani, J., Reconnaissance Vocale et Applications en Aéronautique, Notes de Recherches, 57e Congrès de l'Académie Canadienne-Française pour l'Avancement des Sciences, Montréal, 17 mai 1987

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

MODELISATION D'ENONCES FINALISES DANS UN SYSTEME DE DIALOGUE ORAL HOMME-MACHINE

Guy DEVILLE

**Facultés Universitaires de Namur, B-5000 NAMUR, Belgique.
Centre de Recherche en Informatique de Nancy, BP 239 Campus Scientifique, F-54506 VANDOEUVRE-LES-NANCY.**

RESUME : Cet article présente un modèle de représentation d'énoncés dans un système de dialogue oral homme-machine.

Nous définissons d'abord le cadre expérimental de notre recherche par un bref exposé du système de dialogue oral homme-machine actuellement en cours d'expérimentation au Centre de Recherche en Informatique de Nancy (CRIN).

Nous exposons ensuite la méthodologie qui a inspiré la définition de notre modèle, axée sur la notion de sous-langage.

Le modèle linguistique que nous proposons adopte une approche fonctionnelle du language, notamment inspirée de la théorie de la grammaire des cas. Nous montrons que la réduction des structures prédicatives (verbes, adjectifs) de notre sous-langage, à un nombre fini de 'primitives prédicatives' offre l'avantage de rendre un tel modèle systématique, extensible à d'autres sous-langages d'applications du même type, et facilement utilisable dans une perspective d'automatisation.

Sur base de ce modèle grammatical, on peut concevoir que la représentation sémantique d'énoncés se construit parallèlement à une analyse syntaxique au moyen de règles de transformations spécifiques.

1. INTRODUCTION

Depuis plusieurs années, une des préoccupations centrales des recherches en linguistique informatique concerne la mise en oeuvre d'interfaces orales personnes-machines robustes, conviviales, qui tendent à utiliser des langages les plus proches possible de la langue naturelle. Le modèle linguistique présenté dans cet article, actuellement en cours d'implantation [CARBONNEL 84], [PIERREL 87], [ROUSSANALY 88], [MOUSEL 89] a été défini et mis au point dans un système de dialogue oral qui s'inscrit dans cette perspective [DEVILLE 89].

Ce travail est par essence le fruit d'une approche interdisciplinaire du langage naturel, car il est situé au carrefour des préoccupations tant d'ordre linguistique théorique qu'opérationnel. Dès lors, notre objectif consiste à élaborer une modélisation du langage naturel qui réponde à une triple validation :

- le modèle proposé doit être *linguistiquement* motivé, dans la mesure où il doit être articulé sur des modèles linguistiques éprouvés et satisfaire à une adéquation descriptive maximale sans être uniquement motivé par sa cohérence interne.
- il doit être *empiriquement* validé, c'est-à-dire capable de rendre compte d'énoncés réels produits dans un contexte de dialogue finalisé.
- il doit être enfin *opérationnellement* validé, au sens où il doit offrir la possibilité d'être mis en oeuvre en tant que tel dans le cadre d'un système de dialogue oral homme-machine.

Avant d'exposer en détail ce modèle linguistique, nous allons préciser brièvement le cadre expérimental de notre travail de recherche ainsi que la méthodologie qui a inspiré la définition de ce modèle, axée sur la notion de sous-langage.

2. CADRE EXPERIMENTAL ET METHODOLOGIQUE

2.1. le système DIAL

Le cadre expérimental de ce projet est le système de dialogue homme-machine DIAL actuellement en cours d'implantation au Centre de Recherche en Informatique de Nancy. Le but du projet DIAL est de concevoir et d'implanter un système de dialogue oral homme-machine capable de comprendre et de satisfaire des requêtes provenant d'un grand nombre d'usagers s'exprimant librement, et n'exigeant aucune forme d'apprentissage préalable. De telles spécifications exigent donc le recours au langage naturel (français parlé). Nous verrons cependant que ce *sous-langage*, lié à un domaine d'application spécifique, diffère du langage général par de nombreux aspects, ce qui n'est pas sans conséquences d'un point de vue opératoire. Un tel système devra gérer un dialogue oral *finalisé*, c'est-à-dire un processus convergeant dont l'objectif est la réalisation d'une tâche [MOUSEL e.a. 89]. La tâche retenue ici concerne la demande de renseignements administratifs contenus dans les pages roses de l'annuaire français.

Le système DIAL se présente comme un système multi-expert qui se caractérise tant par l'utilisation de multiples sources de connaissances (structures linguistiques du sous-langage, modèle du dialogue et de l'usager, domaine cognitif lié à l'application, ...) que par une forte interaction entre ces différentes sources de connaissances [MOUSEL e.a. 89]. L'architecture générale du système s'articule autour de cinq processeurs principaux (voir figure 1) :

- APHON, la *composante acoustico-phonétique* construit un treillis de phonèmes à partir du signal acoustique numérisé correspondant à l'énoncé du locuteur.
- PROSO, la *composante prosodique* détecte certains marqueurs prosodiques pertinents sur le signal, tels que frontières lexicales et syntagmatiques, marqueurs d'intonation indiquant la nature globale de l'énoncé (assertion, question, etc.).
- LEX, la *composante lexicale*, construit un treillis de mots à partir du treillis de phonèmes.

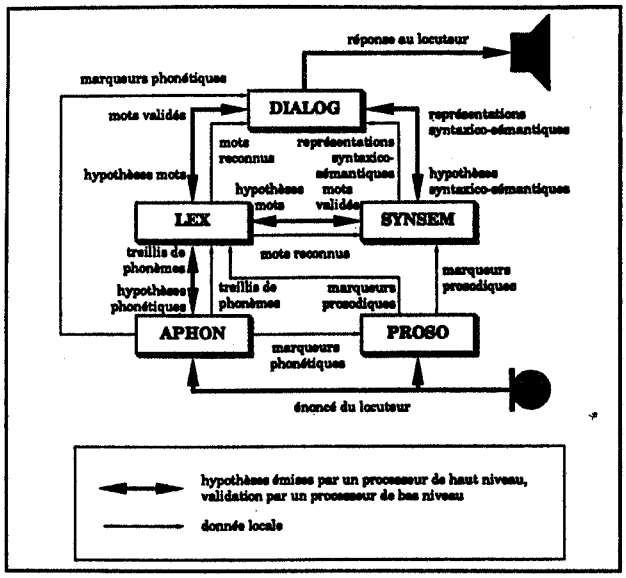


Figure 1. : Architecture du système DIAL

- SYNSEM, la *composante syntaxico-sémantique* produit une structure syntaxique et sémantique de l'énoncé du locuteur à partir du treillis lexical. C'est précisément à ce niveau que notre modèle linguistique est directement implémenté en tant que source de connaissance de SYNSEM.
- DIALOG, la *composante pragmatique*, fournit une réponse au locuteur à partir des structures syntaxico-sémantiques ou du treillis de mots.

Sans perdre de vue l'interaction entre les différentes composantes de DIAL, nous nous focaliserons sur la représentation syntaxico-sémantique des énoncés. Dans le cadre limité de cet article, nous n'exposerons pas le fonctionnement à proprement parler des analyseurs syntaxico-sémantiques du système. Cet approche fonctionnelle de SYNSEM est décrite dans [DEVILLE 89] et [MOUSEL 89].

A l'intérieur du cadre expérimental de DIAL, nous avons adopté dans notre modélisation du langage naturel une approche méthodologique fondée sur la notion de sous-langage.

2.2. La notion de sous-langage

La thèse que nous défendons est que l'étude du langage naturel dans une perspective d'automatisation est déterminée par la notion de 'sous-langage' [FALZON 86], [KITTREDGE & LEHRBERGER 82], [DEVILLE 89]. Nous appellons *sous-langage* un ensemble d'énoncés faisant référence à un domaine sémantique limité et explicitement défini, et utilisé pour une fonction spécifique. De tels énoncés sont engendrés par une grammaire et un lexique spécifique.

Notre approche consistera à montrer que la notion de sous-langage peut s'appliquer de manière pertinente au domaine de la demande de renseignements administratifs, ce qui nous pemettra de définir le langage lié à l'application du système DIAL sur base de critères opérationnels. Un tel langage a été défini empiriquement sur base de deux corpus de dialogues oraux finalisés :

- Une première expérimentation au CRIN a permis d'éliciter 40 dialogues oraux dans un contexte simulant un Centre de Renseignements Administratifs [ROUSSANALY e.a. 86].
- Dans le but de valider la portabilité de notre modèle linguistique vers des langages liés à d'autres applications du même type, nous avons mis au point une expérimentation en vraie grandeur au Ministère belge de l'Emploi et du Travail, où nous avons enregistré une centaine de demandes de renseignements administratifs sur diverses formalités relatives à l'emploi [DEVILLE 87].

Sur base de ces études empiriques, nous montrons que, à chaque niveau de description linguistique, le sous-langage de la demande de renseignements administratifs peut se définir structurellement selon trois modes complémentaires [DEVILLE 89] :

- le mode *restrictif* : en excluant certains traits du langage général, un sous-langage peut être défini comme une forme restreinte de langage. Du point de vue pragmatique, par exemple, l'interprétation d'énoncés finalisés est fonction d'un modèle stéréotypé, réduit de l'usager. Ce modèle stéréotypé n'est pas uniquement dû au domaine d'application spécifique, mais aussi aux croyances et intentions des 'participants' dans une situation communicative coopérative et se rapportant à une monde factuel et sincère. Notons que ce modèle stéréotypé de l'usager devra être intégré aux connaissances de la composante pragmatique d'un système de dialogue [ROUSSANALY 88].
- le mode *déviant* : un sous-langage peut exhiber certains traits spécifiques qui ne se retrouvent pas dans le langage général, et peut dès lors être défini comme une forme déviante de langage. Au niveau sémantique, par exemple, le vocabulaire du langage lié à notre application comporte des termes de jargons qui ne font pas partie du langage général. En d'autres termes, le lexique d'un sous-langage ne constitue pas simplement un sous-lexique du langage général. Par conséquent, si on adopte une approche incrémentale dans la définition du lexique d'un sous-langage, on constate que l'ensemble des mots liés au domaine d'application peut être défini de manière opératoire comme étant la couche périphérique d'une structure stratifiée (le noyau et la couche intermédiaire d'une telle structure sont respectivement constitués de mots 'grammaticaux' -articles, préposition, etc.- et de mots liés à la tâche). On notera qu'une telle couche périphérique doit être définie pour un domaine d'application donné [DEVILLE 89].
- le mode *préférentiel* : une telle approche du phénomène de sous-langage est complémentaire aux modes restrictifs et deviants et peut s'exprimer en termes de préférences statistiques : certains traits du langage général ont une faible probabilité d'occurence dans un sous-langage donné, sans cependant être totalement exclus de ce même sous-langage. Inversement, certains mots, structures et catégories syntaxiques se rencontrent plus fréquemment dans un sous-langage donné que dans le langage général. Le sous-langage de la demande de renseignements administratifs est caractérisé sur le plan syntaxique par une haute fréquence de questions directes et indirectes marquées par le syntagme *est-ce que/est-ce qui* ou uniquement marquées au niveau prosodique. Un mode de description préférentiel se traduira dans la conception de la grammaire du sous-langage : dans notre exemple, le modèle syntaxique donnera une prépondérance à la phrase interrogative en admettant dix-huit constructions différentes uniquement pour ce type de structure.

Après avoir exposé notre cadre expérimental et méthodologique, examinons la composante syntaxico-sémantique du système de dialogue du CRIN, qui intègre notre modèle linguistique en tant que source de connaissance.

3. UN MODELE LINGUISTIQUE POUR LA COMPOSANTE SYNSEM

3.1. Architecture de SYNSEM

Nous avons vu que la fonction de SYNSEM consiste à transformer un treillis phonétique en une double structure syntaxico-sémantique. (voir figure 6). La composante SYNSEM se présente comme un analyseur auquel sont associés trois types de sources de connaissances (figure 2) : (i) la syntaxe du language de l'application, formalisée par une grammaire de type ATN, (ii) un modèle sémantique de ce sous-langage sous la forme d'une grammaire de type fonctionnel et enfin (iii) des règles de correspondances qui crèent une structure sémantique de l'énoncé parallèlement à l'élaboration de la structure syntaxique correspondante.

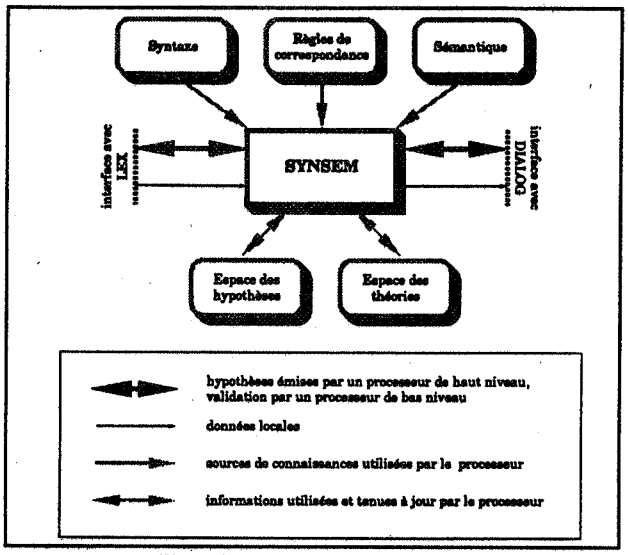


Figure 2. : Architecture de SYNSEM

3.2. Les connaissances syntaxiques de SYNSEM

Si nous distinguons nettement les connaissances syntaxiques et sémantiques, c'est plus dans un souci d'opérativité que de pertinence linguistique. Nous appelons *syntaxe* les contraintes positionnelles et *sémantique* les contraintes combinatoires valables pour un langage donnée [MOUSEL e. a. 89].

Le modèle utilisé pour la description du langage de notre application est une grammaire de type ATN, les Réseaux à Noeuds Procéduraux ou RNP [PIERREL 81]. A chaque noeud d'un tel réseau sont ajoutées des procédures qui gèrent la représentation syntaxico-sémantique de l'énoncé traité et intègrent des traitements particuliers permettant de rendre compte des caractéristiques liées à la parole continue. L'avantage des RNP par rapport aux ATN est la nette distinction entre représentation des connaissances (la grammaire) et leur traitement (l'analyseur) ce qui facilite incontestablement la construction d'un système évolutif.

3.3. Les connaissances sémantiques de SYNSEM

Les connaissances sémantiques de SYNSEM sont formalisées à l'aide d'un modèle fonctionnel du langage qui s'inspire partiellement de [DIK 78]. Examinons les composantes essentielles d'un tel modèle.

3.3.1. la notion de prédication

Nous appelons *prédication* la représentation sémantique sous-jacente d'une expression linguistique - ici, l'énoncé d'un sous-langage -. Une prédication est une structure formée d'un prédicat et d'un nombre adéquat de termes fonctionnant comme arguments de ce prédicat. Ces termes sont spécifiés au moyen d'une fonction sémantique appelée *cas* (figure 3). Les *termes* sont des expressions (noms, groupes prépositionnels, pronoms ou propositions) utilisés pour désigner les entités du domaine conceptuel du sous-langage de l'application. Un *prédicat* est une expression (c'est-à-dire verbes, noms ou adjectifs) qui code les propriétés et les relations d'ordre sémantique entre ses arguments.

Une prédication fait référence à une Configuration du monde liée à un sous-langage, ou Configuration. Ce terme étant défini ici comme un 'Etat de choses' (State of Affairs) dans un domaine conceptuel limité et explicitement défini.

Les prédicats sont dérivés d'un ensemble fini de *primitives prédicatives*, ou primitives. Une primitive code les propriétés combinatoires et sémantiques prototypiques que partage un ensemble de prédicats. En d'autres termes, alors qu'un prédicat et ses arguments font référence à une Configuration particulière, une primitive associée à un profil de cas fait référence à une classe prototypique de Configurations.

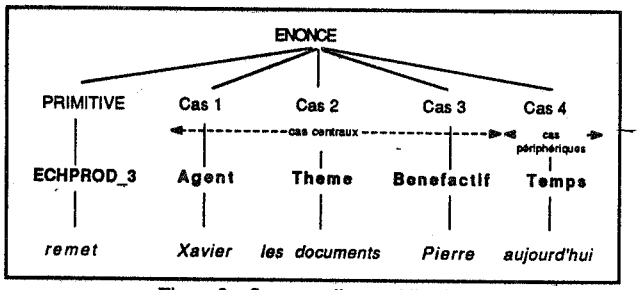


Figure 3. : Structure d'une prédication

Etant donnée l'importance de la notion de primitive prédicative dans notre modèle, nous examinerons plus en détails la typologie de primitives utilisée pour caractériser les prédicats de notre sous-langage.

3.3.2. une typologie de primitive

L'ensemble des primitives prédicatives s'articule autour d'un triple axe de dimensions conceptuelles appelées 'traits de primitives' : les traits *typologiques*, *sémantiques* et de *valence*.

Les traits typologiques de *dynamisme* et de *contrôle* son deux paramètres qui caractérisent trois classes de primitives prédicatives dénotant respectivement trois types de Configurations, les états, les actes et les procès :

1. Un **ETAT** est une configuration *non-dynamique* [-DYN], car il n'implique aucun changement de quelque nature que ce soit. Ce qui signifie que les entités en cause dans une Configuration non-dynamique préservent toutes leurs propriétés ainsi que leur localisation dans l'espace sur l'axe temporel le long duquel cette Configuration est supposée se dérouler. Les phrases suivantes sont des exemples d'états :

   Xavier est Mineur

   Thomas tient son passeport en main.

2. Un **ACTE** est configuration *dynamique et controlée* [+DYN/+CTRL]. C'est une configuration *dynamique* car elle implique nécessairement un changement. En effet, une telle Configuration peut être considérée comme une succession de différents états sur l'axe temporel le long duquel cette Configuration est supposée se dérouler. C'est également une configuration *controlée* car une entité (apellée ici le 'contrôleur') exerce un pouvoir de contrôle et d'initiative sur cette Configuration, et qui, dans une certaine mesure, determine si cette Configuration va avoir lieu ou non. Les phrases suivantes sont des exemples d'états :

   Wivine doit renouveler son passeport.

   Lydie a signé tous les documents.

3. A l'inverse, un **PROCES** est une Configuration *dynamique* mais *non-controlée* [+DYN/-CTRL], car aucune des entités en cause ne peut exercer de manière délibérée une influence sur cette Configuration. Les phrases suivantes sont des exemples de procès :

   Florent a perdu sa carte d'identité.

   Lucie est tombée de son vélo.

En complément des traits typologiques de primitives, nous proposons un ensemble de traits appelés traits *sémantiques* qui permettent de raffiner les trois grandes classes de primitives (actes, états et procès) sur base de critères sémantiques : le trait spatial, attributif et cognitif.

1. Les primitives marquées du trait *attributif* [+ATTR] expriment soit une relation de possession, de transfert ou de perte de possession entre les entités des classes de Configurations respectivement dénotées par ces primitives.

   La primitive **ECHPROD_3**, par exemple, exprime un acte *attributif*, c'est-à-dire un acte [+CTRL +DYN] d'un transfer contrôlé de possession d'une certaine entité entre deux autres entités d'une classe de Configurations dénotée par cette primitive. *Donner, vendre, acheter* et *acquérir* sont des exemples de prédicats dérivés de la primitive ECHPROD_3.

2. Les primitives marquées du trait *spatial* [+SPAT] expriment une relation spatiale entre les entités des classes de Configurations respectivement dénotées par ces primitives.

   La primitive **LOCATION_2**, par exemple, exprime un état *spatial* [-CTRL -DYN -SPAT], c'est-à-dire un ensemble de Configurations caractérisées par une relation non controlée et dynamique de localisation dans l'espace entre une entité et un point de référence spatiale exprimé par une autre entité. *Séjourner, habiter* et *se trouver* sont des exemples de prédicats dérivés de la primitive LOCATION_2.

3. Les primitives marquées du trait *cognitif* [+COGN] expriment une relation de nature physique perceptive, conceptuelle ou émotionelle entre entre les entités des classes de Configurations respectivement dénotées par ces primitives.

   La primitive **PR_COGNT_2**, par exemple, exprime un proces *cognitif* [-CTRL +DYN + COGNT], c'est-à-dire un ensemble de Configurations caractérisées par un comportement non controlé physique perceptif, mental ou émotionnel d'une entité ayant le trait sémantique [+HUMAIN] eu égard à une autre entité. *Voir, entendre, sentir* et *comprendre* sont des exemples de prédicats dérivés de la primitive PR_COGNT_2.

Enfin, à coté des traits *typologiques* et *sémantiques*, les traits *de valence* de primitives déterminent la valence syntaxique d'une classe de prédicats, c'est-à-dire le nombre d'arguments nécessaires à cette classe de prédicats pour préserver leur sens prototytpique tel qu'exprimé dans la définition de leur primitive correspondante.

Ainsi, une primitive *atransitive*, *transitive* ou *ditransitive* désigne un ensemble de Configurations qui impliquent respectivement une relation sémantique unaire, binaire ou ternaire.

3.3.3. Le système casuel

Nous avons vu qu'une prédication est définie structurellement comme un prédicat et un nombre de termes qui fonctionnent comme arguments de ce prédicat. La relation sémantique entre le prédicat et ses arguments est spécifiée à l'aide d'un *cas*.

Plus précisément, un cas est l'expression d'une fonction sémantique prototypique remplie par le terme d'un prédicat eu égard à la primitive dont ce prédicat est dérivé. Nous avons également vu qu'une typologie de primitive a été élaborée sur base d'un ensemble déterminé de traits de primitives. Ces traits ne caractérisent pas seulement chaque primitive selon les dimensions conceptuelles qui se révèlent pertinentes dans un sous-langage donné, mais ils sont également des éléments qui déterminent le nombre et la définition des rôles casuels affectés aux arguments d'un prédicat. Ainsi, par exemple,

- l'**agent** sera défini comme étant l'entité contrôlante d'un acte;
- le cas **bénéfactif** est l'entité non-contrôlante d'une configuration attributive, et qui se trouve dans un état de possession vis-à-vis d'une autre entité, ou qui est l'entité à qui une autre entité est transferrée;
- dans une Configuration spatiale, le cas **direction** désigne le point de référence dans l'espace vers lequel une autre entité se déplace.

La figure 4 donne quelques exemples de primitives du sous-langage de l'application. Ces primitives sont caractérisées par leurs traits typologiques, sémantiques et de valence, ainsi que par le profil de cas qui leur est associé.

EXEMPLES DE PRIMITIVES	TRAITS DE PRIMITIVES									PROFIL DE CAS	EXEMPLE DE PREDICATS
	typolog.		sémant.			valence					
	CTRL	DYN	ATTR	SPAT	COGN	TRS	DITRS				
ACT_2	+	+	-	-	-	+	-			Agent Theme	signer,...
ECHPROD_3	+	+	+	-	-	-	+			Agent Theme Benefact	donner,...
MVMT_2	+	+	-	+	-	+	-			Agent Src/Dir/ Chemin	aller, venir,...
STATUT_1	-	-	-	-	-	-	-			Theme	être mineur,...
PR_COGNT_2	-	+	-	-	+	+	-			Cogniteur Theme	oublier,...

Figure 4. : Exemples de primitives prédicatives avec leur profils de cas

3.4. les règles de correspondance

Avec les modèles syntaxique et sémantique du langage de l'application, une des sources de connaissances de SYNSEM est constituée par les règles de correspondance. Ces règles ont pour fonction de créer une structure sémantique de l'énoncé (en termes de primitive prédicative et de son profil de cas associé) parallèlement à l'élaboration de la structure syntaxique correspondante sur base tant des informations syntaxiques que sémantiques des constituants de cet énoncé. Comme l'illustre l'exemple de la figure 5, cette règle du type condition-action, affecte un certain cas à un groupe nominal d'un énoncé sur base de (i) sa fonction sytactique (sujet), (ii) sa catégorie syntaxique (nom, pronom), (iii) son trait sémantique (+ANIME) (iv) la voix du prédicat dont il dépend (active), et (v) la primitive dont le prédicat est dérivé (+CTRL/+DYN). La règle se lira donc comme suit : *le groupe nominal (nom ou pronom marqué sémantiquement comme ANIME) réalisé syntaxiquement comme sujet d'une phrase dont le prédicat est à la voix active et dérivé d'une primitive ayant les traits de contrôle et de dynamisme (c'est-à-dire un acte) se voit affecter le cas **agent**.*

```
si    un groupe nominal d'une catégorie syntaxique
      de type 'nom', 'pronom sujet' ou 'nom
      propre' réalise un 'groupe sujet'
 et   le trait sémantique de ce groupe est 'animé'
 et   il n'y a pas de marqueur de cas
 et   la voix du prédicat est active
 et   la primitive associée est de type ACT
alors on affecte le cas 'agent' à ce groupe nominal
fsi
```

Figure 5. : Exemple de règle de correspondance

3.5. Implantation de SYNSEM

Le modèle linguistique que nous avons présenté est actuellement en cours d'implantation à l'intérieur de la composante SYNSEM système de dialogue homme-machine DIAL, developpé au Centre de Recherche Informatique de Nancy [MOUSEL 89]. Cette composante se présente actuellement comme un prototype d'évaluation : SYNSEM est paramétrée par les sources de connaissance (ce qui signifie que les modèles linguistiques peuvent être adaptés à d'autres types de sous-langages indépendament des analyseurs) et par les stratégies d'analyse (la sélection de représentations syntaxico-sémantiques de l'énoncé à analyser peut s'opérer au moyen de différentes stratégies).

4. CONCLUSION

L'intérêt et l'originalité du modèle linguistique que nous avons exposé réside dans sa capacité à satisfaire un triple objectif :

- il est *linguistiquement* motivé. Nous avons montré que notre modélisation du langage est articulé sur des dimensions qui s'avèrent linguistiquement pertinentes.
- il est *empiriquement* validé. Nous avons également montré que la notion de sous-langage peut s'appliquer avec pertinence au domaine des requêtes orales d'informations administratives. Ce sous-langage a été défini empiriquement sur base de l'étude de deux corpus de dialogues finalisés. Une telle approche du langage naturel nous permet de valider empiriquement les modèles syntaxique et sémantique correspondants.
- il est *opératif*. Il répond entièrement aux spécifications d'un système de dialogue réel, et son implémentation est en voie d'achèvement.

Ce travail de recherche est le produit d'une approche interdiciplinaire du langage naturel. Notre tâche, en tant que linguiste, consiste à proposer des modèles linguistiques valides dans la perspective du traitement automatique du langage naturel. La tâche de l'informaticien est d'implanter ces modèles, c'est-à-dire de les rendre tout à fait opérationels. Les chercheurs de ces deux disciplines doivent tenir compte des spécificités de cet environement expérimental, les uns du point de vue linguistique et les autres en termes d'opérativité. Dans une telle entreprise, il y aura nécessairement un point où les deux approches du langage divergent.

Ce travail tente de réconcilier les buts divergents de la linguistique et de l'informatique. Nous sommes conscients du fait que cette réconciliation est partielle, mais elle a le mérite de faire un pas en avant, eu égard aux tentatives antérieures.

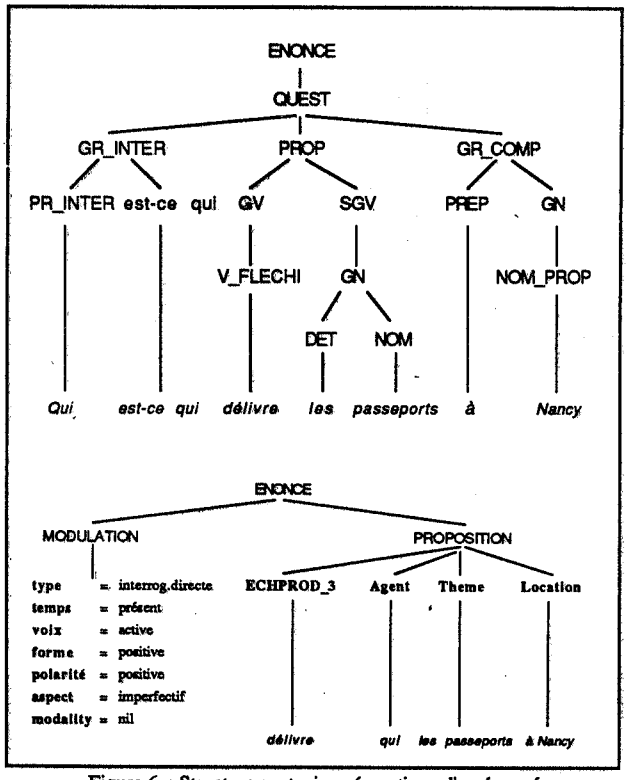


Figure 6. : Structure syntaxico-sémantique d'un énoncé

5. REFERENCES

[CARBONELL e.a. 84]
CARBONELL N., CHARPILLET F., HATON J.P., MANGEOL B., MOUSEL P., PIERREL J.M. & ROUSSANALY A. : Dialogue oral homme machine : bilan du projet MYRTILLE et perspectives. In PIERREL J.M., CARBONELL N., HATON J.P. & NEEL F. (eds.) : *Dialogue homme-machine à composante orale, Actes du Séminaire GRECO Communication Parlée - CNRS*, Nancy, pp. 91-122, Octobre 1984.

[DEVILLE 87]
DEVILLE G. (ed.) : *Corpus de dialogues oraux finalisés en situation réelle : Demande de renseignements auprès du Ministère de l'Emploi et du Travail (Cellule Action-Travail)*, FUNDP Namur, 1987.

[DEVILLE 89]
DEVILLE G.: *Modelization of task-oriented utterances in a man-machine dialogue system.* Thèse de Doctorat en Philosophie et Lettres (Philologie Germanique). Universitaire Instellingen Antwerpen, Antwerpen, 1989.

[DIK 78]
DIK S. : *Functional Grammar.* North-Holland, Amsterdam, 1978.

[FALZON 86]
FALZON P. : *Langages opératifs et compréhension opérative.* Thèse de Doctorat en Sciences Sociales, Université de Paris V - Sorbonne, Paris, 1986.

[KITTREDGE & LEHRBERGER 82]
KITTREDGE R. & LEHRBERGER J. (eds.) : *Sublanguage : studies of language in restricted semantic domains.* de Gruyter, Berlin, 1982.

[MOUSEL 89]
MOUSEL P. : *Syntaxe et sémantique dans un système de dialogue oral homme-machine finalisé en langage naturel.* Thèse de Doctorat en Informatique, Université de Nancy I, Nancy, 1989.

[MOUSEL e.a. 89]
MOUSEL P., PIERREL J.M. & ROUSSANALY A. : *Cooopération entre syntaxe, sémantique et pragmatique dans un système de dialogue oral homme-machine.* Actes du 7ème Congrès AFCET-RFIA, Paris, 1989.

[PIERREL 81]
PIERREL J.M. : *Etude et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu.* Thèse de Doctorat en Informatique, Université de Nancy I, Nancy, 1981.

[PIERREL 87]
PIERREL J.M. : *Dialogue oral homme-machine, Connaissances linguistiques, stratégies et architecture des systèmes.* Hermès, Paris, 1987.

[ROUSSANALY 88]
ROUSSANALY A. : *DIAL : la composante dialogue d'un système de communication orale homme-machine finalisée en langage naturel.* Thèse de Doctorat en Informatique, Université de Nancy I, Nancy, 1988.

[ROUSSANALY e.a. 86]
ROUSSANALY A., MOUSEL P., CARBONNEL N., MANGEOL B., PIERREL J.M.. : Réalisation d'un corpus de dialogues oraux. Application aux renseignements administratifs. *Centre de Recherche en Informatique de Nancy, Rapport Interne n°86-R-083, 1986.*

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Interprétation d'expressions complexes dans un système de dialogue homme-machine

J. Klein, J.M. Pierrel

CRIN/INRIA Lorraine - BP 239 - 54506 Vandœuvre-lès-Nancy

Résumé

La robotisation est un phénomène en pleine expansion, les robots actuels sont programmés pour effectuer des tâches précises et répétitives. L'adjonction d'un système de vision et d'un système de compréhension de commandes en langage naturel permettrait d'utiliser les robots dans des tâches très variées. Dans cet article, nous présentons les éléments de base nécessaires à la reconnaissance et à l'interprétation de commandes exprimées en langage naturel. Nous utilisons une grammaire de cas, à chaque cas est associé un ensemble de grammaires locales définissant les formes possibles que peut prendre le cas (ex : un cas objet peut être représenté par un groupe nominal...). En plus de ces grammaires locales, nous avons défini un ensemble de règles permettant de gérer les coordinations et qualifications de groupes de manière à reconnaître des groupes complexes (ex : le livre rouge et le cahier sur le bureau près de la fenêtre). Nous présentons aussi les mécanismes permettant d'interpréter les groupes complexes reconnus et les moyens de résoudre, en un minimum d'interventions, les ambiguïtés et les échecs d'interprétation.

1. Introduction

En 1984, sous l'initiative d'Olivier Faugeras, cinq laboratoires se sont associés pour définir un projet commun : le développement d'un système de vision pour un robot mobile se déplaçant dans un univers d'intérieur, le projet ORASIS [Mohr-88]. Les cinq laboratoires sont le CERFIA de Toulouse, le LIFIA de Grenoble, les équipes de vision de l'INRIA Rocquencourt et Sophia-Antipolis et le CRIN de Nancy [Thirion-89].
L'idée de départ qui a guidé notre travail, était de fournir à ce robot une interface de communication orale permettant de lui demander de réaliser quelques actions simples sur l'univers dans lequel il évolue.

Le locuteur et le robot ne connaissent l'univers que par l'analyse visuelle qu'ils en ont faite à partir de leur modèle théorique du monde. En aucun cas les objets de l'univers ne seront étiquetés d'une manière unique et commune aux deux interlocuteurs. Pour dialoguer, il leur faudra donc passer par une description des objets dont ils veulent parler. Cette description pourra prendre la forme d'une expression simple (ex : le livre) ou d'une expression plus complexe (ex : le petit livre rouge sur le bureau) et justifie l'utilisation du langage naturel pour ce type d'application qui, en dehors des problèmes de description d'objets, est relativement pauvre aussi bien syntaxiquement que sémantiquement.

L'identification des objets ne se fait donc pas par des noms individuels mais par des expressions, les problèmes qui en découlent, en plus des problèmes classiques de la reconnaissance de la parole continue, sont :

- la reconnaissance et la compréhension d'expressions complexes
- l'instanciation des objets décrits

2. Reconnaissance d'expressions complexes

2.1. Choix de la grammaire

L'utilisation d'une composante orale naturelle dans la commande de robot apporte au locuteur un confort et une rapidité de communication qui sont un atout majeur lors d'interventions urgentes (par exemple commande d'un robot dans une centrale nucléaire en cas d'incidents). Cependant, du côté du système de reconnaissance se posent un certain nombre de problèmes. On peut citer : la reconnaissance de la parole continue, l'analyse d'énoncés elliptiques et anaphoriques et enfin, le non respect de la syntaxe de la langue naturelle [Pierrel-87]. En effet, dans un cas d'urgence, les commandes fournies au robot peuvent subir des altérations syntaxiques imprévisibles tout en restant interprétables. L'analyseur doit être capable de retrouver le sens de tels énoncés, c'est pourquoi il doit disposer d'une grammaire souple autorisant des écarts par rapport aux formes grammaticales de référence. Par exemple, le robot doit pouvoir comprendre chacun des énoncés suivants :

- Ferme la porte
- La porte, ferme-la
- La porte... ferme!
- La porte ...!

Toutes ces raisons nous ont orienté vers le choix d'une grammaire de cas. En effet, dans un langage de type "commande" le verbe est l'élément important de la phrase, c'est lui qui délimite un environnement possible, et sert de pivot à la structure logique profonde de l'énoncé. La compréhension de la phrase est fondée sur la reconnaissance d'unités sémantiques : les cas [Fillmore-68] associés au verbe, eux-même décrits par des unités syntaxiques dont la forme générale est donnée à l'aide de grammaires.

Nous avons défini un ensemble de cas qui n'est pas exhaustif, mais que nous considérons comme suffisant pour notre type d'application. Les principaux cas associés aux prédicats sont :

- AGENT ex : **Pierre** prend un livre
- OBJET ex : Prends **le livre**
- DESTINATION ex : Pose le livre **sur la table**
- SOURCE ex : sors le livre **du tiroir**
- BENEFICIAIRE ex : Donne le livre **à Paul**

On peut compléter cette liste par des cas tels que : manière, temps etc. Contrairement aux applications de compréhension de textes généraux pour lesquelles l'ensemble de cas minimum ne peut pas être défini, pour une application de type commande ou utilisant un langage opératif [Deville-87], il est facile de déterminer l'ensemble des cas nécessaires.

Nous pouvons remarquer que tous les cas n'ont pas la même priorité dans une phrase. Par exemple, si un cas OBJET est absent d'une phrase, cette dernière risque d'être incompréhensible, par contre, si c'est un cas MANIERE qui manque, cela ne nuira en rien à l'interprétation. Nous pouvons donc distinguer trois types de cas en fonction de leur priorité :

- **les cas obligatoires :** ce sont les cas qui doivent figurer explicitement dans l'énoncé sous une forme quelconque, représentation complète ou anaphorique. Ex : le cas OBJET (Prends **le livre,** Pose-**le**)
- **les cas facultatifs :** ce sont les cas qui ne sont pas indispensables à l'interprétation et qui apportent des informations supplémentaires s'ils sont présents. Ex : le cas MANIERE (Pose **doucement** le vase sur la table ...)
- **les cas semi-obligatoires :** ce sont les cas qui ne sont pas forcément présents dans l'énoncé, mais dont l'instanciation est nécessaire à l'interprétation. Dans certaines circonstances, l'absence d'un de ces cas sera pallié par une valeur par défaut (par exemple le cas agent d'une phrase impérative), dans d'autres cas, il faudra faire une recherche dans l'historique du dialogue pour résoudre l'ellipse. Ex : le cas DESTINATION (Pose le livre **sur le bureau,** ...et le cahier)

Nous avons vu que la détection du verbe dirige la reconnaissance, c'est pour cela qu'il fait l'objet d'un traitement particulier de recherche dans l'énoncé. L'étape qui suit la recherche du verbe est celle de l'instanciation des cas. Pour ce faire, nous avons attaché à chaque cas un ensemble de grammaires susceptibles de le modéliser. Ces grammaires que nous appelons "grammaires locales" (par opposition aux grammaires générales) font l'objet du paragraphe suivant.

2.2. Les grammaires locales

Les grammaires locales que nous avons définies décrivent la structuration des éléments importants de l'énoncé mais ne fournissent pas une description syntaxique complète (cf les relatives dans la figure 1.). Leur classification est issue d'une réflexion sur les deux types de référence que l'on rencontre en langage naturel : les références à un concept (ex : ***le petit***, groupe faisant référence au concept livre rencontré plus tôt dans l'énoncé), et les références à une occurrence (ex : ***le*** , groupe faisant référence à une occurrence particulière d'un objet cité plus tôt). La différence entre ces deux types de référence apparaît clairement au niveau syntaxique, c'est pourquoi nous avons défini pour chaque composant syntaxique (GN, GP, GI...) des formes dérivées permettant de prendre en compte les deux types de référence. Par exemple, une forme dérivée d'un groupe nominal (GN) faisant référence à un concept s'appellera GNR, et celle faisant référence à une occurrence s'appellera GNREF. A partir de ces conventions, nous avons établi un ensemble de grammaires locales. La figure 1. nous donne la syntaxe des GN et GP ainsi que celle de leurs groupes dérivés.

GN	--->	art adja nom$^+$ adjp*	
		Ex :	la petite table noire
GNR	--->	art adj$^+$ adj* /	
		Ex :	la petite blanche
		prond$^+$ rel$^+$ GP /	
		Ex :	Celle qui est sur la table
		prond$^+$ rel$^+$ adj$^+$ adj*	
		Ex :	Celle qui est rouge
GNREF	--->	pronp$^+$ / art$^+$ adjn$^+$	
		Ex :	le, il / les deux
GP	--->	prep$^+$ GN$^+$	
		Ex :	sur la petite table rouge
GPR	--->	prep$^+$ GNR$^+$	
		Ex :	sur la petite
GPREF	--->	adv$^+$	
		Ex :	dessus

avec :

GN : groupe nominal		GP:groupe prépositionnel	
GNR : forme dérivée de GN (référence à un concept)		GPR: forme dérivée de GP (référence à un concept)	
GNREF : forme dérivée de GN (référence à une occurrence)		GPREF : dérivée de GP (référence à une occurrence)	
art	: article	rel	: pronom relatif
adj	: adjectif	adv	: adverbe
nom	: nom	+	: indique une présence obligatoire
prep	: préposition	*	: indique un élément pouvant être répété
pronp	: pronom personnel	adja	: adjectif se plaçant avant un nom
prond	: pronom démonstratif	adjp	: adjectif se plaçant après un nom
adjn	: adjectif numérique	/	: ou

Figure 1. Les grammaires locales

Nous avons rapidement présenté les cas utilisés dans notre grammaire ainsi que les formes que peuvent prendre les groupes les instanciant. Il nous reste à définir maintenant l'articulation entre tous ces éléments nous permettant d'obtenir des descriptions complètes d'objets.

2.3. Construction de groupes complexes

Dans une phrase, chaque cas n'est instancié qu'une seule fois, soit par un groupe éventuellement complété par d'autres groupes, soit par une conjonction de groupes (complexes ou non). La première contrainte que nous imposons est d'interdire les conjonctions de phrases, c'est-à-dire que nous n'acceptons que les énoncés ne comportant qu'une seule commande appliquée à un ou plusieurs objets.
C'est pourquoi nous autorisons des énoncés tels que : ***prends le livre sur le bureau et le cahier sur la table,*** car le prédicat PRENDRE nécessite la présence d'un cas objet qui est instancié par la conjonction des groupes décrivant le livre et le cahier.
Par contre, nous n'autorisons pas : ***pose le livre sur le bureau et le cahier sur la table*** , car le prédicat POSER nécessite la présence d'un cas objet et d'un cas destination qui, dans l'exemple sont instanciés chacun deux fois. L'énoncé ci-dessus est en fait une conjonction de deux phrases, la deuxième comportant une ellipse du verbe, ce qui ne respecte

pas la contrainte que nous nous sommes fixée et justifie donc le rejet de ce type d'énoncés.
En résumé, nous autorisons les qualifications et les coordinations de groupes lorsqu'elles sont internes à un cas. La coordination de groupes peut faire apparaître un phénomène de factorisation.

Ex : - Prends le classeur noir et le rouge
- Prends le petit et le grand classeur

Les deux paragraphes suivants présentent la coordination de groupes ainsi que les règles qui la régissent, et la qualification de groupes.

2.3.1. Coordination de groupes

Deux groupes ne peuvent être coordonnés que s'ils sont susceptibles de jouer le même rôle sémantique, donc d'instancier le même cas. Le cas le plus simple que l'on peut rencontrer est celui de la coordination de deux groupes de même type et de même rôle sémantique :

Ex : Le livre rouge et le petit cahier

Livre et cahier sont deux noms représentant des objets, les deux groupes sont des GN, il n'y a donc aucun problème pour les coordonner.

De plus l'étape de coordination permet de résoudre syntaxiquement certaines références :

Ex : Le petit livre rouge et le noir

Après la coordination de ces deux groupes, on obtient : *le petit livre rouge et le livre noir*, ceci nous permet de simplifier l'étape d'interprétation de l'énoncé. Cependant, toutes les coordinations ne peuvent pas se faire aveuglément, par exemple il est impossible de coordonner les deux groupes suivants :

*Ex : * Le livre et le noir*

Pour chaque couple de groupes susceptibles d'être coordonnés, nous avons défini un ensemble de règles que doivent respecter ces groupes, ainsi qu'un ensemble de transformations qu'ils subissent lors de la coordination. La figure 2. présente un sous-ensemble significatif de ces règles.

GN + GN :		
conditions	--->	même rôle sémantique
transformations	--->	-
GN + GNR :		
conditions	--->	le GN contient un qualificatif ou un groupe imbriqué les deux groupes ont le même genre
transformations	--->	transformer le GNR en groupe nominal en lui ajoutant le nom du GN
GP + GP :		
conditions	--->	même rôle sémantique
transformations	--->	si la préposition du 2ème GP est une forme contractée de celle du 1er GP alors remplacer la 2ème par la 1ère

Figure 2. Exemples de règles de coordination

2.3.2. Qualification de groupes

Lorsqu'un objet ne possède pas de caractéristiques permettant de l'identifier, il peut être nécessaire de compléter sa description par sa localisation dans l'univers.

Ex : Le bureau gris près de la fenêtre

Tout comme pour la coordination de groupes, la qualification respecte certaines règles syntaxiques :

- les GP ne peuvent qualifier que des GN, des GNR, des GP ou des GPR,
- les GPR ne peuvent qualifier que des GN, des GNR ou des GP,
- les GP ne peuvent qualifier que des groupes placés devant eux,
- un GP ne peut qualifier un groupe duquel il est séparé par un groupe d'un type autre que GP,
  ex : le livre bleu sur la table et le cahier sur le bureau
  ----> qualification incorrecte

  le livre bleu sur la table et le cahier sur le bureau
  ----> qualification correcte

sauf si ce groupe n'est pas qualifié et est coordonné avec celui faisant office de séparateur, dans ce cas, le GP qualifie simultanément les deux groupes, il est mis en facteur.
ex : le livre et le cahier sur le bureau
Dans ce cas particulier, deux solutions sont envisagées, le groupe est mis en facteur et il qualifie simultanément les deux GN, ou le groupe ne qualifie que le dernier GN.

2.3.3. Conclusion

Ces règles syntaxiques permettent de limiter les combinaisons entre les groupes, elles n'enlèvent rien à la liberté d'expression obtenue par l'utilisation d'une grammaire de cas, mais elles déterminent un ensemble de structures de surface possible. Grâce à ces règles, nous autorisons la description des objets par des groupes tels que :
- le livre et le cahier sur la table près de la fenêtre
- le livre, le classeur ou le cahier sur le bureau
- le livre sur la table près de la fenêtre et le classeur sur le bureau et près de la lampe

3. Instanciation des groupes

A partir de la description d'un objet telle que nous l'avons vue, il faut rechercher dans l'univers l'objet désigné. Plusieurs cas peuvent se produire :
- il y a un et un seul objet correspondant à la description,
- il n'y a pas d'objet,
- il y a plusieurs objets.

Dans le cas où l'on cherche un objet précis et qu'on trouve une occurrence, on est arrivé dans un état de satisfaction. Par contre, si aucune occurrence n'est trouvée, le système doit être capable de faire un diagnostic d'échec pour informer son interlocuteur. Dans le cas où il existe plusieurs occurrences, une stratégie doit être mise en œuvre pour pouvoir lever

l'ambiguïté le plus simplement possible. Le diagnostic d'échec et la résolution d'ambiguïtés font l'objet des deux paragraphes suivants.

3.1. Diagnostic d'échec

En cas d'échec d'instanciation, le locuteur doit être informé sur la cause de cet échec. Les exemples suivants illustrent les différents types d'explications qui peuvent lui être fournies :

- le livre rouge
---> il n'y a pas de livre rouge

- le livre rouge sur le bureau près de la fenêtre
--> il n'y a pas de fenêtre
ou --> il n'y a pas de bureau près de la fenêtre
ou --> il n'y a pas de livre rouge sur le bureau près de la fenêtre

Pour obtenir ces informations, l'analyseur construit l'arbre des occurrences d'objets en commençant par les niveaux les plus profonds. Pour l'exemple : ***le livre rouge sur le bureau près de la fenêtre***, il recherche dans un premier temps toutes les fenêtres, puis tous les bureaux se trouvant près de l'une de ces fenêtres et enfin tous les livres rouges se trouvant sur un des bureaux ainsi déterminés.
Le diagnostic de l'échec se fait à partir du premier ensemble d'occurrences vide.

Dans les cas d'échec, on peut aussi apporter une aide à l'utilisateur dans la mesure du possible :

Ex : le livre sur le bureau près de la fenêtre
---> il n'y a pas de livre sur le bureau près de la fenêtre, mais il y en a un sur le bureau noir.

Le problème qui se pose est l'explosion combinatoire, en effet, il n'est pas question d'indiquer tous les livres qui se trouvent dans l'univers du robot, mais seulement ceux qui respectent le "mieux" la description, à condition qu'ils ne soient pas en trop grand nombre. Pour pouvoir faire ce genre de propositions au locuteur, il faut faire une recherche ascendante, contrairement à l'analyse descendante utilisée pour l'instanciation, c'est-à-dire partir des feuilles possibles de l'arbre d'instanciation et remonter jusqu'à un état d'échec. Si juste avant l'échec, le nombre de solutions possibles est restreint (une ou deux solutions par exemple), une proposition est faite à l'utilisateur.

3.2. Résolution des ambiguïtés

La solution la plus naturelle consiste à choisir une occurrence parmi celles qui sont en conflit et de la soumettre au locuteur en attente d'une confirmation ou d'une infirmation.

Ex : - Prends le livre sur le bureau
- le livre rouge?
- non, le noir

Cependant tous les cas d'ambiguïté ne sont pas aussi simples. En effet, les objets en conflit n'ont pas forcément une caractéristique qui les différencie les uns des autres (comme la couleur dans l'exemple, il pourrait y avoir plusieurs livres rouges). De plus lorsque la description est complexe (groupes imbriqués et coordonnés), il est nécessaire d'adopter une stratégie pour poser une question qui soit la plus pertinente possible.

D'après les maximes de Grice [Grice-75], les informations fournies par le locuteur sont, par hypothèse, suffisantes pour déterminer l'objet, tant en qualité qu'en quantité ; donc s'il y a conflit, c'est suite à une erreur du locuteur. Si l'on reprend l'exemple : ***le livre sur le bureau près de la fenêtre***, et que l'on constate qu'il existe deux bureaux, dont l'un supporte deux livres et l'autre un seul, il paraît évident que le locuteur parle de celui qui est seul sur le bureau. En effet, si le locuteur porte son attention sur le bureau ne supportant qu'un seul livre, il n'éprouve pas le besoin d'apporter des informations supplémentaires. Le but que nous cherchons à atteindre est de sélectionner l'ensemble d'objets sur lequel le locuteur focalise son attention, de manière à ce que la question qui va lui être posée ait un maximum de chances de recevoir une confirmation (ce qui est plus facile à traiter qu'une contestation)

Lorsqu'un objet est décrit de manière complexe, il paraît évident que les objets du niveau le plus profond sont les plus facilement caractérisables, les autres ayant été décrits à partir d'eux!

Nous allons essayer de définir une stratégie permettant de sélectionner l'objet sur lequel le système doit poser une question en cas d'ambiguïté. Nous allons représenter les instanciations des objets grâce à un arbre. La figure 3, nous montre un exemple d'arbre d'instanciation du groupe : le livre sur le bureau près de la fenêtre. Une occurrence est désignée par l'initiale de l'objet suivie d'un numéro(par exemple F1 pour la fenêtre 1).

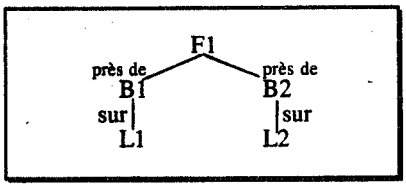
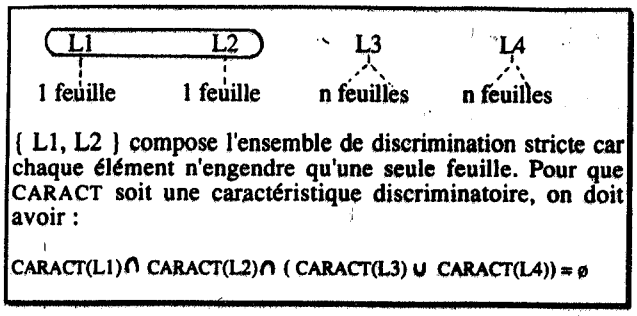


Figure 3. Arbre d'instanciation

La technique de sélection consiste dans un premier temps à sélectionner dans l'arbre d'instanciation, la (ou les) branche(s) optimisant la description, c'est-à-dire, celle(s) qui n'engendre(nt) qu'une seule feuille et qui soi(en)t la plus longue possible. Si les maximes de quantité sont respectées, l'objet désigné par le locuteur doit être une feuille figurant au bout d'une de ces branches. C'est pourquoi la question posée portera sur un objet d'une branche sélectionnée, ceci dans le but d'augmenter les chances de confirmation.

Lorsque les branches sont sélectionnées, il faut déterminer sur quel type d'objet va porter la question (i.e. à quel niveau de profondeur de l'arbre). Pour cela, on recherche à partir des feuilles, s'il existe une caractéristique permettant de faire la discrimination entre les différents objets du même niveau. A chaque niveau testé, nous ne nous intéressons qu'aux occurrences n'engendrant qu'une seule feuille ; donc, la discrimination doit porter sur chacune de ces occurrences d'une part, et sur l'ensemble des autres occurrences, d'autre part. Les occurrences n'engendrant qu'une seule feuille sont regroupées dans ce que nous appelons l'espace de discrimination stricte. La figure 4 illustre, en utilisant l'espace de discrimination, les contraintes que doivent vérifier les valeurs d'une caractéristique pour que celle-ci soit considérée comme discriminante.



{ L1, L2 } compose l'ensemble de discrimination stricte car chaque élément n'engendre qu'une seule feuille. Pour que CARACT soit une caractéristique discriminatoire, on doit avoir :

$$CARACT(L1) \cap CARACT(L2) \cap (CARACT(L3) \cup CARACT(L4)) = \emptyset$$

Figure 4. Contraintes de discrimination

La question posée portera sur une occurrence se trouvant à l'intersection d'une branche sélectionnée et d'un espace de discrimination stricte. La figure 5 illustre différents cas d'instanciation menant à une ambiguïté (à partir du groupe donné dans l'exemple), les branches sélectionnées sont entourées verticalement et les espaces de discrimination stricte, horizontalement ; à chaque niveau, on indique l'ensemble des objets sur lesquels on peut poser une question et les contraintes que doit vérifier la caractéristique retenue.

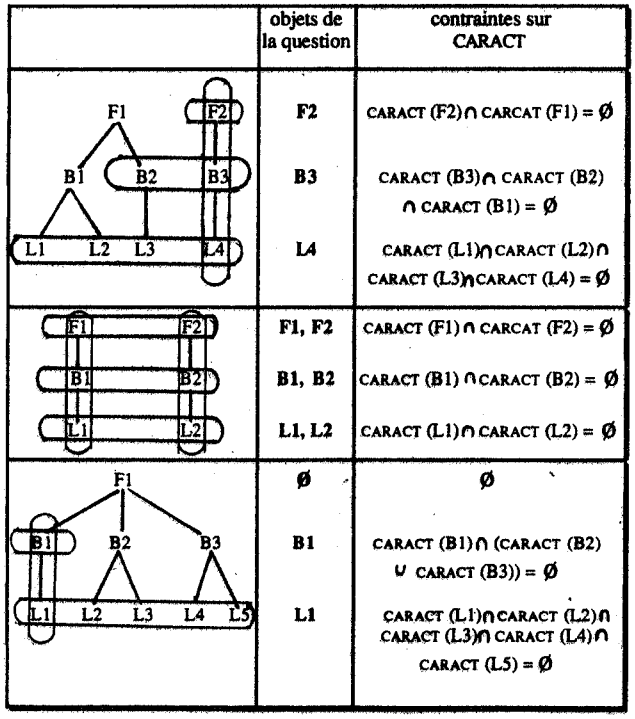


Figure 5. Exemples de cas d'instanciation

4. Implémentation et résultats

L'implémentation du système est faite sur SUN en LELISP 15.2. Dans l'état actuel, le robot est capable de reconnaître des expressions complexes telles que nous les avons décrites. La version actuelle n'est pas encore connectée au module de DAP. Les données du module de reconnaissance correspondent donc à une suite de phonèmes construite à partir du texte représentant la phrase dans lequel on insère quelques erreurs pour simuler les résultats d'un module de DAP (Ex : La bedite taple, le ptit lifre ...).

L'analyseur construit tous les énoncés correspondant aux différentes imbrications et coordinations de groupes. Il sélectionne les énoncés les mieux reconnus puis effectue la recherche des objets dans l'univers. Cette recherche fournit, en plus des listes d'occurrences trouvées, un score d'interprétation reflétant les cas d'ambiguïtés et d'échecs. Les énoncés sélectionnés correspondent à ceux qui ont le meilleur score d'interprétation. Les techniques de diagnostic d'échecs et de résolution d'ambiguïtés sont quant à elles en cours de développement.

5. Conclusion

Pour la résolution des ambiguïtés, nous avons fait l'hypothèse qu'une caractéristique discriminante peut être trouvée afin de poser une question. Il faudrait cependant envisager le cas où cela est impossible (ce qui est le cas lors d'une violation de la maxime de qualité), par exemple pour l'énoncé : le *livre rouge sur le bureau*, s'il existe deux livres identiques sur le bureau, il faut pousser la recherche plus loin en essayant de différencier les objets en conflit grâce à leur position par rapport à d'autres objets. En d'autres termes il faut donner au système les moyens de décrire les objets de manière équivalente à ce qu'il est capable de reconnaître.

Bibliographie

[Deville-87] : G. Deville, H. Paulussen, J.M. Pierrel. *Une grammaire de cas comme modèle de représentation sémantique des énoncés de dialogues oraux homme-machine finalisés.* 6ème congrès AFCET RF-IA, Antibes, Novembre 1987, p. 159-174.

[Fillmore-68] : C. Fillmore. *The case for case.* Universals in linguistics Theory. E. Bach and R.T. Harm eds Rinehart and Winston, 1968.

[Grice-75] : H.P. Grice Logic and conversation. Syntax and semantic 3 : Speech acts, Coles and Morgan, Academic press, New-York, pp 41-58. 1975.

[Mohr-88] : R. Mohr, *Le projet ORASIS.* Actes des premières journées nationales du GRECO-PRC Communication Homme-Machine. Parole, langage naturel et vision. Paris, 24-25 Novembre 1988.

[Pierrel-87] : J.M. Pierrel *Dialogue oral Homme-Machine : connaissances linguistiques, stratégies et architectures des systèmes.* Hermès, Paris. 1987.

[Thirion-89] : E. Thirion *Interprétation et apprentissage géométrique en vision par ordinateur.* Thèse de doctorat de l'institut national polytechnique de Lorraine. 1989.

8 PROSODIE

Président: F. NÉEL
LIMSI-Orsay,France

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

ORGANISATION DE L'ENONCE EN PHASES TEMPORELLES : ANALYSE D'UN CORPUS DE PHRASES REITEREES

V. PASDELOUP

Institut de Phonétique, UA CNRS 261 "Parole et Langage", Université de Provence,
13621 Aix-en-Provence Cedex, France.
Adresse permanente : 8, rue du Château-Landon 75010 Paris

Résumé. - Cette étude, qui est réalisée à partir d'un corpus de 96 phrases réitérées en syllabes ma-ma-ma, se propose de décrire l'évolution de la durée intersyllabique dans la totalité de la structure temporelle. En français, la phrase s'organise en phases de ralentissement de plus ou moins grande amplitude, chaque phase étant suivie par une réinitialisation de la durée syllabique. La phase temporelle est composée d'un petit nombre de syllabes, entre 2 et 5 généralement.

1. INTRODUCTION

L'importance dans la parole du timing des gestes articulatoires a été largement démontrée, tant sur le plan segmental, dans l'organisation temporelle des traits (phénomènes de coarticulation), que sur le plan suprasegmental, dans la structuration temporelle des unités prosodiques. De nombreux travaux soutiennent l'hypothèse que, dans la parole, l'organisation temporelle d'unités des plans segmental et suprasegmental nous informe sur les mécanismes de programmation et de plannification (Butterworth et Goldman-Eisler 1979, Fowler 1980). Ainsi pour Butterworth et Goldman-Eisler (1979), l'organisation macro-temporelle du discours reflète un rythme cognitif sous-jacent : "the temporal rhythm of speech reflects an underlying "cognitive rhythm" ; the hesitant-fluent cycle (...) reflects a planning execution cycle." (1979 : p. 212).

En prosodie, le rôle de la pause et de l'allongement final a suscité de nombreuses études (Goldman-Eisler 1972, Grosjean et Deschamps 1972, Duez 1987) qui mettent en évidence leur fonction démarcative d'unités linguistiques (la taille des pauses et des allongements finaux étant corrélée à l'importance de ces unités linguistiques). La configuration temporelle interne des unités ainsi démarquées ou d'unités de taille inférieure apparaît, tout du moins en français, avoir suscité moins d'intérêt ; certains auteurs s'attachent néanmoins à décrire en français la syllabe longue, située à la finale d'une structure temporelle ou prosodique, non pas comme une syllabe qui se "détache" du groupe de syllabes brèves qui la précèdent, mais comme la culmination d'un mouvement auquel participent tous les éléments de ce groupe : la durée des syllabes inaccentuées tend à progresser jusqu'à la syllabe accentuée dans un mouvement général de ralentissement progressif (Boudreault 1970, Caelen 1981, Duez 1987). L'évolution de la durée intersyllabique dans la totalité de la structure temporelle reste cependant encore peu étudiée en français.

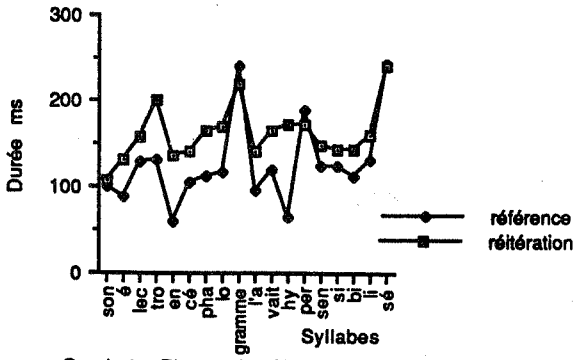
Trois raisons semblent être à l'origine de cette attitude :

- une raison théorique qui a eu des incidences méthodologiques : l'opposition phonologique entre syllabes longue/brève occulte partiellement, lors de l'analyse phonétique, les variations de durée intersyllabique dans le groupe des syllabes brèves. Cette division binaire (syllabes longue/brève), qui est adaptée à une description phonologique, est trop restrictive pour une description phonétique.

- une raison théorique : le français a longtemps été considéré comme une langue syllable-timed dont les syllabes brèves sont isochrones.

- une raison expérimentale : dans de nombreux cas, les variations intersyllabiques de durée sont ambiguës, car interprétables soit comme des différences de durée intrinsèque (microprosodique), soit comme des différences de durée extrinsèque qui relèvent du niveau prosodique.

La réitération de la parole en syllabes ma-ma-ma permet en partie de lever cette ambiguïté quant à la source des variations de durée intersyllabique puisque le principal avantage de cette méthode est la disparition des caractéristiques intrinsèques propres aux phonèmes (fréquence, intensité, durée et timbre) et la conservation de la seule information prosodique (Liberman et Streeter 1978, Nakatani 1978, Larkey 1983). (cf. graphe 1 : observer le cas des syllabes "lec" dans électro-encéphalogramme (syllabe intrinsèquement longue) et "hy" dans hypersensibilisé (syllabe intrinsèquement brève).



Graph 1 : Phrase de référence et sa réitération

Le rythme de la parole qui est un rythme de l'activité comporte certaines des caractéristiques des rythmes biologiques (Fraisse, 1974). La notion de *cycle*, qui revient à des intervalles de temps

plus ou moins égaux et qui est composé d'une série d'états successifs ou *phases*, est une caractéristique élémentaire des rythmes biologiques. Des études de psychomotricité réalisées sur la marche, la mastication ou la succion ont montré que ces activités sont périodiques et basées sur la phasage (Fraisse, 1974). Cependant, même dans l'hypothèse de l'existence d'un rythme biologique sous-jacent dans la parole, un phénomène cyclique ne peut se produire de façon aussi régulière dans l'organisation temporelle de la parole que dans celle de la marche ou de la mastication, en raison des contraintes qui sont imposées par l'ordre de structuration linguistique (entre autres, taille des unités lexicales et syntaxiques). Néanmoins, nous émettons l'hypothèse qu'une certaine régularité existe entre les structures temporelles et que les unités ou les groupements du rythme biologique sous-jacent, telle que la période ou la phase, pourraient correspondre à des unités ou à des groupements de nature linguistique (Pasdeloup, 1988).

2. PROCEDURE EXPERIMENTALE

Un corpus de 16 phrases à réitérer a été sélectionné parmi un corpus lu de 400 productions (40 phrases, 5 locuteurs, 2 répétitions). Ces 16 phrases de référence ont été choisies de telle sorte qu'un ou plusieurs accents secondaires soient réalisés (accent situé sur la première syllabe ou l'antépénultième d'un mot ou sur la dernière syllabe d'un morphème non terminal dans un mot polymorphémique ; exemple : hyper/sensibilisé).

Trois sujets féminins ont enregistré ce corpus en chambre sourde. Chaque phrase a été réitérée deux fois : 96 réitérations en ma-ma-ma : 3 sujets, 16 phrases, 2 répétitions . La réitération d'une phrase en ma-ma-ma étant un exercice difficile, chaque phrase de référence est d'abord répétée normalement avant d'être réitérée en ma-ma-ma. Le sujet effectue cette tâche en plusieurs étapes, sur la première puis sur la seconde moitié de la phrase (ce qui correspond en général au groupe sujet et au groupe verbal) avant de l'effectuer sur toute la phrase ; chacune des étapes est répétée deux fois de suite (pour plus de détails se reporter à Pasdeloup 1990).

3. RESULTATS

3.1. Description de la structure temporelle minimale

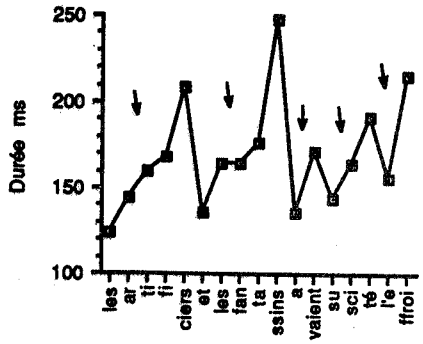
La réitération en syllabes ma-ma-ma accentue, voire rend manifeste, certains phénomènes de structuration temporelle. La phrase s'organise en une succession de phases temporelles (cf. graphe 2). On appelle phase temporelle la structure temporelle minimale récurrente. La phase temporelle se caractérise par un certain type d'organisation interne du débit. En français, la phase temporelle se caractérise :

- par un mouvement de ralentissement progressif du débit qui affecte, en général, toutes les syllabes de la phase temporelle.

- par l'allongement de la dernière syllabe de la phase temporelle qui est la syllabe la plus longue de la phase.

- par le phénomène de réinitialisation de la durée syllabique de la 1ère syllabe de la phase temporelle : après chaque phase temporelle, la durée de la 1ère syllabe de la phase suivante est ramenée à une valeur brève, proche de celle de la 1ère syllabe de la phase précédente et suivante. La 1ère syllabe de chaque phase est, en général, la syllabe la plus brève de la phase. La 1ère syllabe de la 1ère phase (1ère syllabe de la phrase) est souvent extra-brève par rapport aux autres 1ères syllabes de phase où s'effectue la réinitialisation.

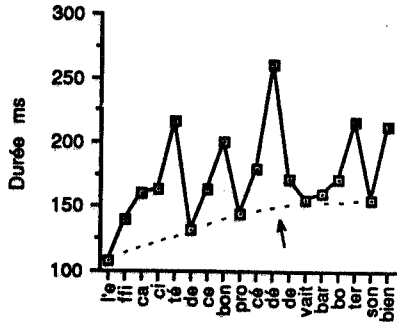
Dans cette perspective, la dernière syllabe de la phase temporelle n'apparaît pas comme une syllabe qui, sur le plan temporel, "se détache" du groupe de syllabes brèves qui la précèdent, mais comme la culmination d'un mouvement auquel participent tous les éléments de la phase.



Graph 2 : Phases temporelles

3.2. Réinitialisation de la durée syllabique

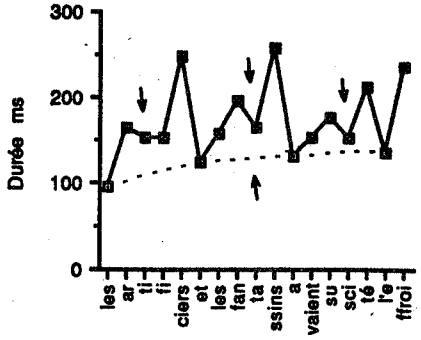
La configuration temporelle d'une phrase s'organise en une succession de ralentissements ou phases temporelles. Chaque ralentissement non final est suivi d'une accélération rapide ou réinitialisation de la durée syllabique qui n'affecte, en général, que la première syllabe de la phase ; la durée de la première syllabe de la phase est ramenée à une valeur brève proche de celle de la première syllabe de la phase précédente ou suivante[1]. Cette durée constitue une valeur étalon et toute modification de cette valeur affecte le macro-débit de l'énoncé. A l'exception de la 1ère syllabe de la phrase qui est souvent extra-brève par rapport aux autres 1ères syllabes de phase où s'effectue la réinitialisation, la durée de cette valeur étalon a tendance, suivant les cas et les sujets, à être quasi-stationnaire ou à évoluer dans des limites restreintes. Dans certains cas, la durée des syllabes où est réalisée la réinitialisation augmente très légèrement et progressivement du début à la fin de la phase (cf. graphe 3). Dans ce cas, la configuration temporelle de la phrase est constituée d'une série de phases de ralentissement, ces phases évoluant elles-mêmes au niveau du macro-débit de l'énoncé dans un mouvement général de ralentissement. Dans d'autres cas, la durée des syllabes où s'effectue la réinitialisation diminue dans la 1ère partie de la phrase (qui correspond généralement au groupe sujet) et augmente dans la seconde partie (qui correspond généralement au groupe verbal).



Graph 3 : Réinitialisations globales

Dans le corpus de phrases réitérées, et plus spécifiquement dans les phrases produites par le sujet 3, on remarque parfois un phénomène de double réinitialisation : entre deux réinitialisations globales se réalise une réinitialisation locale où

la durée de la 1ère syllabe de la phase est ramenée à une valeur brève, moins brève que celle à laquelle est ramenée la durée syllabique lors de la réinitialisation globale (cf. graphe 4). Une phase temporelle est donc divisée en deux sous-phases lorsqu'elle comprend une réinitialisation locale. Le nombre de cas de réinitialisations locales est très inférieur au nombre de cas de rénitialisations globales, mais la proportion varie selon les sujets. Pour les sujets 1, 2 et 3, on observe respectivement 9 réinitialisations locales pour 158 réinitialisations globales, 9 pour 166 et 25 pour 192.



Graph 4 : Réinitialisations locales et globales

Le tableau 1 rend compte des moyennes de durée des syllabes où s'effectuent les réinitialisations globales. Ces durées sont, en moyenne, beaucoup plus brèves lorsqu'il s'agit de la 1ère syllabe de la 1ère phase temporelle (1ère syllabe de phrase) que lorsqu'il s'agit de la 1ère syllabe des autres phases. Par ailleurs, la variabilité intra-sujets est relativement faible en ce qui concerne la durée des syllabes où s'effectue la réinitialisation globale, à l'exception de la durée de la 1ère réinitialisation de la phrase. La durée de la syllabe où s'effectue la réinitialisation globale est donc une valeur relativement stable si on la compare, par exemple, à la durée de la syllabe finale de phase (maximum temporel devant réinitialisation globale). La variabilité inter-sujets apparaît de même relativement faible.

Tableau 1 : Moyennes de la durée des syllabes où s'effectue une réinitialisation globale (moyenne et écart-type)

	Sujet 1	Sujet 2	Sujet 3	Total
réinitialisations globales à l'exception de la 1ère réi. des phrases	144 ms ±13	149 ms ±12	139 ms ±14	144 ms ±13
1ères réinitialisations des phrases	116 ms ±24	123 ms ±26	118 ms ±22	119 ms ±24

3.3. Durée de la dernière syllabe de la phase temporelle

L'allongement de la dernière syllabe de la phase temporelle est, avec le phénomène de réinitialisation de la durée syllabique de la 1ère syllabe de la phase, le deuxième indice acoustique important qui permet de déterminer les limites d'une phase temporelle[2]. Cet allongement n'apparaît pas comme un phénomène isolé mais comme la culmination d'un mouvement général de ralentissement progressif qui affecte toutes les syllabes de la phase temporelle.

Les résultats présentés dans le tableau 2 mettent en évidence une grande variabilité intra-locuteurs dans les valeurs des maxima de durée devant une réinitialisation globale. Cette variabilité semble due, entre autres, au fait que l'organisation hiérarchique des phases temporelles est réalisée essentiellement grâce à l'indice de ralentissement. Cette grande variabilité intra-locuteurs s'oppose à la faible variabilité intra-locuteurs concernant les durées des syllabes où s'effectuent les réinitialisations. Par ailleurs, les moyennes des maxima de durée devant une réinitialisation globale sont supérieures à celles des maxima de durée devant une réinitialisation locale ; cependant, cette comparaison est peu significative étant donné le petit nombre de cas de réinitialisations locales par sujet.

Tableau 2 : Moyennes de durée de la dernière syllabe des phases temporelles (moyenne et écart-type)

	Sujet 1	Sujet 2	Sujet 3	Total
maximum de durée devant une réinitialisation globale	226 ms ±64	219 ms ±79	226 ms ±96	224 ms ±82
maximum de durée devant une réinitialisation locale	163 ms ±8	180 ms ±23	166 ms ±12	168 ms ±15

3.4. Nombre de syllabes par phase temporelle

Les phases temporelles qui débutent par une réinitialisation globale sont composées d'un petit groupe de syllabes dont le nombre se situe majoritairement entre 2 et 5 : dans 90% des cas pour le sujet 1, dans 89% des cas pour le sujet 2, dans 92% des cas pour le sujet 3 et dans 90% des cas pour les trois sujets. Dans plus de la moitié des cas, soit 61% des cas, les phases temporelles sont composées de 2 ou 3 syllabes. Dans ce corpus de 96 phrases, 8 semble être le nombre maximal de syllabes admises dans la constitution d'une phase temporelle. Le nombre de syllabes par phase débutant par une réinitialisation globale est en moyenne de 3.4 (écart-type 1.4) (cf. tableau 3). La variabilité inter-sujets est faible.

Le nombre moyen de syllabes par sous-phase est, pour chaque sujet, inférieur au nombre moyen de syllabes par phase. La différence entre le nombre moyen de syllabes par sous-phase et le nombre moyen de syllabes par phase est environ de 0.6 syllabe pour les trois sujets.

On peut envisager l'hypothèse que cette relative constance dans la taille de la structure temporelle minimale, malgré les contraintes linguistiques telles que la taille variable des unités lexicales et syntactico-sémantiques, est le résultat de contraintes de nature biologique, psychologique et cognitive.

Tableau 3 : Moyennes du nombre de syllabes par phase temporelle et sous-phase (moyenne et écart-type)

	Sujet 1	Sujet 2	Sujet 3	Total
nombre de syllabes par phase débutant par une réinitialisation globale	3.6 ±1.5	3.4 ±1.5	3.3 ±1.3	3.4 ±1.4
nombre de syllabes par sous-phase	3.1 ±0.8	2.8 ±1	2.7 ±0.7	2.8 ±0.8

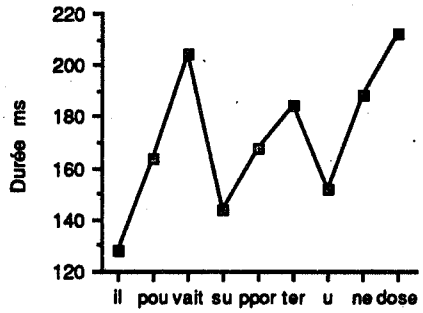
3.5. Organisation interne de la phase temporelle

L'organisation interne de la phase temporelle se caractérise en français par un ralentissement du débit (augmentation de la durée syllabique) qui se réalise sur toutes les syllabes de la phase. On distingue essentiellement deux types de ralentissement :

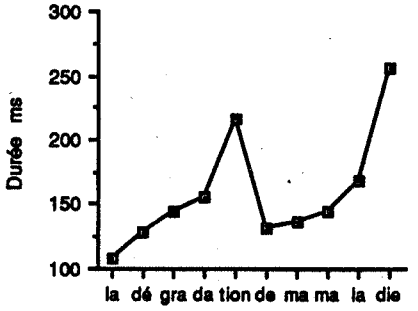
- un ralentissement du débit quasi-linéaire : la durée syllabique augmente régulièrement dans la phase temporelle (cf. graphe 5).

- un ralentissement du débit non-linéaire qui va en s'augmentant sur la ou les dernières syllabes de la phase temporelle : la durée syllabique augmente lentement sur les premières syllabes de la phase temporelle puis plus brutalement sur la ou les dernières syllabes (cf. graphe 6). Ce dernier type de ralentissement semble plus fréquent lorsque l'amplitude du ralentissement (écart de durée entre la première et la dernière syllabe d'une phase) est importante, mais ce n'est pas systématique.

Une accélération du débit (diminution de la durée syllabique) se produit parfois sur les premières syllabes d'une phase avant que le ralentissement s'impose; on remarque cet accélération après un allongement très marqué de la dernière syllabe de la phase précédente, mais aussi à la suite d'une phase affectée d'un faible ralentissement (cf. graphe 3).



Graph 5 : Phases de ralentissement linéaire du débit



Graph 6 : Phases de ralentissement non-linéaire

3.6. Organisation hiérarchique des phases temporelles

La hiérarchie entre les phases temporelles semble indiquée principalement par l'indice de ralentissement du débit et optionnellement par l'indice de réinitialisation. Les phases se hiérarchisent selon le principe de dépendance à droite : une unité de rang inférieur se regroupe avec l'unité de rang supérieur situé à sa droite. Chez deux des trois sujets, la stratégie employée le plus fréquemment pour hiérarchiser les phases consiste à faire varier l'ampleur du ralentissement et à utiliser un seul type de réinitialisation, la réinitialisation globale (cf. graphe 3). Le phénomène d'allongement de la dernière syllabe de la phase permet une hiérarchisation complexe entre les phases temporelles qui peut comprendre de nombreux degrés hiérarchiques. Le troisième sujet, utilise la stratégie précédente ainsi qu'une autre stratégie qui consiste à faire varier, à la fois, le degré d'allongement de la dernière syllabe de la phase et le type de réinitialisation - locale ou globale - : un ralentissement de faible amplitude sera suivi par une réinitialisation locale alors qu'un ralentissement de grande amplitude sera suivi par une réinitialisation globale (cf. graphe 4).

3.7. Configuration temporelle et organisation linguistique

La configuration temporelle de l'énoncé -la segmentation en phases temporelles et leur hiérarchisation- met en évidence la structure accentuelle de l'énoncé et, indirectement, son organisation syntactico-sémantique. L'observation conjointe de l'organisation accentuelle et temporelle des phrases réitérées fait apparaître que la présence d'une syllabe accentuée est souvent associée à celle de la dernière syllabe d'une phase temporelle ou d'une sous-phase et réciproquement : 81% des syllabes accentuées correspondent à la fin d'une phase temporelle ou d'une sous-phase et 90% des syllabes finales de phase temporelle ou de sous-phase sont accentuées. A l'inverse, 95% des syllabes inaccentuées ne correspondent ni à la fin d'une phase ni à la fin d'une sous-phase.

Par conséquent, les phases temporelles correspondent souvent à des groupes accentuels (un groupe accentuel est constitué par une syllabe accentuée précédée généralement par une ou plusieurs syllabes inaccentuées). Une phase temporelle comprend généralement un groupe accentuel, mais elle peut aussi en comprendre deux. Un accent secondaire ne correspond pas systématiquement à la fin d'une phase ou d'une sous-phase, alors qu'un accent primaire correspond, la plupart du temps, à la fin d'une phase : la moitié environ des syllabes qui portent un accent secondaire, soit 48%, correspondent à la fin d'une phase temporelle ou d'une sous-phase (36% correspondent à la fin d'une phase temporelle et 13% à la fin d'une sous-phase suivie par une réinitialisation locale) ; par contre, la quasi totalité des syllabes qui portent un accent primaire, soit 95%, correspondent à la fin d'une phase.

La variabilité inter-sujets est faible en ce qui concerne le nombre de syllabes portant un accent primaire et correspondant à la fin d'une phase temporelle : 96% pour le sujet 1, 93% pour le sujet 2 et 95% pour le sujet 3 (moyenne 95%). Par contre, la variabilité inter-sujets est très forte en ce qui concerne le nombre de syllabes qui portent un accent secondaire et qui correspondent à la fin d'une phase temporelle ou d'une sous-phase : 29% pour le sujet 1, 44% pour le sujet 2 et 71% pour le sujet 3, (moyenne 48%).

4. CONCLUSION

En français, la phrase s'organise temporellement en phases de ralentissement de plus ou moins grande amplitude, chaque phase étant suivie d'une réinitialisation de la durée syllabique, la plupart du temps, globale et parfois locale. La phase temporelle est la structure minimale récurrente. Trois indices acoustiques permettent, en conséquence, de déterminer une phase temporelle en français :

- l'allongement marqué de la dernière syllabe de la phase temporelle ;

- le mouvement de ralentissement du débit qui affecte toutes les syllabes de la phase temporelle ;

- le phénomène de réinitialisation de la durée syllabique de la 1ère syllabe de la phase temporelle.

La taille de la phase temporelle constitue une autre de ses particularités. En effet, une phase temporelle est, dans la plupart des cas, composé d'un petit nombre de syllabes. Dans 90% des cas, le nombre de syllabes par phase se situe entre 2 et 5 et dans 61% des cas les phases sont composées de 2 ou 3 syllabes.

En conclusion, la récurrence de phases temporelles dont les caractéristiques communes sont le processus de réinitialisation de la durée de la 1ère syllabe, le phénomène de ralentissement du débit qui aboutit à l'allongement de la dernière syllabe et le petit nombre de syllabes qui les composent s'apparente à un phénomène cyclique de production. En effet, selon Fraisse (1956 et 1967), une activité motrice qui se base sur le principe de récurrence est plus économique sur le plan de la production motrice. De plus, la très grande stabilité, dans la phrase et pour un même sujet, de la durée de la syllabe où s'accomplit la réinitialisation laisse présumer que le phénomène de réinitialisation est un processus de production controlée. La réalisation d'une série de phases temporelles consisterait à produire, non pas une suite de syllabes de durée croissante, mais une série de "coups d'accélérateur" ponctuels ou impulsions d'accélération du débit, espacés de 2 à 5 syllabes généralement, dans un système dont l'inertie provoquerait un ralentissement du débit.

Par ailleurs, le nombre de syllabes qui constitue une phase temporelle semble ne pouvoir dépasser une certaine limite. Cette limitation semble se réaliser malgré les contraintes imposées par l'organisation linguistique, telles que le nombre variable de syllabes dans les groupes lexicaux et syntactico-sémantiques. On peut donc émettre l'hypothèse que la limitation de la taille de la phase temporelle, entre 2 et 5 syllabes dans 90% des cas, est le résultat de contraintes de nature biologique , psychologique et cognitive. Fraisse (1967 et 1974) remarque d'ailleurs une tendance spontanée chez l'homme à produire des structures rythmiques simples composées d'un nombre limité d'éléments, de l'ordre de 2 à 5. La phase temporelle, qui est d'ailleurs souvent liée à un groupe accentuel, pourrait correspondre à une unité de traitement cognitif de l'information ou "chunk" d'information (Miller, 1956). Cutler et Norris (1988) et Cutler (1990) émettent pour l'anglais, mais pas pour le français, une hypothèse du même type. Ces auteurs soutiennent l'hypothèse d'une stratégie de traitement pré-lexical de l'information basée sur la structure métrique de l'énoncé ; lors du traitement pré-lexical de l'information, l'auditeur effectuerait une pré-segmentation en groupes accentuels.

NOTES

Note 1 : En tenant compte des phénomènes observés dans les phrases réitérées, nous avons arbitrairement décidé que toute diminution de la durée syllabique égale ou supérieure à 5% de la durée de la syllabe précédente serait considérée comme étant associée à un phénomène de réinitialisation globale ou locale. De plus, nous avons considéré qu'une réinitialisation est locale lorsque la durée de la syllabe où s'effectue la réinitialisation est supérieure d'au moins 10% à la durée de la syllabe où s'est effectuée la précédente réinitialisation globale.

Note 2 : Afin de déterminer dans les phrases réitérées quelles sont les syllabes brèves et les syllabes longues, nous avons choisi d'appliquer un seuil différentiel de durée et non un seuil de durée absolue. Une syllabe est donc considérée comme longue lorsqu'elle est allongée d'au moins 25% (Rossi, 1972) par rapport à la durée moyenne des syllabes brèves qui la précèdent et qui appartiennent à la même structure temporelle.

REFERENCES

Boudreault, M. (1970) Le rythme en langue franco-canadienne, Analyse des faits prosodiques, Studia Phonetica, 3, Didier.

Butterworth, B. ; Goldman-Eisler, F. (1979) Recent Studies on Cognitive Rhythm, Of Speech and Time, Eds. A. W. Siegman et S. Fedstein, Lawrence Erlbaum Associates, New-Jersey, 211-224.

Caelen, G. (1981) Structures prosodiques de la phrase énonciative simple et étendue, Thèse de 3ème cycle, Hamburg Phonetische Beiträge, Bd. 34, Buske, Hamburg.

Cutler, A. ; Norris, D. (1988) The Role of Strong Syllabes in Segmentation for Lexical Access, J. Exp. Psychol. : Hum. Perc. & Perf., 14, 113-121.

Cutler, A. ; Exploiting Prosodic Probabilities in Speech Segmentation, Computational and Psychological Approaches to Language Processing, Ed. G. Altman, MIT Press (à paraître).

Duez, D. (1987) Contribution à l'étude de la structuration temporelle de la parole en français, Thèse de Doctorat d'Etat, Université de Provence, Aix-Marseille I.

Fowler, C. (1980) Coarticulation and Theories of Timing Extrinsic, Journal of Phonetics, 8(1), 113-133.

Fraisse, P. (1956) Les structures rythmiques, Studia Psychologica, Publications Universitaires de Louvain.

Fraisse, P. (1967) Psychologie des rythmes humains, colloque "Les rythmes", Lyon 4 Déc. 1967, Journal Français d'Oto-Rhino-Laryngologie, sup. n° 7, SIMEP, 23-33.

Fraisse, P. (1974) Psychologie du rythme, PUF, Paris.

Goldman-Eisler, F. (1972) Pauses, clauses, sentences, Lang. Speech, 15, 103-113.

Grosjean, F., Deschamps, A. (1972) Analyse des variables temporelles du français spontané, Phonética, 26, 129-156.

Henderson, A. ; Goldman-Eisler, F. ; Skabek, A. (1966) Sequential temporal patterns in spontaneous speech, Language and Speech, 9, 207-216.

Larkey, L; (1983) Reiterant speech : An acoustic and perceptual validation, J. Acoust. Soc. Am., 73, 1337-1345.

Liberman, M. et Streeter, L. (1978) Use of nonsense-syllable mimicry in the study of prosodic phenomena, J. Acoust. Soc. Am., 63, 231-233.

Miller, G. (1956) The magical number seven plus or minus two : some limits on our capacity for processing information, Psychological Review, 63, 81-97.

Nakatini, L. et Schaffer, J. (1978) Hearing "words" without words : Prosodic cues for word perception, J. Acoust. Soc. Am., 63, 234-245.

Pasdeloup, V. (1988) Temporal phases in French : an acoustic study of reiterant speech, Proc. of 7th FASE Symposium, Edimbourg, 22-26 Août, 1397-1404.

Pasdeloup, V. ; Modèle de règles rythmiques du français appliqué à la synthèse de la parole, Thèse de Doctorat de 3ème cycle, Université d'Aix-en-Provence, Aix-Marseille I (soutenance prévue en 1990).

Rossi, M. (1972) Le seuil différentiel de durée, Papers in memory of Pierre Delattre, Mouton, La Hague, 435-450.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

RELATIONS PONCTUATION/PROSODIE EN LECTURE ET EN PAROLE SPONTANEE

Isabelle GUAITELLA, Serge SANTI, Christian CAVE

Institut de Phonétique d'Aix-en-Provence, Laboratoire "Parole et Langage",
UA CNRS 261

RESUME:
Dans cette expérience, nous nous proposons de rendre compte des rapports entre les structures prosodiques de la lecture et de la parole spontanée, en relation avec le système de ponctuation usuel. Un test de perception nous permet d'en dégager les principaux phénomènes et de proposer une interprétation de ces relations, en vue, notament, d'une application à la synthèse vocale à partir du texte.

ABSTRACT:
In this paper, we investigate the relationships between prosodic structures in spontaneous and read speech, taking into account the usual ponctuation marks. A perception test allows us to analyse the main structures and to propose an interpretation of the relationships which should prove useful for speech synthesis from text.

INTRODUCTION

La relation entre ponctuation et prosodie est un domaine souvent délaissé par les spécialistes. Rares sont ceux qui considèrent cet aspect de la relation entre écrit et oral (Damourette 1939, Hirst 1987), et plus rares encore ceux qui en font l'objet d'une analyse précise. On peut citer, à ce propos, les travaux de L.Pasques (1978), de L.G.Védénina (1973, 1980), et de I.Fonagy et J.Fonagy (1983). Ces études ont en commun la tentative de définir, au delà des fonctions de la prosodie ou de la ponctuation, le réseau de relations qui unit ces deux moyens d'expression.

En effet les spécialistes de la ponctuation préfèrent généralement se consacrer au signe écrit sans envisager ses relations avec l'oral. Cependant la perspective historique en ponctuation (Catach 1980) apporte beaucoup à l'observation de ses rapports avec l'oral, dans le sens où elle permet de considérer l'évolution d'une ponctuation à l'origine plus rythmique (moyen-ageuse), vers une ponctuation plus syntaxique (à notre époque de culture écrite et de lecture silencieuse).

Au sein de ces études, seuls les travaux de I.Fonagy et J.Fonagy tiennent compte de l'aspect suivant: s'il est intéressant de considérer comment se réalise prosodiquement la ponctuation de l'écrit, il est tout aussi intéressant de se demander comment on peut ponctuer de la parole orale spontanée. On peut donc dégager deux axes d'étude: l'axe écrit et oralisation de l'écrit, et l'axe oral spontané et notation de l'oral.

Nous rappelerons à ce sujet la position de Blanche-Benveniste et Jeanjean (1987) qui n'admettent pas de ponctuer les transcriptions de parole spontanée, car, pour plusieurs sujets qui ponctuent, on obtient des résultats trés différents. En conséquence nous pouvons estimer que les sujets ponctuent probablement en fonction de la prosodie, mais, une fois la ponctuation mise, il semble que ce soit des critères syntaxiques qui découpent le texte (segmentation en phrases...).

I - HYPOTHESES

Un test de perception portant sur la façon dont des sujets ponctuent du texte lu et de la parole spontanée, devrait permettre de répondre à un certain nombre de questions.

* Nous postulons qu'il existe de grandes différences entre la prosodie de lecture et la prosodie spontanée (Guaïtella 1989a,b), et nous pensons que ces différences devraient se manifester à travers la ponctuation réalisée par les sujets. Ainsi on peut supposer que les différences seront rendues par:

- des décalages dans la segmentation (syntaxique) du texte,

- des différences dans la qualité et la quantité d'emploi des divers signes de ponctuation ([,] [.] [;] [:] [...] etc). Nous pouvons d'ores et déjà considérer que le point et la virgule sont plus proches des structures de l'écrit, par le fait qu'ils correspondent à la segmentation du texte en phrases et propositions, et que les autres signes (que nous appellerons "intermédiaires") sont plus proches de la parole spontanée. On a coutume par exemple, de noter les hésitations par les trois points. A ce propos on se référera aux travaux de Varloot (1978) qui donne une bonne description de l'utilisation des signes de ponctuation.

* Nous pensons qu'une connaissance plus précise de l'usage de la ponctuation et de ses rapports avec la prosodie, devrait permettre de considérer que, non seulement la parole orale spontanée, mais aussi l'écrit et son oralisation, dépendent d'autres critères que les critères purement syntaxiques.

Ainsi nous estimons, comme le dit Védénina, que la ponctuation inclut dans sa structure, beaucoup plus que le simple découpage syntaxique.

Une utilisation des possibilités structurelles de la ponctuation, devrait permettre à la synthèse de la parole à partir du texte, de gagner en naturel. En effet cette perspective permettrait d'admettre, même au sein de l'écrit, une relation de la prosodie avec d'autres structures que la syntaxe. De plus une clarification des relations entre parole spontanée et ponctuation, pourrait amener à inclure dans l'"interprétation" du texte lu, une part de naturel évoquant le spontané. En effet, si une intonation de lecture peut paraître suffisante dans de nombreuses applications, la monotonie engendrée par la répétition de schémas figés dans une norme de type "lecture", compte pour beaucoup dans la mauvaise appréciation que portent les utilisateurs de synthèse vocale sur cette dernière. Par exemple, lors de l'interrogation à distance d'une base de données utilisant la synthèse, l'utilisateur préfère avoir l'impression de dialoguer avec une "machine qui parle" plutôt qu'avec une "machine qui lit". Cette impression de vie dans le dialogue ne peut passer que par une prise en compte des phénomènes prosodiques de la parole spontanée dans l'élaboration des messages vocaux synthétiques.

II - EXPERIENCE

Nous avons sélectionné un extrait d'un corpus de parole spontanée constitué d'une interview d'une locutrice enregistrée en chambre anéchoïque (voir annexe). Cet extrait a été transcrit, nous avons supprimé des aspects spécifiques de la parole spontanée (tels que les hésitations) et nous l'avons ponctué d'une façon très normative. Ce texte a ensuite été lu par un autre locuteur enregistré en chambre anéchoïque.

Notre expérience, proche de celle de I.Fonagy et J.Fonagy, en diffère par le fait que nous avons fait entendre des textes au lieu de phrases isolées.

Ces deux textes, spontané et lu, ont été entendus par onze sujets qui devaient les ponctuer, en ayant sous les yeux le texte imprimé très espacé et sans ponctuation. La consigne était de ponctuer les textes comme s'ils étaient en train de faire une dictée à l'école, dictée dont on ne leur indiquerait pas la ponctuation. L'expérience comprenait, en outre, d'autres textes, ce qui a permis d'éviter la mémorisation de la ponctuation mise par le sujet sur le premier texte.

Les sujets ne représentent pas une classe de population homogène, leurs ages s'étalent entre 18 et 65 ans, de ce fait les méthodes par lesquelles ils ont appris à lire sont différentes, ils exercent des professions diverses et appartiennent à des milieux sociaux divers.

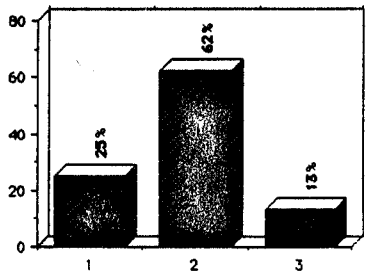
Nous avons analysé les réponses des sujets en segmentant le texte en portions dont les limites correspondent aux endroits où, au moins un sujet, a mis une ponctuation pour le texte lu, ou pour le texte spontané. Ainsi nous avons, pour un endroit du texte donné, la qualité et la quantité de ponctuation mise par les sujets en lecture et en parole spontanée. Il va de soi que certains découpages de portions de textes sont "valables" pour uniquement une des deux versions du texte.

Afin de considérer les rapports entre prosodie et ponctuation, nous avons procédé à une analyse acoustique afin d'obtenir les variations de fréquence fondamentale. Ces analyses ont été effectuées sur un inscripteur à jet d'encre Oscillomink. Nous avons également utilisé un logiciel de traitement du signal implanté sur un mini-ordinateur Masscomp par Robert Espesser, ingénieur CNRS. Pour notre comparaison nous avons tenu compte de la présence ou non d'une pause, et de la variation de pente et d'amplitude de pente sur la ou les syllabes qui précèdent les signes de ponctuation. Pour cette analyse nous avons utilisé une description phonologique de la prosodie (Guaïtella 1990), inspiré par le modèle de Ph. Martin (1982).

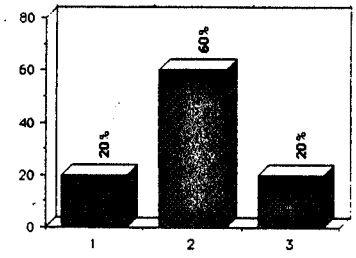
III - RESULTATS

1- Répartition des signes

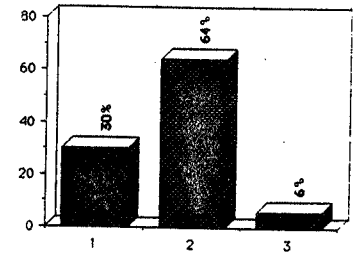
Les tableaux suivants rendent compte de la répartition des différents signes en lecture et en parole spontanée.



Répartition des signes de ponctuation en lecture + parole spontanée



Répartition des signes de ponctuation en parole spontanée



Répartition des signes de ponctuation en lecture

TYPE DE PAROLE →	LECTURE		P. SPONTANEE	
	NOMBRE	POURC.	NOMBRE	POURC.
PONCTUATIONS	326	56 %	258	44 %
VIRGULES	208	58 %	155	42 %
POINTS	99	65 %	53	35 %
INTERMEDIAIRES	18	26 %	51	74 %

Répartition des signes de ponctuation en fonction du type de parole

Nous voyons que les éléments les plus démarcatifs d'unités syntaxiques - le point et la virgule - apparaissent davantage pour la lecture, et que les éléments moins démarcatifs se rencontrent surtout en parole spontanée.

2- Les "lieux de ponctuation"

Afin de faciliter l'explication nous nommerons les endroits du texte où au moins une ponctuation a été mise, les "lieux de ponctuation".

Ces lieux de ponctuation peuvent diverger, d'une part, entre la lecture et la parole spontanée, et, d'autre part, entre les différents sujets. En effet, suivant les lieux de ponctuation, l'accord des sujets est très variable.

2.1- Les lieux d'accord total

Certains lieux correspondent à un accord entre lecture et parole spontanée, et à un accord maximal entre les sujets.

ex: "[...] de se marier avec un français ou une française [...]"

cet extrait a donné deux lieux de ponctuation, dont un n'est pas pertinent en lecture.

TYPE DE PAROLE →	LECTURE				P. SPONTANEE			
TYPE DE PONCTUATION	VIRGULE	POINT	INTER.	TOTAL	VIRGULE	POINT	INTER.	TOTAL
LIEU 32	0	0	0	0	1	0	0	1
LIEU 33	11	0	0	11	5	4	1	10

2.2- Lieux de désaccord

Nous envisageons ici les lieux de désaccord entre lecture et parole spontanée.

2.2.1- Présence de ponctuation en lecture et absence en parole spontanée

ex: [...] et par contre au niveau des mentalités l'empreinte est très forte [...]

TYPE DE PAROLE →	LECTURE				P. SPONTANEE			
TYPE DE PONCTUATION	VIRGULE	POINT	INTER.	TOTAL	VIRGULE	POINT	INTER.	TOTAL
LIEU 21	7	0	0	7	0	0	0	0

2.2.2- Présence de ponctuation en parole spontanée et absence en lecture

ex:"[...]ils ne rêvent que d'une chose c'est de partir pour la france ou de se marier avec un français ou une française[...]

TYPE DE PAROLE →	LECTURE				P. SPONTANEE			
TYPE DE PONCTUATION	VIRGULE	POINT	INTER.	TOTAL	VIRGULE	POINT	INTER.	TOTAL
LIEU 31	0	0	0	0	4	0	3	7

2.3- Justification prosodique

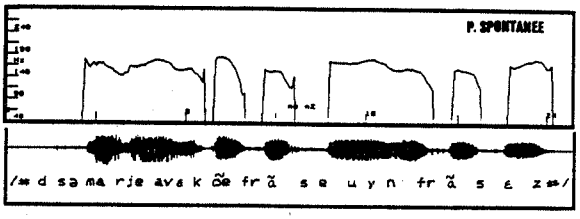
2.3.1- Lieux d'accord

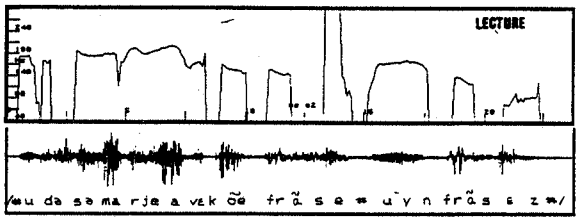
Exemple du lieu 33:

-10 réponses en parole spontanée correspondant à 5 virgules, 4 points, 1 intermédiaire.

-11 réponses en lecture correspondant à 11 points.

Nous avons, en parole spontanée, une pause précédée d'un contour montant non-ample, et, en lecture une pause précedée d'un contour descendant ample .





LIEU 33	P. SPONT.	LECTURE
MONTANT	+	−
AMPLE	−	+
SUIVI D'UNE PAUSE	+	+

2.3.2- Lieux de désaccord

*Exemple du lieu 31:

-7 réponses en parole spontanée, dont 4 virgules et 3 intermédiaires.

-aucune réponse en lecture.

En parole spontanée, nous avons une pause et un contour descendant non-ample alors qu'en lecture nous avons une absence de pause et un contour montant non ample.

LIEU 31	P. SPONT.	LECTURE
MONTANT	−	+
AMPLE	−	−
SUIVI D'UNE PAUSE	+	−

*Exemple du lieu 21:

- aucune réponse en parole spontanée.
- 7 réponses en lecture qui sont 7 virgules.

En parole spontanée, nous avons une absence de pause et un contour ample montant, et, nous avons en lecture, une pause précédée d'un contour ample montant .

LIEU 21	P. SPONT	LECTURE
MONTANT	+	+
AMPLE	+	+
SUIVI D'UNE PAUSE	−	+

Nous observons, à travers ces exemples, une tendance à faire correspondre, en lecture, un point avec un contour ample descendant suivi d'une pause, et une virgule avec un contour ample montant suivi d'une pause. Cependant la tendance la plus générale semble être liée aux phénomènes de pause et d'amplitude du contour.

3- Les bases de la relation entre prosodie et ponctuation

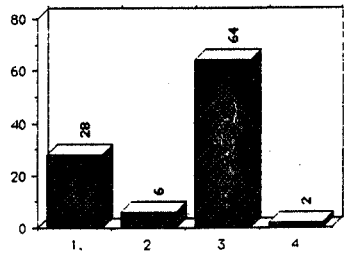
Notre démarche consiste, plutôt que de partir de la ponctuation pour considérer comment elle se réalise prosodiquement, de partir de la prosodie pour voir comment elle se réalise "ponctuativement".

Dans leur ensemble, les lieux de ponctuation se manifestent au niveau prosodique par la présence d'une pause et d'un contour ample.

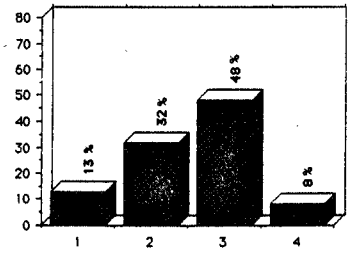
TYPE DE PAROLE →	LECTURE	P. SPONTANEE
* SEULE	14	6
AMP. SEULE	3	15
* + AMP.	31	23
NI * NI AMP.	1	4

AMP. : AMPLITUDE
* : PAUSE

Structure prosodique de l'ensemble des lieux de ponctuation



Pourcentages d'occurence des différents schémas prosodiques en parole spontanée



Pourcentages d'occurence des différents schémas prosodiques en lecture

1 : Pause seule
2 : Amplitude seule
3 : Pause + amplitude
4 : Ni pause ni amplitude

Cependant, lorsque une pause ou un contour ample sont réalisés de façon isolée, c'est plutît la pause qui détermine la mise d'une ponctuation en lecture, contrairement à la parole spontanée où ce sera l'amplitude du contour.

CONCLUSION

Notre étude permet de mettre en évidence les divergences entre prosodie de lecture et de parole spotanée.

Ces divergences sont manifestées par la transcription de l'intonation par le biais du code de ponctuation. Ceci permet, de ce fait, de mieux comprendre le système de ponctuation et sa relation à la prosodie.

Nous pensons que les résultats obtenus peuvent donner lieu à une application future en synthèse de la parole à partir du texte, notament sur les points suivants:

- meilleure compréhension des systèmes des "différentes prosodies"
- possibilité d'intégration, à partir de la ponctuation réelle du texte, ou d'un autre codage graphique restant à définir, du système prosodique à implanter sur le synthétiseur pour que celui-ci produise une prosodie de lecture ou de parole spontanée.

BIBLIOGRAPHIE:

BLANCHE-BENVENISTE C., JEANJEAN C., 1987, Le français parlé, INALF, Didier

CATACH N., 1980, "La ponctuation", Langue Française 45, Larousse, 16-27

DAMOURETTE J., 1939, Traité moderne de ponctuation, Larousse

FONAGY I., FONAGY J., 1983, "L'intonation et l'organisation du discours", Bulletin de la Société de Linguistique de Paris, LXXVIII, 1, Kliencksieck, 161-209

GUAITELLA I., 1989(a), "Variabilité dans la réalisation des contours mélodiques: cas de la structure syntaxique énumérative", Actes du séminaire "Variabilité et spécificité du locuteur", C.I.R.M. Luminy

GUAITELLA I., 1989(b), "Problèmes liés à une analyse comparative de la prosodie en lecture et en parole spontanée", Travaux de l'Institut de Phonétique d'Aix 13

GUAITELLA I., 1990, "Propositions pour une méthode d'analyse de l'intonation en parole spontanée", Actes du Premier Congrés Français d'Acoustique, Lyon (à paraître)

HIRST D., 1987, La représentation linguistique des systèmes prosodiques: une approche cognitive, Thèse de doctorat d'état en phonétique, Université de Provence, Aix-en-Provence

MARTIN Ph., 1982, "Phonetic realisations of prosodic contours in french", Speech Communication 1, North-Holland Publishing Company, 283-294

PASQUES L., 1978, "Ponctuation à l'écrit, arrangement rythmique à l'oral, à propos d'un conte de Marcel Jouhandeau lu par l'auteur", in: Catach et Tournier, La ponctuation recherches historiques et actuelles, actes de la table ronde CNRS "L'histoire de la ponctuation depuis les débuts de l'imprimerie", 189-222

VARLOOT J., 1978, "Faisons le point", in: Catach et Tournier, La ponctuation recherches historiques et actuelles, actes de la table ronde CNRS "L'histoire de la ponctuation depuis les débuts de l'imprimerie", 11-28

VEDENINA L.G., 1973, "La transmission par la ponctuation des rapports du code oral avec le code écrit", Langue Française 19, Larousse, 33-40

VEDENINA L.G., 1980, "La triple fonction de la ponctuation dans la phrase: syntaxique, communicative, sémantique", Langue Française 45, Larousse, 60-66

ANNEXE : Corpus et lieux de ponctuation

1- je pense que c'est pas une empreinte
2- c'est un manque d'empreinte
3- la france
4- a exporté toutes les
5- ressources du pays
6- sans apporter en échange
7- aucune infrastructure économique
8- c'est-à-dire pas de route
9- pas de chemin de fer
10- pas
11- d'usine
12- ou alors
13- ce qu'il y avait a été détruit au départ des français
14- contrairement à d'autres pays
15- ou il y a quand même eu des choses qui ont été construites pour le pays lui-même
16- à madagascar
17- c'est vraiment
18- au niveau zéro
19- et
20- par contre
21- au niveau des mentalités
22- l'empreinte est trés forte
23- c'est-à-dire que les malgaches sont constamment tournés vers la france
24- tout ce qui est bien vient de France
25- tout ce qui les intéresse
26- c'est la france
27- et
28- au niveau d'une certaine couche assez aisée de la population
29- ils ne rêvent que d'une chose
30- c'est de partir pour la france
31- ou
32- de se marier avec un français
33- ou une française
34- en dehors de ça
35- c'est un pays
36- ou
37- dès qu'on est sorti des villes
38- on se retrouve peut-être
39- cinq cent ans en arrière
40- les gens n'ont pas de route
41- pas de voiture
42- vivent dans la jungle
43- en autarcie totale
44- c'est-à-dire
45- qu'ils se suffisent à leurs besoins
46- il y a eu aussi un fort exode rural au départ
47- des français
48- les gens ne savaient plus quoi faire
49- donc ils sont venus vers la ville
50- et
51- une fois qu'ils sont en ville
52- ils stagnent
53- ils meurent de faim
54- il y a les maladies
55- l'empreinte des français
56- de toute façon
57- c'est une empreinte coloniale
58- donc
59- négative

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Les variables temporelles dans le discours spontané de neuf sujets atteints de lésion unilatérale gauche.

Parth M. Bhatt

**Laboratoire de phonétique expérimentale, Département de français,
Université de Toronto, Toronto, Canada.**

Résumé

Cette étude présente les résultats d'une analyse instrumentale de six variables temporelles dans le discours spontané de neuf sujets atteints d'une lésion unilatérale gauche. D'après les données, les sujets atteints de lésion frontale gauche se distinguent des sujets atteints de lésion temporale ou pariétale gauche par le rapport temps d'articulation-temps de locution, la vitesse de parole et le nombre de pauses. Les patients avec lésion temporale se distinguent des patients avec lésion pariétale par la durée moyenne des pauses. La vitesse d'articulation, par contre, ne permet pas de différencier les divers groupes de sujets.

I. Introduction

Cette étude a pour objet d'analyser les variables temporelles du discours spontané de neuf sujets atteints de lésion cérébrale unilatérale de l'hémisphère gauche.

Voici les six variables qui ont servi à l'analyse:

a) le rapport temps d'articulation-temps de locution (RTATL);
b) la vitesse d'articulation;
c) le coëfficient de variation de la vitesse d'articulation;
d) la vitesse de parole;
e) le nombre de pauses;
f) la durée moyenne des pauses.

La plupart des travaux antérieurs sur ce sujet (Barbizet et Lenoir, 1968; Benson, 1967, 1979; Borkowski et coll., 1967; Goodglass et coll., 1964; Goodglass et Kaplan, 1972; Kerschensteiner et coll., 1972; Poeck et coll., 1972; Kreindler, Mihailescu et Fradis, 1980; Quinting, 1971) n'emploient pas cependant de définition phonétique précise des variables temporelles. La vitesse de parole, par exemple, est souvent définie comme le nombre de mots par minute et n'est utilisée que comme un élément parmi d'autres pour distinguer les différentes catégories de sujets aphasiques. La plupart de ces études proposent l'existence de deux catégories générales de troubles aphasiques:

a) les troubles de type non fluent où la production articulatoire est lente et laborieuse, la vitesse de parole étant inférieure à celle des sujets contrôles; ce comportement est habituellement associé aux lésions antérieures (frontales) de l'hémisphère gauche;
b) les troubles de type fluent associés aux lésions postérieures temporales et pariétales de l'hémisphère gauche où la production articulatoire est quasi-normale; ces sujets ont parfois une vitesse de parole plus élevée que les sujets contrôles.

On remarque cependant que la définition de la vitesse de parole utilisée dans ces travaux donne une valeur globale qui ne permet pas de distinguer la part de la durée des pauses de celle de la vitesse d'articulation.

En outre, on ne décrit que de façon très générale le comportement des sujets atteints d'une lésion gauche plus postérieure, temporale (aphasie de Wernicke) ou pariétale (aphasie de conduction) (Hécaen et Angelergues, 1965; Lecours et Lhermitte, 1980).

Parmi les études plus récentes, Deloche, Jean-Louis et Seron (1979) ont étudié la vitesse d'articulation, le RTATL et les pauses chez cinq sujets francophones avec lésion cérébrale qui souffraient chacun d'un trouble langagier différent: aphasie de Broca, aphasie de Wernicke, aphasie de conduction, aphasie anomique et dysarthrie corticale. Ces auteurs ont analysé le comportement de ces sujets dans le cadre d'un exercice descriptif et d'un discours spontané. Leurs résultats indiquent que ces sujets ressemblent aux sujets contrôles en ce qui concerne le nombre de pauses et la vitesse d'articulation. En revanche, la durée moyenne des pauses est plus élevée chez leurs sujets que chez les sujets qui n'ont pas de lésion cérébrale.

Klatt (1980) a étudié la distribution des pauses selon la catégorie grammaticale dans des phrases lues par des sujets aphasiques anglophones. Il a constaté que les verbes étaient associés à la fréquence la plus élevée de pauses, les substantifs à une fréquence moins élevée et les adjectifs à la fréquence la plus faible.

L'objet de cette étude est de préciser et de compléter ce portrait de sujets avec lésion cérébrale en comparant la réalisation des variables temporelles mentionnées ci-dessus par trois groupes de sujets aphasiques.

Il convient de souligner cependant que les résultats présentés ici dérivent d'un corpus qui est encore trop restreint pour que l'on puisse tirer des conclusions absolument définitives.

II. Choix des sujets

Cette enquête porte sur la production spontanée de neuf adultes francophones droitiers provenant du Service de neurochirurgie de l'Hôpital Ste. Anne et de la Clinique des maladies du système nerveux de l'Hôpital de la Salpêtrière. Tous les sujets sont domiciliés dans la région parisienne.

Chaque sujet souffre d'une lésion cérébrale unilatérale vérifiée de l'hémisphère gauche. Les sujets A, B et C (groupe 1) sont atteints d'une lésion du lobe frontal gauche, les sujets D, E et F (groupe 2) d'une lésion du lobe pariétal gauche et les sujets G, H et I (groupe 3) d'une lésion du lobe temporal gauche.

Aucun sujet n'est gaucher et n'a d'antécédents familiaux de gaucherie. Ils étaient tous neurologiquement stables au moment de l'examen. Aucun patient ne montre de syndrome langagier rare ou de localisation lésionnelle inattendue.

Les sujets ont tous été interviewés en milieu hospitalier, soit dans le bureau du médecin traitant, soit dans celui du psychologue clinicien. Dans tous les cas, l'échantillon de parole est un extrait d'environ 300 syllabes de discours spontané tiré d'un examen clinique de langage. Le patient répondait à des questions portant sur l'histoire de sa maladie, ses difficultés langagières, son métier, ou bien ses projets de vacances.

III. Les variables temporelles

1. Le rapport temps d'articulation-temps de locution

La figure 1 donne les résultats de l'analyse instrumentale de la durée totale, de la durée de parole et du rapport temps d'articulation-temps de locution.

C'est chez les sujets avec lésion frontale (46.23%) que le rapport temps d'articulation-temps de locution est le plus faible. On remarque cependant que ce sont les sujets A (39,55%) et C (37,73%) qui montrent la réduction la plus sévère du rapport temps d'articulation-temps de locution, alors que le sujet B réalise un pourcentage voisin à celui des autres locuteurs atteints de lésions cérébrales (61,41%).

Les sujets des autres groupes réalisent entre 60% et 81% de parole dans leur échantillon, valeurs légèrement inférieures à celles obtenues par Grosjean et Deschamps (1972, 1973).

La comparaison des moyennes du rapport temps d'articulation-temps de locution indique qu'il existe une différence très significative entre les groupes 1 (lésion frontale) et 3 (lésion temporale). La comparaison des moyennes des autres groupes révèle une différence assez importante entre les groupes 1 (lésion frontale) et 2 (lésion pariétale) et les groupes 2 (lésion pariétale) et 3 (lésion temporale).

	Temps de locution	Temps d'articulation	Rapport TA-TL
Sujet A	22 553,00cs	8 922,00cs	39,55%
Sujet B	15 141,00cs	9 297,50cs	61,41%
Sujet C	21 277,00cs	7 602,00cs	37,73%
Moyenne	19 650,33cs	8 607,17cs	46,23%
Ecart type	3 229,55	727,10	10,76
Sujet D	9 780,50cs	6 734,50cs	68,86%
Sujet E	13 421,00cs	8 631,50cs	64,31%
Sujet F	12 560,00cs	7 631,50cs	60,76%
Moyenne	11 920,50cs	7 665,83cs	64,64%
Ecart type	1 553,50	774,83	3,32
Sujet G	7 888,00cs	6 111,00cs	77,47%
Sujet H	11 516,50cs	9 430,50cs	81,89%
Sujet I	9 580,50cs	6 437,50cs	67,19%
Moyenne	9 661,67cs	7 326,33cs	75,52%
Ecart type	1 482,44	1 493,83	6,16

Figure 1. Temps de locution, temps d'articulation, rapport temps d'articulation-temps de locution

Comparaison des moyennes (test de T)
1. Rapport temps d'articulation-temps de locution
Groupe 1 / Groupe 2 t = -2,832 p (4 d.d.l.) = 0,0473
Groupe 1 / Groupe 3 t = -4,093 p (4 d.d.l.) = 0,0150
Groupe 2 / Groupe 3 t = -2,693 p (4 d.d.l.) = 0,0545

2. Vitesse d'articulation

La figure 2 donne les résultats obtenus sur la vitesse d'articulation, le coëfficient de variation de la vitesse d'articulation et la vitesse de parole.

La vitesse d'articulation a été calculée séparément pour chaque groupe accentuel, et non selon la méthode de calcul global préconisée par Grosjean et Deschamps (1972, 1973). La vitesse d'articulation est définie ici comme étant la valeur moyenne de l'ensemble des groupes accentuels.

Les trois groupes se comportent de façon relativement homogène sur le plan de la réalisation de cette variable. La comparaison des moyennes de la vitesse d'articulation ne permet de déceler aucune différence importante entre les trois groupes. Il semblerait alors, qu'abstraction faite des pauses, les trois groupes de locuteurs ont un débit articulatoire global assez semblable.

	Vitesse d'articulation (syll/sec)	Coëfficient de variation	Vitesse de parole (syll/mn)
Sujet A	3,59	0,518	77,40
Sujet B	2,99	0,505	102,60
Sujet C	4,20	0,397	82,80
Moyenne	3,59	0,470	87,60
Ecart type	0,49	0,050	10,83
Sujet D	4,63	0,412	183,00
Sujet E	4,07	0,771	135,00
Sujet F	3,72	0,424	124,20
Moyenne	4,14	0,540	147,40
Ecart type	0,37	0,170	25,56
Sujet G	4,46	0,338	196,80
Sujet H	3,74	0,491	163,80
Sujet I	4,36	0,396	184,20
Moyenne	4,19	0,410	181,60
Ecart type	0,32	0,060	13,60

Figure 2. Vitesse d'articulation et vitesse de parole

Comparaison des moyennes (test de T)

2. Vitesse d'articulation

Groupe 1 / Groupe 2 t = -1,552 p (4 d.d.l.) = 0,1957
Groupe 1 / Groupe 3 t = -1,776 p (4 d.d.l.) = 0,1504
Groupe 2 / Groupe 3 t = -1,770 p (4 d.d.l.) = 0,8681

3. Coëfficient de variation de la vitesse d'articulation

Groupe 1 / Groupe 2 t = -0,6842 p (4 d.d.l.) = 0,5314
Groupe 1 / Groupe 3 t = 1,331 p (4 d.d.l.) = 0,2541
Groupe 2 / Groupe 3 t = 1,249 p (4 d.d.l.) = 0,2798

4. Vitesse de parole

Groupe 1 / Groupe 2 t = -3,731 p (4 d.d.l.) = 0,0203
Groupe 1 / Groupe 3 t = -9,365 p (4 d.d.l.) = 0,0007
Groupe 2 / Groupe 3 t = -2,046 p (4 d.d.l.) = 0,1102

3. Le coëfficient de variation de la vitesse d'articulation

Les résultats de la figure 2 montrent également que les trois groupes se ressemblent en ce qui concerne le coëfficient de variation de la vitesse d'articulation (l'écart-type divisé par la moyenne globale). La comparaison de moyennes ne fournit aucun résultat significatif du point de vue statistique. Ce résultat confirme les observations des études antérieures qui ont trouvé que la vitesse d'articulation ne varie pas de façon inattendue chez les sujets aphasiques (Howes, 1964, 1967: Howes et Geschwind, 1964).

4. Vitesse de parole

Les locuteurs atteints de lésion unilatérale gauche sont très hétérogènes du point de vue de la vitesse de parole (nombre de syllabes par minute incluant la durée des pauses, Figure 2). La comparaison des moyennes des trois groupes montre l'existence de différences hautement significatives entre, d'une part, les groupes 1 (lésion frontale) et 2 (lésion pariétale), d'autre part, les groupes 1 (lésion frontale) et 3 (lésion temporale). Les sujets avec lésion frontale (87,60 syllabes par minute) ont une vitesse de parole bien inférieure à celle des deux autres groupes. Il convient de noter que les trois moyennes sont nettement inférieures à celles qu'ont obtenues Grosjean et Deschamps (1972, 1973).

5. Nombre de pauses

On a considéré comme pause tout arrêt du signal oscillographique, fréquentiel et d'intensité, sauf dans les cas de réalisation d'une occlusion consonantique. On n'a pas retenu le seuil temporel de 25cs établi par Goldman-Eisler (1968) car on a observé des pauses (par exemple entre deux voyelles) au-dessous de ce seuil qui paraissaient tout à fait pertinentes du point de vue perceptif.

Les résultats de l'analyse des pauses sont donnés dans la figure 3. Les locuteurs atteints de lésion frontale gauche produisent un nombre très élevé de pauses: A (98), B (100) et C (139). Ces valeurs sont bien au-dessus de celles des deux autres groupes.

La comparaison des moyennes des trois groupes révèle l'existence de différences significatives entre les groupes 1 (lésion frontale) et 2 (lésion pariétale) et les groupes 1 (lésion frontale) et 3 (lésion temporale).

Ce résultat confirme l'observation de Benson (1967) qui a obtenu un résultat analogue avec des sujets aphasiques anglophones.

6. Durée moyenne des pauses

Les résultats de la figure 3 indiquent également que les locuteurs atteints de lésion frontale gauche se comportent de façon peu homogène en ce qui concerne la durée moyenne des pauses. Les locuteurs A (139,09cs) et C (98,38cs) réalisent des pauses de durée moyenne assez longue, alors que le sujet B (48,43cs) se rapproche, comme pour le rapport temps d'articulation-temps de locution, de la moyenne des autres sujets avec lésion gauche.

Du fait de cette hétérogénéité, la comparaison des moyennes ne permet pas de dégager de différence statistiquement importante entre le groupe 1 et les deux autres groupes. En revanche, elle indique l'existence d'une différence hautement significative entre les groupes 2 (lésion pariétale) et 3 (lésion temporale). La durée moyenne des pauses est d'ailleurs l'unique variable temporelle où les sujets avec lésion pariétale se distinguent clairement de ceux qui ont une lésion temporale.

	Nombre de pauses	Durée moyenne des pauses
Sujet A	98,00	139,09cs
Sujet B	100,00	48,43cs
Sujet C	139,00	98,38cs
Moyenne	112,33	95,30cs
Ecart type	18,87	37,08
Sujet D	47,00	64,80cs
Sujet E	63,00	76,02cs
Sujet F	71,00	69,41cs
Moyenne	60,33	70,08cs
Ecart type	9,98	4,60
Sujet G	38,00	46,76cs
Sujet H	50,00	41,72cs
Sujet I	56,00	56,12cs
Moyenne	48,00	48,20cs
Ecart type	7,48	5,97

Figure 3. Nombre de pauses et durée moyenne des pauses

Comparaison des moyennes (test de T)

5. Nombre de pauses
Groupe 1 / Groupe 2 t = 4,219 p (4 d.d.l.) = 0,0135
Groupe 1 / Groupe 3 t = 5,489 p (4 d.d.l.) = 0,0054
Groupe 2 / Groupe 3 t = 1,712 p (4 d.d.l.) = 0,1620

6. Durée moyenne des pauses
Groupe 1 / Groupe 2 t = 1,169 p (4 d.d.l.) = 0,3073
Groupe 1 / Groupe 3 t = 2,172 p (4 d.d.l.) = 0,0956
Groupe 2 / Groupe 3 t = 5,028 p (4 d.d.l.) = 0,0073

III. Conclusion

Les résultats présentés ci-dessus permettent d'affirmer l'existence de quatre variables temporelles où les différences de comportement des sujets atteints de lésion unilatérale gauche correspondent au critère de localisation intrahémisphérique du siège lésionnel:

a) le rapport temps d'articulation-temps de locution (différence hautement significative entre les groupes 1 et 3 et différence assez significative entre les groupes 1 et 2 et 2 et 3);
b) la vitesse de parole (différence significative entre les groupes 1 et 2 et 1 et 3);
c) le nombre de pauses (différence significative entre les groupes 1 et 2 et 1 et 3);
d) la durée moyenne des pauses (différence significative entre les groupes 2 et 3);

Ces résultats confirment la similarité du comportement des sujets avec lésion pariétale ou temporale en qui concerne les variables temporelles. La durée moyenne des pauses est le seul critère qui permet de différencier ces deux groupes.

D'autre part, ces résultats confirment de nouveau l'existence de différences importantes entre le comportement des sujets atteints d'une lésion frontale et celui des sujets atteints d'une lésion gauche plus postérieure. Il convient de souligner cependant que ces différences reposent surtout sur la réalisation des pauses. Dans ce contexte, les variables les plus importantes

sont le nombre de pauses et le pourcentage relatif de pauses (voir aussi Bhatt, 1989). L'expression spontanée des sujets avec lésion frontale gauche est ponctuée d'un grand nombre de pauses dont la durée globale dépasse celle de la parole. En outre, le débit articulatoire des sujets avec lésion frontale est tout à fait comparable à celui des autres sujets avec lésion cérébrale.

Ces constatations permettent à leur tour d'affirmer que la discontinuité de la production verbale (phénomène déterminé par le nombre de pauses et non par le débit articulatoire ou par la durée moyenne des pauses) constitue l'élément-clé du comportement langagier des sujets avec lésion frontale gauche.

Référence bibliographiques

BARBIZET, J. et G. LENOIR (1968) Etude dynamique d'échantillons verbaux chez des sujets normaux et chez des malades atteints de lésions cérébrales: étude du débit verbal, L'année psychologique, 68, 431-449.

BHATT, P. (1989) Temporal variables and fundamental frequency following unilateral left anterior or posterior lesion, Proceedings of the 13th International Congress on Acoustics, Vol. 2, P. Pravica, G. Drakulic and B. Totic, Dir., Sava Centar: Belgrade, 441-44.

BENSON, D.F. (1967) Fluency in aphasia: correlation with radioactive scan localization, Cortex, 3, 373-94.

BENSON, D.F. (1979) Aphasia, alexia and agraphia, New York: Churchill.

BORKOWSKI, J., A. BENTON et O. SPREEN (1967) Word fluency and brain damage, Neuropsychologia, 5, 135-140.

DELOCHE, G., J. JEAN-LOUIS et X. SERON (1979) Study of the temporal variables in the spontaneous speech of five aphasics, Brain and Language, 8, 241-250.

GOLDMAN-EISLER, F. (1968) Psycholinguistics: experiments in spontaneous speech, New York: Academic Press.

GOODGLASS, H. et E. KAPLAN (1972) The assessment of aphasia and related disorders, Philadelphia: Lea and Febiger.

GOODGLASS, H., F. QUADFASEL et W. TIMBERLAKE (1964) Phrase length and the type of severity of aphasia, Cortex, 1, 133-52.

GROSJEAN, F. (1980) Linguistic structures and performance structures: studies in pause distribution, Dans Temporal variables in speech, H. Dechert et M. Raupach, Dirs., The Hague: Mouton, 91-106.

GROSJEAN, F. et A. DESCHAMPS (1972) Analyse des variables temporelles du français spontané, Phonetica, 26, 129-156.

GROSJEAN, F. et A. DESCHAMPS (1973) Analyse des variables temporelles du français spontané, Phonetica, 28, 191-226.

HÉCEAN, H. et R. ANGELERGUES (1965) Pathologie du langage: l'aphasie, Paris: Larousse.

HOWES, D. (1964) Application of the word frequency concept to aphasia, Dans Disorders of language, A.V.S. De Reuck et M. O'Connor, Dirs., London: Churchill, 47-75.

HOWES, D. (1967) Some experimental investigations of language in aphasia, Dans Research in verbal behavior and some neurophysiological implications, K. Salzinger et S. Salzinger, Dirs., New York: Academic Press, 181-199.

HOWES, D. et N. GESCHWIND (1964) Quantitative studies of aphasic language, Dans Disorders of communication, D. Rioch et E. Weinstein, Dirs., Baltimore: Williams and Wilkins, 229-244.

KERSCHENSTEINER, M., K. POECK et E. BRUNNER (1972) The fluency non-fluency dimension in the classification of aphasic speech, Cortex, 8, 233-247.

KLATT, H. (1980) Pauses as indicators of cognitive functioning in aphasia, Dans Temporal variables in speech, H. Dechert et M. Raupach, Dirs., The Hague: Mouton, 113-120.

KREINDLER, A., L. MIHAILESCU et A. FRADIS (1980) Speech fluency in aphasics, Brain and Language, 9, 199-205.

LECOURS, A. et F. LHERMITTE (1980) L'aphasie, Paris: Flammarion.

POECK, K., M. KERSCHENSTEINER et W. HARTJE (1972) A quantitative study on language understanding in fluent and non-fluent aphasia, Cortex, 8, 299-304.

QUINTING, G. (1971) Hesitation phenomena in adult aphasic and normal speech, The Hague: Mouton.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

EVALUATION D'UN MODELE D'ENONCIATION PAR LES DONNEES PROSODIQUES

GENEVIEVE CAELEN-HAUMONT

Institut de la Communication Parlée UA368
INPG et Université Stendhal
46 Avenue F. VIALLET 38031 GRENOBLE Cédex

RESUME

Le but de cette communication est d'une part de proposer dans le domaine sémantique un modèle original de l'énonciation, reposant sur la prise en compte d'une hiérarchisation des unités informatives, et d'autre part d'en opérer l'évaluation grâce à une confrontation statistique avec les données d'un corpus constitué de textes lus par 12 locuteurs et par l'intermédiaire de 14 paramètres dérivés de Fo. L'un d'entre eux, également original, se présente comme un paramètre pertinent tant pour les scores obtenus que pour le nombre de ses réalisations dans les énoncés.

INTRODUCTION

Pour tenter de comprendre la diversité des réalisations prosodiques, de nombreux travaux ont été consacrés à l'étude des relations entre les données prosodiques et les niveaux linguistiques. Le domaine syntaxique a sans doute été le plus largement exploité; pour le français rappelons brièvement les travaux de DELATTRE (9), DI CRISTO (10) (21), MARTIN (16) ou pour l'anglais, GROSJEAN et al. (13), OLLER (17), KLATT (15) ou COOPER (6) ... Par ailleurs et pour nous restreindre au français, la prosodie en relation avec le lexique a été analysée par VAISSIERE (24)(25); dans le domaine sémantique, le registre de l'énonciation a été exploré par ROSSI (20)(21)(22), G. CAELEN-HAUMONT (2) de même que celui de la complexité des signifiés (3)(4).

En ce qui nous concerne, nous pensons que le locuteur en fonction des impératifs de la situation ou du contexte, met en oeuvre une stratégie de base faisant référence à des domaines linguistiques tels que la syntaxe, la sémantique ou la pragmatique, et cette stratégie de base, dans un système de communication socio-culturel donné, est modulée en fonction des contraintes de l'expression orale . Ces stratégies de base, au sein de ces macro-structures linguistiques peuvent être très diverses en fonction des locuteurs. C'est pourquoi un ensemble de 6 modèles linguistiques a été défini pour tenter de circonscrire un maximum de stratégies utilisées par les locuteurs. Le but de cette communication est évaluer l'un d'entre eux par la confrontation des données de 12 locuteurs.

Ces données sont issues de la lecture d'un texte de 50 mots (dont 30 lexicaux) lu par 12 locuteurs (6 femmes et 6 hommes) selon des consignes précises. Nous n'envisagerons ici que les réalisations de la première consigne qui demandait une réalisation naturelle et intelligible. Ces données extraites de 36 fichiers spectres segmentés et étiquetés ont été regroupées sur LSI 11-73 au sein de fichiers d'observations obtenus de manière semi-automatique par la procédure de gestion de bases de données ARCANE (1).

1. LES MODELES LINGUISTIQUES

Dans une première approche du problème, nous avions pensé pouvoir utiliser directement le modèle de la quantification de la complexité des signifiés que nous avions défini auparavant (3). La confrontation avec les données expérimentales nous a révélé un certain nombre de problèmes, à savoir: 1/ les stratégies que développent les locuteurs s'inscrivent généralement dans le cadre de la phrase et non du texte : autrement dit la stratégie utilisée est contextuelle 2/ le modèle doit tenir compte des influences provenant d'autres niveaux (syntaxe, énonciation, structure syllabique ...) et les neutraliser 3/ les paramètres prosodiques significatifs en relation avec le modèle ne sont pas nécessairement les mêmes pour tous les locuteurs.

Il était manifeste que cette première approche était trop rigide. Le but principal de notre travail est en fait d'identifier et de quantifier l'importance du registre linguistique sous-jacent propre à la stratégie du locuteur. Pour ce faire, alors que l'énoncé réalisé est toujours la résultante (et dans une proportion variable) d'un ensemble de contraintes de types pragmatiques, sémantiques, syntaxiques, phonétiques et prosodiques, nous ne devons développer qu'un seul registre linguistique par modèle théorique pour apprécier correctement son importance. Les autres registres exerçant leur fonction (et parfois à plus d'un titre) au sein de l'énoncé, nous ne pouvons cependant pas nous attendre à ce que les valeurs issues des énoncés réalisés et des modèles soient entièrement superposables.

La méthode a donc évolué et nous avons adopté une stratégie très souple. Dans ce but, 6 modèles ont été définis, se rattachant chacun à un domaine linguistique spécifique et souvent à un sous-domaine à savoir : 1/ hiérarchie syntaxique (S1) 2/ dépendance syntaxique

(S2) 3/ énonciation (E1, E2) 4/ pragmatique (P1) 5/ complexité des signifiés (C1). Le modèle E2 se distingue simplement du modèle E1 dans la mesure où l'on privilégie l'unité informative (le rhème) dans la quantification.

Par ailleurs, entre le niveau des modèles sous-jacents et les réalisations prosodiques, un autre niveau a été introduit, celui des règles d'ajustement prosodique de ces modèles. Dans l'état actuel, ces règles sont descriptives et non prédictives.

Dans le domaine de la quantification, la nature des paramètres prosodiques, celle de l'unité et le type de codage offrent le cadre de l'analyse. Pour les paramètres prosodiques, les paramètres dérivés de Fo ont été pour le moment les seuls à être utilisés. Parmi les candidats possibles, nous avons pensé en premier lieu que la valeur absolue de l'écart minimum/maximum offrait sans doute le support le mieux adapté à la recherche des corrélations avec les niveaux supérieurs du langage; cependant de manière à contenir la variabilité des réalisations, nous avons également retenu le Fo moyen, et la valeur maximale.

En ce qui concerne la nature de l'unité susceptible d'offrir un support de manifestation à ces divers modèles, diverses possibilités ont été envisagées au sein du mot lexical. Pour cette évaluation, 4 cibles ont été retenues, la voyelle de la syllabe finale, la syllabe finale, le mot lexical et une portion selon le cas plus ou moins large du mot lexical en fonction de la localisation des deux types d'extremum, la nature du segment retenu variant selon le cas de la voyelle finale au mot en passant par une ou plusieurs syllabes mais en incluant toujours tout ou partie de la syllabe finale. Ces diverses cibles ont toutes été sélectionnées de manière semi-automatique en positionnant à l'écran dans le fichier spectre étiqueté les frontières requises, ces dernières étant toujours placées sur les segments phonétiques réputés voisés.

Pour le codage, pensant qu'il serait plus facile de superposer les courbes issues des modèles et celles des réalisations des locuteurs, nous avons décidé de projeter toutes les valeurs dans un espace à 4 dimensions, remettant à plus tard la possibilité de choisir un espace de projections supérieur. Dans la mesure où les cadres de la phrase et du texte servent de référence aux locuteurs, nous avons exploré ces deux niveaux de codage, ce qui a donc donné lieu à deux versions distinctes.

2. UN MODELE D'ENONCIATION POUR LA PROSODIE

Dans cette communication, nous restreindrons notre évaluation à un modèle particulier dans le champ de la sémantique, à savoir celui de l'énonciation, domaine qui opérant sur les unités informatives, semble offrir un cadre tout à fait approprié aux fonctions prosodiques.

2.1. STRUCTURE THEME/RHEME

Comme on le sait, les notions introduites à l'origine par les linguistes de l'Ecole de Prague DANEŠ (8), FIRBAS (13) ont donné lieu à une abondante littérature, notamment en France PERROT et al. (18), SLAKTA (23), COMBETTES (5), HAGEGE (14), PERROT (19). Bien que les concepts secondaires et la terminologie soient encore à ce jour en débat, il semble que chacun s'accorde sur le contenu des notions appelées thème/rhème (ou thème/prédicat ou thème/propos) qui sont bien entendu à distinguer soigneusement des notions syntaxiques.

A la suite de DAHL (7), on considère le thème comme la partie de la "représentation sémantique d'une phrase" dans laquelle "on définit ou nomme un ensemble ou un individu" et le propos (ou le rhème) comme la partie dans laquelle "on dit quelque chose à propos de cet individu ou cet ensemble". Ces notions ont été introduites dans le domaine de la prosodie en particulier par ROSSI (20)(21)(22), CAELEN-HAUMONT (2) et l'analyse vise à mettre en relation -et pour reprendre la terminologie de DUCROT (11)-, les structures du niveau de la phrase, unité abstraite sous-jacente et les structures de l'énoncé prosodique conçu comme le résultat de l'énonciation, observable et repérable dans le temps et l'espace.

2.2. STRUCTURE SUPPORT/APPORT

Prenant en considération les écarts de variation des paramètres de Fo dans les mots lexicaux, nous avons remarqué qu'ils semblaient d'une part répondre à une hiérarchisation et d'autre part à une certaine forme de cyclicité, phénomènes que les structures énonciatives dans leur développement actuel ne pouvaient pas prendre en compte. Le domaine de l'énonciation nous a paru pouvoir fournir un cadre bien adapté à la description de tels phénomènes.

En ce qui concerne la justification théorique d'un tel modèle et pour plus de commodité, nous appuierons nos réflexions sur la première phrase de notre texte qui comporte 15 mots lexicaux. Selon la théorie de l'énonciation, la structure thématique de cette phrase serait la suivante :

"D'éminents biologistes et d'éminents zoologistes américains
<- ->
THEME

ont créé pour des vers géants un nouveau phylum
<- ->
RHEME

dans l'actuelle classification des nombreuses espèces vivantes"
<- ->
THEME

Si l'on considère les unités énonciatives ainsi analysées, on s'aperçoit qu'en fait la structure est composite. Le rhème par exemple est constitué d'éléments qui au regard de l'apport d'informations ne sont pas sur le même plan[1] : ainsi le syntagme "pour des vers géants" au sein du rhème a une fonction thématique puisque c'est bien à propos de ces vers géants que quelque chose est dit.

Par ailleurs, il est bien difficile d'admettre que le deuxième thème se trouve sur le même plan que le premier, les rôles de support de l'information n'étant pas interchangeables dans cet exemple. Et si même ce deuxième

[1] 1 Pour ROSSI (21), thème et rhème peuvent être l'objet d'une hiérarchisation mais cette hiérarchisation n'est envisagée que sur l'axe syntagmatique.

thème se rattachait au rhème précédent, comme certaines analyses peuvent l'accréditer, nous nous ramènerions au cas précédent, à savoir une sous-unité thématique dans une unité rhématique.

Considérant maintenant le problème de savoir quel est l'élément le plus porteur d'information au sein de ce qui est dit -le rhème-, on s'accorde à penser, me semble-t-il, qu'il s'agit d'un élément inférieur à l'unité que constitue le rhème, à savoir le syntagme "un nouveau phylum".

Il semble donc qu'une analyse énonciative purement linéaire ne peut pas rendre compte de ces plans et niveaux différents, et que seule une méthode hiérarchique peut assurer la gestion d'une telle structure.

Dans l'analyse que nous présentons ici (Graphe n°1), la phrase n'est plus considérée comme le cadre de l'actualisation virtuelle de la structure énonciative opérant sur les syntagmes, mais comme la structure de surface dérivée d'une structure profonde. A ce titre, et comme en syntaxe générative, elle reformule l'extension de ses unités et les relations qui s'établissent entre elles.

Les concepts de base étant différents, la terminologie a évolué vers la dichotomie support/apport. Il convient de préciser que ce n'est pas parce que le principe d'opposition est binaire (support/apport) qu'au niveau de la forme ou du contenu la dérivation est nécessairement binaire.

Comme on peut le juger, cette méthode est intéressante à plus d'un titre, non seulement au niveau de la théorie linguistique puisqu'elle semble rendre compte de faits jusque là délaissés, mais aussi sur le plan de la réalisation concrète dans la mesure où répondant à la sollicitation des données, elle offre le cadre d'une quantification hiérarchisée très simple.

3. EVALUATION DU MODELE

3.1. LES ANALYSES STATISTIQUES

3.1.1. DESCRIPTION GENERALE

Ce modèle linguistique, de même que les 5 autres modèles qui ont été par ailleurs définis (et sur lesquels je ne m'étendrai pas davantage dans le cadre de cette communication), fournit donc la référence sous-jacente aux réalisations des énoncés. L'unité voyelle de la syllabe finale lexicale n'étant jamais le cadre des scores les meilleurs, a été rapidemment éliminée et il est donc resté en lice 14 paramètres (3 paramètres + 1 x 2 unités syllabes + mots x 2 modes de codage texte + phrase).

La tâche suivante a consisté à sélectionner les modèles les plus producitifs et les énoncés/paramètres/codage les plus proches de ces derniers, la proximité étant évaluée simplement par le nombre de points superposables. C'est sur les meilleures correspondances que porte l'évaluation qui suit.

L'évaluation du modèle peut s'effectuer en plusieurs points 1° par le comptage du nombre de fois où ce modèle a servi de meilleure référence aux réalisations du locuteur 2° par l'utilisation de procédures statistiques 3° par la superposition des courbes 4° par le nombre de règles nécessaires à l'ajustement du modèle 5° par le nombre d'unités affectées par ces règles d'ajustement. Compte-tenu de l'espace imposé pour cette communication, ces trois derniers critères ne pourront pas être utilisés.

En ce qui concerne le comptage, sachant que pour une phrase donnée et un locuteur donné, on peut trouver des modèles différents avec un score identique, on s'est intéressé au pouvoir explicatif de chaque modèle par rapport au nombre de phrases. Comme on le lit sur le tableau n°2 puis n°3, c'est le domaine sémantique qui pour cette première consigne de lecture est la référence la plus souvent choisie (75% des énoncés). Les modèles E1 et E2 constituant deux versions très proches l'une de l'autre, le modèle de l'énonciation recueille à lui seul 53% des meilleurs scores, soit 19 sur un total de 36.

Il est maintenant intéressant de savoir si une unité morphosyntaxique est privilégiée dans les réalisations locuteurs. Le tableau n°2 nous informe que sur l'ensemble de tous les modèles qui obtiennent les scores les meilleurs de coïncidence avec les réalisations des locuteurs, et en tenant compte des ex-aequo, 39% des énoncés utilisent l'amplitude de Fo dans les limites de la syllabe finale, contre 61% dans le reste du mot.

En ce qui concerne le codage des projections dans l'espace 1-4, on constate que l'unité de réalisation pour une majorité de phrases -donc de locuteurs- est le texte ; au niveau formel même si les phrases successives se réfèrent à des modèles linguistiques différents, dans la majorité des cas toutefois (84%), elles sont traitées comme sous-ensembles d'une unité plus vaste, le texte.

Une autre point d'intérêt concerne le type de paramètre : y-a-t-il un paramètre privilégié parmi les paramètres retenus ou la distribution est-elle régulière entre tous les paramètres?

Selon les fichiers d'observations, les divers paramètres prosodiques limités aux paramètres dérivés de Fo sont calculés dans le cadre de la syllabe finale lexicale ou du mot lexical. Etant donné la variabilité de réalisation du /ə/ en position finale de syllabe (présence/absence, dans/hors système linguistique), nous avons pris la décision, quitte à affaiblir nos corrélations, de ne pas inclure ce segment lorsqu'il se trouve à la frontière du mot. Il convient aussi de préciser que ces divers paramètres bien que différents ne sont pas indépendants les uns des autres : ceci n'a pas d'importance, notre propos étant de savoir quelle est la meilleure formulation de ces paramètres.

Si donc T1 (cf. tableau N°3) représente la valeur absolue de l'écart minimum/maximum en syllabe finale ou dans les limites du mot, T3 le Fo moyen, T4 ce que nous avons appelé "l'intonème" (ou valeur absolue de l'écart minimum /maximum à l'intérieur du mot et incluant au moins la voyelle finale), T5 la valeur maximale de Fo, nous constatons en observant les pouvoirs explicatifs des divers paramètres relativement aux phrases, que la différence minimum /maximum dans l'unité est porteuse de signification (T1 + T4) dans 64% des énoncés de phrases dont 58% pour T1, contre 42% pour T3 et 31% pour T5. Ces résultats sont très intéressants et viennent confirmer notre intuition globale : l'écart maximum/minimum en valeur absolue au sein du mot lexical est un lieu privilégié d'interaction entre la prosodie et les niveaux supérieurs du langage.

3.1.2. LES COINCIDENCES MODELES PREDICTIFS / REALISATIONS

. La deuxième phase d'évaluation du modèle est celle des coïncidences avant l'application des règles d'ajustement. Nous ne retiendrons pas la méthode des corrélations statistiques ni celle du calcul du carré de la différence de 2 courbes, car l'influence qui provient d'une autre composante linguistique par exemple, superpose sa structure à la précédente en l'évacuant : lorsque les valeurs de prédiction ne sont pas réalisées, il importe peu dans ces conditions que les valeurs soient proches ou non. Nous n'utiliserons donc que le calcul des pourcentages des valeurs de coïncidence entre les modèles prédictifs et les réalisations locuteurs.

De manière générale, les 12 scores avant application des règles d'ajustement, évoluent donc entre 50 et 86% pour les modèles énonciatifs. Sur ces 12 scores, 75% sont égaux ou supérieurs à 60%. Les scores de coïncidence entre les valeurs prédictives des modèles et les valeurs de réalisation évoluant de 58% à 70%, celui du modèle énonciatif atteint 63%.

CONCLUSION

Notre objectif global est de tenter de décrire les réalisations prosodiques d'un corpus de lecture à l'aide d'un ensemble de modèles linguistiques syntaxiques, sémantiques et pragmatique souvent originaux et d'évaluer la nature et l'importance de cette référence sous-jacente. Dans cette communication, nous avons cherché à évaluer les modèles de type énonciatif qui semblent le plus souvent se trouver en situation de correspondance la meilleure avec les réalisations concrètes des énoncés. Nous avons proposé un modèle de l'énonciation inédit.

Par ailleurs, nous avons mis en évidence l'aspect significatif d'un paramètre prosodique jusque là à ma connaissance non utilisé -ou alors qualitativement- à savoir la valeur absolue de l'écart de la variation mélodique dans le mot lexical ou la syllabe finale.

Quant à l'évaluation de ce modèle, on peut dire qu'il constitue l'un des modèles le plus largement utilisé parmi les locuteurs. Ses performances, avant l'application des règles d'ajustement sont satisfaisantes, si l'on considère que le modèle théorique ne repose sur la prise en compte que d'une seule composante linguistique.

Il est intéressant de remarquer que ce modèle peut constituer la seule référence linguistique du locuteur pour les trois phrases (locuteur RI), ce qui laisse à penser que ce champ sémantique spécifique est susceptible de faire système, de même que l'utilisation d'un même type de paramètre, à savoir la valeur absolue de l'amplitude de Fo dans la syllabe ou le mot (locuteurs BR, CA, CO, PE, RI).

REMERCIEMENTS

Nous tenons à remercier B. COMBETTES, Professeur à NANCY II, pour son aimable disponibilité et ses conseils amicaux prodigués notamment dans les domaines de la sémantique.

REFERENCES

[1] CAELEN J., CAELEN-HAUMONT G., VIGOUROUX N., BARRERA C., MALET J.(1986)
ARCANE : Acquisition et Recherche de Connaissances Acoustico-phonétiques dans un Noyau Evolutif, Actes des 15èmes JEP, GALF-CNRS, Aix-en-Provence, 207-211.

[2] CAELEN-HAUMONT G. (1978, 1981)
Structures prosodiques de la phrase énonciative simple et étendue, Thèse de 3ème cycle, Toulouse. Hamburger Phonetische Beitrage, Band 34, Hamburg Buske.

[3] CAELEN-HAUMONT G. (1986 a)
Propositions pour un modèle sémantique simplifié de la complexité des signifiés, Actes des 15èmes JEP, GALF-CNRS, Aix-en-Provence, 201-205.

[4] CAELEN-HAUMONT G.(1986b)
Grammatical Components and Macro-Prosody: Quantitative Analysis Toward Statistical Correlations, Proc. Montreal Symposium on Speech recognition, Montreal Canada, 82-84.

[5] COMBETTES B. (1977)
Ordre des éléments de la phrase et linguisque du texte Revue Pratiques n°13, 91-101.

[6] COOPER W.E. (1975)
Syntactic Control of Speech Timing
PH. D. Thesis, MIT (unpublished).

[7] DAHL O. (1974)
Topic, Comment, Contextual Boundness and Focus Buske, Hamburg.

[8] DANES F. (1968)
Some Thoughts on the Semantic Structure of the Sentence, Lingua 21, 55-69.

[9] DELATTRE P. (1969)
Syntax and Intonation; a study in disagreement — Study of sounds, XIV, 21-40.

[10] DI CRISTO A. (1975)
Recherches sur la structuration prosodique de la phrase française
Actes des 6èmes JEP, GALF-CNRS, Toulouse, 95-116.

[11] DUCROT O.(1984)
Le dit et le dire, Les Editions de Minuit, Paris.

[12] FIRBAS (1974)
Some Aspects of the Czechoslovak Approach to Problems of Functional sentence perspective, in Papers on Functional Sentence Perspective, DANES F. ed., Mouton, La Haye.

[13] GROSJEAN F., DOMMERGUES J-Y. (1983)
Les structures de performance en psycholinguistique L'Année Psychologique, 83, 513-536.

[14] HAGEGE C. (1976)
Du thème au rhème. Pour une théorie cyclique. la Linguistique, 14, 3-38.

[15] KLATT D. H. (1975)
Vowel Lengthening is Syntactically Determined in a Connected Discourse, J. Phonetics 3, 129-140.

[16] MARTIN P. (1977)
Résumé d'une théorie de l'intonation
Bulletin de l'Institut de phonétique de Grenoble, 6, 57-87.

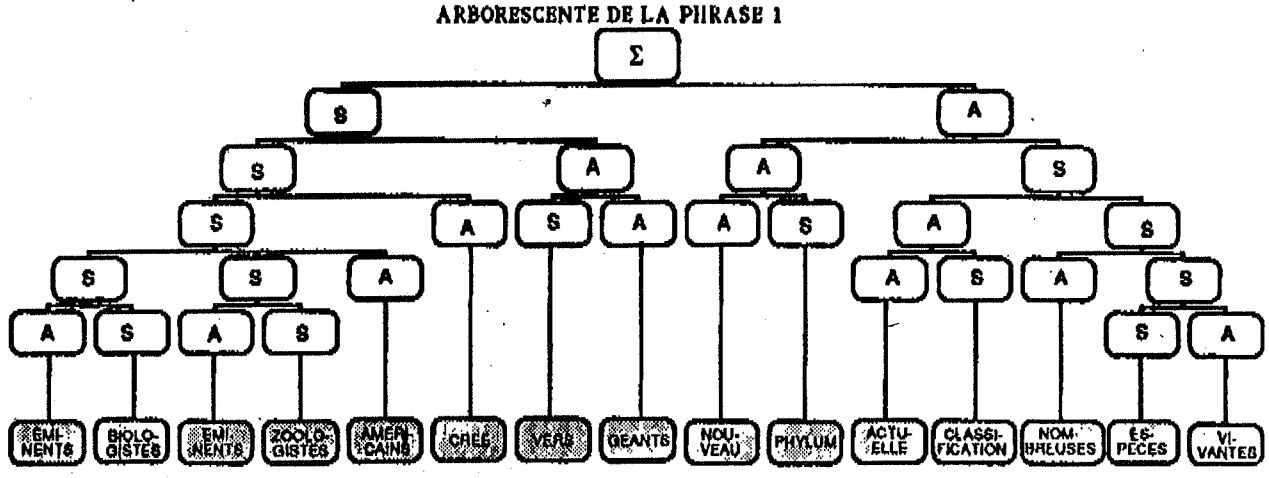
[17] OLLER D. K. (1973)
The effect of position in Utterance on speech Segment Duration in English, J. Acoust. Soc. Am, 54, 1235-1247.

[18] PERROT J. et LOUZOUN M. (1974)
Message et apport d'information : à la recherche des structures, Langue Française, 21, 122-35.

[19] PERROT J. (1978)
Fonctions syntaxiques, énonciation, information B.S.L. LXXIII, 85-101.

[20] ROSSI (1973)
L'intonation prédicative en français dans les phrases transformées par permutation, Linguistics, 103, 64-94.
[21] ROSSI M., DI CRISTO A., HIRST D., MARTIN P., NISHINUMA Y. (1981)
L'intonation, de l'acoustique à la sémantique
Klincksieck, Paris.
[22] ROSSI M. (1987)
Peut-on prédire l'organisation du langage spontané?
Etudes de Linguistique Appliquée, 66, Didier, 20-48.
[23] SLAKTA D. (1975)
L'ordre du texte
Etudes Linguistique Appliquée, 19, Didier, 30-42.
[24] VAISSIERE (1975)
On French Prosody
MIT-QPR, 212-23.
[25] VAISSIERE (1980)
La structuration acoustique de la phrase française
Ann. Sc. Norm. Sup., Pisa, III, 102, 529-560.

GRAPHE N°1 : STRUCTURE ENONCIATIVE ARBORESCENTE DE LA PHRASE 1



SYNTAXE		SEMANTIQUE			PRAGMATIQUE
39%		75%			56%
S1	S2	E1	E2	C1	P1
14%	25%	36%	17%	22%	56%

TABLEAU N°1 : POURCENTAGES DE REPARTITION DES MODELES EN FONCTION DES ENONCES DES LOCUTEURS

UNITES MORPHOSYNTAXIQUES RETENUES	
MOT LEXICAL	SYLLABE LEXICALE FINALE
61%	39%
DOMAINE D'EXTENSION DU CODAGE	
TEXTE	PHRASE
84%	16%
PARAMETRES DE Fo : % PAR RAPPORT AUX 36 ENONCES			
T1	T3	T4	T5
58%	42%	8%	31%

TABLEAU N°2 : SCORES OBTENUS PAR LES UNITES DE REFERENCE, LE TYPE DE CODAGE ET LES DIVERS PARAMETRES DE Fo

LOC	PH1	PH2	PH3	LOC	PH1	PH2	PH3
BR	T1 S2 67%	T4 P1 75%	T4 P1 71%	PA			T5 E1 71%
CA	T1 P1 60%	T1 S2 75%	T1 S2 P1 57%	PE	T1 E1 60%	T1 E2 75%	T1 E2 P1 T3 E1 71%
CO	T1 E1 53%	T1 E1 S1 E2 50%	T1 P1 66%	PI		T1 E1 S2 63%	T1 P1 66%
FA			T5 E1 71%	RI	T1 E1 53%	T1 E1 E2 P1 63%	T1 E1 86%
FO	T1 E1 60%			SE	T1 S1 C1 60%		
IN		T1 P1 63%	T1 E2 86%	UR	T1 E1 S2 P1 53%	T1 E1 63%	

TABLEAU N°3 : POUVOIRS EXPLICATIFS DU PARAMETRE T1-T4 ET DU MODELE ENONCIATIF E1-E2

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

LA PAROLE RESPIRE...: L'ORGANISATION DES DUREES DES GROUPES DE SOUFFLE ET DE PAUSE, COMME UN DES INDICES DU TEMPO DE LA PAROLE, EN COMORIEN.

REY V.

INSTITUT DE PHONÉTIQUE- AIX EN PROVENCE

RESUMÉ

Notre description de l'organisation prosodique d'un énoncé s'intéresse à décrire des "habitudes prosodiques", à l'aide du paramètre de durée.

Au regard des données, la durée des pauses peut être un indice de segmentation de la parole continue, en unités que nous appellons Groupes de Souffle.

A l'intérieur des Groupes de Souffle, l'organisation des segments dégagent des unités plus petites, appellées Mots Prosodiques. Les locuteurs produisent majoritairement des GS ne comprenant qu'un seul Mot Prosodique. Dans les cas où les GS ont au moins deux Mots Prosodiques, il s'avère qu'il existe un principe d'alternance des durées. Ce principe d'alternance est d'après nous, un indice du tempo de la parole, en comorien.

INTRODUCTION

S'intéresser à l'organisation prosodique d'un énoncé consiste à décrire une structure, organisée sur un axe temporel.

Les recherches expérimentales dans ce domaine, se réfèrent souvent à trois paramètres acoustiques, la durée, la fréquence fondamentale et l'intensité. Ces paramètres sont des indices spécifiques, décrivant la substance prosodique. Ce présent travail concerne exclusivement la durée.

Notre démarche est essentiellement empirique, et gage qu'il est possible de construire des unités à partir des documents acoustiques. Les unités discrètes émanent de l'organisation de la substance elle-même.

Nous posons comme postulat que, même si l'expressivité est une variable majeure dans l'organisation des durées, dans la parole continue, elle doit se réaliser dans un cadre prosodique donné, relevant des contraintes physiologiques d'une part, mais aussi, et c'est cela qui nous intéresse, des habitudes prosodiques des locuteurs. Le mot "habitude prosodique" signifie que nous dégageons des tendances: le locuteur pourra toujours produire une performance particulière.

Enfin, nous avons choisi de travailler sur une langue africaine, pour deux raisons au moins: ABRY & CREISSELS & SONKO (1976) constatent la rareté des travaux concernant les phénomènes prosodiques de ces langues, et ceci est toujours d'actualité, à notre connaissance. Les recherches sur ces langues (à tons ou à accent) peuvent apporter de nouvelles orientations sur les problématiques des phénomènes prosodiques.

Nos questions sont donc: comment est gérée la dimension temporelle de la structure prosodique d'un énoncé? Cette gestion est-elle un indice du tempo de la parole en comorien?

LE CORPUS

Notre corpus a deux caractéristiques: il s'agit d'une histoire, exprimée dans une langue africaine, le comorien (langue parler dans l'archipel des Comores). Ce parler est une langue bantoue (G40), langue à accent à liberté limitée.

De plus, cette histoire d'une durée de cinq minutes, fut racontée de mémoire par trois locuteurs, originaires des Comores. Ces personnes s'etaient entraînées auparavent, et ont raconté l'histoire en utilisant exactement les mêmes mots. Les données sont donc comparables.

L'enregistrement s'est effectué en chambre anéchoïque, sur un magnétophone Nagra III. Puis nous avons procédé à une segmentation des phrases, sur l'oscillogramme.

LA MÉTHODOLOGIE

Le problème principal est le critère de segmentation: une phrase est une unité qui ne se retrouve pas forcément dans la parole: elle peut, pour des raisons d'expression et de style subir un découpage.

Nous définissons une première unité appellée Groupe de Souffle (GS). Cette unité est construite de la manière suivante: elle est bornée par une pause d'une durée d'au moins 800 ms et le rapport de durée entre ce GS et cette pause est significativement différent (supérieur au seuil d'audibilité de 20% (ROSSI 1972)).

A l'intérieur d'un GS, nous définissons d'autres unités: ce sont des mots ou suite de mots, séparés par des pauses. Nous appellons ces unités "Mots Prosodiques" (MP) car ils correspondent aux unités décrites par VAISSIERE (1980).

Nous avons donc des GS, construits de façon mécanique, à l'intérieur duquel il y a des mots prosodiques (groupes de mots séparés par des pauses).

Il est tout à fait possible, qu'à l'intérieur du mot prosodique, il y ait d'autres segmentations marquées par les autres paramètres acoustiques. Ce sera l'objet de recherches futures.

La durée nous permet donc de construire deux unités, les GS et les MP.

Dans un deuxième temps, nous avons regardé l'organisation des durées, à l'intérieur des GS. Nous nous sommes posée deux questions:

Combien y-a-t-il de Mots Prosodiques à l'intérieur d'un GS?

Comment sont organisés les mots prosodiques et les pauses à l'intérieur d'un GS? Pour répondre à cette question, nous avons systématiquement fait, à l'intérieur d'un même GS, le rapport de durée entre les Mots Prosodiques d'une part et d'autre part, entre les Pauses. Quand les durées étaient différentes entre les Mots Prosodiques, nous avons appellé [l] la durée la plus longue et [c] la durée la plus courte; et [l'] et [c'] respectivement pour les pauses. Nous dirons alors, que l'organisation des durées est régie par un principe d'alternance. Quand la différence des durées n'étaient pas significativement perceptibles, nous les avons appellées de la même manière.

Le tableau ci-dessous illustre le cas d'un GS avec deux MP (les chiffres 1 et 2 renvoient à l'ordre d'apparition dans la phrase):

Tableau 1: Un exemple de formalisation des données

	durée (en ms.)	formalisation
MP1	964	l(964>>676)
MP2	676	c
P1	80	c' (80<<900)
P2	900	l'

RÉSULTATS

* La production des GS avec un ou deux MP sont les plus fréquents.

Chaque GS correspond à une unité de sens (exclamation, une phrase, un mot). Il n'y a pas donc contradiction entre la segmentation en GS et le sens.

Les trois locuteurs n'ont pas le même nombre total de GS (53, 68 et 56): la segmentation de l'histoire est différente selon les stratégies des locuteurs.

Tableau 2: Organisation des Groupes de Souffle et des Mots Prosodiques, selon les locuteurs.

Locuteurs	1	2	3	Total
GS avec:				
1 MP	17	30	25	72
2MP	16	24	18	58
3MP	15	10	8	33
4MP	4	2	5	11
plus de 4 MP	1	2	0	3
Total de GS	53	68	56	177

* Tout comme la segmentation de l'histoire en GS, la segmentation en un ou plusieurs mots prosodiques dépend de plusieurs facteurs, dont entre autres, le mode d'expression. En l'état actuel des connaissances, on ne peut donc pas savoir a priori combien de mots prosodiques on aura dans un GS.

Toutefois, on peut dégager certaines régularités, dans l'organisation des durées: à l'intérieur des GS, on observe toujours un principe d'alternance, soit des durées de chaque mot prosodique (comme bref-long), soit des durées de chaque pause (comme bref-long).

Peut-on prévoir la forme de l'alternance?

Nous avons regardé pour chaque GS comprenant au moins deux MP, l'organisation des durées.

Les GS avec deux MP révèlent les observations suivantes:

Il s'avère que les trois locuteurs ont tendance à produire les mêmes types d'alternance:

Tableau 3: Illustration du principe d'alternance des durées des Mots Prosodiques et des Pauses, dans des Groupes de Souffle comprenant deux Mots Prosodiques.

locuteurs	1	2	3	Total
cc/c'l'	3	4	4	11
cl/c'c'	0	1	0	1
cl/c'l'	5	7	5	17
lc/c'l'	7	9	6	22
ll/c'l'	1	3	2	6
lc/l'l'	0	0	1	1
Total				58

Deux configurations sont exceptionnelles: cl/c'c' et lc/l'l', représentées par un seul cas.

Les configurations les plus fréquentes révèlent deux phénomènes:

- si la durée du premier MP est longue, alors on a une tendance majoritaire à obtenir une double alternance: la durée du deuxième MP est courte, et les durées des pauses s'inversent (4e ligne: lc/c'l'). Dans quelques cas (~10%), la durée du deuxième MP est longue, et l'alternance s'effectue alors seulement dans l'organisation des durées des pauses (5e ligne: ll/c'l').

- si la durée du premier MP est courte, deux possibilités se présentent: soit la durée du deuxième MP est courte également, on retrouve alors le phénomène d'alternance dans l'organisation des durées des pauses (1e ligne: cc/c'l'); soit la durée du deuxième MP est longue, et l'organisation des durées des pauses est isomorphe à l'organisation des durées des MP (3e ligne: cl/c'l').

Les GS comprenant 3 mots prosodiques illustrent également une organisation alternée des durées des mots prosodiques et des pauses respectivement. Cependant l'ensemble des possibilités d'alternance est beaucoup plus importantes.

Si l'on essaie de prévoir cette alternance, il s'avère que le seul critère qui paraisse pertinent, est l'organisation des durées des deux derniers MP (de gauche à droite). En appellant "x", la durée du premier MP, "y", celle de la pause subséquente, nous obtenons le tableau suivant:

Tableau 4: Illustration du principe d'alternance des durées des Mots Prosodiques et des Pauses, dans des Groupes de Souffle comprenant trois Mots Prosodiques.

locuteurs	1	2	3	Total
xlc/yl'l'	0	1	0	1
xlc/yc'l'	6	3	3	12
xcl/yc'l'	4	3	3	10
xcc/yc'l'	3	1	1	5
xll/yc'l'	0	1	1	2
xcl/yl'l'	1	1	0	2
xll/yl'l'	1	0	0	1
Total				33

Ainsi, quelles que soient la durée du premier mot prosodique ("x") et la durée de la première pause ("y"), il semble que dans l'ensemble des combinaisons possibles, les locuteurs privilégient trois structures temporelles:

- xc/yc'1'
- xcc/yc'1'
- xlc/yc'1'

Ces trois combinaisons sont celles repérées dans les GS comprenant deux MP.

Il ne s'agit pas bien sûr d'un principe absolu, et figé, mais d'une tendance, révélant une régularité dans l'organisation de l'alternance des durées des Mots Prosodiques d'une part et des pauses subséquentes, d'autre part.

Les GS comprenant au moins quatre mots prosodiques sont plus rares (≈ 8%). Si le phénomène d'alternance est également observé, nous n'avons pas pu, en raison de la faible représentativité, dégager un principe de régularité.

CONCLUSION

Au regard des données, la durée des pauses peut être un indice de segmentation de la parole continue, en unité que nous appellons Groupe de Souffle.

A l'intérieur des Groupes de Souffle, l'organisation des segments dégagent des unités plus petites, appellées Mots Prosodiques. Les locuteurs produisent majoritairement des GS ne comprenant qu'un seul Mot Prosodique.

Dans les cas où les GS ont au moins deux Mots Prosodiques, il s'avère qu'il existe un principe d'alternance des durées. Les schèmes les plus fréquents de notre corpus sont:

{(MP1bref/MP2long- P1bref/P2long), (MP1bref/MP2bref-P1bref/P2long), (MP1long/MP2bref-P1bref/P2long)}, avec MP, mot prosodique, P, pause, 1 et 2, place des MP et des P dans le GS.

Ce principe d'alternance est d'après nous, un indice du tempo dans la parole, en comorien.

BIBLIOGRAPHIE

1. ABRY C., CREISSELS D., SONKO B., 1976. Les réalisations tonales des nominaux spécifiques en mandingue occidental: étude phonologique instrumentale et perceptive, Bulletin de l'Institut de Phonétique de Grenoble 5, 55-174.

2. BRUCE G., 1984, On the Phonetics of Rhythm: Evidence from Swedish. ASA. Meeting in Norfolk, 6-10.

3. BUTCHER A. , 1973, Aspects of the Perception and Production of Pauses is Speech. Arbeitsberichte 1, University Kiel.

4. COOPER W., PACCIA-COOPER J.M., 1980, Syntax and Speech. MIT Press, Cambridge Massachussets.

5. DUEZ D., NISHINUMA Y., 1985, Some evidence on Rhythmic Patterns of Spoken French. Perilus, Stockolm, 30-40.

6. FONAGY I., 1983, La Vive Voix, Essai de Psycho-Phonétique. Paris: Payot.

7. FRAISSE P., 1956, Les Structures Rhythmiques, Etude Psychologique. Paris: Editions Erasme.

8. GROSJEAN F., COLLINS M. , 1979, Breathing Pausing and Reading: Phonetica. 36, 98-114.

9. GROSJEAN F., GROSJEAN L., LANE H. , 1979, The Patterns of Silence: Structures in Sentence Production. Cognitive Psychology. 11, 58-81.

10. KLATT D.H., 1981, Lexical Representation for Speech Production and Perception. The Cognitive Representation of Speech, MYERS T., LAVER J., ANDERSON J., eds. North-Holland Publishing Compagny, Amsterdam, 11-31.

11. LINDBLOM B., RAPP K., 1973, Some Temporal Regularities of Spoken Swedish. Papers from the Institute of Linguistics University of Stockholm.

12. PHILIPPSON G., 1988, L'accentuation du Comorien, Essai d'analyse métrique. Etude Océan Indien, 9.

13. REY V., 1989, Approche Phonologique et expérimentale des faits d'accent en Shingazidja (Parler de la Grande Comore). Thèse Nouveau Régime, Aix en Provence.

13. ROSSI M., 1972, La perception de la durée et ses implications phonétiques, Travaux de l'Institut de Phonétique d'Aix en Provence 1, 151-166.

15. SCOTT D.R., 1982, Duration as a cue to the Perception of a Phrase Boundary: JASA . 71(4), 996-1007.

16. VAISSIERE J., 1980, La structuration acoustique de la phrase française, Annali della scuola Normale Superiore di Pisa. Série III, 10 (2), 350-360.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

UTILISATION DE REGLES PROSODIQUES EN RECONNAISSANCE DE LA PAROLE

M.K. NASRI, G. CAELEN-HAUMONT, J. CAELEN

ICP/INPG
46, Av. F. Viallet
38031 Grenoble Cedex

RESUME

Cet article décrit l'utilisation des règles prosodiques pour segmenter le signal de la parole en phrases, mots (mots lexical et mots grammaticaux) et mettre des marques sur leur frontières.

Les paramètres prosodiques (F0, énergie, durée) sont estimés sur les noyaux vocaliques. Les paramètres prosodiques sont ensuite normalisé en utilisant le codage linéaire.

Le superviseur dans un système de reconnaissance automatique de la parole est guidé par les marques prosodiques. Les règles sont inférées par l'apprentissage statistique des paramètres prosodiques, et elles sont ensuite validées linguistiquement par un expert.

1. INTRODUCTION

Le contour prosodique d'une phrase dépend de nombreux facteurs parmi lesquels:
(a) le contexte linguistique qui influe sur le style d'élocution,
(b) le type de communication (discours naturel, dialogue, texte lu, etc.),
(c) le contexte textuel à travers les modalités de phrase --énonciatif/interrogatif-- ou la structuration de l'énoncé,
(d) la variabilité inter-locuteur (facteurs socio-linguistiques, débit d'élocution),
(e) la stratégie utilisée dans la communication (répartition des accents, découpe en mots, etc.),
(f) la situation pragmatique incluant le(s) destinaire(s) du discours.

Le contour prosodique est donc très variable d'un locuteur à l'autre, pour une même situation de dialogue. Néanmoins il se réfère au système linguistique de la langue: c'est pourquoi il représente un champ d'investigation intéressant la reconnaissance automatique de la parole.

La prosodie apporte en effet, des informations à divers niveaux dans le processus de la reconnaissance:

- phonétiques, à travers la microprosodie (on sait que les voyelles ouvertes sont plus énergétiques que les voyelles fermées ou que la micromélodie des occlusives sonores est différente de celle des consonnes nasales ou que le VOT (Voice Onset Time) est un indice intéressant pour les occlusives, etc.),
- lexicales, les accents de mots ne peuvent pas être placés aléatoirement, ils dépendent en premier lieu de la position du mot dans le groupe prosodique ainsi que du nombre de syllabes,
- syntactico-sémantiques, les marqueurs sémantiques (culmination dans le processus de la focalisation du sens par exemple) ou syntaxiques (distribution des pauses, allongement de syllabes par exemple) contribuent à la ponctuation orale de la phrase sur le plan de sa structure.

Il serait donc utile de repérer automatiquement ces marqueurs en reconnaissance, soit (a) pour prédire la présence de frontières de mots ou de groupes de mots en vue de contraindre la reconnaissance, notamment les accès lexicaux, soit (b) pour vérifier l'adéquation d'une hypothèse de structure de phrase (en termes de rythme, frontières de syntagme, frontières de mots) par rapport au contour prosodique.

Plusieurs questions se posent au sujet de la prosodie et de son utilisation en reconnaissance:

1. existe-t-il des schémas prosodiques "profonds" utiles à la reconnaissance qu'un traitement adéquat des paramètres prosodiques pourrait faire apparaître ?
2. comment doit-on utiliser les informations prosodiques disponibles dans une stratégie ascendante (bottom-up) et descendante (top-down) ? Ou en d'autres termes doit-il y avoir plusieurs stratégies au sujet de la prosodie ?
3. comment faire coopérer les sources de connaissances prosodiques avec les autres sources de connaissances dans un système de reconnaissance ?

Cet article tente de répondre à ces questions en ne considérant ici que la stratégie ascendante d'analyse prosodique. A l'issue de cette étape on doit disposer d'hypothèses de frontières (mots, groupes) servant de repères au superviseur du système de reconnaissance. Le cas de la vérification (stratégie descendante) sera traitée dans un prochain article.

2. LES PARAMETRES PROSODIQUES

Pour utiliser les informations prosodiques, il convient de disposer de paramètres locaux (à court terme) et globaux (squelettes de contours à long terme) c'est-à-dire des informations caractéristiques de la structure de surface et de la structure profonde, voire d'une structure intermédiaire.

Pour la structure profonde il y a lieu en fait de distinguer une structure de phrase et une structure de texte: ce deuxième cas n'est pas envisagé dans cet article pour des textes très longs (les valeurs prosodiques retenues sont en effet normalisées phrase par phrase).

2.1. Détection des noyaux vocaliques

La structure prosodique de la phrase en français [Caelen-Haumont, 81], [Rossi, 81], [Vaissière, 83], [Bailly, 83] peut-être dégagée à l'aide des trois paramètres Fo (pitch), E (intensité ou énergie) et D (durée) calculés sur les syllabes. Mais la localisation des syllabes en analyse ascendante est une gageure: du fait qu'elles dépendent plus du niveau phonologique et lexical que du niveau acoustique, leur début et leur fin ne sont pas toujours clairement repérables et donc la mesure précise de leur durée n'est pas envisageable. Sur la courbe mélodique, l'instant de prélévement des valeurs de Fo est également sujet à discussions: par exemple, aux 2/3 de la voyelle [Rossi, 72] ou à l'instant de maximum de stabilité. En ce qui concerne la courbe d'intensité des problèmes de

mesure se posent également: (a) l'échelle (dB, linéaire, pondération perceptive, etc.), (b) la bande passante, (c) la fenêtre temporelle.

Devant tous ces niveaux de difficulté, une stratégie raisonnable consiste à localiser les centres des voyelles --qui émergent suffisamment bien sur la courbe d'intensité-- puis à s'étendre de part et d'autre pour obtenir une zone de stabilité suffisante pour le calcul des paramètres prosodiques moyens. Nous appelons ces zones "noyaux vocaliques".

Principe de détection:

Sur le plan acoustique il est possible de détecter des ruptures sur les courbes prosodiques calculées pour chaque trame du signal. Ces ruptures ne correspondent pas toujours à des frontières de syllabes mais le segment le plus intense compris entre deux ruptures peut souvent être assimilé à un noyau vocalique (fig. 1). Dans ces conditions, en regroupant deux segments successifs "intense" et "non-intense" on peut définir une unité prosodique proche de la syllabe pour laquelle la localisation acoustique ne pose pas de problèmes majeurs: nous appelons cette unité une pseudo-syllabe. Le problème est de savoir si ces pseudo-syllabes ont un sens et si elles peuvent être utiles vis-à-vis de la fonction démarcative de la prosodie.

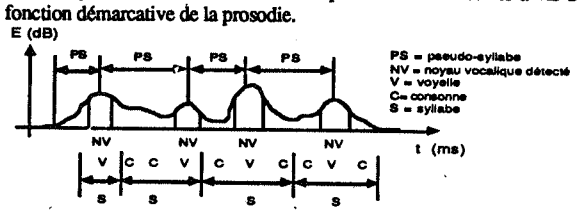


Fig. 1: Schématisation des notions de noyau vocalique (NV) et de pseudo-syllabe (PS).

On remarque que la notion de pseudo-syllabe peut s'éloigner notablement de la notion de syllabe (S) en particulier lorsque la structure syllabique devient complexe (CCV, CVC, VCC, etc.). En fait la durée de cette pseudo-syllabe est fortement corrélée au rythme des voyelles dans la phrase.

Méthode de détection des noyaux vocaliques :

Le principe de la méthode peut se résumer de la manière suivante :

1. Calcul de la densité spectrale du signal (sur des trames de 10 ms)

2. Calcul des énergies dans les trois bandes suivantes:
   a. E_V: énergie vocalique dans la bande 60-3000 Hz
   b. E_h: énergie haute fréquence dans la bande 3000-5000 Hz
   c. E_g: énergie basse fréquence dans la bande 60-900 Hz

3. Utilisation de l'énergie haute fréquence et de l'énergie basse fréquence pour une décision voisé/non-voisé (si E_h-E_g > seuil)

4. Lissage itératif --de type Hamming-- de la courbe de l'énergie E_V dans les régions de voisement (7 lissages successifs).

5. Calcul de la dérivée du 1er ordre de la courbe lissée dans les régions de voisement.

6. Recherche des frontières primaires sur la courbe de la dérivée par détection des pics passant au-dessus d'un seuil.

7. Recherche itérative des frontières secondaires:

   a. Comparaison de la durée entre deux noyaux vocaliques successifs avec la durée moyenne entre deux noyaux (cette dernière est déterminée à partir d'une étude statistique de durée entre les noyaux). Si la durée est plus grande que la durée moyenne on calcule la dérivée de la courbe énergétique la moins lissée entre ces deux noyaux vocaliques et on recherche des frontières d'un noyau (comme en 6.).

   b. Comparaison de la durée entre les frontières d'un noyau vocalique avec la durée maximale d'un noyau vocalique. Si cette durée est plus grande alors on recherche les frontières d'un noyau vocalique sur la courbe la moins lissée de la dérivée de l'énergie.

   c. Comparaison de la durée d'un noyau vocalique avec la durée minimale d'un noyau vocalique. Si cette durée est plus petite que la durée minimale alors on élimine les frontières de ce noyau.

Résultats obtenus et discussion:

Le taux de détection global des noyaux vocaliques est de 98.5% (nombre de noyaux correctement détectés/nombre de noyaux étiquetés) sur les phrases phonétiquement équilibrées de BDSON (la base de données des sons du français) et sur un 2 ème corpus réalisé par V. Auberger pour trois locuteurs (2 hommes et 1 femme). Un noyau vocalique est déclaré bien détecté lorsque l'étiquette de voyelle (posée par un expert au centre du phonème) est à l'intérieur du segment correspondant à ce noyau vocalique.

Les erreurs de détection des noyaux vocaliques peuvent être rangées en deux catégories:

a. cas de sous-segmentation
   succession de deux voyelles V/V (40% du total des erreurs)

b. cas de sur-segmentation
   succession semi-voyelle voyelle Sv/V (30%)
   les voyelles longues en fin de syntagme (10%)
   les voyelles longues en finale ayant une énergie basse (20%)

Les conséquences des erreurs de détection dans les séquences CV, VV, CCV sont moindres qu'il n'y paraît: en effet cette détection n'est faite que pour préparer la phase de squelettisation des courbes prosodiques définies ci-après et les erreurs ne sont pas cumulables avec celles du niveau suivant. On remarque que les erreurs relatives les plus importantes concernent la sous-segmentation V/V c'est-à-dire le cas où un seul noyau est détecté au lieu de deux: dans ce cas ce noyau est souvent plus long et risque de ce fait de porter un accent de durée. Par contre dans le cas de sur-segmentation d'une semi-voyelle/voyelle en deux noyaux, on peut assimiler ce problème à une diphtongue --bien qu'il n'en existe pas en français-- de manière à rattraper a posteriori ce type d'erreur. Dans les deux derniers cas de sursegmentation des voyelles longues, cela n'est pas nuisible car les deux noyaux détectés restent encore suffisamment longs pour garder une marque finale d'allongement.

2.2. Calcul des paramètres prosodiques

Les paramètres prosodiques calculés sur les noyaux vocaliques et retenus comme pertinents, sont:

NV(n) = noyau vocalique numéro n,
Ec(n) = énergie codée sur 4 niveaux,
Fo(n) = valeur de Fo sur le NV(n),
Fc(n) = fréquence fondamentale Fo(n) codée sur 4 niveaux,
D(n) = durée entre deux noyaux vocaliques,
∂D(n) = D(n)-D(n-1) (∂D < 0/∂D > 0 => accélération/ralentissement),
∂Dc(n) = ∂D(n) codée sur 16 niveaux localement dans la phrase,
DLD(n) = différence entre la ligne de déclinaison de la durée et la durée D(n),
DLDc(n) = DLD(n) codée sur 8 niveaux,
DLFc(n) = La différence entre la ligne de déclinaison de F_0 et la valeur locale F0(n). (par exemple DLF < 0 => accent mélodique),

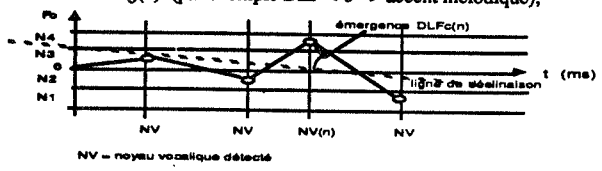


Fig. 2: La ligne de déclinaison sur un paramètre est calculée par ajustement linéaire sur les valeurs de ce paramètre sur les noyaux vocaliques successifs. Les valeurs d'émergence telles que DLFc s'en déduisent par différence. Les niveaux N1 à N4 sont normalisés d'une phrase à la suivante et répartis linéairement sur le support de variation du paramètre. Les notions de niveaux sont inspirées de Delattre [Delattre, 1966].

La visualisation de ces paramètres sous forme de courbes squelettisées conduit à la figure 3:

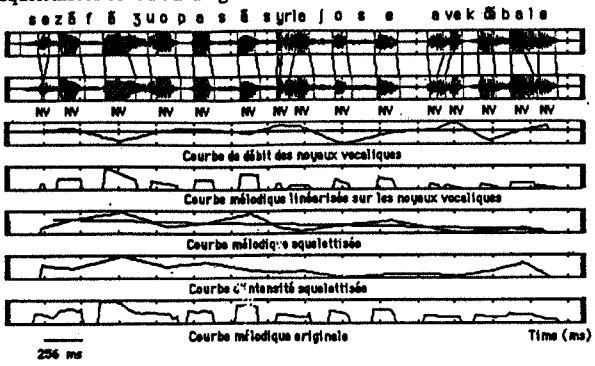


Fig. 3: Paramètres prosodiques et noyaux vocaliques de la phrase "Ces enfants jouent au passant sur la chaussée avec un balai" (phrase tirée d'un corpus de V. Auberger). De haut en bas: le signal et les frontières phonémiques détectées par un phonéticien, les noyaux vocaliques calculés, les courbes prosodiques lissées. Dans cette figure on a :

- la courbe de débit calculée, à partir des durées des pseudo-syllabes:

$$Rps(n) = (Dps(n)-Dps(n-1))/Dps(n-1).$$

Rps(n) = Le rythme dans la pseudo-syllabe numéro n.
Dps(n) = La durée de la pseudo-syllabe numéro n.

- la courbe mélodique linéarisée, calculée par l'ajustement linéaire des valeurs de F0 sur le noyau vocalique.

- la courbe mélodique squelettisée est la courbe qui relie entre elles, les valeurs moyennes de F0 sur les noyaux vocaliques, en les plaçant aux centres de ces noyaux.

La courbe d'intensité squelettisée est la courbe qui relie entre elles, les valeurs moyennes de l'énergie sur les noyaux vocaliques, en les plaçant aux centres de ces noyaux.

2.3. Signification et utilisation de ces paramètres

Il faut bien sûr se poser la question de la signification de ces paramètres puisqu'ils sont calculés sur des noyaux vocaliques délimitant des pseudo-syllabes et non des syllabes. Intuitivement, on peut penser que les accents prosodiques se distribuent principalement sur les voyelles et que donc les courbes définies ci-dessus contiennent une information prosodique pertinente. De façon moins naïve nous avons conduit une expérimentation pour mettre en correspondance ces courbes prosodiques avec des courbes de durée (réelle) de syllabes délimitées manuellement sur le signal (Fig. 4).

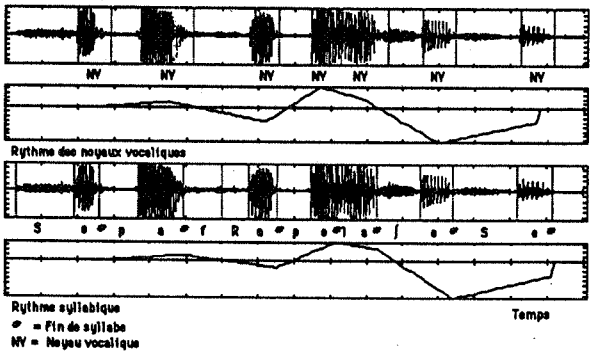


Fig. 4. On peut voir dans cette figure de haut en bas : le signal avec les noyaux vocaliques détectés automatiquement ; le rythme des noyaux vocaliques ; le signal segmenté et étiqueté à la main ; le rythme syllabique.

Malgré la différence de notion entre syllabe et pseudo-syllabe les deux courbes de durée (non pondérées) sont corrélées, d'une part parce qu'il y a proportionnellement peu de syllabes de type VC ou VCC en français et d'autre part parce que la mesure de la durée prise entre les centres des noyaux vocaliques n'est pas assujettie à la détection des frontières. La différence devient nettement plus importante toutefois, si l'on pondère la courbe réelle en fonction des durées intrinsèques des phonèmes composant la syllabe. Ce phénomène est identique pour l'intensité et la mélodie. Mais dans une stratégie ascendante il s'agit là d'un faux problème puisque le décodage acoustico-phonétique n'étant pas encore achevé il est impossible de pondérer ex nihilo les paramètres prosodiques.

On peut maintenant se poser la question de la pertinence de paramètres non pondérés. Du point de vue de la reconnaissance ascendante il est clair que si le module de décodage acoustico-phonétique ne fournit pas de résultats suffisamment fiables, la pondération des paramètres tenant compte de la nature des phonèmes ne peut être envisagée puisque ces phonèmes ne sont pas (ou mal) reconnus. Cependant cette pondération peut être envisagée dans une perspective descendante pour la vérification prosodique à un niveau profond. En outre, l'utilisation des paramètres bruts fournit un autre avantage: celui du découplage de la décision prosodique et de la décision phonétique.

La véritable question est plutôt de savoir quel est le rendement de ces paramètres prosodiques non pondérés dans une tâche de repérage de mots ou de groupes dans la phrase.

3. REPERES PROSODIQUES DANS LA PHRASE

La tâche de repérage prosodique est utile à plus d'un titre et particulièrement dans notre système (DIRA = Dialogue Intégré et Reconnaissance Automatique) de reconnaissance de dialogues oraux que nous décrivons ci-après. Précisons toutefois que le vocabulaire envisagé dans les applications de DIRA n'excède pas 3000 mots et que le dialogue est à buts finalisés (syntaxe, sémantique et pragmatique limitées).

3.1. Architecture du système DIRA

Ce système est un système multi-experts distribué organisé autour d'une architecture de blackboard. Il est contrôlé par un superviseur qui active les experts selon la progression de la reconnaissance le long de la phrase. Le superviseur est un générateur de plans dont la base de connaissances contient un ensemble de règles décrivant la méta-stratégie. Le système se présente donc comme une société d'agents gouvernés par un planificateur.

Ces agents, au nombre de quatre, communiquent leurs informations à travers le blackboard. Le superviseur se présente comme un cinquième agent qui gère le flux des données dans ce blackboard (à l'aide de sémaphores), planifie la stratégie et fixe les points de rendez-vous (points de synchronisation sur les ilôts de confiance). Les agents restent autonomes dans l'exécution de leur tâche dès que celle-ci est définie et activable.

Les agents sont:

- le Décodeur Acoustico-Phonétique (D.A.P.) qui propose (ou vérifie) des macro-traits et des traits phonétiques à partir du (ou sur le) signal d'entrée,
- l'Analyseur Lexical (A.L.) qui par des accès variés au lexique propose (ou vérifie) des mots,
- l'Analyseur Syntactico-Sémantique (A.S.S.) qui contrôle la cohérence des groupes syntagmatiques au niveau syntaxique et sémantique ou prédit le ou les prochains mots possibles,
- le module de compréhension (C.) qui contrôle les groupes de sens, gère le dialogue et construit les informations pour l'interface de communication avec l'application.

Pour activer au moment opportun ces différents agents le planificateur a donc besoin de savoir où en est la reconnaissance,

c'est-à-dire où il en est lui-même dans la phrase. Par exemple si l'on se trouve entre deux mots il posera le problème de "liaison" à ces agents ou si l'on se trouve en fin de syntagme il demandera une vérification syntaxique à son agent A.S.S. Il est donc de la première importance d'avoir des points de repères a priori dans la phrase pour guider le travail du planificateur. On peut donc dire que dans le système DIRA la prosodie guide le superviseur. C'est cette fonction que nous décrivons maintenant.

3.2. Règles prosodiques dans DIRA

Le superviseur dispose d'un module prosodique qui le guide vers quelques hypothèses sur la position des mots, sans information a priori sur la nature des mots ou des phonèmes qui composent ces mots. Ce module est écrit en Prolog et fonctionne sur une base de règles dont le but est de positionner les marqueurs suivants:

DP	= début_de_phrase
FP	= fin_de_phrase
AI	= accent_d'intensité
AM	= accent_mélodique
MG	= mot_grammatical
DML	= début_mot_lexical
FML	= fin_mot_lexical

Un marqueur est positionné à l'aide d'une étiquette et d'un coefficient de fiabilité qui peut être assimilé à une probabilité de localisation qui permettent au superviseur d'ordonner les hypothèses.

Début et fin d'une phrase:

Ph1. SI (∂D(n) ≥ 500 ms)
ALORS NV(n) <-- 'DP'
NV(n-1) <-- 'FP' + 'FML'

Commentaire: Un grand allongement de la durée entre deux noyaux vocaliques indique un début de phrase à hauteur du deuxième NV et donc la fin de la phrase précédente pour le premier NV. La fin de phrase s'accompagne en général d'une fin de mot lexical. Le cas d'une seule phrase est traité par la détection de 'EOF' (fin de fichier).

Ph2. SI (∂D(n) ≥ 300 ms) ET (Fc(n-1) < 2) ET (Ec(n-1) < 2)
ALORS NV(n) <-- 'DP'
NV(n-1) <-- 'FP' + 'FML'

Commentaire: Dans cette règle on tolère un allongement moins important que dans la règle précédente mais on impose deux niveaux bas, pour l'énergie et la fréquence fondamentale (ceci ne fonctionne que dans les cas des phrases énonciatives).

Hypothèse de présence d'un mot grammatical:

Ces règles sont applicables entre les deux frontières de la phrase 'DP' et 'FP' positionnées précédemment et permettent d'étiqueter certains noyaux vocaliques 'HMG' (Hypothèse de Mot grammatical). Les mots grammaticaux sont souvent monosyllabiques: dans le cas contraire on ne fait pas la distinction entre début et fin de mot grammatical.

MG1. SI (Fc(n)-Fc(n-1)≤-2) ET (∂D(n)<0) ET (∂D(n)-∂D(n-1) ≤ -50 ms)
ALORS NV(n) <-- 'HMG'
NV(n-1) <-- 'HFML'

Commentaire: Le fondamental baisse fortement (saut de 2 niveaux au moins) et s'accompagne d'une accélération de la durée d'au moins 50 ms: il y a présomption dans ce cas d'un mot grammatical (noté HMG) et donc le noyau précédent peut recevoir une fin de mot lexical (notée HFML).

MG2. SI (Fc(n)-Fc(n+1)≤-2) ET (∂D(n)<0) ET (∂D(n)-∂D(n+1) ≤ -50 ms)
ALORS NV(n) <-- 'HMG'
NV(n+1) <-- 'HDML'

Commentaire: C'est le même raisonnement que précédemment mais appliqué au NV qui suit le mot grammatical: on compare les paramètres des noyaux NV(n) et NV(n+1) pour positionner le début du mot grammatical et du mot lexical suivant.

MG3. SI (Fc(n)-Fc(n-1) ≤ -3) ET (∂D(n) > 0) ET (∂D(n) < 40 ms) ET (Ec(n)-Ec(n-1) ≤ -3) ET (NV(n+1)≠ 'FP' +'FML')
ALORS NV(n) <-- 'HMG'
NV(n-1) <-- 'HFML'

Commentaire: Cette règle se distingue des deux précédentes par un ralentissement accompagné d'un fort abaissement de la fréquence fondamentale et de l'énergie du noyau NV(n-1) au noyau NV(n) à condition que ce dernier ne soit pas une fin de phrase.

MG4. SI (∂DC(n) < -2) ET (∂DC(n+1) < -2) ET (DLDc(n) > 0) ET (DLDc(n+1) > 0)
ALORS NV(n) <-- 'HMG'
NV(n+1) <-- 'HMG'
NV(n-1) <-- 'HFML'
NV(n+2) <-- 'HDML'

Commentaire: Cette règle a pour but de détecter la succession de deux mots monosyllabiques grammaticaux ou d'un mot bisyllabique grammatical lorsque: il y a deux accélérations successives (∂DC(n)<0 et ∂DC(n+1)<0) avec deux durées courtes (en-dessous de la ligne de déclinaison). Ces deux noyaux sont notés 'HMG' et les noyaux adjacents reçoivent une marque de début et fin de mot lexical.

MG5. SI (DLDc(n) = 8) ET (∂DC(n) = -8)
ALORS NV(n) <-- 'HMG'

Commentaire: L'accélération est maximale et la durée extrêmement courte.

Hypothèse de fin d'un mot lexical:

La série des règles ci-après (ML1 à ML5) traite des fins et des débuts de mots lexicaux. Le cas des accents d'expressivité sera envisagé ultérieurement.

ML1. SI (∂D(n)>=40 ms) ET (DLFc(n)<0) ET (Fc(n)=4) ET (Ec(n)>=3)
ALORS NV(n) <-- 'HFML'

Commentaire: Le noyau présente un point de ralentissement (∂D(n)>0) et un point d'accent mélodique avec une énergie haute.

ML2. SI (∂D(n) < -120 ms) ET (Ec(n) = 4) ET (Fc(n) = 4) ET (DLFc(n) ≥ 0)ET((Ec(n)-Ec(n-1) ≥ 1)OU(Fc(n)-Fc(n-1) ≥ 1))
ALORS NV(n) <-- 'HFML'

Commentaire: Une accélération avec un accent d'énergie et un accent de fréquence fondamentale et une montée de l'énergie ou de la fréquence fondamentale par rapport au noyau précédent, un noyau qui émerge au-dessus de la ligne de déclinaison en fréquence indiquent un accent de fréquence: on étiquette alors le noyau vocalique comme une fin de mot lexical (cette règle est très efficace en début de phrase). Contrairement à ce qu'on pourrait attendre, la fin de mot lexical se présente ici avec une accélération du rythme: cela est dû au fait que les paramètres de durée ne sont pas pondérés --donc la dernière syllabe du mot n'est pas toujours la plus longue. De plus dans le cas où plusieurs mots lexicaux se succèdent, il peut y avoir une répartition des accents secondaires qui contrecarrent les accents primaires.

Hypothèse de début d'un mot lexical:

ML3. SI (∂D(n)<0) ET (Fc(n)=3) ET (Fc(n+1)=4) ET (Ec(n)>=3)
ALORS NV(n) <-- 'HDML'

Commentaire: Le noyau présente un point d'accélération et un point de montée de Fo avec une énergie haute.

ML4. SI (∂D(n)<-150ms) ET (Fc(n)>=3)ET(Ec(n)>=3) ET (DLFc(n)<0) ET((Ec(n)-Ec(n-1)>=1) OU (Fc(n)-Fc(n-1)>=1))
ALORS NV(n) <-- 'HDML'

Commentaire: Le noyau présente un point de forte accélération avec une

énergie forte, un Fo élevé et la valeur de Fo émerge au dessus de la ligne de déclinaison; l'énergie et la fréquence fondamentale sont plus hauts que dans le noyau précédent.

ML5. SI (∂D(n)<-180) ET (DLFc(n)<0) ET (Ec(n)=4) ET (Fc(n)=4)
ALORS NV(n) <-- 'HDML'
Commentaire: Le noyau présente un point d'accélération très fort, un point d'accent d'énergie et un point d'accent mélodique.

3.3. Résultats et discussion

La fig. 5 montre, sur trois phrases, le positionnement des marques prosodiques effectué à l'aide des règles décrites ci-dessus.

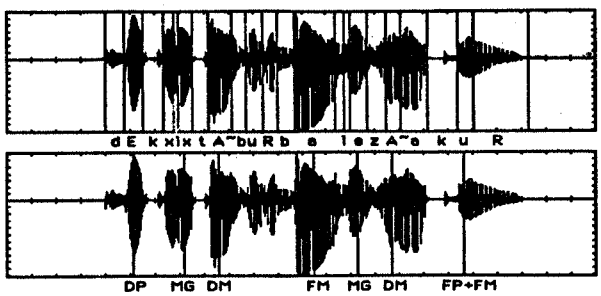


Fig. 5. Positionnement des marqueurs prosodiques sur la phrase : "des que le tanbour bat les gens accourent" (phrase extraite du corpus PEQ de BDSON). On remarque que la répartition des marques est à peu près régulière pour cette phrase.

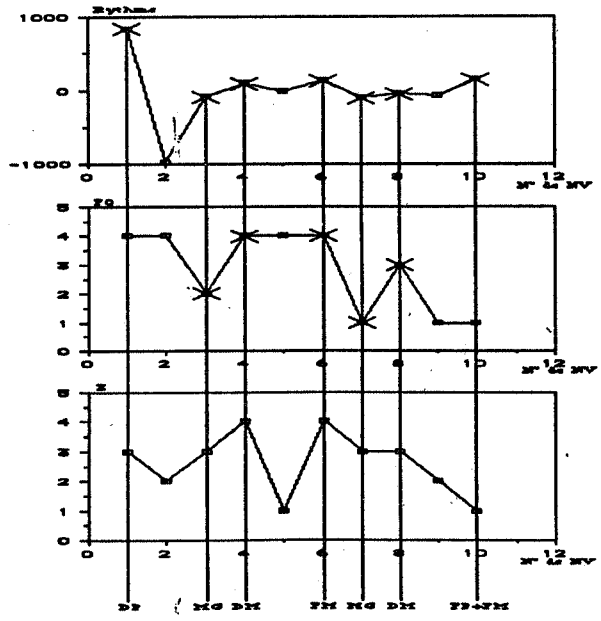


Fig. 6: cette figure présente de haut en bas les paramètres suivants : la courbe de rythme, la courbe de F0, la courbe de E.

Dans la figure (Fig.6) les paramètres qui sont responsables de l'étiquetage des noyaux vocaliques sont indiqués par une croix.

Les règles responsables de l'étiquetage sont :

DP : Ph1 ; MG : MG2 ; DM : MG2 ; FM : MG1; MG : MG1 ou MG2 ; DM : MG2 ; FP+FM : Ph1.

Ces règles ont été appliquées sur 120 phrases lues par 3 locuteurs. Les résultats sont présentés dans le tableau 1:

Etiquette	FP+FM	DP	HMG	HDML	HFML	TOTAL
Correct	120	120	150	156	174	720
Incorrect	0	0	24	24	18	66
Taux	100%	100%	86.2%	86.6%	90.6%	91.6%

Tableau 1: Résultats de détection des frontières prosodiques sur 120 phrases et 3 locuteurs.

Les valeurs indiquées ci-dessus sont des taux de confiance pour les frontières détectées. Cela ne signifie donc pas que 91.6% des frontières sont détectées, mais que parmi celles qui le sont 91.6% sont correctes.

Le taux de détection (ou taux de productivité des règles) est d'environ 40% ce qui est très suffisant pour le rôle démarcatif de la prosodie assigné au système de reconnaissance. De plus, les frontières sont réparties assez uniformément le long de la phrase (fig. 5) ce qui assure une régularité de guidage au superviseur.

4. CONCLUSION

Dans cet article nous avons présenté une méthodologie de détection de frontières prosodiques à l'aide de règles fonctionnant sur les variations des paramètres prosodiques non pondérés (rythme, Fo, Energie). Les événements que l'on peut prédire dans une stratégie ascendante, avec une bonne fiabilité sont:

1. Début d'une phrase et fin de phrase.
2. Début d'un mot lexical.
3. Fin d'un mot lexical.
4. Mots grammaticaux, monosyllabiques, bisyllabiques.

Les résultats obtenus permettent d'adopter ce type de règles (et de paramètres) dans une stratégie ascendante pour guider le superviseur d'un système de reconnaissance automatique de la parole, en lui donnant des points de repère lexico-syntaxiques dans la phrase. Entre ces points de repères on peut prédire le nombre de syllabes. Par conjonction de ces deux types d'informations on limite les lexèmes candidats au cours de la reconnaissance.

5. BIBLIOGRAPHIE

[Bailly, 83] G. BAILLY (1983)
Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération.
Thèse de Doctorat d'Ingénieur, INP Grenoble.

[Caelen-Haumont, 81] G. CAELEN-HAUMONT (1981)
Structures prosodiques de la phrase énonciative simple et étendue
Thèse de 3ème cycle, HamburgerPhonetischeBeiträge, Band 34, Hamburg Buske.

[Cooper, 75] W.E. COOPER (1975)
Syntactic control of speech timing
PH. D. thesis, MIT (unpublished).

[Cooper, 80] W.E. COOPER, J. PACCIA-COOPER (1980)
Syntax and speech
Harvard University Press, Cambridge.

[Delattre, 66] P. DELATTRE (1966)
Les dix intonations de base du français
French Review 40(1), pp. 1-14.

[Harris, 52] J. D. HARRIS (1952)
Pitch discrimination
J. Acoust. Soc. Am., Vol. 24, pp.750-755.

[Klatt, 75] D. H. KLATT (1975)
Vowel lengthening is syntactically determined in a connected discourse
J. Phonetics, Vol. 3, pp. 129 -140.

[Klatt, 76] D. H. KLATT (1976)
Linguistic uses of segmental duration in English : Acoustic and perceptual evidence
J. Acoust. Soc. Am., Vol. 59, 5, pp. 1208-1221.

[Lea, 75] LEA W.A. , MEDRESS M.F. & SKINNER T.E. (1975)
A prosodically guided speech understanding strategy.
IEEE Trans. on Acoust. Speech and Sig. Proc., ASSP. Vol. 23, L pp. 30-38.

[Lea, 84] LEA W. A., CLERMONT F. (1984)
Algorithms for acoustic prosodic analysis.
PROC. ICASSP Vol. 3, pp. 42.7.1-42.7.4.

9 PRODUCTION ET SYNTHÈSE

Président: E KELLER
UQAM-Montréal,Canada

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

UN POSTE "VISAGE-PAROLE".
Acquisition et traitement de contours labiaux.

Med-Tahar LALLOUACHE

Institut de la Communication Parlée. URA CNRS 368
INPG/ENSERG, Université STENDHAL, BP 25 X
38 040 GRENOBLE Cedex FRANCE

RESUME

Pour l'étude de la production de la parole - et en vue de la synthèse audio-visuelle - on connaît toute l'importance de la saisie simultanée des paramètres physiologiques et du signal de parole correspondant. Notre travail a consisté à mettre en place une chaîne d'acquisition et de traitement de labio-films vidéo couleur. L'extraction automatique de huit paramètres de la face et du profil nous a déja permis d'exploiter systématiquement ces mesures pour l'évaluation de modèles en production et perception de la parole. A partir de ces différentes validations, une base de données visage-parole est en cours de constitution en vue de la synthèse **à visage parlant** pour les sons du français.

1. INTRODUCTION

L'intérêt de l'analyse des informations faciales dans la parole n'est plus à démontrer, que ce soit pour la perception visuelle [1,2], ou l'animation du visage [3,4], et tout particulièrement pour la synthèse audio-visuelle [5,6,7] ; sans compter les bénéfices attendus pour la reconnaissance de la parole [8].

Les procédures d'acquisition et de traitement d'images que nous proposons, et dont nous avons testé la faisabilité, entrent dans le cadre d'un ensemble de recherches valorisant ces divers aspects [Abry & Schwartz in 9]. Procédures et traitements sont plus spécifiquement orientés vers la constitution d'une **base de données visage-parole**.

2. POUR L'ACQUISITION COULEUR

2.1. Un bilan d'expériences acquises

Une longue expérience de l'acquisition d'images labiales par enregistrement noir et blanc (35 mm ou vidéo) nous a permis d'en mesurer les limites [10,11]. L'extraction automatique de contours géométriques est dans ce cas basée sur un seuillage après traitement de l'image achrome. Or, **aucun éclairage, maquillage, filtrage n'a jamais pu supprimer les tâches erratiques dues à la brillance sur fond sombre ou à des ombres portées sur fond clair**. La nécessité d'utiliser l'obturateur des caméras nous force à employer un éclairage relativement violent et de ce fait nous empêche de prévoir la solution d'un éclairage diffus.

Ainsi des lèvres maquillées en blanc impliquent un noircissement des dents [10,11]. On ne peut donc dans ce cas éviter l'un des deux "bruits" suivants : brillance par réflexion sur les dents (surtout due à la salive), lors d'un éclairage de face direct; ou présence d'ombres portées par les lèvres, surtout en protrusion, lors d'un éclairage de haut et/ou de bas.

A l'inverse l'utilisation d'un maquillage noir pour les lèvres nécessite une image intéro-labiale de forte intensité lumineuse; mais l'éclairage indispensable introduit inévitablement des ombres dans cet espace intéro-labial, lors de certaines productions vocaliques ouvertes.

Dans les deux cas une détection automatique des contours internes du vermillon nécessite un traitement **ad hoc** dépendant du visème enregistré et donc n'est ni systématique ni fiable à 100 % : une vérification d'expert est toujours requise.

Ceci est inacceptable dans la mesure où l'on entreprend une étude dynamique des mouvements labiaux, avec un nombre important d'images à traiter (15 minutes d'enregistrement vidéo correspondent à 45.000 trames en PAL). Dans le même temps, ce qui reste contraignant est la **précision** des mesures exigées : on ne peut tolérer que des erreurs inférieures à 0.5 mm. Ceci est important pour l'identification des stimuli visuels, quand on sait qu'une modification de l'ordre du mm en protrusion des lèvres, entre deux stimuli, peut conduire à une bascule quasi-complète de l'identification [+rond] en [-rond] [12].

Nous avons pu montrer par ailleurs que les mesures manuelles [11,14,15] entraînaient une erreur non négligeable, tout spécialement sur les petites aires aux lèvres, pour lesquelles on connaît la sensibilité articulatoire, acoustique et visuelle.

C'est pour cet ensemble de raisons que nous avons envisagé et retenu les possibilités offertes par une acquisition **couleur** du matériau test. Bien entendu, il va de soi que nos arguments en faveur de la couleur ne sont pas inspirés par des considérations qui ne relèveraient que de l'esthétique (avec cependant des implications concrètes pour les tests de perception visuelle). Nos arguments sont donc essentiellement fondés sur les contraintes d'un traitement finalisé : celui des images vidéo du visage parlant vu de face - et on verra que le traitement de profil n'impose pas le même jeu de contraintes.

Dans le cas où l'image d'une scène est colorée, il est certain que le paramètre luminance n'est pas le plus significatif. L'exploitation de l'information de chrominance s'avère d'emblée plus puissante. On a ainsi pour chaque point trois paramètres (teinte, saturation, luminance) au lieu d'un, ce qui rend évidement plus sélective la recherche de l'information désirée.

Il existe la possibilité d'enregistrer les images, puis de les traiter ultérieurement sur carte couleur. C'est ce que nous n'envisageons pas pour l'instant, vu la faible bande passante du canal chrominance sur les magnétoscopes standards.

Notre solution sera d'extraire, en temps réel à partir d'un signal RVB, provenant d'une caméra, l'information désirée. puis de la transposer dans le canal luminance - la bande passante du signal luminance étant 4 à 5 fois plus large que celle du signal chrominance - afin de l'enregistrer sur

magnétoscope. Cette solution conserve une haute qualité à l'information désirée. Ces objectifs sont réalisables avec l'utilisation du **chroma-key**.

2. 2. Le chroma-key

Ce générateur de découpage couleur universel, utilisé généralement dans les unités de programmation vidéo, est conçu pour obtenir divers effets spéciaux, tels que le découpage des signaux couleurs, le découpage externe, l'effacement, etc.

Dans le mode de fonctionnement par découpage de signaux couleurs, une couleur spécifique appelée "couleur à découper" est extraite d'une image ("image à découper") et une autre image appelée "image à insérer" est incrustée dans la zone d'où la couleur à découper a été extraite. La couleur à effacer peut se choisir à volonté dans une palette. En outre un arrière-fond de couleur simple, produit par le générateur de couleur incorporé peut être choisi comme "image à insérer". La teinte, la saturation et la luminance de cette couleur de fond peuvent être réglées de façon continue.

Pour avoir un découpage très précis, il est impératif que la caméra utilisée soit de très haute résolution et possède les sorties RVB nécessaires. Rappelons que ces signaux contiennent une plage de fréquence, en chrominance, beaucoup plus large que celle du signal vidéo-composite.

En principe n'importe quelle couleur peut être choisie pour la zone de découpage. Mais celle qui est utilisée pour les sujets vivants est le bleu. La raison en est que les tons de la peau ne contiennent virtuellement pas de bleu (ce qui n'est pas le cas pour le vert et le rouge) : dès lors ils sont reproduits sans distorsion au cours du découpage.

3. LA PRISE D'IMAGES EN PAROLE

3.1. La prise de vues

La prise de vues (fig. 1) est effectuée dans des conditions optimales d'éclairage et de maquillage avec des points de repère cutanés. Les lèvres sont en bleu, les dents ne sont plus noircies.

Le fait de ne plus noircir les dents (outre la suppression d'une gêne pour les sujets) est un autre avantage de la prise de vue couleur, puisqu'on connaît l'importance des dents dans les tests de perception visuelle [16]. Pour conserver ces images de référence, nous utilisons, lors de l'enregistrement, deux magnétoscopes U-MATIC SP. La bande enregistrée via le **chroma key** servira pour les mesures ; la seconde enregistrée directement à partir des deux caméras (face et profil) sera utilisée pour les tests perceptifs. On verra par la suite qu'il est aisé de faire la correspondance temporelle entre les images des deux bandes.

Le locuteur est installé dans une chambre sourde, à égale distance des deux caméras (1.5 m). Sa tête est fixée dans un casque de spéléologue, celui-ci étant solidaire du dossier de la chaise. Ce casque peut être réglé en hauteur et profondeur. Le sujet est porteur d'une monture de lunettes fixées derrière la tête au moyen d'une bande velcro ; ainsi les lunettes deviennent solidaires des légers hochements de tête, encore possibles lors de la phonation. Ces légers mouvements, qui subsistent une fois la tête maintenue, sont en fait de quelques mm. Mais ces mm sont, on le sait, précieux dans la mesure de la protrusion, car la dynamique de celle-ci, dans la parole, n'est que d'une dizaine de mm tout au plus (et lorsqu'il s'agit de sujets hyper-articulateurs !). Par conséquent, sur la branche droite de la monture est fixée une réglette qui servira de référence aux mesures de profil.

La prise de vues proprement dite se fait au moyen de deux caméras tri-CCD, synchronisées, présentant une résolution supérieure à 500 lignes TV, sur chaque canal (R,V,B) et un rapport S/B de 55 dB. Ces caméras sont munies de zooms (x10) et d'obturateur au 1/250ème, 1/500ème et 1/1000ème de seconde.

Les sorties R,V,B de la caméra prenant la face sont dirigées vers le **chroma key**. La couleur des lèvres est découpée et un signal de luminance nulle lui est substituée. Pour qu'il ne subsiste, après incrustation, aucune ambiguité entre le niveau de gris des lèvres et d'éventuelles ombres à l'intérieur de la fente labiale, le niveau de noir de la caméra est légèrement relevé. Les lèvres à la sortie du **chroma key** paraissent ainsi noires et sans relief. Cette image est mélangée (avec un effet de volet vertical) à l'image provenant de la caméra profil, puis enregistrée sur un premier magnétoscope.

Sur le second magnétoscope, les deux images, face et profil, sont enregistrées sans modification.

Les deux bandes vidéos sont ensuite "time-codées" grâce à un générateur de code image. Celui-ci permet d'insérer, dans les lignes vidéo de la suppression trames, un numéro à chaque image. Ce numéro peut être aussi incrusté, en mode graphique, dans un coin de l'image, pour le repérage visuel.

3.2. La synchronisation son-image

Parallèlement à l'enregistrement vidéo, le son est acquis grâce à deux microphones placés à 20 cm du sujet. Un des deux microphones est relié directement au canal audio 1 des deux magnétoscopes; le second passe par le synchronisateur. Le rôle de ce dernier est d'envoyer sur le canal 2 un bip (3 kHz sur la durée de 2 trames), déclenché conjointement, par un bouton-poussoir et par un front montant de la synchro-trames. Ce bouton poussoir est actionné manuellement avant chaque phrase porteuse. En synchronisme avec le bip, un pavé de LED (temps de réponse de quelques µs) est allumé dans le champ de l'une des caméras.

Ce double repérage son et image permet à un éditeur de signal de regénérer la synchro-trames et, connaissant le numéro de la première trame où les LED sont allumées , de déduire le numéro de n'importe quelle trame correspondant à une position temporelle dans le fichier son.

4. LE POSTE VIDEO-PAROLE

Ce poste est dédié à l'étude des images du visage parlant. Son but est donc en premier lieu l'analyse automatique des images vidéo, celles du visage pour l'instant. En second lieu, ce poste sert aux tests de perception audio-visuelle, ainsi qu'aux doublages audios.

4.1. L'unité de traitement

Il s'agit d'un Goupil G40 doté d'un disque dur de 72 MO (compatible IBM PC AT). L'interface signal vidéo-PC est constitué par une carte MATROX (PIP 1024) entièrement programmable (4 plans mémoire 512x512x8 bits, 256 niveaux de gris, numérisation temps réel) . Un logiciel (PC SCOPE vers. 2.O) pilote la carte et permet de disposer d'une bibliothèque de fonctions de base en traitement d'images. L'interface signal audio-PC est constituée par une carte OROS (AU 20). Le signal est échantillonné à 16 kHz (16 bits).

Les magnétoscopes sont pilotés par le PC via les deux ports RS 232 à travers une carte d'interface. Cette interface permet, en plus du pilotage de toutes les fonctions du magnétoscope, de relire le code de chaque image et de le transmettre au PC.

Les images à traiter sont repérées en fonction du signal de parole grâce à un éditeur spécialement conçu (inspiré sur certains points d'EDISIG [17]). Une fois les images à traiter sélectionnées, il suffit de donner, au logiciel de gestion du magnétoscope et de traitement, le numéro de la première et de la dernière image, pour que le processus d'acquisition et de traitement démarre. Celui-ci traite 100 images en 12 minutes. Ce temps comprend durée du traitement et durée de repositionnement de la bande vidéo (en fait 50% du temps). Ce dernier temps pourrait être réduit considérablement par l'emploi d'un lecteur vidéo-disque, compatible avec le poste.

4.2. Le traitement

Le signal vidéo est numérisé à partir du magnétoscope (sortie N/B, qui supprime les bruits de la voie de chrominance). Les images sont saisies à la volée (3 de suite : en tout 6 trames).

Le logiciel extrait les trames paire et impaire de chaque image, et pour chaque trame les lignes manquantes sont calculées par interpolation. Ce qui nous donne des images 512x512 pixels.

Puis on procède automatiquement aux mesures latérales et frontales (fig. 2).

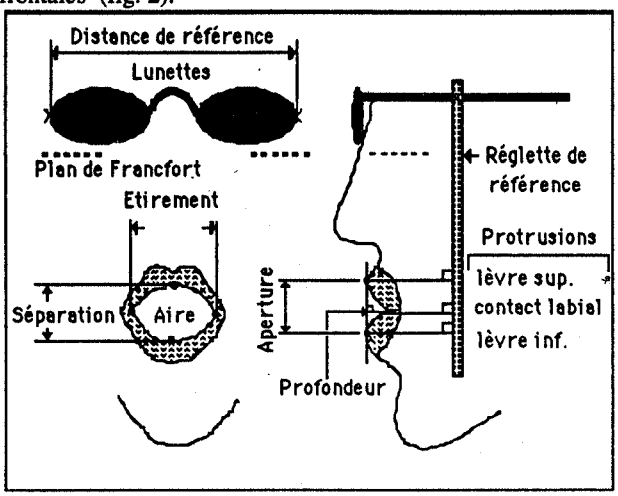


Fig. 2 : Mesures latérales et frontales.

De face :
A : l'étirement intéro-labial ; **B** : la séparation intéro-labiale ; **S** : l'aire intéro-labiale ;
De profil :
P1 : la protrusion de la lèvre supérieure ; **P2** : la protrusion de la lèvre inférieure ; **C** : la position du contact labial ; **D** : l'aperture extéro-labiale (ou aperture du pavillon labial) ; **L** : la profondeur du pavillon labial.

4.2.1. Le traitement de face

Il est basé sur un seuillage global, l'image étant pratiquement binaire à la sortie du **chroma-key** : l'histogramme de la fenêtre d'analyse (encadrant la bouche) vue de face est explicite (Fig. 3).

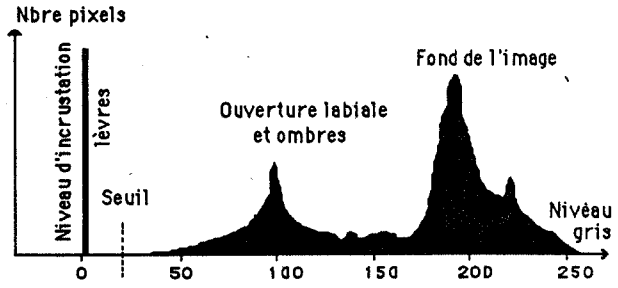


Fig. 3 : Histogramme de la fenêtre d'analyse cadrant la bouche.

Sur l'image seuillée on applique, par morphologie mathématique, une transformation d'ouverture (érosion suivie d'une dilatation par un élément structurant symétrique). Le rôle de cette transformation est de lisser les contours. Un suivi de contour en image binaire [18] permet d'extraire les contours interne et externe et de les sauvegarder sous forme de chaînes de caractères (codage de Freeman [19]). Une seconde étape consiste à extraire les points pertinents du contour interne. Les deux points des commissures des lèvres, qui sont les deux points les plus éloignés du contour et permettent de mesurer l'étirement (A). L'aperture (B), qui est mesurée à partir du centre de gravité de la fente, méthode qui tient compte de l'asymétrie de la forme (cf. fig. 4). Enfin l'aire (S) est mesurée par un simple comptage de pixels.

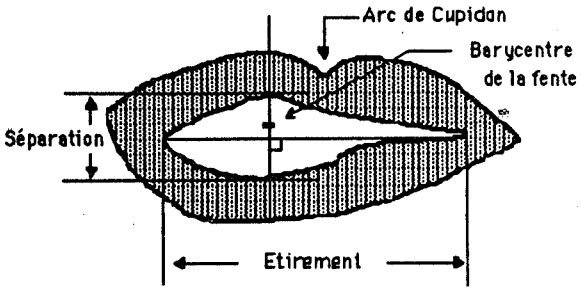


Fig. 4 : Détermination des paramètres de face sur un cas, parmi d'autres, de formes labiales assymétriques.

4.2.2. Le traitement de profil

Du moment que l'espace intéro-labial (lequel est si changeant qu'il nous a donné le plus de fil à retordre en prises de vues pour traitement) n'est pas visible sur l'image de profil, et comme l'éclairage nécessaire est moindre, la détection sera basée sur un gradient de Sobel, qui nous donne les meilleurs résultats [20]. Cette détection sera suivie d'une stratégie de poursuite de contour, laquelle aura aussi pour effet de les amincir (en effet les opérateurs gradients ne garantissent pas des contours minces).

Une seconde étape sera la détection des points , pour nous, intéressants sur le contour de profil : les deux protrusions, qui sont les deux points du vermillon, les plus éloignés de la réglette ; le contact labial qui en est le plus rapproché (cf. fig. 2). Pour avoir des mesures précises, on prend comme référence non pas la réglette mais son axe d'inertie principal dont l'épaisseur est de 1 pixel (rappelons que la protrusion maximale est de 9 mm pour notre sujet).

5. PRECISION DES MESURES

Pour nos premiers essais, nous avons relevé nos différentes précisions de mesures (moyennes). Pour l'étirement labial, la précisison est de 3% (direction de la mesure dans le sens horizontal, là ou l'on a le maximum de résolution). Pour la séparation, elle est de 10%. Pour l'aire, de 4%. Enfin pour les mesures de profil (tous les points sont mesurés dans la même direction), notre précision est de 3%. Evidemment, toutes ces précisions dépendent du grossissement choisi lors de la prise de vue.

Ces résultats sont éloquents, surtout si on les compare à nos précédents résultats acquis avec une seule caméra N/B, sans obturateur (sans compter que, pour le profil, un miroir avait été utilisé, avec les erreurs induites par ce procédé [11]).L' imprécision des mesures pouvaient atteindre 40 % surtout quand les gencives du sujet étaint visisbles.

L'imprécision sur la séparation s'explique en grande partie par le fait que celle-ci est verticale (c.-à-d. dans le sens où la résolution de l'image est de moitié), et parce qu'il est difficile de maquiller la muqueuse intérieure des lèvres (visible lors de la réalisation de sons protrus comme [ʃ]). Un fort grossissement, est donc nécessaire pour les petites fentes labiales : il faut donc au maximum grossir les lèvres dans le cadrage choisi.

5. RESULTATS

Les premiers résultats obtenus sur un corpus, constitué avec un locuteur (sélectionné après de nombreux essais), ont été utilisés dans la mise en œuvre d'un test de perception, pour évaluer les apports des informations auditives et visuelles (pour ces résultats cf. [21]).

La précision temporelle de nos mesures semble permettre largement de mettre en évidence les phénomènes remarquables, comme l'anticipation, dans la parole.

Ainsi, est-il possible de lire, sur la figure 5, l'initiation du changement d'aire aux lèvres de la voyelle [e] vers la voyelle [ø] et ceci au moins dès le second tiers de l'émission acoustique de la voyelle non arrondie. On notera que ce phénomène d'anticipation - déjà observé chez le même locuteur pour la transition des voyelles hautes [i] -> [y] - n'est pas symétrique, puisque l'initiation du changement d'aire dans [y] -> [i] ne se produit qu'à partir de la fin de l'émission acoustique de la voyelle arrondie[21].

Plus remarquable encore est le décours des événements pour la séquence [ikstsky]. On y lit clairement sur la protrusion supérieure (P1) - alors que l'inférieure (P2), la séparation (B), l'étirement (A) et l'aire (S) semblent davantage dépendants des positionnements de la mandibule pour produire les dentales [s,t] - que l'initiation du geste vocalique se produit juste avant le début du segment qui précède la frontière (de mot et/ou de syllabe), comme dans la séquence précédente monoconsonantique [edø]. Ce type de résultat - reproduit sur plusieurs locuteurs - pourrait remettre en cause le modèle **look-ahead** proposé jusqu'à présent pour le français [22; cf.23, pour une évaluation récente des différents modèles d'anticipation].

Dans l'évolution des paramètres contribuant au contrôle de l'aire aux lèvres, on remarquera l'aspect plutôt "créneau" des gestes d'étirement (A), alors que le contrôle de la séparation (B), est beaucoup plus "mouvementé". On notera portant que les irrégularités observées sur ce paramètre se retrouvent régulièrement sur les autres répétitions de cette séquence, aux mêmes instants de mise en place des [s] et [t]. La séparation est le paramètre qui est le mieux corrélé avec l'évolution de l'aire aux lèvres (S). La relation du produit de l'étirement et de la séparation à l'aire (S=k.A.B) donne statistiquement des valeurs autour de 0.75 pour le coefficient k (cf. fig. 6); ce qui rejoint, sur des échantillons de 1.5 s environ, les valeurs données en moyenne, dans la littérature, pour les positions cibles des lèvres [24,25,26].

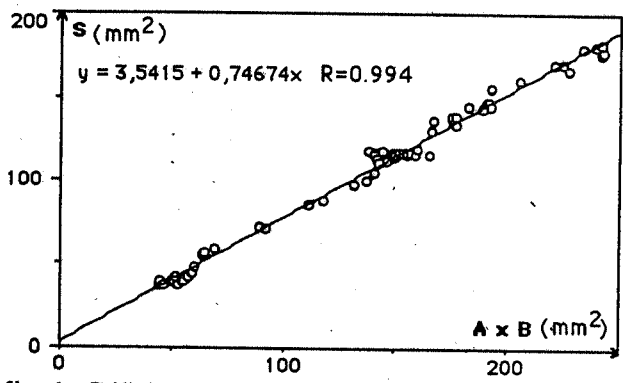


fig. 6 : Prédiction de l'aire à partir de l'étirement et de la séparation des lèvres

REMERCIEMENTS

Tout d'abord à nos locuteurs qui ont accepté de se laisser maquiller et bronzer sous 1000 watts halogènes : Danièle Larreur, Marie Cathiard et Jean-Luc Schwartz. A Christine Delattre et Argyro Tseva pour leur talent de maquilleuses; à Alain Arnal pour son indispensable concours technique, son efficacité et sa disponibilité; à Jean-Pierre Charras du LTIRF qui nous a fait bénéficier de son savoir en traitement d'image; enfin à Christian Abry, Louis-Jean Boë, Pierre Escudier et Jean-Marc Dolmazon qui ont tout fait pour la mise en place de ce poste "Visage-Parole".

[1] **DODD B. CAMPBELL & R. (1987) Eds.** Hearing by eye : the psychology of lipreading. Lawrence Erlbaum Associates, London.

[2] **MASSARO & D.W. (1987)** Speech perception by ear and eye : a paradigm for psychological inquiry. Lawrence Erlbaum Associates, London.

[3] **PARKE F.I. (1974)** A parametric model for human faces. Doctoral Dissertation, University of Utah.

[4] **MAGNENAT-THALMANN (1989)** The problematics of facial animation. In N. MAGNENAT-THALMANN & D. THALMAN (Eds), State-of-the-art in computer animation, Proceedings of computer animation '89. Springer-Verlag.

[5] **PARKE F.I. (1975)** A model for human faces that allows speech synchronized animation. Computer and Graphics, 1, 3-4

[6] **BROOKE N.M. & SUMMERFIELD Q. (1983)** Analysis, synthesis and perception of visible articulatory movements. Journal of Phonetics, 11, 63-76.

[7] **NAHAS M. HUITRIC H. & SAINTOURENS M. (1988)** Animation of a B-Spline figure. The Visual Computer, 3, 272-276.

[8] **BROOKE M. & PETAJAN E.D. (1986)** Seeing speech : investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. Conference Publication n° 258, Inter. Conf. on Speech Input/Output; Techniques and Applications (24-26 March).

[9] **CATHIARD M.A. (1988/1989)** La perception visuelle de la parole : aperçu de l'état des connaissances. Bull. Inst. Phonétique de Grenoble, 17-18, 109-193.

[10] **ABRY C. BOE L.-J. CORSI P. DESCOUT R. GENTIL M. & GRAILLOT P. (1980)** Labialité et phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales. Publication de l'Université des Langues et Lettres de Grenoble.

[11] **LALLOUACHE M.T. (1987)** Détection automatique du contour des lèvres et extraction des paramètres constitutifs. Rapp. D.E.A. ENSER Grenoble.

[12] **CATHIARD M.-A. (1988)** Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français. Mémoire de maîtrise, Labo. de Psychologie, Grenoble II.

[14] **TOUSSIGNANT B. (1979)** Détermination acoustique du conduit vocal et modélisation du mouvement des lèvres. CNET, note technique DAS/ETA/69.

[15] **LALLOUACHE M.T. & WORLEY C. (1988)** Saisie édition et traitement d'images et de signaux articulatoires : lèvres et mâchoire. Journal d' Acoustique, 1, 215-220.

[16] **McGRATH M. SUMMERFIELD Q. & BROOKE N.M. (1984)** Roles of lips and teeth in lipreading vowels. Proceedings of the Institute of Acoustics, 6, 401-406.

[17] **BENOIT C. (1984)** EDISIG : encore un éditeur de signal ?!! 13èmes JEP-GCP du GALF, 211-213.

[18] **ROSENFELD A. & KAK A. (1976)** Digital Picture Processing. Academic Press.

[19] **FREEMAN H. (1961)** On the encoding of arbitrary geometric configurations. IEE Trans. on Electronic Computers, EC-10, 260-68.

[20] **BASSEVILLE M. (1979)** Détection de contours : méthodes et études comparatives. Ann. Télécomm., 34, n° 11-12, 559-579.

[21] **ESCUDIER P. BENOIT C. & LALLOUACHE M.T.** (1990) Identification visuelle de stimuli associés à l'opposition [i]/[y]. 1er Congrés de la Soc. française d'acoustique, Lyon (10-13 avril).

[22] **BENGUEREL A.P. & COWAN (1974)** Coarticulation of upper lip protrusion in French. Phonetica, 30, 41-55.

[23] **PERKELL J.S. (1989)** Testing theories of speech production : implications of some detailed analysis of variable articulatory data. Nato Advanced Study Institute "Speech Production and Speech Modelling", Bonas, France (17-29 juillet), preprint [à paraître].

[24] **FROMKIN V.A. (1964)** Lips positions in American English vowels. Language and Speech, vol. 7, 215-225.

[25] **LINDBLOM B.E.F & SUNDBERG J.E.F. (1971)** Acoustical consequences of lip, tongue, jaw and larynx movements. J. Acoust. Soc. Am., Vol. 50, 1166-1179.

[26] **ABRY C. & BOE L.-J. (1986)** Laws for lips. Speech Communication, 5, 97-104.

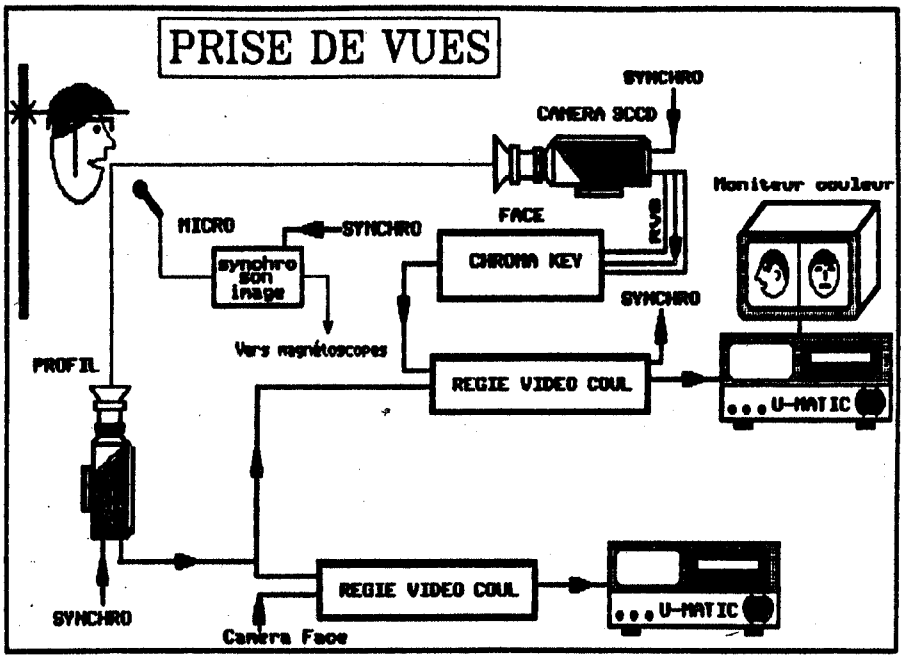


Fig. 1 : Dispositif de prises de vues.

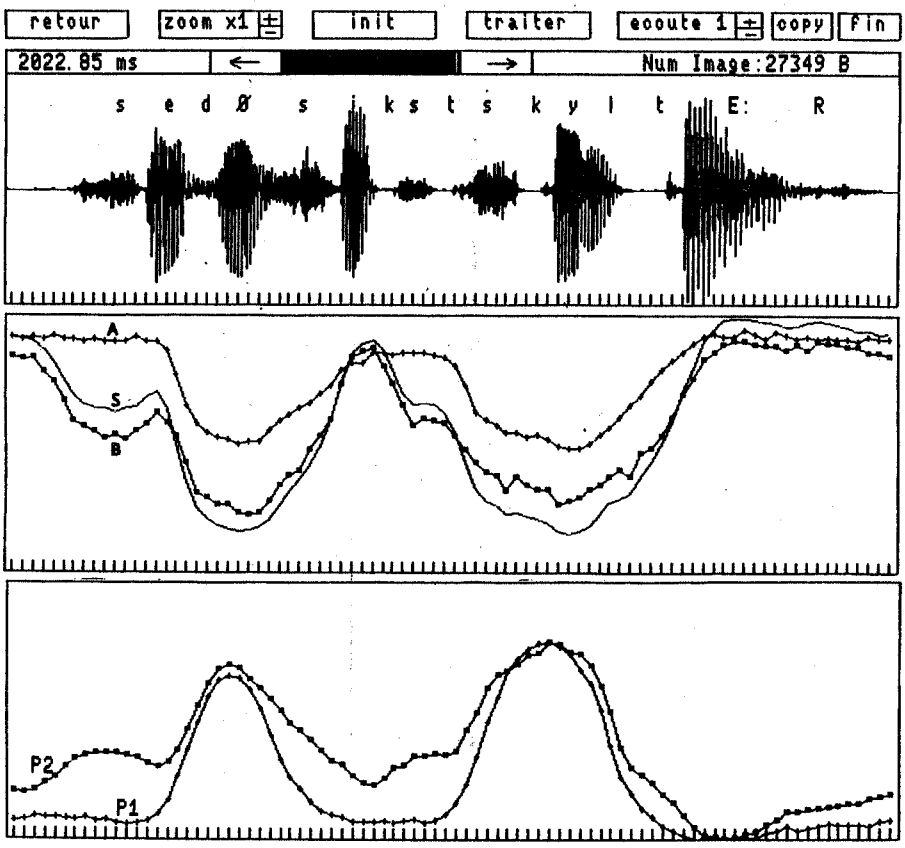


Fig. 5 : Visualisation de 5 fonctions temporelles extraites automatiquement sur la géométrie des lèvres : **A** (étirement), **B** (séparation), **S** (aire), **P1** (protrusion sup.), **P2** (protrusion inf.). Le signal de parole choisi est une réalisation de "Ces deux Sixtes sculptèrent", séquence qui offre, dans le français non-marqué de ce locuteur, la possibilité de produire une suite de 5 consonnes entre [i] et [y].On notera la relative simplicité du geste de protrusion de la lèvre supérieure (**P1**) et les dates de son déclenchement (cf. texte).
N.B. : Les graduations temporelles (20 ms) correspondent à la synchro-trames regénérée. Le num(éro) im(age) est celui de la première trame ici visualisée, la durée affichée en ms correspondant à sa date par rapport au bip de référence.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Relations entre les trois premiers formants et la géométrie du conduit vocal

Bernard Delyon, François Delyon

IRISA, campus de Beaulieu, 35042 RENNES Cedex, France
Centre de Physique Théorique, Ecole Polytechnique, 91128 Palaiseau, France

Résumé

On propose ici une étude de la production des voyelles où se trouvent mêlés des aspects quantitatifs et qualitatifs. Les voyelles sont séparées en trois groupes acoustiques, selon les modes de vibration qui entrent en jeu, puis chaque groupe est étudié séparément.

Abstract

A two-step scheme is proposed for relating vocal-tract shapes and formant frequencies. Vowels are split into three acoustic groups and each group is studied separately. This approach combines qualitative and quantitative considerations.

1 Introduction

Dans les vingt dernières années, de nombreux auteurs ont essayé de construire des "formules d'inversion" permettant de passer directement du spectre d'un signal vocalique à la fonction d'aire du conduit vocal (voir la synthèse de G.Fant in [3]). Considéré sous l'angle mathématique, il est bien connu que ce problème contient de nombreuses singularités ([8]); elles expliquent fort bien les défauts des modèles proposés, motivés par un souci unificateur et refusant l'introduction de connaissances a priori. Nous avons été conduits à traiter ce problème comme une étude de cas (trois configurations possibles pour les voyelles orales). De plus, nous avons abandonné l'ambition (peu réaliste) de revenir à une fonction d'aire mais plutôt d'obtenir une vision qualitative des phénomènes (même si cette dernière pourrait facilement conduire à des fonctions d'aires réalistes et par chance peu vérifiables).

On présente ici une étude des voyelles françaises par le biais de trois modèles de conduit vocal illustrés par la figure 1 (les deux derniers sont basés sur des idées classiques ([1] p.53), différentiables à la simple vue de la répartition des trois premiers formants:

1. Perturbation du tuyau droit (/ɛ/, /œ/, /ø/, /a/, /ɔ/):
   les trois premiers formants sont quasiment en progression arithmétique (1,3,5).
2. Résonateur simple (/i/, /e/, /y/):
   le premier formant est bas et le second élevé.
3. Résonateurs couplés (/u/, /o/):
   les deux premiers formants sont bas.

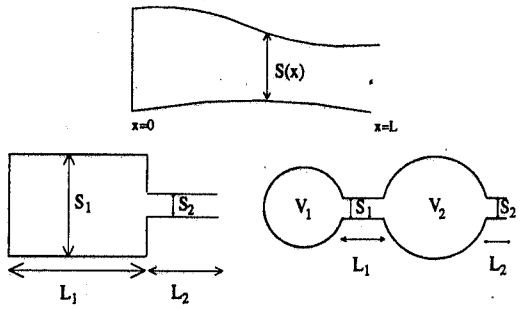


Figure 1 : L_i=longueur du tube i, S_i=section du tube i, V_i=volume de la cavité i

2 Perturbation du tuyau droit (/ɛ/, /œ/, /ø/, /a/, /ɔ/)

On considère dans cette section le cas où le conduit vocal peut être considéré comme "proche" du tuyau droit, au sens où si $S(x)$ désigne la section du conduit vocal en x ($x = 0$ à la glotte et $x = L$ aux lèvres) alors $\frac{d}{dx}\log(S(x))$ est petit, ce qui revient en gros à dire que S reste du même ordre de grandeur le long du conduit vocal. On utilisera une méthode simple de petites peturbations.

Le système differentiel vérifié par la pression et la vi-

tesse le long du conduit vocal, pour une onde de fréquence f est ([6]):

$$\begin{aligned} \frac{d}{dx}(v(x)S(x)) &= kS(x)p(x) \\ \frac{d}{dx}(p(x)) &= -kv(x) \end{aligned} \quad (1)$$

où

x = abscisse le long du conduit vocal

k =nombre d'onde $= 2\pi f/c$ (c = vitesse du son),

pression $= \rho\, c\, p(x)\cos(\omega t)$ (ρ = densité de l'air, $\omega = 2\pi f$)

vitesse $= v(x)\sin(\omega t)$

Nous allons considérer la nouvelle variable α, l'angle entre la pression et la vitesse:

$$\tan(\alpha) = \frac{v}{p} = \frac{-1}{kp}\frac{dp}{dx} \quad (2)$$

Un calcul facile permet d'obtenir

$$\frac{d\alpha}{dx} = k + \frac{S'}{2S}\sin(2\alpha); \quad (3)$$

le n-ième formant f_n est la fréquence résonante pour laquelle la pression (ou la vitesse, ou α) s'annule n fois le long du conduit vocal; il est caractérisé par les conditions de bords (relations de résonance: vitesse nulle à la glotte et pression nulle aux lèvres):

$$\alpha(0) = 0 \quad (4)$$

$$\alpha(L) = (n - 1/2)\pi \quad (5)$$

En intégrant entre 0 et L l'équation (3), on obtient pour le n-ième formant:

$$(n - 1/2)\pi = k_n L + \frac{1}{2}\int_0^L \frac{S'}{S}\sin(2\alpha)dx. \quad (6)$$

L'approximation au premier ordre quand S'/S est petit conduit à remplacer $\alpha(x)$ par sa valeur dans le cas du tube droit ($S' = 0$), c'est-à-dire par $k_n^0 x$ (où $k_n^0 = (n - 1/2)\pi L$), et l'on obtient, après une integration par parties, la formule approximée:

$$\begin{aligned} \log(f_n) - \log(f_n^0) &\simeq \int_0^1 \log(S(xL))\ \cos((2n-1)\pi x)dx, \\ f_n^0 &= (2n-1)cL/4. \end{aligned} \quad (7)$$

Cette équation exprime une relation approchée entre le déplacement des formants et la transformée en cosinus du logarithme de la fonction d'aire (voir aussi [7]). Il faut noter que $\log(S)$ est soumis à des variations bien moins importantes que S. Il est clair que l'information récoltée ici est tout à fait insuffisante pour reconstruire le conduit vocal, puisqu'on obtient la transformée en cosinus pour les fréquences impaires seulement; il est toutefois possible de calculer le produit scalaire de $\log(S)$ avec certaines fonctions, combinaisons linéaires de ces cosinus, en faisant les mêmes combinaisons linéaires des logarithmes de formants.

En utilisant les trois premiers formants, on peut donc obtenir les produits scalaires p_1, p_2, p_3 de $\log(S)$ par les fonctions $\cos(x)$, $\cos(3x) - \cos(x)$, $6\cos(x) + 5\cos(3x) + 3\cos(5x)$, $(0 < x < \pi)$, representées par la figure 2.

Ces trois fonctions ont été choisies de sorte que p_1, p_2, p_3 représentent respectivement les indices ouvert/fermé, avant/arrière, labialisé/non-labialisé, et on a:

$$p_1 = -\log(f_1/f_1^0)$$
$$p_2 = -\log(f_2/3f_1)$$
$$p_3 = -6\log(f_1/f_1^0) - 5\log(f_2/f_2^0) - 3\log(f_3/f_3^0)$$

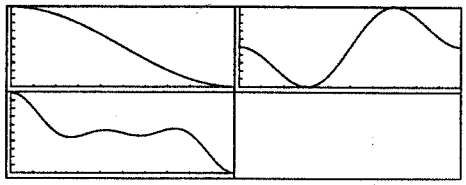


Figure 2 : fonctions $\cos(x)$, $\cos(3x) - \cos(x)$, $6\cos(x) + 5\cos(3x) + 3\cos(5x)$, $(0 < x < \pi)$.

On obtient alors les plans (p_2, p_1) et (p_3, p_1) des figures 3 et 4. Notons que le trait avant/arrière semble mieux représenté par $\log(f_2/3f_1)$ que par l'habituel $\log(f_2)$, et que la valeur 0 se trouve bien à sa place. Le second plan est à mettre en relation avec le trait bémolisé/diésé (cf [4] p.161).

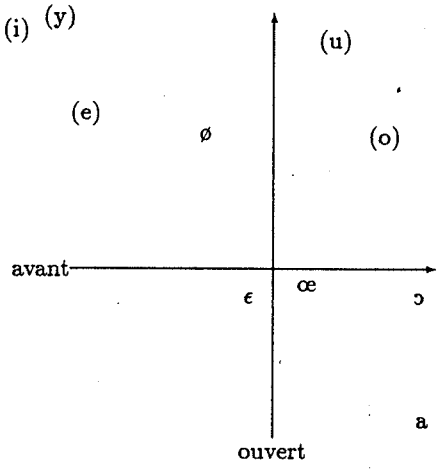


Figure 3 : plan (p_2, p_1) pour les valeurs de formant tirées de [5] et f_1^0=500Hz; on a également représenté les voyelles qui ne font pas partie du modèle.

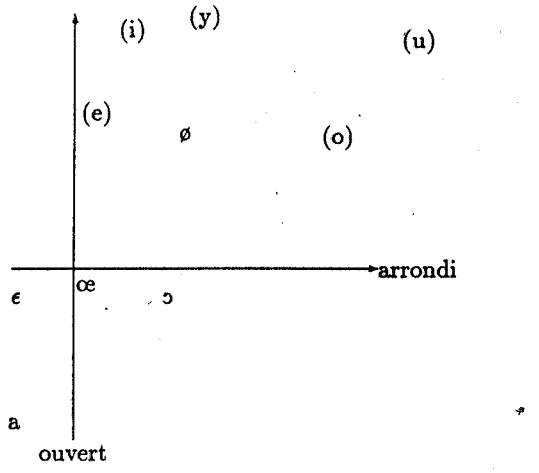


Figure 4 : plan (p_3, p_1) pour les valeurs de formant tirées de [5] et f_1^0=500Hz; on a également représenté les voyelles qui ne font pas partie du modèle.

3 Résonateur simple (/i/, /e/, /y/)

On se réfère à la figure 1. S_1 est supposée grande devant S_2 et que L_1 est supérieure à L_2. En raison du saut de section, l'équation pour l'angle α possède une singularité en $x = L_1$ et on a:

$$\frac{d\alpha}{dx} = k \quad \text{dans un tuyau} \tag{8}$$

$$\frac{\tan(\alpha(L_1^+))}{\tan(\alpha(L_1^-))} = \frac{S_1}{S_2} \tag{9}$$

à la jonction des deux tuyaux.

La deuxième équation résulte simplement de la continuité du débit et de la pression. Ces équations, écrites pour f_1, f_2, f_3, donnent les trois équations:

$$\tan(k_i L_1)\tan(k_i L_2) = \frac{S_2}{S_1}, \quad i = 1, 2, 3. \tag{10}$$

On peut alors déduire les trois inconnues L_1, L_2, $\frac{S_1}{S_2}$ par une méthode de Newton. On obtient en fait des résultats quasiment identiques avec l'heuristique très simple suivante:

la figure 5 montre le parcours typique de l'angle α (et même, du vecteur $(p(x), v(x))$, $x = 0, .., L$), pour les trois premiers formants.

Il convient de remarquer comment le saut de tangente projette le vecteur (p, v) sur l'un des deux axes dans le cas des deux derniers formants: il y a un point très proche de la jonction qui se trouve soit à vitesse nulle (pour f_2), soit à pression nulle (pour f_3).

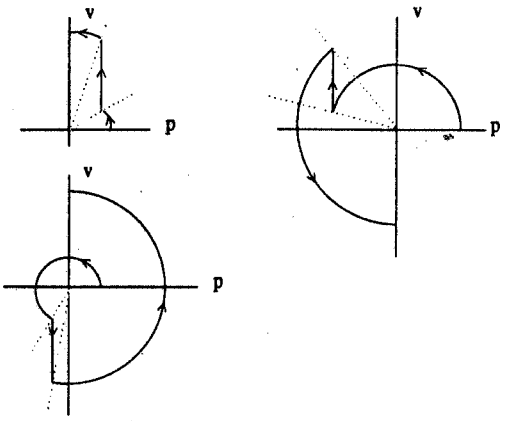


Figure 5 : parcours du vecteur $(p(x), v(x))$ pour les formants f_1, f_2, f_3 le long du conduit vocal, dans le cas du résonateur simple.

Ceci signifie, en d'autres termes que f_2 et f_3 correspondent à la résonance propre de chacun des tuyaux (fermé ou ouvert aux deux bouts), tandis que f_1 est la basse fréquence du résonateur de Helmholtz. D'où:

$$\begin{aligned} L_1 &\simeq \frac{c}{2f_2} \\ L_2 &\simeq \frac{c}{2f_3} \\ \frac{S_1}{S_2} &\simeq \frac{1}{\tan(k_1 L_1)\tan(k_1 L_2)} \end{aligned} \tag{11}$$

Le calcul fait pour les valeurs moyennes des formants masculins tirées de [5] pour les trois voyelles /y/, /i/ et /e/ donne les résultats mis dans le tableau 1.

	/y/	/i/	/e/
L_1	9,4	8,2	8,4
L_2	7,8	5,5	6,2
S_1/S_2	3,5	6	3,3

Tableau 1 : valeurs de L_1, L_2 (en cm), S_1/S_2 (cf fig1) obtenues pour les voyelles associées au résonateur simple.

Le tableau 2 suivant donne les valeurs des coefficients d'écart entre les formants masculins et les formants féminins ($K_i = f_i(homme)/f_i(femme)$; valeurs tirées de [5]); on observe bien que les formants de "longueur" f_2 et f_3 ont un K_i proche de 1,2 tandis que le formant de Helmholtz f_1 est plus stable (ce phénomène est moins net pour le /e/, voyelle qui se rapproche du premier modèle):

	/y/	/i/	/e/
K_1	1,02	0,99	1,14
K_2	1,14	1.19	1,2
K_3	1,2	1,14	1,18

Tableau 2 : valeurs des coefficients d'écart entre les formants masculins et feminins pour trois voyelles.

4 Résonateurs couplés (/u/, /o/)

Ne seront ici considérés que les deux premiers modes de vibration, associés à f_1 et f_2. C'est un cas où ces deux fréquences sont assez basses et l'on supposera la pression constante dans chacun des résonateurs et la vitesse constante dans les tubes. En les notant p_i et v_i, on a les relations classiques (loi de Newton et conservation de la masse):

$$\begin{aligned} L_1\rho\frac{dv_1}{dt} &= p_1 - p_2 \\ L_2\rho\frac{dv_2}{dt} &= p_2 \\ -\frac{V_1}{c^2}\frac{dp_1}{dt} &= S_1 v_1 \rho \\ -\frac{V_2}{c^2}\frac{dp_2}{dt} &= S_2 v_2 \rho - S_1 v_1 \rho \end{aligned} \tag{12}$$

Après élimination des vitesses, on obtient:

$$\frac{d}{dt^2}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = M \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \tag{13}$$

et

$$M = 4\pi^2 \begin{pmatrix} -(q_1V_1)^{-1} & (q_1V_1)^{-1} \\ (q_1V_2)^{-1} & -(q_2V_2)^{-1} - (q_1V_2)^{-1} \end{pmatrix} \tag{14}$$

où

$$q_i = \frac{4\pi^2}{c^2}\frac{L_i}{S_i} \tag{15}$$

Les valeurs propres de cette matrice sont les carrés des pulsations ($\omega_i = 2\pi f_i$), changés de signe, si bien que

$$f_1^2 + f_2^2 = \frac{1}{q_1V_1} + \frac{1}{q_1V_2} + \frac{1}{q_2V_2} \tag{16}$$

$$f_1^2 f_2^2 = \frac{1}{q_1V_1q_2V_2} \tag{17}$$

La dernière équation suggère de définir un coefficient global de constriction (élevé si les volumes sont grands et les constrictions fortes):

$$c = -\log(f_1) - \log(f_2) \tag{18}$$

On peut également obtenir en faisant le quotient de l'équation (16) par la racine de (17):

$$\frac{f_1}{f_2} + \frac{f_2}{f_1} = \left(\frac{q_2V_2}{q_1V_1}\right)^{1/2} + \left(\frac{q_2V_1}{q_1V_2}\right)^{1/2} + \left(\frac{q_1V_1}{q_2V_2}\right)^{1/2}. \tag{19}$$

Posons

$$\lambda = \left(\frac{q_2V_2}{q_1V_1}\right)^{1/2} \tag{20}$$

alors

$$\frac{V_1}{V_2} = f(\lambda) \tag{21}$$

où

$$f(\lambda) = \frac{1}{\lambda}\left(\frac{f_1}{f_2} + \frac{f_2}{f_1}\right) - \frac{1}{\lambda^2} - 1 \tag{22}$$

λ étant inconnu, la valeur la plus vraisemblable de $\frac{V_1}{V_2}$ est celle pour laquelle $f'(\lambda) = 0$. On trouve finalement

$$\frac{V_1}{V_2} \simeq \left(\frac{f_1}{f_2} - \frac{f_2}{f_1}\right)^2/4 \tag{23}$$

On définira le coefficient de dissymétrie:

$$\begin{aligned} d &= \log\left(\left(\frac{f_2}{f_1} - \frac{f_1}{f_2}\right)^2/4\right) \\ &= \log\left(\left(\frac{f_2}{f_1} + \frac{f_1}{f_2}\right)^2/4 - 1\right) \end{aligned} \tag{24}$$

ainsi nommé par référence aux équations (23) et (19).

En utilisant les formants moyens tirées de [5] on obtient pour c et d les valeurs données dans le tableau 3. Comme on pouvait l'espérer, c a tendance à discriminer (/u/, /o/) de /ɔ/ tandis que d discrimine /u/ de (/o/, /ɔ/), ce qui confirme bien l'interprétation des paramètres.

Notons qu'il est aisé de vérifier à partir de (13) que f_1 correspond à un mode de résonance où les deux cavités vibrent en phase ($p_1 > 0, p_2 > 0$), tandis que f_2 est associé à un mode avec opposition de phase ($p_1 > 0, p_2 < 0$).

c	F	H
/u/	1,4	1,4
/o/	0,9	1,2
/ɔ/	0,3	0,6

d	F	H
/u/	0,2	0
/o/	-0,9	-0,5
/ɔ/	-0,8	- 0,8

Tableau 3 : valeurs des coefficients c et d pour trois voyelles et les deux sexes. On a ajouté ɔ bien qu'il ne fasse pas partie du modèle.

5 Conclusion

L'approche proposée ici a consisté relier les *caractéristiques* géométriques et acoustiques du conduit vocal par l'introducton a priori de trois modèles simplifiés. On voit

ainsi que l'on peut obtenir de manière assez élementaire des résultats expliquant la présence des différents formants et les modes de vibration associés. Ces résultats sont en assez bon accord avec les données de [2].

Références

[1] L.J.BOE, *Introduction à la Phonétique Acoustique*, Institut de Phonétique de Grenoble, 1972

[2] A.BOTHOREL, P.SIMON, F.WIOLAND, J-P.ZERLING, *Cinéradiographie des Voyelles et Consonnes du Français*, Institut de Phonétique de Strasbourg, 1986.

[3] G.FANT, *The relations between area functions and the acoustic signal*, Phonetica, 1980.

[4] J.S.LIÉNARD, *Les processus de la communication parlée*, Masson, 1977.

[5] F.LONCHAMP, *Les sons du Français*, Institut de Phonétique de Nancy.

[6] J.D.MARKEL ET A.H.GRAY, *Linear Prediction of Speech*, Springer, 1976.

[7] MERMELSTEIN, *Determination of The Vocal-Tract Shape from Measured Formant Frequencies*, J.Acoust.Soc.Am., vol.41, May 1977.

[8] *The Quantal Nature of Speech*, Journal of Phonetics, vol. 17,number 1/2, Jan./April 1989.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

ANTICIPATION ET RETENTION DANS LES MOUVEMENTS VOCALIQUES DU FRANCAIS

Mhania GUERTI et Gérard BAILLY

Institut de la Communication Parlée INPG / ENSERG - Université STENDHAL
Unité Associée au CNRS n°368 46, avenue Félix Viallet 38031 Grenoble cedex

ABSTRACT

The present study concerns the acoustic effects of the reduction of vocalic gestures (undershoot) in order to produce coarticulation rules for a synthesis-by-rule system. We present here the data obtained on vocalic gestures V_1V_2 without any consonantal perturbation.

We defined acoustical macrosensibility functions as the average deviation suffered by a certain vowel from its intrinsic formantic values (standard vocalic triangle of the speaker). These functions depend on the acoustic distance between the two vowels at the first order. However, our data evidence some non-linearities due to the underlying articulatory gestures which explain the dissymetry between some forward assimilation macrosensitivity functions and backward ones (preparation of the gesture versus inertia of the articulators).

We conclude with a perception test which evidences the need of the modelization of the coarticulation phenomenon even in case of simple vowel-to-vowel movements.

INTRODUCTION

Les voyelles sont traditionnellement décrites en termes de caractéristiques spectrales statiques, souvent référencées sous le terme de cibles acoustiques. Ainsi la représentation de ces cibles dans l'espace de leurs 2 ou 3 premiers formants constitue une topologie de référence très utile pour l'étude comparative des langues d'un point de vue phonologique [1]. Cependant si la structure de ces triangles vocaliques reste relativement invariante, elle subit des translations, homothéties, voire rotations importantes entre locuteurs. De plus, ces cibles sont souvent non atteintes, lorsque ces voyelles sont coarticulées avec des consonnes dans la parole naturelle. Ce phénomène est connu sous le nom de réduction de transitions formantiques "undershoot" [2], [3]. Son extention dépendra de variables multiples tels que la vitesse d'articulation, le contexte consonantique, les stratégies individuelles...

La présente étude concerne l'effet acoustique de cette réduction des gestes articulatoires en vue de son utilisation dans un système de synthèse par règles du Français [4]. En effet, si la synthèse par unités stockées permet de rendre compte d'une manière intrinsèque de ces phénomènes sans pouvoir en assumer les déformations dues au débit (degré de coarticulation constant), il reste à construire des modèles de coarticulation suffisamment sophistiqués afin d'en rendre compte aisément par règles.

Nous exposons les données obtenues sur des gestes vocaliques V_1V_2 (sans perturbations consonantiques). Ces derniers font apparaître une tendance générale à une influence acoustique proportionnelle à la distance spectrale inter-cibles mais présentent aussi des non-linéarités importantes dues souvent au geste articulatoire sous-jacent.

1. MOTIVATIONS

Dans deux études complémentaires: étude spectrographique et modélisation de fonctions d'aires [5], [6] Öhman met en évidence la nature très complexe du phénomène de la coarticulation dans des logatomes VCV (C= occlusive sonore) pour le Suédois et l'Anglais Américain et l'existence de deux contrôles gestuels du conduit vocal : geste consonantique superposé à un substrat vocalique. Confirmé par d'autres études [7], ce phénomène permet au locuteur de concilier anticipation (préparation du geste) et rétention (inertie, économie du geste) en maintenant les traits perceptifs pertinents. Cette négociation locuteur-auditeur se traduit d'un point de vue acoustique par une assimilation régressive et / ou progressive des configurations vocaliques adjacentes. Reste à quantifier les importances relatives de ces phénomènes en fonction du contenu spectral des voyelles concernées et du débit.

Nous présentons ici une étude des déformations du substrat vocalique en fonction du contenu spectral à débit relativement constant. Ces déformations seront représentées grâce à des diagrammes de MSA: macro-sensibilités acoustiques (fig.1).

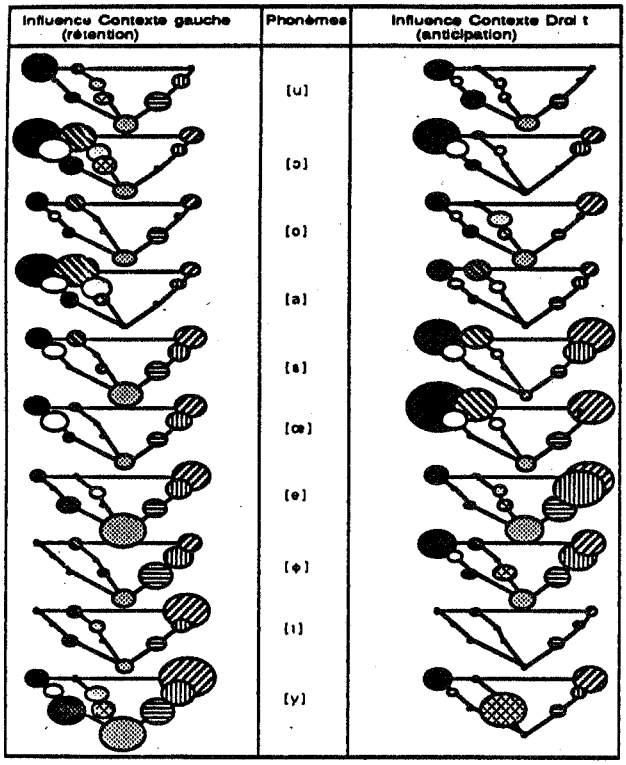


Figure 1: Les MSA des voyelles orales Françaises

2. PROCEDURE EXPERIMENTALE

2.1. Corpus

Le corpus est constitué de diphones V_1V_2 (voyelles orales du Français) d'un dictionnaire de diphones à formants [8]. Ces diphones sont extraits de logatomes porteurs, maintenant un contexte constant et relativement neutre d'un point de vue articulatoire et acoustique[$\phi C_1V_1V_2C_2\phi$] avec C_1, C_2 = [p] ou [t] selon le caractère labialisé ou non de la voyelle adjacente. Nous avons au total 10*10=100 réalisations. Nos résultats vont donc porter sur un locuteur professionnel ayant reçu pour consigne de garder un débit constant et une intonation neutre.

2.2. Méthodologie

Une première détection des formants a été effectuée manuellement sur des tracés spectrographiques. Puis un alignement automatique a permis d'ajuster finement des suivis sur un treillis de candidats obtenus par analyse LPC à 16 coefficients. Certaines courbes ont été post-éditées à l'aide d'une tablette d'entrée graphique liée à un calculateur.

3. RESULTATS ET MODELISATION ACOUSTIQUE DE LA COARTICULATION

3.1. Hypothèses de travail.

Afin d'étudier comparativement les phénomènes d'assimilation progressive et régressive (notées respectivement: A_p et A_r), nous supposons que les deux phénomènes sont additifs et se réalisent par rapport à une référence fixe prise égale à la cible intrinsèque (notée I), ceci pour un débit suffisamment lent (cibles non réduites). De plus, l'influence est supposée fonction de la configuration formantique du contexte vocalique. Ce qui nous conduit à représenter les cartographies suivantes:

Anticipation (influence de V_2 sur V_1):

$$F(V_1) = fA_r [F_I (V_1) , F_I (V_2)]$$

Rétention (influence de V_1 sur V_2):

$$F(V_2) = fA_p[F_I (V_2) , F_I (V_1)]$$

Nous avons estimé tout d'abord les configurations formantiques intrinsèques des cibles acoustiques $F_I(v)$. Ensuite nous proposons une évaluation qualitative et quantitative des fonctions fA_r et fA_p.

3.2. Détermination des cibles acoustiques intrinsèques

Les fréquences formantiques cibles du locuteur ont été déterminées par analyse statistique [9] et corrigées, si nécessaire, manuellement afin de respecter la structure phonologique du Français (tab.1).

Phonèmes	Valeurs intrinsèques			Coefficients Anticipation									Coefficients Rétention								
				F1			F2			F3			F1			F2			F3		
	F1	F2	F3	α1	α2	α3	α1	α2	α3	α1	α2	α3	α1	α2	α3	α1	α2	α3	α1	α2	α3
[a]	622	1370	2390	0.05	0.00	0.13	-0.06	0.24	-0.07	-0.01	0.01	-0.01	0.13	-0.09	0.21	-0.13	0.07	0.00	-0.04	0.00	0.14
[i]	225	1995	2961	-0.01	-0.03	-0.02	0.01	-0.02	0.00	-0.04	0.09	-0.23	0.21	0.22	0.15	0.01	-0.02	-0.08	0.01	-0.05	0.07
[u]	230	752	2108	0.17	0.00	-0.39	0.00	0.02	0.12	0.02	0.02	0.06	0.04	-0.02	-0.06	-0.24	0.06	0.42	0.06	-0.05	0.19
[y]	230	1650	2069	0.12	0.02	-0.09	0.03	0.24	0.29	0.02	0.06	-0.04	0.22	0.27	-0.20	0.08	-0.17	0.55	0.01	0.01	-0.01
[e]	320	1879	2421	0.20	-0.13	0.11	0.11	0.13	0.07	0.03	0.16	0.22	0.24	0.02	0.18	0.00	-0.11	0.16	0.08	-0.02	0.33
[ɛ]	440	1697	2332	0.18	-0.07	0.25	-0.15	-0.05	0.24	-0.02	0.15	0.27	0.14	-0.11	0.44	-0.11	-0.02	0.18	-0.05	0.13	0.10
[ø]	330	1454	2156	0.29	0.19	-0.33	-0.07	0.16	0.02	0.05	0.01	0.27	0.01	0.15	-0.12	-0.05	0.12	-0.04	0.07	-0.10	0.19
[œ]	439	1419	2313	0.42	0.06	-0.26	-0.13	0.09	0.37	0.06	0.05	0.11	0.07	0.12	-0.03	-0.09	0.20	0.08	0.02	0.06	0.16
[o]	337	797	2260	0.13	-0.02	0.66	-0.08	-0.05	0.26	0.03	0.05	-0.11	-0.03	-0.01	0.26	-0.15	0.01	-0.04	0.08	0.00	0.23
[ɔ]	456	990	2453	0.18	0.06	-0.26	0.02	0.15	0.17	0.09	-0.03	-0.32	-0.04	0.24	-0.12	-0.13	0.05	0.07	0.08	0.06	0.00

Tab.1: Valeurs formantiques intrinsèques et coefficients αk d'assimilation régressive et progressive.

3.3. Macro-sensibilités acoustiques

Afin de vérifier le comportement qualitatif du modèle de coarticulation décrit plus haut, nous avons estimé des valeurs de macrosensiblités acoustiques des voyelles en fonction du contexte vocalique à partir de mesures de déviations de leurs réalisations formantiques par rapport aux valeurs intrinsèques. La mesure adoptée est la suivante :

$$Msa(V) = \frac{\sum_{j=1}^{3} \frac{|F_j(V)-F_{Ij}(V)|}{F_{Ij}(V)}}{3}$$

3.4. Commentaires

Pour chaque voyelle contextuelle considérée, un cercle de diamètre proportionnel à la MSA, est centré sur la cible correspondante (cf.fig. 1).

Ces cartographies permettent une vue d'ensemble des comportements des différentes voyelles en fonction du contexte. Nous allons les commenter brièvement afin de vérifier que les classes de comportement sont en accord avec une classification articulatoire. Nous retenons essentiellement trois caractéristiques acoustico-articulatoires: labialisation, aperture et position de la constriction linguale.

3.4.1. tendances générales

L'examen général des MSA confirme les études précédentes [10], soit une tendance à la proportionnalité des MSA à la distance inter-cibles: plus les deux points d'articulation sont loin est plus la déformation sera importante. Cela peut se traduire soit par un tir court [9] ("undershoot": trajectoire acoustique plus courte que celle réunissant les cibles) soit par un tir long ("overshoot") dépendamment de la topologie de la correspondance articulatoire-acoustique. Ces phénomènes sont accentués lors d'un débit plus rapide [11]. Cependant, la voyelle coarticulée doit rester assez proche de la cible acoustique pour qu'il n'y ait pas confusion avec une autre voyelle. Nos MSA montrent la très grande stabilité de la voyelle [i], probablement due à la proximité de [y] et [e] (sauf pour le [u], car la baisse seule du F_2 n'est pas suffisante pour provoquer la confusion avec [y] ou [e]).

Les voyelles périphériques sont plus influentes que les voyelles centrales ce qui semble exclure a priori un contrôle articulatoire basé sur un conduit neutre. Ce phénomène est maximal pour la voyelle [i].

3.4.2. Non-linéarités

Dans le cas des réalisations de la coarticulation anticipatrice de [a] et [ø], il est intéressant de constater que le caractère le plus influent est [+fermé]. Cette influence décroît proportionnellement au degré d'aperture. Ces voyelles auront tendance à diminuer le F_1 de [a] et [ø]. Cette dernière est résistante à l'influence de [œ], car elles peuvent être considérées toutes deux comme ouvertes. Dans le cas de la rétention, [a] et [ø] subissent les influences conjointes de l'antériorité [+avant] et de la fermeture [+fermé]. En ce qui concerne l'influence du degré d'aperture, le comportement est identique au précédent mais ici, [a] et [ø] sont très sensibles à l'antériorité de la voyelle précédente.

4. MODELISATION

Cette étude suggère que les phénomènes de coarticulation acoustiques obéissent à des contraintes de type articulatoire (minimisation des distances) et perceptif (respect des distances perceptives) [12]. Nous proposons un modèle de coarticulation permettant au système perceptif de calculer les cibles intentionnelles à partir de la connaissance des cibles intrinsèques tout en maintenant les hypothèses d'influence contextuelle non linéaires. La configuration acoustique réalisée est obtenue par la relation suivante:

$$Log(F_j(V)) = Log(F_{Ij}(V)) + \sum_{k=1}^{3} \alpha_{kAt}(Log(F_{Ij}(V_{CAt})-Log(F_{Ij}(V)))$$

avec At=Ar ou Ap

Les coefficients α_{kAt} ont été déterminés algébriquement en minimisant l'écart quadratique moyen entre données réelles et modèle. Celui-ci conduit à une erreur moyenne de 1.5% alors que les modèles sans coarticulation ($F_j(V)=F_{Ij}(V)$) et les linéaires (sans Log) donnent respectivement 7% et 3%.

Diphone	F1 Calc	F1 Reel		F2 Calc	F2 Reel		F3 Calc	F3 Réel	
aa	622	620	(0.32,0.32,-0.2)	1370	1409	(2.80,2.80,1.4)	2390	2351	(1.66,1.66,-0.8)
ia	550	537	(2.39,15.72,-12.6)	1601	1612	(0.63,15.01,-17.3)	2562	2550	(0.49,6.27,-7.0)
ua	561	588	(4.59,5.78,-7.6)	1489	1456	(2.25,5.94,7.5)	2444	2408	(1.50,0.75,1.5)
ya	520	489	(6.48,27.20,-18.8)	1575	1546	(1.86,11.41,-13.9)	2436	2391	(1.91,0.04,1.0)
ea	555	599	(7.34,3.75,-7.2)	1524	1489	(2.37,8.02,-10.0)	2457	2440	(0.70,2.07,-2.5)
ɛa	580	596	(2.77,4.27,-5.4)	1453	1495	(2.80,8.39,-7.6)	2414	2501	(3.46,4.44,2.8)
ɸa	557	554	(0.55,12.27,-10.7)	1491	1525	(2.22,10.16,10.1)	2415	2482	(2.68,3.73,2.5)
œa	588	594	(0.97,4.71,-5.0)	1435	1475	(2.67,7.12,6.2)	2412	2434	(0.90,1.81,1.4)
oa	595	578	(2.93,7.52,-5.6)	1424	1478	(3.64,7.34,6.0)	2430	2439	(0.36,2.03,1.9)
ɔa	618	612	(0.93,1.55,-1.1)	1392	1363	(2.12,0.48,0.6)	2428	2496	(2.69,4.25,3.0)
	2.9	8.3	ya [-18.8]	2.3	7.7	ia [-17.3], ɸa [10.1]	1.6	2.7	ia [-7.0], ɔa [3.0]

dc la plus influente a la moins influente

i [12.27] y [11.26] ɸ [7.75] e [6.55] u [5.55] ɛ [5.30]

o [4.49] œ [4.19] ɔ [1.56] a [0.81]

Tab.2: Présentation des résultats d'optimisation pour le [a] en A_p (assimilation progressive).

Des résultats d'optimisation pour le [a] en assimilation progressive Ap ont été présentés (tab.2). Pour chaque formant du [a], les valeurs calculées et mesurées sont figurées. Les trois nombres entre parenthèses représentent l'erreur quadratique: du modèle, celle des données mesurées et cibles

intrinsèques et enfin le pourcentage d'undershoot ou d'overshoot (nombres négatifs ou positifs) de la cible. La dernière ligne montre l'erreur quadratique moyenne des données fournies par le modèle, celle donnée par les cibles intrinsèques, ainsi que les voyelles qui réalisent les maxima de déviation. Suit un classement de la voyelle la plus influente à la moins influente avec les MSA correspondantes.

5. TEST PERCEPTIF

Pour évaluer la pertinence d'une telle modélisation, nous avons voulu effectuer un premier test global d'intelligibilité en comparant deux stimuli de synthèse, l'un généré à partir des cibles intrinsèques, l'autre prenant en compte l'effet de la coarticulation selon le modèle présenté ci-dessus.
Afin de s'affranchir des problèmes liés à l'accent, nous avons synthétisé des substrats vocaliques $[V_1V_2V_1V_2V_1]$ symétriques avec un F_0 constant (fig.2).

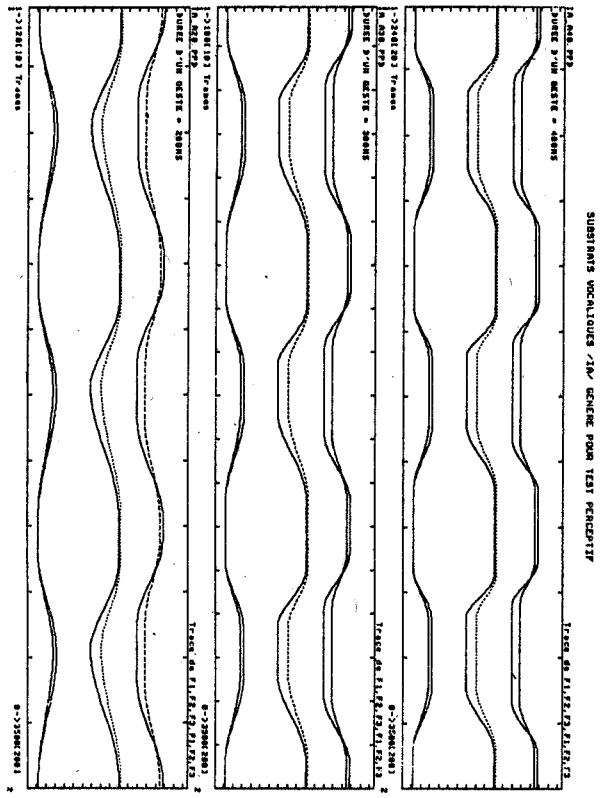


Fig.2: Exemple de stimulus de geste vocalique symétrique utilisé dans le test perceptif.

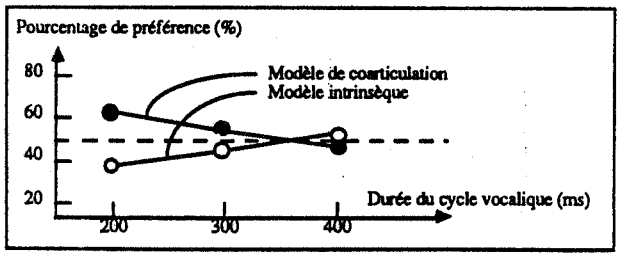


Fig.3: Résultats du test perceptif représentant le pourcentage.

La figure 3 présente pour diverses durées de cycle vocalique (200-300-400 ms), le pourcentage moyen de préférence forcée entre ces deux types de stimuli. Une tendance très nette de préférence des cibles coarticulées aux cibles intrinsèques lorsque le débit augmente se dégage de ce schéma.

6. CONCLUSIONS

Cette étude montre tout l'intérêt que doivent porter les systèmes de synthèse par règles aux phénomènes de coarticulation en ne se limitant pas aux influences du substrat vocalique sur le geste consonantique (variabilité du locus) mais aussi au geste vocalique lui-même qui présente déjà des phénomènes d'undershoot et d'overshoot sans contraintes de débit. Ce qui d'ailleurs justifie le succès rencontré jusqu'ici par les méthodes de synthèse par unités stockées qui incorporent intrinsèquement tous ces phénomènes. Néanmoins, il reste à démontrer la capacité de cette dernière méthode à gérer ces phénomènes à des débits variables.

D'autre part, cette étude montre des réalisations acoustiques dont beaucoup s'expliquent par une négociation entre contraintes articulatoires et perceptives. De ce fait, les règles de synthèse devront intégrer des contraintes articulatoires et des heuristiques perceptives.

Nous pensons que les effets de la coarticulation fournissent des indices à l'auditeur qui facilitent l'identification des segments et par conséquent participent à l'amélioration de l'intelligibilité ainsi qu'au naturel de la parole synthétique.

Il reste à étudier comment ces phénomènes de coarticulation vocalique se réalisent à différents débits lors de perturbations consonantiques [14].

BIBLIOGRAPHIE

[1] Lindblöm B. (1986). "Phonetic universals in vowel systems", in *Experimental Phonology* (Ohala J.J. Ed), New-York, Academic Press, 13-44.

[2] Lindblöm, B. (1963). "Spectrographic study of vowel reduction", JASA, 35, 1773-1781.

[3] Gay, T. (1978). " Effect of speaking rate on vowel formant movements", JASA, 63, 223-230

[4] Bailly G., Tran A. (1989). "Compost: a rule-compiler for speech synthesis, EuroSpeech, Paris, 136-139.

[5] Öhman, S.E.G. (1966). "Coarticulation in VCV utterances: Spectrographic measurements", JASA, 39, 151-168.

[6] Öhman, S.E.G. (1967). " Numerical model of coarticulation", JASA, 41, 310-320.

[7] Recasens, D. (1984). "Vowel to vowel coarticulation in Catalan VCV sequences", JASA, 76(6), 1624-1635.

[8] Bailly G., Murillo G., Al Dakkak O., Guérin B. (1988), "A text-to-speech system for French using formant synthesis", Speech '88, 7th FASE Symposium, Edimburgh, 255-260.

[9] Lonchamp, F. (1989). "Les sons du Français: Analyse acoustique descriptive", Collection Technique et Scientifique des Télécommunications. Ed. Masson, 79-130.

[10] Majid, R. (1986). "Modélisation articulatoire du conduit vocal: Exploration et exploitation. Fonctions de macro-sensibilité paramétriques et voyelles du Français", Thèse de Dr. Ingénieur, INP- Grenoble.

[11] O'shaugnessy, D. (1986). "The effects of speakings rate of formant transitions in French synthesis by rule", ICASSP, V. 3, 2027-2030.

[12] Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context", JASA, 85(5), 2135-2153.

[13] Benguerel, A.P. & Adelman, S. (1976). "Perception of coarticulated lip-rounding", Phonetica 33,113-126.

[14] Broad, D. J., & Clermont, F. (1987). "A methodology for modeling vowel formant countours in context", JASA, 81(1), 155-165.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Efficacité de la Prédiction Non Linéaire de Vecteurs dans le Codage de la Parole à Très Bas Débit

Yan Ming Cheng and Douglas O'Shaughnessy

INRS-Télécommunications
3 Place du Commerce, Ile-des-Soeurs, Verdun, Québec H3E 1H6 Canada

Resumé

Nous présentons ici une technique de prédiction non linéaire de vecteurs au lieu de la technique conventionelle de quantification vectorielle pour le codage de la parole. L'application de cette technique conduit à un vocodeur au débit de 450 bits/s qui à la sortie produit une parole plus naturelle que le codage de la prédiction linéaire.

I. Introduction

L'évolution du codage de la parole, reduisant les débits de transmission de quelques centaines kbits/s à quelques centaines bits/s, peut être regardée comme une considération avec plus de details de la chaine de communication verbale d'homme-à-homme. En considérant que la parole, qu'on utilise pour se communiquer, est seulement une onde acoustique, des techniques numériques conduisent à un débit de transmission d'environ 100 kbits/s. S'appuyant sur le fait que les signaux de parole sont auto-corrélatifs, la technique d'*Adaptive Predictive Coding* qui transmet séparément les informations corrélatives et non corrélatives de la parole d'une manière très efficace, obtient un débit d'environ 32-64 kbits/s. En considérant que la parole est produite selon un mecanisme special des organes d'humains et doit être reçue par le système auditif d'humain, le concept de filtre-excitation et la pondération perceptive rendent possible un débit d'environ 16-32 kbits/s. Récemment, au lieu de représenter la parole comme des signaux à une dimension, sa représentation vectorielle avec le concept de filtre-excitation et la pondération perceptive peuvent permettre, via quantification vectorielle, un débit d'environ 4-16 kbits/s avec une haute qualité à la réception et d'environ 250-800 bits/s avec bonne intelligibilité.

Dans des recherches courantes, on regarde très naturellement, comme dans le cas des signaux uni-dimensionnels, si dans la séquence les vecteurs de parole représentés sont corrélatifs ou dépendants, si telles corrélations ou dépendance peuvent être utilisées pour la réduction du débit, et comment on peut l'utiliser efficacement. Cette communication prétend fournir une étude préliminaire pour résoudre les questions ci-dessus. L'organisation de cette communication est la suivante: la section II présente une vue de la représentation vectorielle de la parole; la théorie de la prédiction non linéaire de vecteurs est présentée dans la section III; les résultats primitifs et discussions sont dans la section IV; et la section V conclut cette étude.

II. Représentation Vectorielle de la Parole

La parole peut être représentée par ses signaux acoustiques et aussi par ses caractéristiques, qui sont souvent décrites sous la forme de vecteurs. Le flux du signal de la parole correspond à une trajectoire de vecteur caractéristique de la parole. Cette trajectoire peut être considérée temporellement comme une série de vecteurs à l'état stable liés l'un à l'autre par le segment vecteuriel de transition. Donc, la propriété temporelle de cette trajectoire est d'alterner entre l'état stable et la transition. La durée de chaque état stable et de chaque transition est très variable. Phonétiquement, l'état stable correspond à un phonème caractérisé par la configuration articulatoire, et la transition correspond à la partie centrale d'un diphone, c'est-à-dire, soit le phonème caractérisé par les mouvements articulatoires soit la connection entre deux phonèmes.

Nous étudions ici deux sortes de représentations de séquences de vecteurs de la trajectoire analogique de la parole, qui ne perdent pas les informations utiles. Première représentation: l'échantillonage uniforme de la trajectoire, et nommée désormais Séquence des Vecteurs à Ecart Uniforme (SVEU). La SVEU est souvent obtenue par l'analyse traditionnelle ou trame-par-trame de la parole. Pour avoir une représentation assez fidèle, l'intervalle entre deux vecteurs consécutifs ou la longeur de trame doit être suffisament court. En conséquence, un vecteur d'un état stable est probablement repété plusieurs fois. Deuxième représentation: l'échantillonage non uniforme de la trajectoire, et nommée désormais Séquence des Vecteurs à Ecart Non Uniforme (SVENU). Chacun des vecteurs présente un état stable ou une transition entière. La SVENU peut être obtenue par exemple par une technique développée récemment, la décomposition temporelle [1]. La figure 1 montre un schéma illustratif des deux représentations vectorielles. Si on ne tient pas compte de la différence des interprétations physiques, les codages de ces deux séquences ne doivent pas avoir une différence significative. Ce n'est plus vrai, si l'on inclut l'aspect physique. L'exemple suivant peut grossièrement expliquer la différence pour les codages de ces deux séquences. Si l'on prend trois vecteurs consécutifs

n'importe où dans la SVENU, il n'est pas nécessaire d'avoir le codage du vecteur central, si l'on a déja codé les vecteurs d'avant et d'arrière. La raison en est que, si les deux vecteurs sont à l'état stable, le vecteur central est certainement déterminable d'une manière ou d'autre à partir de ces deux états; et si les deux vecteurs d'avant et d'arrière sont les transitions, le vecteur central est déterminé par exemple par l'effet de coarticulation. Si l'on applique cette stratégie à la SVEU, il peut avoir plusieurs possibilités pour le vecteur central. C'est-à-dire que le coût de son codage va être élevé par rapport au cas précédent. On peut considérer qu'une telle stratégie présente la prédiction non linéaire de vecteurs, d'une manière non rigoureuse. Dans la section suivante, nous discutons en détail la théorie du codage.

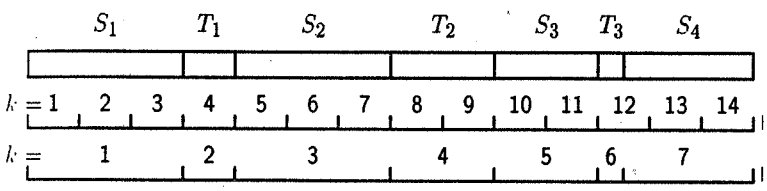


Fig. 1 Illustration des deux séquences. De haut en bas, la première ligne est la parole caractérisée alternativement par l'état stable (S) et transitoire (T); la deuxième ligne est la segmentation trame-par-trame pour obtenir SVEU; et la troisième ligne est la segmentation à trame variable pour obtenir la SVENU. k est l'indice des trames dans ce schéma.

III. Théorie de la Prédiction Non Linéaire de Vecteurs

Dépendance linéaire et non linéaire [2]

Nous essayons de définir, d'abord, le terme "prédiction non linéaire de vecteurs" dans le cadre du codage de la parole. Pour ce but, on doit introduire les notions de dépendance linéaire et non linéaire, car dans une séquence les vecteurs dépendants linéairement peuvent (resp. non linéairement) être prédits linéairement (resp. non linéairement). Supposons que $\vec{\mathbf{x}}_k$ est un vecteur au temps k. On dit la séquence avoir une dépendance linéaire de l'ordre q, si:

$$a_0\vec{\mathbf{x}}_k + a_1\vec{\mathbf{x}}_{k+1} + a_2\vec{\mathbf{x}}_{k+2} + \cdots + a_p\vec{\mathbf{x}}_{k+q} = 0;$$

autrement dit, si n'importe quel vecteur est égale à une combinaison pondérée des $q-1$ vecteurs voisins. Cette dependance est aussi couramment appelée corrélation. On dit la séquence avoir une dépendance non linéaire, si:

$$p(\vec{\mathbf{x}}_0, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{k+1}) \neq p(\vec{\mathbf{x}}_0)p(\vec{\mathbf{x}}_1)\cdots p(\vec{\mathbf{x}}_k),$$

où $p()$ est la *probability density function* (pdf). Dans la plupart des applications, nous pouvons considérer cette séquence comme un processus Markovien, dans le sens que la pdf de cette séquence est:

$$\begin{aligned} p(\vec{\mathbf{x}}_0, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k) &= p(\vec{\mathbf{x}}_0, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{k-1})p(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_0, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{k-1}) \\ &= p(\vec{\mathbf{x}}_0, \vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_{k-1})p(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_{k-1}) \\ &= p(\vec{\mathbf{x}}_0)p(\vec{\mathbf{x}}_1|\vec{\mathbf{x}}_0)\cdots p(\vec{\mathbf{x}}_k|\vec{\mathbf{x}}_{k-1}); \end{aligned}$$

c'est-à-dire que la probabilité d'une séquence est déterminée par la probabilité transitoire. Telle dépendance est aussi couramment appelée la dépendance statistique ou probabiliste.

Codage (Codeur/Décodeur) Conventionnel de Vecteur

Supposons qu'un dictionnaire à M mots ou vecteurs $\vec{\hat{\mathbf{x}}}^i$ $(0 \leq i \leq M-1)$ est exprimé par l'ensemble $\hat{\mathbf{X}}$. La fonction du codeur est de trouver l'indice d'un mot, i, dans le dictionnaire, tel que la distorsion introduite entre les vecteurs $\vec{\mathbf{x}}$ et $\vec{\hat{\mathbf{x}}}^i$ soit minimale. Le dictionnaire est construit selon une procédure d'apprentissage à partir d'une base de données qui contient toutes les statistiques de la parole. La fonction du décodeur est de reproduire le vecteur $\vec{\mathbf{x}}$ par $\vec{\hat{\mathbf{x}}}^i$ à partir d'un indice donné. Donc le vecteur $\vec{\hat{\mathbf{x}}}^i$ s'appelle aussi le vecteur de reproduction. Cette procédure du codage est connue sous le nom de quantification vectorielle. L'erreur totale dûe à ce codage pour transmettre une séquence de vecteurs est minimale avec une probabilité de 1.

Prédiction Non Linéaire de Vecteur dans le Codage

Supposons qu'à l'instant $k-1$, la reproduction du vecteur $\vec{\mathbf{x}}_{k-1}$ est $\vec{\hat{\mathbf{x}}}^i_{k-1}$ dans le codage conventionel. Si la séquence est non linéairement dépendante et une chaine Markovienne, le vecteur $\vec{\mathbf{x}}_k$ tombe, avec une probabilité très proche de 1, dans un sous-ensemble de $\hat{\mathbf{X}}$ à L vecteurs de reproduction $(L \leq M)$, qui ont les plus grandes probabilités transitoires. C'est-à-dire qu'on peut prédire de manière probabiliste le prochain vecteur. Au lieu d'un dictionnaire à M vecteurs, on en utilise un à L vecteurs. Une réduction du débit, $(\log_2 M - \log_2 L)$/vecteur, est obtenue par une telle prédiction. Le codage de la prédiction non linéaire garantit que l'erreure totale introduit est minimale avec une probabilité très proche de 1.

Pour concevoir le codeur/décodeur de la prédiction non linéaire, on doit connaître la probabilité du prochain vecteur, quand on sait le vecteur courant. Sous le langage formel, on doit estimer la matrice de transition du processus Markovien pour le dictionnaire $\hat{\mathbf{X}}$. Si le processus est non variable temporellement, ce qui est toujours le cas dans notre application, l'apprentissage de la matrice de transition peut se faire sur une séquence suffisamment grande et valable, aux sens phonétique, phonologique etc... des signaux de la parole. Et ensuite, on choisit une longueur L, par le critère de l'efficacité, qui va être discuté plus loin. Pour chaque vecteur de reproduction, $\vec{\hat{\mathbf{x}}}^i$, on trouve L vecteurs de reproduction qui ont les plus grandes probabilités transitoires, afin de construire un sous-dictionnaire associé

au vecteur $\vec{\hat{\mathbf{x}}}^i$. En tout, il y a M sous-dictionnaires dans le codeur et aussi dans le décodeur. La procédure du codage dans le codeur ne fait qu'une recherche exhaustive pour trouver le vecteur de reproduction $\vec{\hat{\mathbf{x}}}_k$ dans le sous-dictionnaire qui est associé au vecteur de reproduction $\vec{\hat{\mathbf{x}}}_{k-1}$; et celle du décodage est similaire. Ce codage est un cas particulier de Quantification Vectorielle aux Etats Finis (QVEF) [3].

Estimation de la matrice de transition

Comme mentionné ci-dessus, le clé pour concevoir la prédiction non linéaire de vecteur est d'estimer correctement la matrice de transition. Mais, le simple exemple suivant indique comment ce travail est pratiquement difficile. Supposons que le débit de vecteurs soit environ 15/seconde (noter que c'est le cas dans la pratique pour SVENU; pour SVEU, le débit est plus élevé), et qu'on ait un dictionnaire à 1024 vecteurs. Donc, il y a 1048576 ($= 1024 \times 1024$) probabilités transitoires dans la matrice de transition. Pour que cette matrice soit significative, il faut environ dix fois plus de vecteurs dans la séquence d'apprentissage, c'est-à-dire 10485760 vecteurs; autrement dit, environ 194 heures de parole. Il est très difficile de réaliser une telle quantité de travail compte tenu de la puissance des ordinateurs actuellement et des moyens de stockage.

Nous proposons ici une estimation de la matrice de transition, qu'il est possible de réaliser par une séquence courte ou "incomplète" d'apprentissage, en tenant compte des caractéristiques de la parole. Regardons chaque vecteur à n-composantes. L'ensemble des vecteurs peut être représenté par un nuage dans un espace à n-dimensions. Chaque point ou particule dans le nuage correspond à un vecteur. Le système phonétique humain partitionne cet espace d'une manière inconnue. Une autre façon présentant cette partition est de considérer que, si un point dans l'espace appartient à une classe du système phonétique, les points voisins appartiennent aussi à cette classe. Donc, si une passe $(\vec{\mathbf{x}}_k, \vec{\mathbf{x}}_{k-1})$ a lieu, chaque vecteur dans un rayon limité du vecteur expéditeur, $\vec{\mathbf{x}}_{k-1}$, devra effectuer une passe à même vecteur de destination, $\vec{\mathbf{x}}_k$. Appliquant cette conception à l'estimation de la matrice de transition, si une passe, $(\vec{\hat{\mathbf{x}}}^j_k, \vec{\hat{\mathbf{x}}}^i_{k-1})$, a lieu (c'est-à-dire que la passe va du vecteur de reproduction $\vec{\hat{\mathbf{x}}}_{k-1}$ au temps $k-1$ au vecteur de reproduction $\vec{\hat{\mathbf{x}}}_k$ au temps k), les vecteurs de reproduction dans le rayon limité du vecteur $\vec{\mathbf{x}}_{k-1}$ effectuent des passes à la même destination $\vec{\mathbf{x}}_k$. La probabilité transitoire du i^{ieme} vecteur de reproduction au j^{ieme} vecteur de reproduction est égale au rapport du nombre des passes entre ces deux vecteurs de reproduction sur le nombre total des passes partant du i^{ieme} vecteur de reproduction. Théoriquement, le rayon est un paramètre détèrminant la qualité de cette approximation, et il peut être réellement mesuré par la distorsion [4]. Mais, la partition du système phonétique n'a pas une densité uniforme. Elle est en générale plus dense où se situent les voyelles, et moin dense où se situent les consonnes. Ainsi, l'utilisation directe de la mesure de distorsion pour calculer le rayon est fastidieuse en pratique. Par contre, si l'on utilise les S vecteurs les plus proches du vecteur $\vec{\mathbf{x}}_{k-1}$ pour la mesure de rayon (S étant un nombre constant), on résout bien le problème de la densité non uniforme. La valeur appropriée de S devra être définie experimentalement. Nous le présenterons plus tard.

Efficacité du Codage

L'efficacité du codage est premièrement le rapport entre les bits utilisés pour transmettre les vecteurs et les quantités d'information contenue dans la séquence de vecteurs. Les informations du codage conventionnel et du codage de la prédiction non linéaire sont, respectivement, mesurées par l'entropie du dictionnaire (H_0) et par l'entropie de la matrice de transition (H_1) (pour leurs définitions, voir [5]). Plus l'entropie est proche du nombre de bits utilisés, plus le codage sera efficace. Un deuxième critère d'efficacité est que les entropies doivent être supérieures ou égales à la quantité d'information de la source, dans notre cas, l'information de la parole. Au niveau phonétique, telle quantité d'information est d'environ 5-6 bits. Les travaux précédents montrent que le codage conventionnel doit utiliser un nombre de bits par vecteur bien supérieur à cette limite. Désormais, on ne considère que l'entropie de la matrice de transition; de plus, on va montrer que son entropie est très proche de cette limite. Dans le cadre de la parole, des probabilités fortes d'auto-transition de vecteur évoquent une basse entropie de la matrice de transition. Dans un cas extrême, les probabilités des auto-transitions de vecteur sont égales à un, et celle des trans-transitions de vecteur, à zero (c'est-à-dire que la matrice de transition est unitaire); l'entropie de la matrice de transition est ainsi égale à zero. Cette analyse sert de base pour critiquer qualitativement l'efficacité selon le deuxième critère.

IV. Résultats et Discussions

Dans cette section, nous montrons et discutons les résultats de l'estimation de la matrice de transition, et comparons les efficacités des SVEU et SVENU dans le cadre du codage de la parole à très bas débit.

Matériel de la parole

Nous utilisons 180 phrases d'anglais équilibrées phonétiquement prononcées par plusieurs locuteurs anglophones. Les signaux de la parole sont échantillonnés à 10 kHz. SVEU provient de l'analyse LPC (*Linear Predictive Coding*) avec la fenêtre de Hamming de 25.6 ms et un recouvrement de 12.8 ms. SVENU provient de *Decomposition Temporelle à Court Terme* appliquée sur un modèle qui extrait conjointement des coefficients du filtre du conduit vocal et aussi de l'excitation glottique. Nous n'entrons pas en détail dans cette procédure car manque d'espace (voir [6,7] pour la curiosité).

Résultats pour l'estimation de la matrice de transition

Dans ce paragraphe, nous ne montrons que l'estimation de la matrice de transition du cas SVENU, car celle du cas SVEU n'est pas appropriée pour démontrer la capacité de notre algorithme à cause de sa très basse entropie. Nous adoptons un dictionnaire à 1024 vecteurs de reproduction ($M = 1024$). La figure 2 montre l'entropie (H_1) en fonction de la variation

du rayon S, d'un à quinze. L'augmentation de l'entropie avec l'augmentation de S est très importante au début, et ensuite devient très faible. Ceci prouve que notre algorithme est capable d'effectuer une telle estimation sur une séquence "incomplète" de vecteurs. Dans l'intervalle de $S \in [5, 9]$, l'entropie est presque égale à l'information au niveau phonétique. Pour montrer que la matrice estimée est appropriée pour construire le codage de la prédiction non linéaire de vecteur, nous examinons le rapport de signal-sur-bruit (SNR) de vecteur à la sortie du codeur,

$$SNR = \sum_{k=0}^{\infty} \frac{\|\vec{\mathbf{x}}_k\|}{\|\vec{\mathbf{x}}_k - \hat{\vec{\mathbf{x}}}_k\|}.$$

Deux grandeurs de sous-dictionnaire, $L = 64$ et 128, sont utilisés pour les investigations. La figure 3 montre SNR contre la variation de S. Il est clair que $S = 5$ ou 7 donne un choix à SNR maximal pour les deux grandeurs des sous-dictionnaires. Ces résultats montrent encore une fois que notre algorithme augmente significativement le SNR du codage. Le sous-dictionnaire à $L = 128$ a un SNR de 0.5 dB plus que celui à $L = 64$.

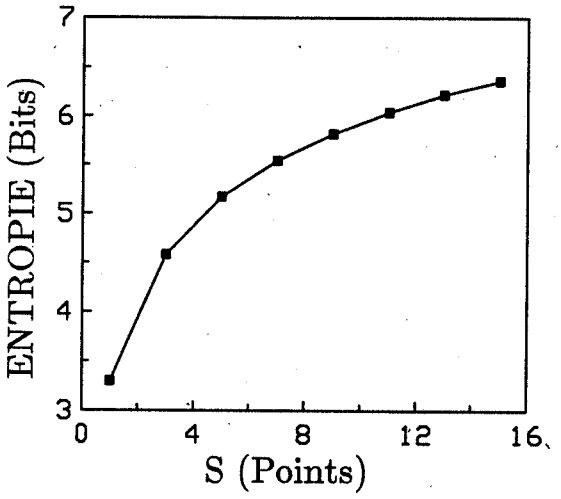


Fig. 2 L'entropie (H_1) du codeur à la prédiction non linéaire de vecteur en fonction du rayon (voir texte).

Résultats de l'efficacité du codage

Pour satisfaire le premier critère d'efficacité, il suffit de prendre pour grandeur des sous-dictionnaires, un nombre de bits très proche de l'entropie de la matrice de transition. Donc, ce critère sera toujours atteint dans le cas de la prédiction non linéaire de vecteur. Le deuxième critère mesure la structure intrinsèque de la séquence des vecteurs, et est indépendant de la grandeur des sous-dictionaires. Nous adoptons une comparaison qualitative de l'efficacité entre SVEU et SVENU, car une comparaison quantitative exigerait une quantité énorme de calcul. Selon la théorie de la section précédente, l'entropie est inversement

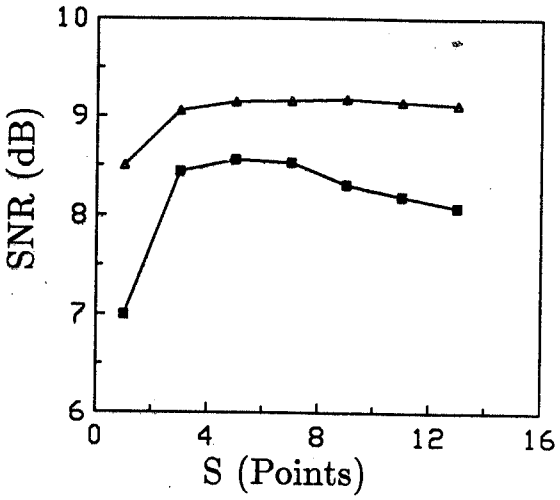


Fig. 3 SNR à la sortie du codeur en fonction du rayon (voir le texte). La ligne avec les triangles est $L = 128$, celle avec les carrés noirs est $L = 64$.

proportionnelle à la probabilité d'auto-transition de vecteurs. En plus, l'auto-transition des vecteurs peut être comfirmée lorsque la distorsion entre deux vecteurs consécutifs n'excède par un certain seuil. La distance normalisée [2] d'Itakura-Saito est empruntée comme mesure de distorsion. La figure 4 trace les probabilités d'auto-transition, qui sont égales au rapport du nombre des paires de vecteurs consécutifs dont la distorsion est inférieure à un seuil donné sur le nombre des paires totales, pour SVEU et SVENU. La probabilité pour SVENU augmente presque linéairement avec le seuil de distorsion. Celle pour SVEU monte très rapidement au début et tend ensuite vers la saturation, c'est-à-dire que, dans la plupart des cas, la matrice de transition correspondante a une entropie très basse. Les valeurs absolues des probabilités d'auto-transition pour SVEU sont bien élevées par rapport à SVENU, c'est-à-dire que l'entropie de SVEU est bien basse par rapport à celle de SVENU. Tenant compte des résultats du paragraphe précédent, où l'entropie de SVENU est égale ou un peu supérieure à l'information phonétique, l'entropie de SVEU doit être inférieure, d'une manière significative, à l'information phonétique; autrement dit, une telle séquence n'est pas appropriée pour le codage présenté.

V. Conclusion

Nous avons présenté dans cette communication la technique de la prédiction non linéaire de vecteur pour le codage de la parole. Ceci nous permet de transmettre les coefficients du filtre du conduit vocal à un débit très proche de l'information phonétique. Actuellement, nous avons implémenté cette technique dans un vocodeur. 450 bits/s et un delai d'environ 200 ms suffisent pour le codage des filtres du conduit vocal et de l'excitation glottique [8], qui assurent à la parole d'être plus naturelle qu'à la

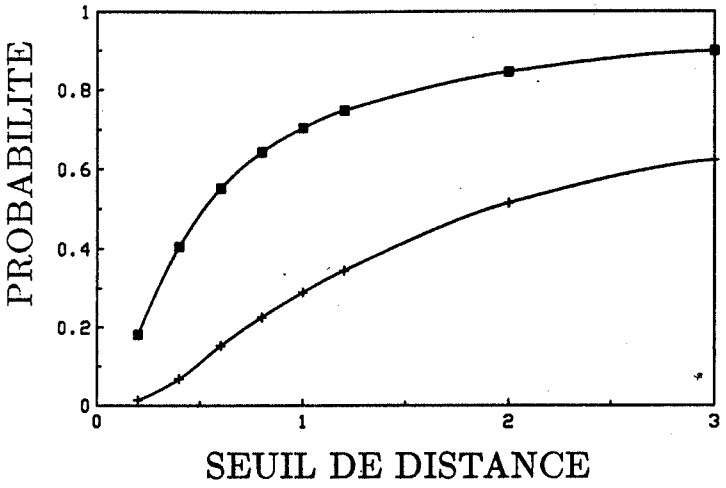


Fig. 4 La probabilité d'auto-transition en fonction du seuil de distance d'Itakura-Saito. La ligne avec les carrés noirs correspond au cas SVEU, et celle avec les croix, SVENU.

sortie du codage LPC. Mais surtout, la prédiction non linéaire de vecteur économise environ 100 bits/s par rapport à la technique conventionelle de la quantification vectorielle dans ce contexte.

Remerciement

Nous voulons remercier sincèrement Mlle Odile Rougé, stagiaire de l'ENST à Paris, qui a mis en français cette article.

Références

1. B.S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition," *Proc. Int. Conf. IEEE ASSP*, Boston, pp. 81-84, 1983.
2. J. Makhoul, S. Roucos and H. Gish, "Vector Quantization in Speech Coding," *Proc. IEEE*, Vol. 73, No. 11, pp. 1551-1588, November, 1985.
3. J. Foster, R.M. Gray and M.O. Dunham, "Finite-State Vector Quantization for Waveform Coding," *IEEE Trans. IT-31*, No. 3, pp. 348-359, 1985.
4. R.M. Gray, A. Buzo, A.H. Gray Jr. and Y. Matsuyama, "Distortion Measures for Speech Processing," *IEEE Trans. ASSP-28*, No. 4, pp. 367-376, 1980.
5. N. Abramson, *Information Theory and Coding*, McGraw-Hill, 1963.
6. Y.M. Cheng and D. O'Shaughnessy, "Performance of Short-Term Temporal Decomposition in Very-Low-Rate Speech Coding," submitted to *IEEE Trans. on ASSP*, 1989.
7. Y.M. Cheng and D. O'Shaughnessy, "On 450-600 BPS Natural-Sounding Speech Coding," submitted to *IEEE Trans. on ASSP*, 1989.
8. Y.M. Cheng and D. O'Shaughnessy, "A 450 bps Vocoders with Natural-Sounding Speech," *Proc. Int. Conf. IEEE ASSP*, Albuquerque, 1990.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

CODEUR CELP à DEBIT VARIABLE: APPLICATION au CODAGE des DIPHONES.

S. White[1,2,3],P. Mabilleau[1,2] et E. Moulines[3]

1:Université de Sherbrooke, Sherbrooke, Canada

2:CCRIT, Ministère des Communications, Laval, Canada

3:CNET, LAA/TSS/RCP Lannion,France

RESUME

Nous proposons dans cet article deux structures pour réaliser un codeur de type CELP à débit variable. Une de ces deux structures a été implantée pour réaliser un codeur dont le débit moyen est de 18 kbits/s. Il a été adapté pour coder des unités acoustiques pour la synthèse de parole à partir du texte. Des tests d'écoute informels démontrent que l'opération de codage est pratiquement transparente.

ABSTRACT

We present in this paper two structures which allow time varying bit rate in CELP coder. One of these has been implemented to realize a coder with an average bit rate of 18 kbits/s. It has been adapted to code an acoustic units database used by a speech synthesizer. Informal listening tests showed that high quality speech is obtained by this coding process.

1-INTRODUCTION

De plus en plus de systèmes nécessitent de stocker du signal de parole. Les contraintes pour des applications du type stockage/restitution sont particulières et souvent moins sévères que celles imposées pour la transmission. En effet, la plupart des applications en transmission impliquent que le codage se réalise en temps réel et à débit fixe. Nous proposons de nous affranchir de cette dernière contrainte. En effet, rien n'oblige le débit à être constant pour le stockage. La production de parole étant un processus non stationnaire, il est raisonnable de penser que certains segments exigeront de stocker plus d'informations (et donc un débit binaire plus grand) pour être codés avec la même qualité.

Une idée très proche du débit variable a été présentée depuis quelques années par le biais d'articles proposant d'allouer dynamiquement le débit binaire des codeurs CELP [Kronn,88][Taniguchi,89][Jayant,89]. Ces auteurs exposent différentes stratégies pour modifier l'allocation du débit entre les paramètres du codeur. Généralement, il s'agit de partager le débit entre l'excitation et les coefficients LPC. La stratégie que nous proposons est de conserver constant le nombre de bits accordés aux coefficients LPC et de faire varier le nombre de bits alloués à l'excitation.

Le propos de cet article est double. Nous présentons d'abord deux structures qui permettent à un codeur de type CELP une variation temporelle de son débit. Une application d'une de ces structures au codage d'unités acoustiques pour la synthèse du français viendra ensuite.

2-STRUCTURES du CELP à DEBIT VARIABLE

Deux moyens possibles pour réaliser un codeur CELP à débit variable sont illustrés sur la figure 1.

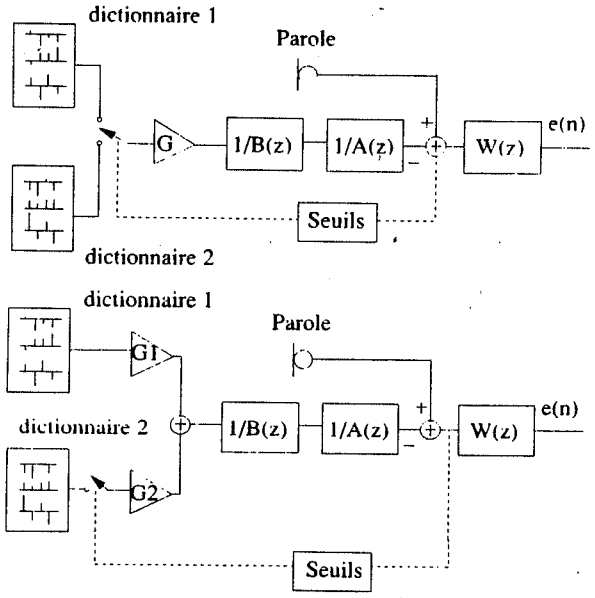


Figure 1: Deux structures pour réaliser un codeur CELP à débit variable

(en haut) : structure "parallèle".
(en bas) : structure "série".

1/B(z) : filtre prédicteur long-terme.
1/A(z) : filtre prédicteur court-terme.
W(z) : A(z)/A(z/g) filtre de pondération et g le facteur de pondération.

La première structure est dite "parallèle". La figure 1 donne un exemple avec deux dictionnaires d'excitation. Posons leurs cardinaux respectifs égaux à N_1 et N_2. L'algorithme du choix du dictionnaire se résume comme suit. (P.1): La recherche du code optimal s'effectue d'abord dans le dictionnaire 1. Le rapport signal à bruit obtenu après l'application du premier dictionnaire RSB_1 est calculé sur la longueur du code. Si RSB_1 est plus grand qu'un certain seuil que nous nommons S_1 ($RSB_1 > S_1$) alors l'indice du code et son gain sont stockés. L'algortihme s'arrête. On passe au prochain bloc de signal à coder. (P.2): Si $RSB_1 < S_1$ alors la recherche du code optimal se fait dans le dictionnaire 2. Si le rapport signal à bruit après l'application de ce deuxième dictionnaire (RSB_2) apporte une amélioration plus grande qu'un seuil S_2 ($RSB_2 - RSB_1 > S_2$) alors c'est l'indice et le gain de ce second code qui sont stockés. Evidemment, un bit supplémentaire doit être stocké pour indiquer au décodeur dans quel dictionnaire retrouvé le code. Le débit moyen est réduit si la condition suivante est satisfaite:

$$\log_2(N_1+N_2) > 1 + P_1 \times \log_2(N_1) + P_2 \times \log_2(N_2) \quad (1)$$

où P_1 et P_2 sont les probabilités d'application des dictionnaires 1 et 2 respectivement. On peut généraliser cette condition pour N dictionnaires:

$$\log_2\left(\sum_{i=1}^{N} N_i\right) > \log_2(N) + \sum_{i=1}^{N} P_i \times \log_2(N_i) \quad (2)$$

où P_i est la probalité d'application du ième dictionnaire et N_i son cardinal. La seconde structure est dite "série" i.e. que la recherche du code optimal est séquentielle. (S.1) La première étape de l'algorithme est identique à (P.1). (S.2) Si $RSB_1 < S_1$ alors le signal d'erreur résultant est calculé et la recherche du code optimal se poursuit dans le dictionnaire 2 (un algorithme semblable est proposé dans [Davidson,Gersho,88] sauf que nous optimisons les gains séparément). Si le rapport signal à bruit s'améliore au delà du second seuil S_2 ($RSB_2 - RSB_1 > S_2$) alors l'indice et le gain de ce second code sont stockés EN PLUS de ceux du premier dictionnaire. Puisqu'il y a un gain par code stocké dans la structure "série", il faut inclure le débit accordé aux gains dans la condition qui assure la réduction du débit. Elle devient:

$$b_1 + b_2 + \log_2(N_1+N_2) > 1 + b_1 + \log_2(N_1) + P_2 \times (\log_2(N_2) + b_2) \quad (3)$$

où b_i est le nombre de bits accordés pour quantifier le ième gain. En généralisant pour N dictionnaires:

$$\sum_{i=1}^{N} b_i + \log_2\left(\sum_{i=1}^{N} N_i\right) > \log_2(N) + \sum_{i=1}^{N} (P_i \times (\log_2(N_i) + b_i)) \quad (4)$$

où $P_1 = 1$ puisque le dictionnaire 1 est toujours appliqué. La différence fondamentale entre la structure "série" et "parallèle" réside donc dans le choix des dictionnaires. Dans la première, le codeur choisit le dictionnaire 1 OU le dictionnaire 2. Dans la deuxième, le choix s'effectue entre le dictionnaire 1 seulement OU la combinaison des dictionnaires 1 et 2.

3-IMPLANTATION de la STRUCTURE "SERIE"

Nous avons choisi d'implanter la structure "série" avec deux dictionnaires. La fréquence d'échantillonnage des unités acoustiques est de 16 kHz (bande élargie) puisque le système était destiné au codage de parole haute qualité. Le tableau 1 résume les paramètres que nous avons choisis pour le codeur. Pratiquement, nous nous sommes servis des techniques exposées dans [Adoul,87] pour accélérer la simulation.

Tableau 1: paramètres choisis pour le codeur

Paramètres	
Prédiction à court terme:	
fréq. de renouvellement	12 ms
fenêtre d'analyse	Hamming 16 ms
estimation des filtres	autocorrélation
ordre du filtre	18
préaccentuation	non
Prédiction à long terme:	
analyse	boucle fermée
fréq. de renouvellement	2 ms
période minimale:maximale	3 ms:11 ms
nombre de coefficients	1
Excitation:	
dictionnaire 1	1984 codes
dictionnaire 2	1984 codes
facteur de pondération g	0.8

Nous avons fait quelques expériences [White, 90] qui nous ont montré le peu d'intérêt à utiliser des dictionnaires de natures différentes. Les deux dictionnaires sont donc identiques. Ils sont composés de toutes les combinaisons possibles de deux impulsions unitaires signées sur 32 échantillons soit 2 ms à 16 kHz ((32 X 31/ 2) *4 = 1984 codes ce qui exigent 11 bits). La nature de ce dictionnaire évoque les idées de base du codage multi-impulsionel [Atal,Remde, 82]. En fait, ce dictionnaire réalise l'optimisation simultanée de la position de deux impulsions sur la trame (en l'occurence 2 ms) tout en imposant un gain unique (et non un gain par impulsion comme dans le cas du codage multi-impulsionel).

4-CHOIX des SEUILS S1 et S2

En appliquant toujours les deux dictionnaires, un codeur sans quantification avec les paramètres du tableau 1 permet d'obtenir une qualité très satisfaisante et un RSB segmental total de 17.5 dB sur une base de données de 30 secondes de parole (2 hommes et 2 femmes). Avant de choisir les seuils, nous avons observé sur cette même base de données que l'application du dictionnaire 2 dans les segments non voisés n'améliorent pas la qualité perçue. En codant toujours avec les deux dictionnaires dans les segments voisés et avec le premier seulement dans les segments non voisés, la différence n'est que très faible avec la qualité obtenue par l'application des deux dictionnaires sur tout le signal. On ne retient donc l'utilisation du débit variable que pour les segments voisés. Nous avons mené des tests informels qui

nous ont montré que des seuils de $S_1 = 20\,dB$ et $S_2 = 1\,dB$ n'entrainaient qu'une faible dégradation par rapport à la qualité obtenue en appliquant toujours les deux dictionnaires dans les segments voisés. Pour ces seuils, le probabilité d'application du dictionnaire 2 n'est que de 25.3 % pour les phrases de la base de données ce qui implique une réduction importante du débit. Si les gains des codes sont quantifiés avec 3 bits et l'indice du mot de code est stocké sur 11 bits, la réduction de débit est de 37.3 %.

Tableau 2 : Réduction du débit

Sans la stratégie du débit variable: (Application des deux dictionnaires sur tout le signal)
(1/2 ms)*(3+11 bits)*2 = 14 000 bits Avec la stratégie du débit variable:
(1/2 ms)*(3+11 bits)*(0.747*2+0.253)= 8 771 bits
Nombre de bits de réduction = 5 229 bits

5-INTEGRATION du CODEUR et du SYSTEME de SYNTHESE

Le codeur a été utilisé pour réduire le volume mémoire requis pour stocker les unités acoustiques d'un système de synthèse à partir du texte en français. Ces unités, principalement des diphones, sont au nombre de 1200: la durée de l'enregistrement est de l'ordre de 3 minutes, ce qui représente environ 5 Moctets à 16 kHz de fréquence d'échantillonnage. Ces unités ont été employées avec les techniques TD-PSOLA (Time Domain Pitch Synchronous Overlapp and Add) [Moulines, 90], qui modifient les paramètres prosodiques. Ces techniques requièrent l'utilisation de marqueurs disposés de façon synchrone de la fréquence fondamentale dans les segments voisés. Nous avons cherché à exploiter ces informations dans le codeur. Puisque ces informations sont nécessaires au traitement TD-PSOLA, le codeur pouvait en disposer sans augmentation du débit. Premièrement, les marqueurs de fréquence fondamentale donnent (de façon approchée) la position de l'instant de fermeture de glotte à l'intérieur de chaque période fondamentale. La période "locale" peut donc être calculée en soustrayant les positions relatives de deux marqueurs consécutifs. Nous avons observé que [White, 90], en limitant la recherche du délai de la prédiction à long-terme autour de la période "locale", la qualité perçue n'était pas affectée de façon significative même si le rapport signal à bruit segmental diminue sensiblement. Ceci permet de réduire le nombre de bits alloués au codage du délai de prédiction à long-terme puisque seule la différence entre la période "locale" et le délai optimal du prédicteur est stockée. De plus, le phénomène de doublement de la fréquence fondamentale, qui est reconnu pour dégrader la qualité, est évité. Deuxièmement, les marqueurs de voisement indiquent les segments des diphones qui sont voisés. Nous gardons la même stratégie de n'appliquer que le premier dictionnaire sur les segments non voisés déterminés par ces marqueurs. Finalement, nous avons synchronisé l'analyse LPC en centrant la fenêtre d'analyse sur les marqueurs de fréquence fondamentale. Il y a donc un modèle LPC PAR période fondamentale. Cette opération facilite la resynchronisation des modèles LPC lors de la synthèse TD-PSOLA [Charpentier, Moulines, 89]. Cependant, il serait prohibitif de stocker exactement tous ces modèles LPC. Seulement certains modèles sont stockés et les autres sont reconstitués par interpolation. Dans les segments non voisés, des marqueurs sont disposés de façon asynchrone environ toutes les 10 ms. Les modèles sont stockés exhaustivement pour ces segments.

6-QUANTIFICATION et RESULTATS

Les filtres LPC ont été quantifiés à l'aide des lignes spectrales [Sugamura, Itakura, 86]. Un nombre fixe de bits quantifie chaque ligne spectrale. Il s'agit d'une quantification séquentielle où la dernière ligne spectrale (la Pième dans le cas d'un filtre d'ordre P) est codée en premier, suivie de la P-1ième etc...("backward sequential adaptive uniform quantization").

Les gains des codes sont quantifiés de façon relative au gain précédent. C'est le rapport entre les gains successifs qui est quantifié tel que l'exprime la formule (5):

$$\hat{gain}1 = \hat{gain}1_{-1} \times q_i \quad (5)$$

où $\hat{gain}1$ est la version quantifiée de $gain1$, $\hat{gain}1_{-1}$ est le gain quantifé du code précédent et finalement q_i le quantificateur du rapport $(gain1/\hat{gain}1_{-1})$. La distribution de ces rapports (voir figure 2) est centrée autour de 1.0. Ceci exprime la forte corrélation qui existe entre les gains successifs et justifie le choix de quantifier le rapport.

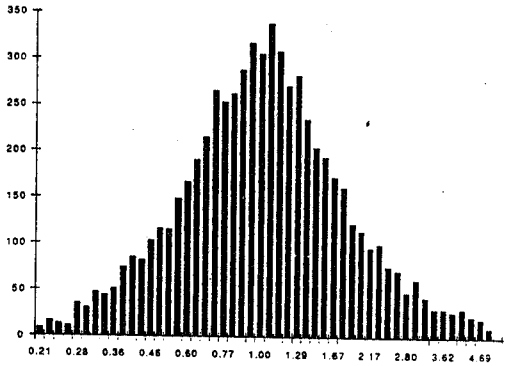


figure 2: Distribution des rapports $(gain1/\hat{gain}1_{-1})$

Le gain du dictionnaire 2 est aussi codé de façon relative. C'est cependant le rapport entre le $gain2$ et le $\hat{gain}1$ qui est quantifié. Il existe également une forte corrélation entre ces deux gains. Le dictionnaire 2 modélise l'excitation une fois retirée la contribution du dictionnaire 1. Le signal qui reste à modéliser possède généralement une énergie moins grande ce qui explique que les $gain2$ soient majoritairement plus faibles que les $\hat{gain}1$.

Le tableau 3 détaille le débit moyen du codeur. Avec des seuils $S_1 = 20\,dB$ et $S_2 = 1\,dB$, le pourcentage d'application du dictionnaire 2 est de 33.0 % pour les diphones versus 25.3 % pour les phrases.

Tableau 3: Distribution des bits

Paramètres	Renouvellement (ms)	Bits/ param. (bits)	total (bits/s)
Prédiction court-terme: Filtres LPC (après interpolation)	15	68	4 533
Prédiction long-terme: Délai pour...			
les segments voisés	2 (82 %)	4	1 640
les segments non voisés	2 (18 %)	7	630
Gain	2	3	1 500
Excitation: Indice(s) et gain(s)... du dictionnaire 1	2 (67 %)	14	4 690
des dictionnaires 1 et 2	2 (33 %)	28	4 620
Choix des dictionnaires	2 (82 %)	1	410
Débit moyen total			18 023

La qualité obtenue avec des phrases possédant les mêmes informations que les diphones (marqueurs de fréquence fondamentale et de voisement) est très satisfaisante. Un auditeur naif a peine à distinguer la version originale de la version codée. Des tests sont en cours pour savoir si cette distinction est aussi difficile à faire entre les phrases synthétisées obtenues à l'aide des diphones codés et non codés.

7-CONCLUSION

Nous avons présenté deux structures une nommée "série" et l'autre "parallèle" permettant la variation temporelle du débit d'un codeur CELP. Ces structures ont l'avantage d'être très souples. Une variation presque continue du débit est possible en ajustant les seuils qui déterminent les dictionnaires d'excitation utilisés pour coder un segment du signal de parole. Nous avons ensuite développé un codeur avec une structure dite "série" pour le codage d'unités acoustiques pour un système de synthèse de parole. Des informations propres à ce système ont été intégrées dans l'algorithme de codage notamment les marqueurs de fréquence fondamentale et de voisement. Cette application nous a permis de démontrer que la structure "série" permet de réduire considérablement le débit tout en maintenant une qualité très proche de l'original. Nous avons également montré que le codeur CELP peut être employé en bande élargie.

REFERENCES

[Adoul,87]: Adoul, J.P., Mabilleau, P., Delprat, M. et Morissette, S., (avril 87), "Fast CELP Coding Based on Algebraic Codes", IEEE Proc. International Conference on ASSP, 1957-1960.

[Atal,85]: Atal, B.S., Schroeder, M.R., (mars 85), "Code Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates", IEEE Proc. International Conference on ASSP, v. 3, 937-940.

[Atal,Remde,82] Atal, B.S., Remde, J.R., (avril 82), "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates", IEEE Proc. International Conference on ASSP, 614-617.

[Charpentier,Moulines,89]: Charpentier, F., Moulines, E., (1989), "Pitch Synchronous Waveform Processing Techniques for Text-to-Specch Synthesis Using Diphones", Eurospeech 89, v. 2, 13-19.

[Davidson,Gersho,89]: Davidson, G., Gersho, A., (1988), "Multiple-Stage Vector Excitation Coding of Speech Waveforms", IEEE Proc. International Conference on ASSP, 163-166.

[Jayant,89]: Jayant, N.S., Chen, J.-H., (1989), "Speech Coding with Time Varying Allocation to Excitation and LPC Parameters", IEEE Proc. International Conference on ASSP, 65-67.

[Kroon,88]: Kroon, P., Atal, B.S., (1988), "Strategies for Improving the Performance of CELP Coders at Low Bit Rates", IEEE Proc. International Conference on ASSP, 151-154.

[Moulines,90]: Moulines, Eric, (1990), Algorithmes de codage et de modifications des paramètres prosodiques pour la synthèse de parole à partir du texte., thèse de doctorat, Ecole Nationale Supérieure des Télécommunications.

[Sugamura,Itakura,86]: Itakura, N., Sugamura, N., (juin 86), "Speech Analysis and Synthesis Methods Developed at ECL in NTT-From LPC to LSP", Speech Communications, v. 5, 6, 199-215.

[Taniguchi,89]: Taniguchi, T., Unagami, S., Gray, R., (1989), "Multi-mode coding: Application to CELP", IEEE Proc. International Conference on ASSP, 156-159.

[White,90]: White, S., (1990), Codeur CELP à débit variable: Application à un système de synthèse par diphones, mémoire de maitrise, Université de Sherbrooke, Canada.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

VERS UNE PRODUCTION AUTOMATIQUE DE TEXTES PHONETIQUES POUR L'ARABE STANDARD A PARTIR DE SA GRAPHIE

SAROH A., BRUSSET J., TIHONI J.

Laboratoire IRIT-CERFIA,UA au CNRS 824
Université Paul SABATIER,118, route de NARBONNE
31062 Toulouse Cedex/ France

RESUME

Le présent travail s'inscrit dans le cadre général du traitement automatique de la parole. Nous nous intéressons dans cet article aux possibilités du système SYAMSA (SYstème d'Analyse Morpho-Syntaxique de l'Arabe), en particulier au rôle du lexique dans un tel système et aux interactions du niveau lexical avec les autres niveaux de traitement (analyse morpho-phonologique et morpho-syntaxique). Dans cet article nous décrivons les différentes composantes intervenant dans la génération phonétique de l'arabe standard (AS) à partir de sa graphie.

ABSTRACT

This present work is in keeping with the general pattern of automatic speech processing. We are interesting in this paper in SYAMSA system possibilities, particularly in the lexicon role in such system and the interactions beetween lexical level and the others processing levels (morpho-phonological and morpho-syntactic analysis). In this paper, we are describing the different components intervening in the standard arabic phonetic generation from its spelling form.

1 INTRODUCTION

L'étude de la phonétique arabe a été entreprise par SIBAWAYHI en l'an 180 de l'Hégire. Actuellement pour les besoins du traitement automatique de la parole cette étude connait un regain d'intérêt, citons à cet égard les travaux de MORADI [Moradi 85], [Moradi 87] portant sur la synthèse basée sur un dictionnaire des diphones et ceux de RAJOUANI [Rajouani et all 87], [Es-sakalli 87] concernant la prosodie.

Le sujet de cet article porte sur l'étude de la génération phonétique pour l'arabe standard à partir de sa graphie, ceci dans le but d'orienter le système SYAMSA [Brusset 87], [saroh 88], [saroh 89a], [saroh 89b] vers le traitement automatique de la parole.

La phonétisation automatique de l'arabe repose en particulier sur l'emploi d'un lexique et de l'analyseur morpho-phonologique. Par ailleurs ce sont les phénomènes d'interactions entre mots (liaisons, élisions, ...) et d'assimilations qui suggèrent l'utilisation des règles phonologiques.

2 LE SYSTEME SYAMSA

L'organisation de la base de données est de type relationnel [Delobel 83]. La composante syntaxique se présente comme un système expert (SE) qui applique les règles de la grammaire arabe aux énoncés soumis en entrée. La figure 1 donne l'architecture du système.

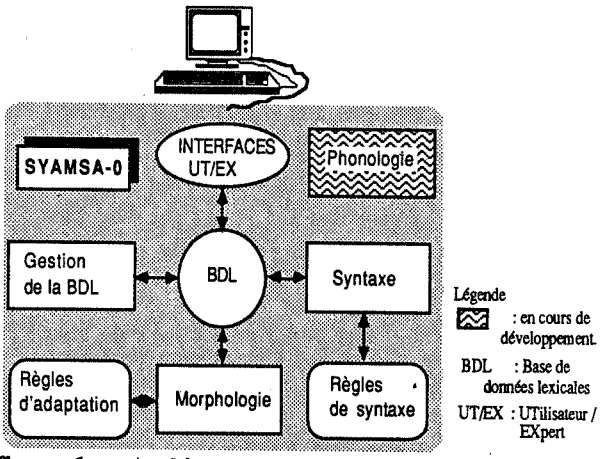


figure 1 : Architecture du système SYAMSA

La version actuelle comporte quatre parties: la base de données lexicales, la composante morphologique qui se compose de l'analyseur et du générateur, la composante phonologique et la composante syntaxique.

L' objectif principal de nos travaux est de développer une base de données lexicales (BDL) de l'arabe écrit et parlé pour l'analyse morpho-phonologique et morpho-syntaxique. L'intérêt de cette base réside dans le fait qu'elle rassemble le maximum d'informations de natures diverses sur chaque racine à savoir la représentation phonétique, le schème, le préfixe, le suffixe et les valeurs grammaticales. Cette base doit être suffisamment complète pour générer la majorité des mots arabes "voyellés". Autour d'elle, nous avons développé une composante morphologique et une composante syntaxique, celle-ci sous forme d'un SE dont la base de règles représente la grammaire arabe, la base de faits étant édifiée à partir de la BDL.

2.1 La base de données lexicales

L'organisation de la BDL est celle d'une BD relationnelle qui décrit les racines, les règles de dérivation et les règles de

flexion. A la différence du lexique de Debili [Debili 85], [Debili 86], notre BDL est fondée sur les mêmes principes que BDLEX [Pérennou 86], [Pérennou 87], développée au laboratoire IRIT-CERFIA pour le français parlé et écrit.

Le stock initial contient 2.000 racines de AL-SABIL [Reig 83]. Chaque entrée lexicale contient la représentation graphique et phonétique de la racine, des informations de nature morphologique à savoir le type de la racine —trilitère, quadrilitère— sa catégorie —saine, hamzée, redoublée, assimilée, concave, défectueuse— et les liens de chaînage spécifiés par l'expert qui permettent d'associer telle racine à telle règle de dérivation. Des mécanismes morphologiques —dérivationnels et flexionnels— permettent de générer, à partir des racines, 200.000 formes fléchies qui leurs correspondent (formes conjuguées pour les verbes, déclinaisons pour les substantifs et les adjectifs). Ces formes générées, sont placées dans un dictionnaire dans lequel, en regard de chaque mot, on trouve l'ensemble de ses valeurs grammaticales hors contexte ainsi que sa représentation phonétique. Elles sont regroupées en quatre catégories grammaticales : les verbes (ex كَتَبَ kataba), les particules variables (ex كَانَ kâna), les noms (ex مَنْزِلٌ manizilun) et les adjectifs (ex جَمِيلٌ jamîlun).

2.2 Composante morphologique

La composante morphologique fonctionne en mode génératif [Saroh 89a] et mode analytique [Saroh 89b].

• En mode génératif: sont opérationnelles les conjugaisons à tous les modes et à tous les temps avec la prise en compte du caractère défectif (en temps et en personne). De même, est opérationnelle la dérivation qui se base sur des règles dérivationnelles et flexionnelles. Dans le système nous distinguons deux types de dérivations : étymologique (voir exemple fig. 2) et commutative, la première est caractérisée par trois sous-types de dérivations: déverbative, qui consiste à générer un verbe à partir d'un autre, déverbale qui consiste à générer selon des schèmes fixes des formes nominales à partir des formes verbales et dénominative qui, d'un nom ou d'un adjectif permet de tirer un autre nom, adjectif ou verbe. La deuxième consiste à générer des formes à partir des racines ayant le même sens général, la même matière graphique ou phonique des consonnes qui les construisent et qui ne diffèrent que par l'interchangeabilité de l'ordre de ces consonnes. Notons que, dans le cas de la dérivation des formes issues d'une racine "anormale" ou "semi-normale", il est nécessaire de faire appel aux règles morpho-phonologiques (règles d'adaptation) pour résoudre le problème de gémination et d'assimilation dans une forme.

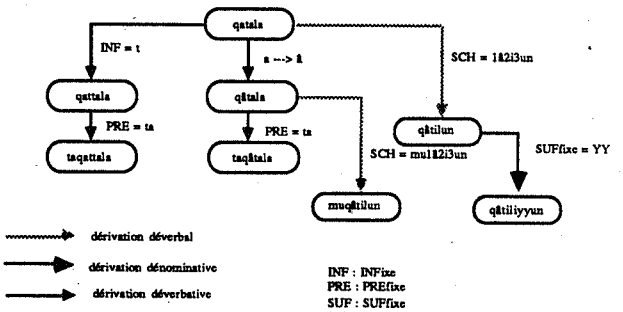


figure 2: Exemple de dérivation étymologique

• En mode analytique: le système d'analyse traite les mots à morphologie régulière —l'essentiel du lexique des mots arabes—, les mots à morphologie irrégulière —les mots issus d'une racine "anormale" et "semi-normale" qui sont en nombre limité— et les mots outils comme les prépositions, les conjonctions, etc. Rappelons que le problème principal à ce niveau est celui de l'agglutination des mots par des prépositions, des conjonctions, des articles, des pronoms, etc. Pour résoudre ce problème nous avons mis au point un algorithme qui met en évidence les différentes unités lexicales qui constituent le mot, ceci en se basant sur deux sous-lexiques l'un contenant les préfixes, l'autre contenant les suffixes. Ils sont organisés en arborescences factorisant les préfixes (resp les suffixes) qui par le jeu de l'agglutination peuvent former un seul préfixe (resp un seul suffixe) en commun. Cette analyse peut être considérée comme le résultat des opérations qui permettent d'associer à chaque mot en entrée l'ensemble de ses valeurs grammaticales hors contexte (VGHC) ainsi que sa représentation phonétique sous jacente correspondante.
Exemple: L'analyse du mot وبالمدرسة wabialmadrasati renvoit les informations suivantes:

graphique	phonétique	VGHC
و ب ال مدرسة	wa bi al madrasati	CONJonction PREPosition DETErminant Nom, féminin, singulier, génitif

2.3 Composante syntaxique

Par la position centrale qu' elle occupe dans l'analyse, la composante syntaxique [Boubaker 86], [saidi 85] constitue un élément important dans le système, elle est décrite par un système expert couplé à la base de données lexicales, ce qui a permis l'analyse d'un nombre important de phrases. Elle valide la structure de la phrase ainsi que le contrôle en genre, en nombre et en flexion en utilisant la méthode ascendante, qui, partant des traits morpho-syntaxiques attachés à chaque mot de la phrase, remonte, règle par règle, jusqu'à la résolution complète.

Au coeur de la base de connaissances se trouve une grammaire descriptive de la langue arabe, elle est décrite sous forme de règles de production représentant des phrases de structure simple. Cette grammaire assigne à une phrase une description syntaxique sous la forme d'un arbre dont chaque noeud est valué par un symbole lexical auquel est associé un ensemble de traits morpho-syntaxiques. A l'aide des informations véhiculées par les traits, nous avons exprimé des contraintes dans chaque règle, ces contraintes se traduisent en terme de rapports entre les mots. Ces rapports expriment l'accord en genre, en nombre et en cas (cas casuel). Outre ces contraintes, il en existe d'autres telles que celle qui exprime la flexion d'un mot en fonction de sa position ainsi que de sa fonction dans la phrase. Une contrainte peut être que, par exemple, une fois définis le genre, le nombre et le cas casuel du sujet المبتدأ, il doit y avoir égalité entre respectivement le genre, le nombre et le cas de ce sujet avec le genre, le nombre et le cas de l'attribut الخبر qui le suit. Pour exprimer ces contraintes d'accord, nous avons défini des fonctions booléennes au niveau de chaque règle d'analyse, qui permettent de vérifier l'égalité des traits morphologiques de certains mots. Les traits morphologiques sont:

• pour les noms et les adjectifs : le genre, le nombre et le cas,

• pour les verbes : le genre, le nombre et la désinence.

Pour exprimer les contraintes en flexion, nous avons défini des fonctions à valeur constante, qui fixent la flexion des différents mots qui constituent la phrase.

Afin de définir un formalisme adéquat pour représenter la grammaire, nous avons représenté une règle par la structure suivante:

règle = <S [S_1, S_2, ..., S_n], {g, n, c}, S_k, <$Flx_1(S_1)$, $Flx_2(S_2)$, ..., $Flx_n(S_n)$ >>

où
S : représente un symbole syntaxique —partie conséquente de la règle—
S_1, S_2, ..., S_n: représentent des symboles syntaxiques ou lexicaux—partie prémisse de la règle—
g, n, c :fonctions booléennes, qui expriment les contraintes de l'accord, g dénote des contraintes en genre, n des contraintes en nombre, c des contraintes en cas casuel.
S_k:représente le symbole syntaxique dont les traits morpho-syntaxiques remontent au niveau supérieur de l'analyse c'est-à-dire que S hérite les mêmes valeurs que S_k.
Flx_1,..., Flx_n:représentent des fonctions à valeurs constantes —nominatif (N), génitif (G), et accusatif (A)—, elles expriment les flexions associées aux différents symboles.

Exemple
<PN [SUJ, ATT], {g(SUJ,ATT), n(SUJ, ÅTT), c(SUJ,ATT)}, _, <Flx(SUJ)=N, Flx(ATT)= N>>
SUJ: SUJet ATT: ATTribut

Cette règle ne peut être appliquée que dans le cas où le SUJet et l'ATTribut ont le même genre, le même nombre et le même cas casuel.

Nous avons classé ces règles de production par rapport à la classification de Chomsky, parmi celles des grammaires à contexte libre.

Le déroulement des différentes tâches à exécuter pour analyser une phrase se fait selon l'algorithme suivant:

Début
- lecture puis segmentation de la phrase en mots,
- analyse morphologique de chaque mot,
- analyse syntaxique de la phrase,
—contrôle de la structure grammaticale, contrôle de l'accord en flexion, en désinence, en genre et en nombre—
- édition de l'arbre syntaxique de la phrase (voir exemple en annexe)

Fin.

Après l'analyse syntaxique, en fonction de la catégorie syntaxique, de la ponctuation, des blancs séparant deux mots, les règles positionnent des marqueurs de pause ou de liaison. Ces marqueurs jouent un rôle important dans le déclenchement de certaines règles phonologiques. Elles traduisent également au niveau production phonétique un arrêt partiel ou total.

Exemple **SI** CS(UMi) = DETE
Alors Marqueur_liaison(UM_i, UM_{i+1}, ' -')
Cette règle se lit
"*si la catégorie syntaxique de l'unite morphologique courante est un déterminant alors positionner un marqueur de liaison (-) entre celle-ci et celle qui la suit*"

2.4 Composante phonologique

Une fois le texte transcrit dans sa représentation phonologique et les marqueurs placés, le module phonologique sera déclenché pour assurer les transformations notamment les problèmes de jointure ou ceux prenant en compte les contextes phonétiques non résolu au niveau morphologique, ceci en se basant sur les règles phonologiques (de liaison, de gémination, d'élision, d'assimilation, etc) qui seront appliquées d'une manière linéaire. Pour représenter cette composante, nous avons l'intention d'adapter le système GEPH (GEnérateur PHonologique) [Pérennou 88], [Pérennou 89] développé pour le français dans le cadre du GRECO-PRC pôle communication Homme-Machine à la langue arabe. Ceci nous conduit à adopter la même structure de données et la même syntaxe pour les règles.

Nous avons regroupé les phonèmes consonantiques dans les classes suivantes: occlusive, fricative, liquide, vibrante et semi-voyelle, chaque classe regroupe des phonèmes dont les valeurs de certains traits distinctifs sont communes. De plus quelques traits ont été ajoutés pour particulariser les phonèmes.
Exemple : Classe semi-voyelle

SEMI-VOYELLE

TRAITS	Sonant	Emphatique	nasal	Bilabial	labio dental	Apico dental	Alvéodental	Apico alvéolaire	Pré-dorso alvéolaire	Alvéopalatal	Palatal	Vélaire	Uvulaire	Pharyngal	Glottal
w / و	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-
j / ي	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-

Les problèmes traités par la composante phonologique portent sur les phénomènes d'emphases, de gémination, d'assimilation, de transformation de Hamza. Pour rendre compte de ces faits, nous avons mis au point les règles suivantes:

•• Règles d'emphases
Le système vocalique arabe comprend trois voyelles brèves s'opposant aux trois voyelles longues. Leurs timbres se trouvent modifiés lorsqu'elles sont adjacentes à une consonne emphatique (/ص ṣ/, /ض ḍ/, /ط ṭ/, /ظ ẓ/). Pour rendre compte de ces faits, il convient d'ajouter au système vocalique six voyelles emphatiques ainsi que deux règles (EMP_1, EMP_2) qui permettent de passer d'une voyelle (brève ou longue) non emphatique à une voyelle emphatique comme illustré dans le tableau ci-dessus.
(EMP1) V-----> V' / C___ (EMP2) $\overline{V}$ -----> $\overline{V}$' / C___

REGLE	NON-EMPHATIQUE	EMPHATIQUE
EMP_1	a	a'
	u	u'
	i	i'
EMP_2	â	â'
	û	û'
	î	î'

•• Règles d'assimilation
Le système traite des phénomènes qui peuvent se produire entre consonnes contigües ou voisines. L'un des phénomènes principaux est l'assimilation. Nous distinguons deux types d'assimilations progressive et régressive, ces assimilations peuvent être totale ou partielle (accomodement).
Exemple: parfois la nasale /ن n/ est en assimilation totale avec les phonèmes /ر r/, /ل l/ et /م m/.
Ceci se traduit par la règle suivante

$$n \text{ ----------> } \begin{bmatrix} m \\ r \\ l \end{bmatrix} / \_\_ \begin{bmatrix} m \\ r \\ l \end{bmatrix}$$

Exemple
min+rabbi ## -----> ## mirrabbi
an + mâ ## -----> ## ammâ
an + lâ ## -----> ## allâ
Le phénomène d'assimilation intervient également au niveau du déterminant al, son / l/ s'assimile à la première consonne du mot qui le suit si cette consonne appartient à l'ensemble des phonèmes suivants (/ش š/, /ص ṣ/, /س s/, /ز z/, /ل l/, /ر r/, /ن n/, /ض ḍ/, /ط ṭ/, /ظ ẓ/, /د d/, /ذ d/, /ث t/, /ت t/) ces derniers sont nommés des consonnes solaires, pour les distinguer nous avons ajouté à la matrice des traits distinctifs: le trait solaire qui n'a toutefois pas d'existence articulatoire, ce qui nous amène à la règle suivante:

l -----> X / a___ (+X) X ∈ {solaires}

Exemple

al + samâ?u ## ------> ## as+samâ?u##

•• Règles de transformation de HAMZA

Le Hamza codé ? est sujet à diverses altérations déterminées par la combinatoire de contextes possibles. Elle est régie par un ensemble de règles au nombre de dix-sept qui permettent son affaiblissement, son passage soit à des voyelles longues /آ â/, /إِيـا î/, /أُوْ û / soit aux semi-voyelles / y / ou / w /, sa chute, etc. Parmi les règles, nous trouvons

Hamza_1 ? -----> ∅/ C+ ___V avec V ∈ [a, i, u]

Cette règle facultative, peut servir à un allégement qui sera obtenu par le retranchement du hamza, sa voyelle étant reportée sur la consonne non vocalisée qui le précède.

Exemple

man #?abû-ka ## -----> ## man # abu-ka

Hamza_9 ?V _____> ∅ / _____ +?V ?V +_____

Selon SIBAWAYHI la réalisation de deux hamza vocalisés en joncture "n'est pas de la langue arabe", donc l'allégement s'impose ce qui entraîne la chute de l'un des deux hamza.

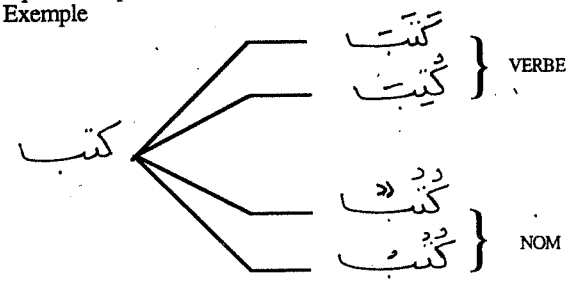
Exemple

fa-qad # ga?a # ?abû-ka ## -----> ## fa-qad # ga?a # abû-ka##

3 CONCLUSION

La réalisation du système SYAMSA a permis de cerner les problèmes relatifs à une BDL et de connaissances de l'AS tel qu'il est écrit ou parlé. Le premier enjeu essentiel de ce projet est de fournir des outils linguistiques aux industriels des langues. Par ailleurs certains points n'ont pas été évoqués dans cet article et méritent un traitement plus approfondi notamment ceux concernant le problème de la phonétisation des textes non vocalisés car l'absence de voyelles dans un mot de la langue arabe peut entrainer une ambiguïté dans sa prononciation, par conséquent elle peut affecter son sens sémantique.

Exemple



Le traitement de ce problème n'est possible que dans un cadre de collaboration entre les spécialistes en informatique, en linguistique et en phonétique arabe.

REFERENCES

[Boubaker 86] O. Boubaker: Un système expert en morphologie, syntaxe et sémantique pour l'étude de la langue arabe; Thèse 3ème cycle, UPS, Toulouse, 1986.

[Brusset 87] J. Brusset, A. Saroh: Création d'une base de données lexicales de l'arabe écrit utilisable par un système morpho-syntaxique; 4ème colloque internationnal de linguistique arabe et informatique, C.E.R.E.S., Tunis, 9-12 novembre, 1987.

[Debili 85] F. Debili, L. Zouari: Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatisée d'un dictionnaire arabe; COGNITIVA, 4-7 juin, 1985.

[Debili 86] F. Debili: Analyse morphologique de l'approche sans dictionnaire à l'approche sans grammaire; 8ème séminaire Tuniso-Français d'informatique Tunis, 5-7 mai, 1986.

[Delobel 83] C. Delobel, M. Adiba: Base de données, des modèles réseaux et hiérarchique au modèle relationnel TSI, vol 2, N° , 1983, pp. 43-62.

[Es-skalli 87] L. Es-skalli, A. Rajouani, M. Najim, M. Zyoute, D. Chiadmi: Elément d'un modèle intonatif de la phrase affirmative en arabe., 16ème Journée sur la parole, Hammamet, Tunisie, 5-9 oct 1987.

[Moradi 85] A. Moradi, M. Najim, A. Rajouani : Unlimited vocabulary synthesis system for arabic language Proc. of the 4 th inter. conf. on digital processing of signal in communication. conghbrough 22-26 apr. 1985.

[Moradi 87] A. Moradi : Validité et limites du diphone en tant qu'unité de synthèse pour la langue arabe standard, 16ème Journée sur la parole, Hammamet, Tunisie, 5-9 oct 1987.

[Pérennou 86] G. Pérennou, M. De Calmès: BDLEX : une base de données et de connaissances du français parlé; Séminaire GRECO/ GALF, Toulouse, 16-17, 1986.

[Pérennou 87] G. Pérennou, M. De Calmès: BDLEX Lexical Data and knowledge Base of Spoken and Written French; European Conference on Speech Technology, Edinburgh, 1987.

[Pérennou 88] G. Pérennou : Le projet BDLEX de base de données et de connaissances lexicales et phonologiques, Premières journées nationales du GRECO-PRC, actes du GRECO-PRC, paris 24-25 nov 1988.

[Pérennou 89] G. Pérennou, J. Tihoni, M. De Calmès, I. Ferrané : Idiolecte et phonologie incidence sur la transcription automatique adaptée au locuteur par le système GEPH, séminaire "Variabilité et spécificité des locuteurs: étude et application", Luminy-Marseille 20-21 juin.

[Rajouani 87] A. Rajouani, M. Najim, D. Chiadmi, M. Zyoute : Synthesis by rule of arabic language to be published in proc. of European Conference on Speech Technology, Edinburg, 1-3 sep 1987.

[Reig 83] D. Reig: Dictionnaire Arabe-Français, Français-Arabe(AS-SABIL), (éd) Collection Saturne, Larousse, 1983.

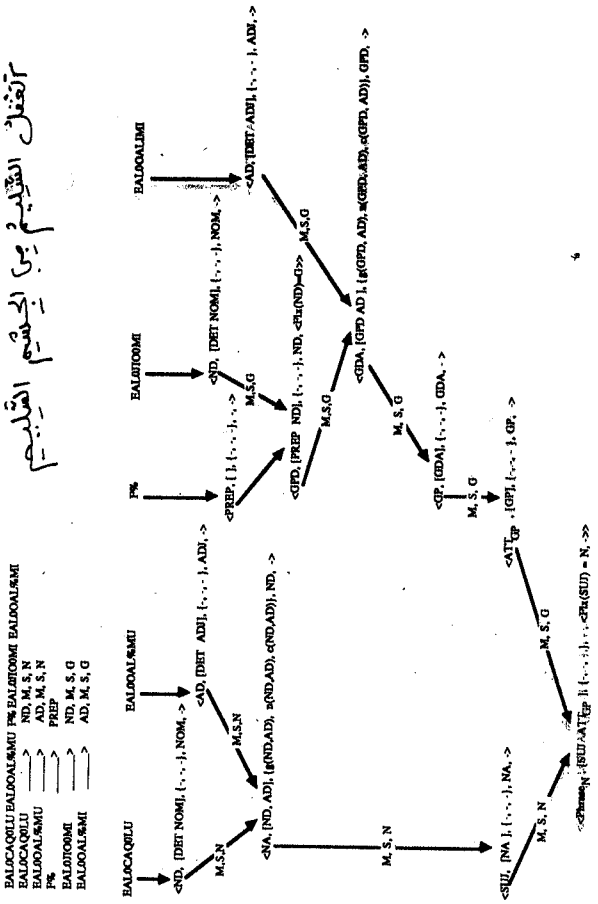
[Saidi 85] M.L. Saidi: Etude et réalisation d'un système expert appliqué à l'analyse morpho-syntaxique de phrase en langue arabe; Thèse 3ème cycle, U.P.S., Toulouse, 1985.

[Saroh 88] A. Saroh: Conception et réalisation du système SYAMSA (SYstème d'Analyse Morpho-Syntaxique de l'Arabe; Rapport IRIT-CERFIA, UPS, Toulouse, 1988.

[Saroh 89a] A. Saroh, J., Brusset : Morphological analysis system of written arabic in SYAMSA (SYstem for Morpho-Syntactical Analysis of Arabic), III International Congress Expert Systems, Florence, Italie, nov 2-5, 1989.

[Saroh 89b] Saroh, A., Brusset, J. : Un modèle de générateur morphologique de l'arabe, 7ème congrès, AFCET, RF-IA, Paris, 27 nov - 1 déc, 1989.

ANNEXE



XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

RESEAUX CONNEXIONISTES POUR LA TRADUCTION ORTHOGRAPHIQUE - PHONETIQUE : APPLICATION A L'ESPAGNOL ET AU FRANCAIS

Susanna GONZALEZ (*) et Jean-Pierre TUBACH (**)

(*) Université Polytechnique de Madrid (Département d'Informatique)
et Télécom Paris
(**) Télécom Paris, Département Signal (CNRS URA 820)

RESUME

Cet article traite de l'utilisation de réseaux connexionistes pour le passsage de textes alphabétiques à leur transcription phonétique. Aucune règle explicite n'est utilisée, un apprentissage étant effectué sur un texte dont la transcription est connue.

Les réseaux utilisés sont des perceptrons multicouches, et l'algorithme d'apprentissage est la rétropropagation du gradient. Nous faisons d'abord le point sur des réalisations précédentes pour l'Anglais (NETtalk et NETspeak), puis décrivons les mécanismes que nous avons utilisés pour l'espagnol et pour le français.

L'évaluation des résultats fait apparaître que des textes assez courts permettent déjà un apprentissage valable.

1 Introduction

Les systèmes de traduction orthographique - phonétique développés durant les années 70 et la première moitié des années 80 étaient fondés sur l'utilisation de règles explicites, founies par des experts humains (Fervers et al., 1976, Divay et Guyomard, 1977, Catach et Meissonier, 1979, Prouts, 1979, Goyer et al., 1979, Lety, 1980, Aubergé, 1985 pour le français ; Klatt, 1976, Hunnicutt, 1980 pour l'anglais ; Santos et Nombela, 1982 pour l'espagnol; etc...)

Les réseaux connexionistes, les perceptrons multicouches en particulier, peuvent fournir une alternative intéressante par leur possibilité d'apprentissage automatique de la transformation.

Nous supposerons connus dans la suite les principes de base des perceptrons multicouches (voir par exemple une présentation didactique dans Lippmann, 1987)

2 NETtalk et NETspeak

2.1 NETtalk

NETtalk fut le premier réseau connexioniste pour le passage d'un texte alphabétique à sa transcription "phonétique". Proposé par Sejnowski et Rosenberg, 1986, il a été expérimenté par ces auteurs pour la langue anglaise.

Il s'agit d'un perceptron multicouches, avec un seul niveau caché. L'apprentissage des poids de connexion est réalisé par l'algorithme de rétropropagation du gradient (Rumelhart et al., 1986; Le Cun, 1987).

La couche d'entrée reçoit à un moment donné les représentations de sept lettres du texte alphabétique (fenêtre glissante), la couche de sortie doit alors fournir la représentation du "phonème" correspondant à la lettre centrale. La transcription phonétique d'un mot ne comportant pas le même nombre de caractères ("phonèmes") que la forme orthographique, des caractères "nuls" (sans équivalent dans l'autre forme) sont positionnés manuellement dans les données d'apprentissage.

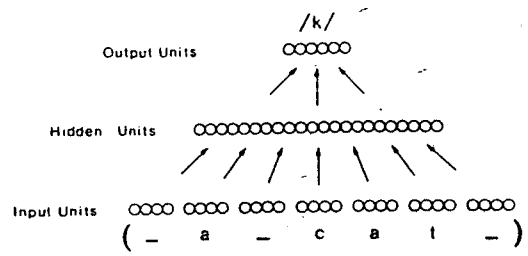
Chaque caractère est représenté sur 29 éléments binaires de la couche d'entrée (26 lettres de l'alphabet, séparateur de mots et ponctuation). Il y a donc en tout 7x29 = 203 éléments binaires dans la couche d'entrée.

La représentation d'une unité "phonétique" sur la couche de sortie est un ensemble de 21 traits articulatoires binaires (liés au lieu d'articulation, au voisement etc...) et de 5 unités additionnelles représentuatant l'accentuation (stress) et les limites de syllabes.

Il n'y a donc pas vraiment une sortie en termes de phonèmes, d'où les guillements dont nous avons entouré ce mot depuis le début du paragraphe, et l'évaluation précise est délicate, du fait de la redondance du système de traits : connecté à un synthétiseur utilisant ces traits, le système produit une parole compréhensible, sans que l'on connaisse ses performances précisément pour ce qui est du problème de la traduction orthographique phonétique. Il n'est d'ailleurs pas fourni de résultats détaillés d'évaluation.

Il y a 80 à 120 unités cachées, suivant les versions, dans la couche centrale du réseau. La connectivité est totale entre couches adjacentes, mais il n'y a pas de connexions directes entre la couche d'entrée et celle de sortie.

La figure qui suit est empruntée à NETtalk, et schématise le fonctionnement du réseau.



Plus techniquement, la sortie de chaque unité du perceptron est une fonction non linéaire de la somme pondérée de ses entrées

$$Oj = f(I_j) = \frac{1}{(1+e^{-I_j})} \qquad (1)$$

où O_j est la sortie de la $j^{ème}$ unité, et I_j la somme pondérée de ses entrées

$$I_j = \Sigma_i \, w_{ji} \, O_i \qquad (2)$$

où w_{ji} est le poids de connexion de la $i^{ème}$ à la $j^{ème}$ unité.

Lors de l'apprentissage, les poids sont modifiés après que toutes les données d'apprentissage ont été présentées, selon les formules suivantes:

$$\Delta w_{ji}^{(n)}(u+1) = \alpha \, \Delta w_{ji}^{(n)}(u) + (1-\alpha) \, \delta_j^{(n+1)} \, o_i^{(n)} \qquad (3)$$

où Δw_{ji} est le changement du poids w_{ji}
α est un paramètre de lissage (typiquement 0.9)
δ_j est la dérivée de l'erreur par rapport à l'entrée de l'unité i.

$$w_{ji}^{(n)}(t+1) = w_{ji}^{(n)}(t) + \varepsilon \Delta w_{ji}^{(n)} \qquad (4)$$

où t est le nombre de mises à jour de poids
et ε un taux d'apprentissage (typiquement 1.0)

L'apprentissage de NETtalk a été effectué sur deux ensembles de données : 1000 mots fréquents du Miriam Webster's Pocket Dictionary, et la transcription d'un enregistrement d'enfant (1024 mots).

2.2 NETspeak

NETspeak est une nouvelle implémentation de NETtalk, proposée par McCulloch et al., 1987. Les principales modifications portent sur la représentation des lettres et des phonèmes, et sur les paramètres d'apprentissage.

Les lettres ont été groupées en cinq classes "phonétiques" grossières (voyelles, plosives non voisées, plosives voisées, fricatives et autres (liquides, nasales, diphtongues)), correspondant à leur prononciation la plus courante. Il s'agit là d'une introduction de connaissances a priori, pour tenter de diminuer la complexité de la tâche d'apprentissage. Chaque lettre est ainsi codée par 11 éléments binaires (5 pour la classe et 6 pour le numéro dans la classe).

Les phonèmes sont représentés par 25 éléments binaires (17 traits phonologiques, 5 pour l'accentuation et les limites de syllabes et 4 pour la ponctuation).

Comme dans NETtalk, la connectivité est totale entre couches adjacentes. Il y a 77 unités dans la couche cachée.

En ce qui concerne l'apprentissage, les formules proposées par Rumelhart sont utilisées pour la modification des poids de connexion :

$$\Delta_p w_{ji}(t+1) = \eta \, \delta_{pj} \, o_{pi} + \alpha \, \Delta_p w_{ji}(t) \qquad (5)$$

où $\Delta_p w_{ji}$ est le changement au temps t du poids de connexion des unités i et j lorsque le "pattern" p est présenté en entrée,
η est le taux d'apprentissage ($\eta = 0.05$),
α est un terme de moment ($\alpha = 0.95$)
Les poids sont mis à jour, conformément à

$$w_{ji}(t+1) = w_{ji}(t) + \Delta_p w_{ji}(t) \qquad (6)$$

La mise à jour des poids est calculée, comme l'indique l'indice p dans les formules, après présentation de chaque caractère, ce qui est plus coûteux mais plus efficace que de le faire seulement après une présentation complète de l'échantillon d'apprentissage.

L'apprentissage a été fait sur "la plupart" (sic) des 16280 mots du Teacher's Word Book.

Après stabilisation de l'apprentisage, les résultats sur des données de test sont de 95% de bits de sortie exacts, menant à une meilleure hypothèse correcte au niveau phonème dans 87% des cas (mais à une coïncidence parfaite dans 53% des cas seulement). C'est le chiffre de 87% qu'il faut retenir.

3 Réseau pour l'Espagnol

Nous décrivons ici un travail effectué par S. Gonzalez, 1989, pour son mémoire de fin d'études à l'Université Polytechnique de Madrid, sous la direction de Thierry Michaux.

La traduction orthographique - phonétique pour l'espagnol est nettement moins complexe que dans le cas de l'anglais. En règle quasi générale, une lettre produit un phonème et un seul, les exceptions portant sur les paires de lettres ll, rr et ch, qui produisent un seul phonème. Elles sont traitées spécialement, de façon à introduire un peu de connaissance a priori. L'accentuation des mots est elle aussi très régulière.

L'architecture est la même que dans NETtalk et NETspeak : perceptron multicouches avec une couche cachée, fenêtre glissante de sept caractères en entrée.

La différence la plus importante porte sur les entités de sortie : il s'agit bien ici de phonèmes. Ils sont codés sur 41 entités binaires, soit 37 pour les phonèmes proprement dits et 4 pour l'accentuation et les pauses.

En entrée, chaque caractère est codé sur 35 unités binaires : 26 lettres de l'alphabet, plus 4 pour "n tilda", et "ch", "ll", "rr" (déjà mentionnés), et 5 supplémentaires pour la ponctuation.

L'algorithme d'apprentissage est celui de NETspeak, mais les paramètres d'aprentissage les mieux adaptés se sont révélés différents : 0.25 pour le taux d'apprentissage, et 0.5 pour le paramètre de moment.

L'échantillon d'apprentissage était extrait d'un journal espagnol. Il a suffi d'un texte de 1553 caractères pour obtenir 100% de sorties correctes sur l'échantillon d'apprentissage, et 95% sur un texte de test différent, de même nature. Ceci illustre bien la relative facilité du problème dans le cas de la langue espagnole.

4 Réseau pour le Français

Ce travail a été effectué par S. Gonzalez au cours d'un séjour à Télécom Paris, sous la direction de J.P. Tubach.

Le cas du français est plus complexe que celui de l'espagnol : groupes de lettres produisant un seul phonème, lettres produisant plusieurs phonèmes. (Ce dernier cas, qui s'observe également en anglais, (ex. ox --> /oks/) ne semble pas avoir été traîté dans NETtalk et NETspeak).

Ces différencees de longueur entre entrée et sortie ont été traitées de la façon suivante pour l'échantillon d'apprentissage :

- lorsque plusieurs lettres contribuent à la production d'un seul phonème, des caractères "+", en nombre voulu, sont insérés manuellement dans la transcription souhaitée. Par exemple : eau --> o++ .

- lorsqu'une lettre ne contribue à la formation d'aucun phonème, par exemple le p final de "beaucoup", un "@" est inséré dans la transcription souhaitée.

Ainsi on aura : beaucoup --> bo++ku+@

- pour le cas des lettres qui produisent deux phonèmes, on a ajouté dans la liste des unités de sortie des entités spéciales correspondant à ces paires de phonèmes (ks, gz, ji, ij, wa).

Les textes phonétiques traîtés (voir plus loin) utilisent une transcription assez fine, comportant des diacritiques (allongement, dévoisement, sonorisation, etc...). Ces symboles ont été pris en compte, mais ne sont pas comptés en cas d'erreur.

Il y a finalement 51 unités binaires de sortie, représentant 36 phonèmes, les 5 paires de phonèmes mentionnées ci-dessus, et 10 unités supplémentaires pour les 2 caractères spéciaux "+" et "@" , les séparateurs et les diacritiques (dont la marque d'accent). Dans l'évaluation des résultats, les 2 caractères spéciaux ne sont pas pris en compte, et les erreurs sur les diacritiques ne sont pas comptées.

Les caractères en entrée sont codés sur 33 unités binaires, correspondant aux 26 lettres de l'alphabet, et aux caractères français : les lettres accentuées sont codées comme leurs homologues non accentuées avec un bit supplémentaire pour l'accent, et ç est un caractère en soi.

La couche d'entrée comporte donc 7x33 = 231 unités.

On a utilisé 90 unités dans la couche cachée. L'algorithme d'apprentissage est le même que pour l'espagnol.

Les données d'apprentissage et de test ont été empruntées au corpus de textes phonétiques rassemblés à partir de plusieurs sources par Tubach et Boe, 1985, dont la version avec texte alphabétique a été publiée dans Tseva et al., 1988.

On a utilisé trois textes, T1 de 3734 caractères, T2 de 2648 caractères, et T3 de 2353 caractères. Il s'agit du début du corpus "G", représentatif d'un français parlé cultivé assez formel (conférence à France Culture).

Après apprentissage sur T1 seul, on obtient 89% de sorties correctes au niveau phonétique sur T1, et 65% sur T2 pris comme test. L'apprentissage est donc encore insuffisant.

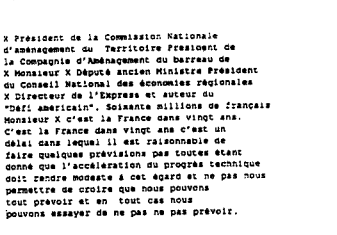
On a ensuite effectué un apprentissage sur T2 partant des poids de l'apprentissage précédent (ce qui est moins couteux mais moins efficace qu'un apprentissage sur l'ensemble T1 + T2, puisque les itérations ne sont pas faites sur la totalité des données d'apprentissage). On obtient alors 98% sur T2 (et 97 % sur T1), et 94% sur T3 pris comme test, donc de meilleurs résultats que ceux rapportés (pour l'anglais) par les méthodes connexionistes évoquées précédemment.

Le temps total d'apprentissage est de 8 heures sur un VAX Server 3600 (3 MIPS) , dont l'architecture n'est pas particulièrement adaptée à ce type de traitement. La programme était écrit en C, sans souci particulier d'optimisation de performance. Au cours de ces apprentissages, T1 a été présenté 210 fois, puis T2 170 fois

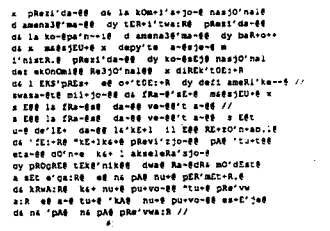
Nous présentons enfin des exemples de données d'apprentissage et de résultats. (Codification phonétique : voir Tubach et Boe, 1985).

Dans T1 (données d'apprentissage)

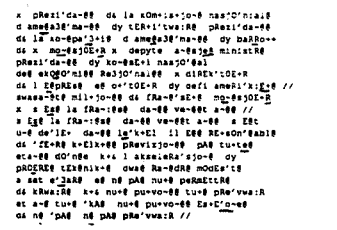
Texte alphabétique solution



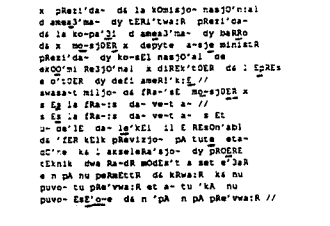
Texte phonétique



Sortie du réseau "éditée"



Sortie du réseau



Dans T3 (données de test)
Texte alphabétique
solution

```
Restaca a l'intérieur de chaque région à
équilibrer l'attraction de cette métropole
avec les autres  villes du dispositif que
l'on appelle l'armature urbaine.
Après cette exégèse dûe à X sur les
"Antimémoires" écoutons  ce qu'X a répondu
à la question de X portant précisément sur ce  titre.
X après dix ans de silence vous revenez à
la littérature pour publier un livre Antimémoires
mais pourquoi ce titre  "Antimémoires".
```

Texte phonétique

```
REstcA'Ra  a l e-@teRjOE+R dE S+ak+@ Re'3jo-@
a ekilI'bRe@  *l atRakajo-@ dE sEt+@ metRO'pOl@
avEK lez o+tRa@ 'vil+@@  dy dispozi'tif
k+A l o-n a'p+El+@  l aRmatyR@ yR'bE+n@ //
apRE@ sEt+@ EGZe'3E:z@  dy@ a x  syR lez
a-@time'mwa:R@@  aku'to-@@  as k+ x  a
Repo-@'dy  a la k+Eatjo-@ dE a  pORta-@@
"pResize'ma-@@  syR as 'titR.@ //
x  apRE@ diz a-@@ dE si'la-:@s@  vu+@
R6v@'ne@  a la lit+eRa'ty:R@  pu+R
pyblI?'e@  u-@ livR@ "a-@time'mwa:R@@
mE+@ pu+Rk+wa s@ 'titR.@  a-@time'mwa:R@@ //
```

Sortie du réseau
"éditée"

```
REstRa  a l e-@teRjOE+R dE k+ak+@ Re'3jo-@
a ek+lli:'bRe@  l at+Rakajo-@ dE sEt+@
metRO'pOl@  avEK lez o+tRa@ 'vil+@@  dy
dispEzi'tif  k+A l o-@ a'p+El+@  l aRmatyR@
yR'bE+n@ // apRE@ sEt+@ E//e'3Es@
dy@ a x  syR lez a-@time'mwa:R@@
aku'to-@@  as k+ x  a Re'po-@dy
a la k+Eatjo-@ dE x  pORta-@@ pResize'ma-@@
syR as 'ti'tR@ // x  apRE@ di//
a-@@ dE si'la-:@s@  vu+@ RAva-'ne@  a la
lit+eRa'tyR@  pu+R pyp'lje@  u-@ 'li:vR@
a-@time'mwa:R@@ mE+@ pu+Rk+wa: s@ 'titR@
a-@time'mwa:R@@ //
```

Sortie du réseau

```
REstRa  a l e-teRjOER dE kak Re'3jo-  @
ekili:'bAe  l atRakajo- dE sEt metRO'pOl
avEK lez otRa 'vil  dy dispEzi'tif  ks l
o- a'pEl  l aRmatyR yR'bEn // apRE sEt E//
e'3Es  dy a x  syR lex a-time'mwaiR
aku'to-  as k x  a Re'po-dy  a la kEatjo-
dE x  pORta- pResize'ma-  syR s 'ti'tR //
 x  apRE di// a- dE si'la-:s  vu R6VE'ne
a la liteRa'tyR  puR pyp'lje  u- 'li]VR
a-time'mwaiR mE puRkwa: s''titR  a-time'mwaiR //
```

Les erreurs sont soulignées. On remarque que bon nombre d'entre elles correspondent en fait à des prononciations possibles (e muet prononcé ou non, en particulier), d'autres à des mots à prononciation exceptionnelle (monsieur).

5 Conclusion

Il serait bien sûr intéressant de pouvoir traîter directement à l'apprentissage des solutions phonétiques ne comprenant pas les symboles spéciaux rétablissant l'égalité des nombres de caractères en entrée et en sortie, solution de facilité jusqu'à présent utilisée pour l'anglais comme pour le français. Ce n'est pas impossible dans le principe, et il y a là le sujet d'une étude à venir.

Quoi qu'on ne puisse encore affirmer que cette approche connexioniste ait montré définitivement sa supériorité, elle représente incontestablement un champ d'investigation intéressant, avec déjà des résultats non négligeables.

On peut aussi indiquer, dans le sens inverse, que le problème de la traduction orthographique phonétique (et, pourquoi pas, phonétique orthographique) fournit aux techniques connexionistes un domaine d'application susceptible de les faire progresser, d'autant que sa complexité est variable d'une langue à l'autre.

Le tableau final récapitule les caractéristiques principales des réseaux référencés ou développés dans cette étude.

	input units	output units	hidden units	connections
NETtalk	203	26	120	8398
NETspeak	77	25	77	3850
l'Espagnol	245	41	100	14745
Le Français	231	51	90	16371

Remerciements

Le travail relatif au français a été financé en partie par le projet Esprit Basic Research Action SPRINT (SPeech Recognition using Integrated Neural Techniques).

Bibliographie

AUBERGE V.
"Passage automatique du texte orthographique vers le texte phonétique" JEP SFA, Paris, 1985

CATACH N., MEISSONIER V.
"Pour une meilleure formalisation de la conversion automatique graphème phonème" JEP GALF, Grenoble, 1979 pp 173-182

DIVAY M., GUYOMARD M.
"Conception et réalisation d'un programme de transcription graphème - phonétique du français". Thèse de 3eme cycle, Univ. de Rennes, 1977

FEVERS H., LE ROUX J., MICLET L.
"Programme de transcription orthographique phonétique en langue française". Rapport ENST, Paris, 1976.

GONZALEZ S.
"Desarrollo de un modelo conexionista y applicacion ; NETtalk para el Espanol" Publié par la Faculta Informatica, Univ. Polytechnique de Madrid, 1989.

GOYER P., DEGRYSE D., GUERIN B.
"TROPS, un système de transcription orthographique phonétique et de synthèse en français". JEP GALF, Grenoble, 1979, pp 212-217.

HUNNICUTT S.
"Grapheme to phoneme rules : a review". Speech Technology Laboratory QPSR 2-3, 1980, pp 38-60

KLATT D.H.
"Structure of a phonological rule component for a synthesis by rule program". IEEE Trans ASSP Vol 24, 1976, pp 391-398

LE CUN Y.
"Modèles connexionistes de l'apprentissage". Thèse, Université Paris VI, 1987.

LETY M.
"Transcription orthographique phonétique : un système interpréteur". Thèse DDI, Univ. Scientifique et médicale de Grenoble, 1980.

LIPPMAN R.P.
"An introduction to computing with neural nets". IEEE ASSP Magazine, April (1987)

MC CULLOCH N., BEDWORTH M., BRIDLE J.
"NETspeak, a reimplementation of NETtalk". Computer Speech and Language, Vol 2, 1987, pp 289-301

PROUTS B.
"Traduction phonétique du texte écrit en français". JEP GALF, Grenoble, 1979

RUMELHART D.E., HINTON G.E., WILLIAMS R.J.
"Learning internal representations by error backpropagation". In D.E. Rumelhart & J.L. McClelland (Eds.), Parallel Distributed Processing : Explorations in the microstructure of cognition. Vol 1 : Foundations. MIT Press (1986)

SANTOS J.M., NOMBELA L.R.
"Text to speech conversion in Spanish. A complete rule synthesis system". IEEE ICASSP, Paris, 1982, pp 1593-1596.

SEJNOVSKI T., ROSENBERG C.R.
"NETtalk : a parallel network that learns to read aloud". Johns Hopkins University. Technical report JHU/EECS-86/01, (1986)

TSEVA A., TUBACH JP, BOE LJ
"Corpus alphabétique phonétique" Travaux de l'Institut de Phonétique de Grenoble, 2 volumes, 1988

TUBACH J.P BOE L.J
"Un corpus de transcriptions phonétiques : constitution et exploitation statistique" Document technique ENST 85D001 (1985) ; 2ème édition en préparation, (1990)

10 DIALOGUE ORAL ET RECONNAISSANCE

Président: R. DE MORI
CRIM-Montréal,Canada

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

L'AMORCAGE SEMANTIQUE EN COMPREHENSION

J. Caelen et K. Nasri

ICP / INPG
46, Av. F. Viallet, 38031 Grenoble Cedex

RESUME : Cette communication tente de démontrer l'intérêt d'introduire un module de compréhension partielle et progressive dans un système de reconnaissance de la parole en amont du module d'analyse pragmatique et en aval de l'analyse syntaxico-sémantique. Partant de considérations psycholinguistiques sur le phénomène d'amorçage sémantique, on aboutit à la conclusion qu'un tel modèle pourrait permettre une validation rapide d'un segment de phrase sans instanciation nécessaire de tous les actants de la phrase. Cette notion est mise en oeuvre dans le système DIRA à partir d'une analyse en constituants et d'une analyse fonctionnelle. Les résultats montrent d'une part, que le facteur de branchement lexical est réduit d'un facteur 10 en moyenne au cours de la reconnaissance et d'autre part, que les relations prédicatives entre les mots "amorcés" ne sont pas obligatoirement dépendantes de l'application.

1. INTRODUCTION

Le système DIRA est un système multi-experts organisé autour d'une architecture de tableau noir (blackboard) et muni d'un superviseur (Fig. 1).

Ces experts sont:
- le **Décodeur Acoustico-Phonétique** (DAP) qui propose (et vérifie) des macro-traits et des traits phonétiques à partir du (ou sur le) signal d'entrée [Caelen, 88],
- l'**Analyseur Lexical** (A.L.) qui par des accès variés au lexique prédit (et vérifie) des mots,
- l'**Analyseur syntaxico-Sémantique** (ASS) qui contrôle la cohérence des groupes syntagmatiques au niveau syntaxique et sémantique, et prédit le ou les prochains mots possibles [Reynier et al, 89],
- le **module de Compréhension en Constituants Majeurs et Mineurs** (C^2M^2) qui contrôle les groupes de sens,
- le **module de Dialogue** (D) qui gère le dialogue et construit les informations pour l'interface de communication avec l'application.
- l'**Analyseur Prosodique** (P) qui pose des marqueurs de débuts et fins de mots sur les pseudo-syllabes [Nasri et al, 89].

Le superviseur gère les drapeaux dans le blackboard, planifie la stratégie et fixe les points de rendez-vous (synchronisation sur les ilôts de confiance). La stratégie developpée dépend beaucoup des compétences des experts mis en oeuvre, c'est pourquoi il faut s'interroger sur le rôle exact de chacun d'entre eux. Le but de cet article est de discuter des niveaux linguistiques et plus particulièrement du module de compréhension C^2M^2.

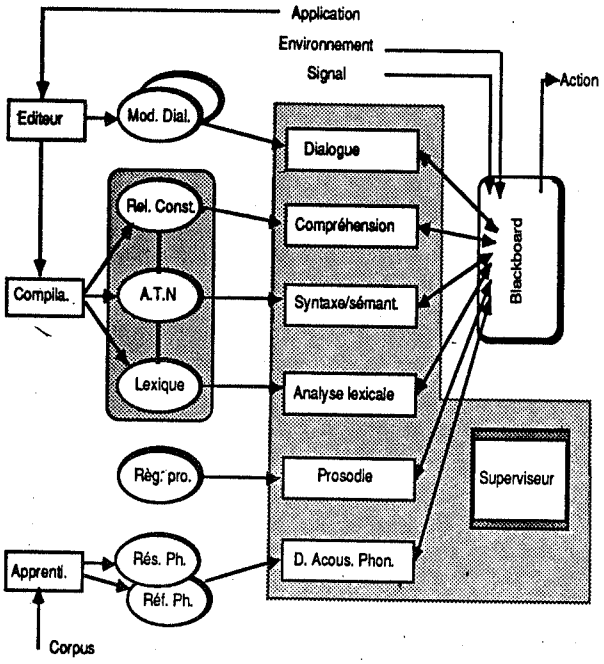


Fig 1: Architecture du système DIRA. Légende: Mod.Dial.: modèle de dialogue, Rel. Const. : relations entre les mots des constituants, ATN : réseau syntaxico-sémantique compilé, Lexique : lexique de l'application, Règ.Pro.: règles prosodiques, Rés. Pho. : réseaux phonétiques, Réf. Ph. : références phonétiques.

2. DE L'AUTONOMIE DES MODULES LINGUISTIQUES

Dans l'élaboration d'une architecture pour un système de compréhension de la parole et de dialogue, une première question est bien évidemment de déterminer quelles sont les composantes de ce système. Cette question n'est pas évidente pour les niveaux linguistiques : quelles sont en effet les relations qu'entretiennent le module de dialogue —par sa composante pragmatique— et l'analyseur syntaxico-sémantique (dans une architecture telle que ci-dessus) ? Faut-il un module de compréhension distinct de ces composantes ou doit-il être inclus dans la composante pragmatique ? Peut-on faire uns distinction utile entre compréhension partielle et compréhension complète ?

Classiquement, on distingue 3 niveaux : la syntaxe, la sémantique et la pragmatique souvent plus ou moins fondues les unes dans les autres. Certaines architectures (par ex. DIAL [Mousel et al, 89]) intégrent la compréhension dans le dialogue, au niveau de l'interprétation des messages, en utilisant la pragmatique de l'application pour instancier les schémas correspondants aux cas du verbe de la phrase : cela sous-tend l'idée que la tâche de compréhension reste sous la dépendance du contexte applicatif et que la compréhension n'a de sens qu'une fois que tous les actants ont été instanciés. Dans d'autres architectures (par ex. DIALORS [Luzzati, 89]) ce principe est généralisé en éliminant l'analyse syntaxique globale de l'énoncé et en privilégiant la sémantique et la pragmatique : les schémas sont instanciés directement à partir de la forme de surface de l'énoncé en fonction des contraintes fortes venant de l'application.

Ces deux options typiques, sous-tendent l'idée que la "compréhension" n'est possible qu'après une analyse pragmatique complète de l'énoncé. Nous pensons au contraire qu'une certaine compréhension partielle est déjà possible avant l'instanciation des schémas et peut suivre le même processus général de gauche à droite que les autres analyseurs. La question est donc : que peut-on comprendre dans un énoncé, sans l'appui du contexte de l'application ? ou plus précisément : y a-t-il une part de l'énoncé qui est "compréhensible en soi" ? Il est évident que la réponse à cette question conditionne l'architecture du système tant il est vrai que ce qui est *compréhensible en soi* pourrait être traité de manière plus précoce et indépendamment de la pragmatique —on éviterait de surcroît certaines compilations inutiles lors d'un changement d'application. Si cette hypothèse est correcte, elle conduirait à mettre en oeuvre un module spécifique de "compréhension partielle" placé entre l'analyse syntaxico-sémantique et l'analyse pragmatique.

Pour discuter plus à fond ce problème, examinons quelques hypothèses avancées par les psycholinguistes et les psychologues.

La signification d'un énoncé est véhiculée [Chomsky, 80] par :

(a) le sens attaché aux items lexicaux,

(b) les relations entre ces items, établies sur l'axe de l'énoncé,

(c) certaines connaissances spécifiques au domaine auquel se rapporte l'énoncé,

(d) la situation référentielle qui regroupe les entités perceptives et/ou conceptuelles de l'énoncé,

(e) la situation d'énonciation.

La question ci-dessus replacée dans le cadre de la signification, relève donc de la question de l'autonomie des traitements psycholinguistiques. La composante (b) relève de compétences proprement psycholinguistiques qui sont au centre de ce débat. A l'opposé les composantes (c) et (d) relèvent de mécanismes cognitifs généraux qui entretiennent des rapports étroits avec les connaissances pragmatiques. L'idée défendue par certains psycholinguistes est que les connaissances non spécifiquement psycholinguistiques interviennent relativement tardivement, après qu'une *forme logique* ait été assignée à chaque phrase. C'est bien là l'idée de l'hypothèse d'autonomie du traitement psycholinguistique qui conduirait à l'introduction d'un module spécifique entre l'analyseur syntaxico-sémantique et le dialogue. Ce module aurait alors à traiter de la composante (b) pour fournir la *forme logique* de l'énoncé. D'un autre côté les composantes (a) et (c) semblent liées par le biais des significations lexicales contextuelles au domaine —"morceau" prend un sens différent dans "morceau de musique" et "morceau de boeuf". La question de l'autonomie du traitement psycholinguistique est donc indissociable de celle de *l'insertion lexicale.*

Les items lexicaux apparaissant dans les énoncés ont en outre, la propriété de référer implicitement ou explicitement à des situations. Pour l'attachement des significations aux items, plusieurs modèles généraux semblent possibles :

(a) - apprentissage à partir de situations réelles

- généralisation de ces situations par typicalisation et construction d'archétypes (ou schémas)

- rappel de l'archétype comme contexte par défaut lors de la compréhension d'une phrase

- mise en correspondance de l'item et de la situation : le lien produit est appelé *sens*,

ou,

(b) - apprentissage sur des situations réelles

- mise en correspondance de ces situations à travers le langage

- généralisation des items lexicaux et des expressions langagières en informations abstraites et symboliques

- compréhension de la phrase par manipulation des symboles par un processus général (General Problem Solver) en vue du "calcul" d'un résultat appelé *sens*.

L'argument suivant plaide pour la solution (b) : l'être humain peut "comprendre" une phrase totalement nouvelle. A quelle situation la rattache-t-il alors ? Inversement, le choix du sens correct d'un mot ambigü est fait d'une manière étonnament rapide par l'humain en fonction du contexte. D'où l'idée d'organiser l'analyse comme en (a) non pas autour de règles générales mais d'après les interactions idiosyncrasiques des mots avec leurs contextes. De ceci il ressort que les deux stratégies sont possibles et que donc, la notion de contrôle devient essentielle pour répartir les rôles entre ces divers processus.

21. Quelques hypothèses et modèles

(a) l'hypothèse interactionniste lexique/sémantique

[Bever, 70] suppose qu'une des stratégie employée par l'être humain pour comprendre une phrase consiste à supposer que la séquence *Nom-Verbe-Nom* dans la structure de surface de la phrase correspond à la suite *Acteur-Action-Objet*. La machine perceptuelle de Bever, qui modélise la compréhension d'une phrase, a comme entrée la structure de surface de cette phrase et comme sortie, sa structure interne représentée par les relations entre les constituants majeurs. Après avoir expérimenté sur des sujets, [Fraser et al, 63] ont trouvé que les enfants interprètent la phrase "The cow was kissed by the horse" comme "The horse kissed the cow": à deux structures de surface différentes (formes active et passive) correspondent bien une seule structure interne. Pour des adultes le passage d'une forme à l'autre n'est pas aussi directe et dépend beaucoup du degré de cohérence entre syntaxe et sémantique. La question est donc de savoir quelle est la relation entre la structure *Nom-Verbe-Nom* (acteur-action-objet) et l'analyse syntaxique, puis comment le sujet intègre les résultats des différentes analyses.

Dans le cadre de cette théorie, [Forster, 79] a présenté un modèle de compréhension —extension de la stratégie de Bever— qui stipule que les constituants majeurs dans la phrase sont fonctionnellement reliés par des contraintes sémantiques. A côté des étages qui correspondent aux analyseurs lexical, syntaxique, sémantique, le modèle de Forster dispose d'un GPS (General Problem Solver). Dans certains cas, ce GPS peut dériver une interprétation de la phrase seulement par ré-arrangement des items lexicaux avant que l'analyse syntaxique de la phrase soit terminée. Pour soutenir sa théorie Forster cite les résultats d'une expérience faite par Ratcliff où des sujets sont capables de classer des phrases syntaxiquement incorrectes, en phrases acceptables ou non: cela signifie que dans la compétition syntaxe/compréhension les sorties des modules d'analyse sont confrontées avec des poids différents

—ce mécanisme n'est cependant pas encore complétement élucidé à l'heure actuelle. Il en déduit que le GPS doit être capable de donner une interprétation correcte pour une phrase où on a la possibilité (syntaxique) de changer les positions des constituants et interdire toute interprétation dans le cas contraire: c'est le phénomène de "lexical priming" (amorçage sémantique) c'est-à-dire de facilitation de l'identification d'un mot par la présentation préalable ou simultanée d'un autre mot relié sémantiquement.

[Kiger et al, 83] ont démontré l'existence du "backward priming" (amorçage rétrograde) temporaire qui influence la tâche de décision lexicale. Un mot présenté avec une latence de 50 ms après un mot cible peut faciliter la classification de la cible dans le cas où les mots sont fortement reliés sémantiquement. Des combinaisons de "forward et backward priming" temporaires peuvent augmenter les effets d'amorçage pour accélérer l'accès lexical et faciliter la classification des réponses des sujets (en diminuant les temps d'évaluation des représentations sémantiques de la phrase).

Les expériences qui sont faites dans ce cadre tentent de montrer que les adultes utilisent des stratégies de type Nom-Verbe-Nom proposées par [Bever, 70] et que la présence des relations entre les mots (amorçage sémantique) influencent la compréhension de la phrase bien avant la fin de l'analyse syntaxique (en tout cas pour des phrases incorrectes).

(b) l'hypothèse autonomiste

Sans nier les effets d'amorçage sémantique, les partisans de l'hypothèse autonomiste considèrent qu'il s'agit de phénomènes internes au lexique mental sans incidence sur les mécanismes proprement dits de compréhension [Fodor, 83]. Des expériences de "saturation sémantique" dans lesquelles les mots sont répétés un grand nombre de fois, mettent en évidence un ralentissement des phénomènes d'amorçage lexical. Ceci suggère qu'il faut distinguer entre diffusion de l'activation à l'intérieur du réseau responsable des phénomènes d'amorçage et accès aux significations lexicales. Fodor suppose que ces significations lexicales sont traduites dans un format compatible avec la forme logique de l'énoncé.

(c) l'hypothèse propositionnelle

Dans cette variante de l'hypothèse autonomiste, on suppose que les connaissances sémantiques sont mises sous forme prédicative, par exemple l'entrée lexicale "tuer" = causer(x, changer(état(y)), z) avec x=animé, y=mort, z=animé. Il reste à instancier les variables en recherchant dans l'énoncé les items adéquats.

Selon [Le Ny, 79] la compréhension d'une phrase comporte donc :

1. la décomposition sémantique des mots lexicaux qui permet de faire apparaître des prédicats élémentaires ainsi que les arguments qui devront être recherchés dans l'énoncé,
2. le ré-arrangement des arguments autour des prédicats élémentaires pour aboutir au sens global de l'énoncé.

Dans cette perspective la compréhension est guidée par l'analyse de la signification individuelle des mots lexicaux à travers une forme logique sans véritable analyse syntaxique. Néanmoins, celle-ci sera nécessaire chaque fois que des caractéristiques de surface devront être prise en compte.

Cette hypothèse selon laquelle les significations lexicales pourraient intervenir sous forme de règles (ou propositions) a été également défendue par Fodor. Mais pour lui, la compréhension comporterait deux étapes distinctes quasiment inversées par rapport à la précédente :

1. une analyse syntaxique permettant d'associer une forme logique à l'énoncé,
2. l'application sur cette forme logique des règles définissant le domaine.

(d) l'hypothèse inférentielle

Cette approche [Wilson, 79] donne une grande importance au contexte dans l'interprétation des énoncés. Cependant, le processus de déduction dont dépend l'attribution d'une signification pertinente, démarre après qu'une forme logique ait été assignée à l'énoncé.

(e) l'hypothèse procédurale

[Dowty et al, 81] ont appliqué la théorie des modèles procéduraux au langage naturel. Aux noms propres sont associés des individus, aux noms communs et aux adjectifs des ensembles d'objets présentant telle ou telle caractéristique, aux verbes transitifs des ensembles de couples d'individus ou d'objets, etc. Ce type de définition en extension peut être remplacé par des fonctions permettant d'associer une valeur de vérité (Vrai/Faux) au mot lexical considéré (tout individu du monde réel, tout couple d'individus, etc.). L'ensemble des valeurs sémantiques associées aux mots lexicaux constituent le modèle par rapport auquel les énoncés sont interprétés. Des règles sémantiques, permettent de déterminer la valeur de vérité d'une phrase à partir des valeurs sémantiques assignées à ses constituants —par exemple <Si X est un nom propre et Y est un adjectif la phrase (X est Y) est vraie si la valeur sémantique de X appartient, au sens de la théorie des ensembles, à la valeur sémantique de Y>. La valeur sémantique d'un syntagme nominal complexe déterminé par l'intersection de la valeur sémantique de ces constituants, et la valeur sémantique d'une phrase sont aussi l'intersection des valeurs sémantiques de leurs constituants. On peut dire qu'un modèle mental associé à un énoncé est décrit comme une représentation mentale de la situation dénotée par l'énoncé. Ceci suppose que l'auditeur soit capable de construire cette représentation —d'où l'idée de *sémantique procédurale*.

[Winograd, 72] a réalisé le programme SHRDLU qui permet de commander un robot en langage naturel et de dialoguer avec lui. Interpréter un énoncé selon ce programme, c'est construire et activer un programme qui exécutera la commande correspondante. Le nom du programme est fourni, directement par le verbe principal de la phrase. L'essentiel du travail réside dans la recherche des référents des syntagmes nominaux qui sont passés comme des arguments au programme.

2.2. Pour une structure logique intermédiaire

Après cette présentation rapide de quelques modèles, il semble qu'une majorité d'auteurs attribue une réalité à l'existence d'une *forme logique* comme étape nécessaire et/ou concurrente à la syntaxe dans le processus de compréhension. Le problème de l'*insertion lexicale* n'a pas d'adhésion consensuelle —elle est souvent placée par les auteurs *après* le calcul de la forme logique (mais parfois avant *et* après).

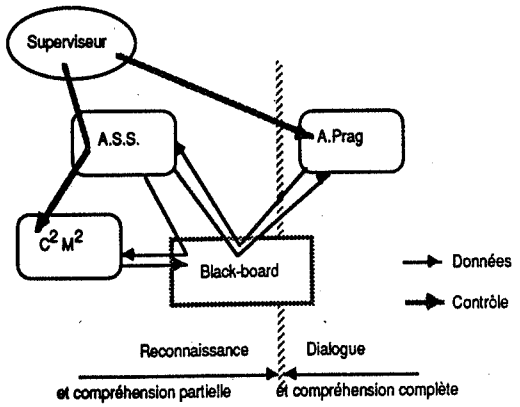
Donc dans l'état actuel des choses, une idée pour le traitement automatique de la parole est d'insérer un module spécifique entre la syntaxe et la pragmatique. Ce qui est le plus facilement réalisable est de mettre en oeuvre le modèle de Fodor (cf. hypothèse c) qui préserve l'autonomie pshycholinguistique et qui s'exprime sous forme de règles, pour simuler l'amorçage sémantique entre les constituants de la phrase.

De manière plus séquentielle —et simplificatrice— on peut donc concevoir deux étapes dans l'analyse : la première qui s'appuie sur les marqueurs sémantiques principaux de la phrase (mots, groupes) et qui produit un premier type de compréhension par "mise en relation des éléments de sens" et une deuxième qui ne permet une véritable compréhension qu'en contexte, à l'aide de schémas par exemple. C'est ce double point de vue qui est mis en oeuvre dans le système DIRA :

- le module C^2M^2 (Compréhension en Constituants Majeurs et Mineurs) s'occupe des processus concernant l'amorçage sémantique lié à la forme logique de l'énoncé,
- le module Dialogue s'occupe, entre autres, de l'instanciation des schémas au niveau pragmatique.

Une autre série d'arguments non linguistiques cette fois, plaide en faveur de l'introduction d'un module de compréhension en aval du module de dialogue; ce sont des arguments (1) liés à l'architecture choisie dans le système DIRA —réduire le facteur de branchement lexical typique d'une stratégie gauche-droite avec retour arrière— et (2) liés aux difficultés propre à la reconnaissance du langage oral —construire une structure sémantique s'appuyant le plus possible sur les mots lexicaux, ce qui évite, dans une certaine mesure, de reconnaitre parfaitement les mots grammaticaux.

Une solution est donc d'introduire dans le système, sous le contrôle du superviseur, un module, le C^2M^2 (Compréhension en Constituants Majeurs et Mineurs) qui reste sous la dominance de l'ASS (Analyseur syntaxico-Sémantique) (Fig. 2) :



avec :
ASS = Analyseur Syntaxico-Sémantique,
A.Prag = Analyseur Pragmatique,
C^2M^2 = Compréhension en Constituants Majeurs et Mineurs
Fig. 2 : Flux de données et contrôle entre les modules linguistiques de haut niveau dans le système DIRA

3. LA COMPREHENSION EN CONSTITUANTS MAJEURS ET MINEURS DANS LE SYSTEME DIRA

Pour rester compatible avec l'ensemble de la stratégie choisie pour le système DIRA, la compréhension doit se faire au fur et à mesure de la reconnaissance c'est-à-dire dans le sens gauche-droite, tantôt en vérification tantôt en prédiction. La vérification doit se faire sur des séquences de mots —constituant des phrases incomplètes— présentes dans le black-board à un instant donné. La prédiction doit permettre de faire des hypothèses sur quelques structures de phrase et mots possibles —par *amorçage sémantique*— utilisables par l'analyseur syntaxico-sémantique.

3.1. La composante pragmatique

Nous ne détaillerons pas cette composante qui présente toutes les caractéristiques habituelles que l'on rencontre dans les modules de dialogue, à savoir la base de connaissance statique décrivant les objets de l'univers, les sites géographiques, les actions possibles sur ces objets (l'application est un univers de robotique), les faits et les connaissances dynamiques telles que les historiques. En particulier, cette composante a pour tâche d'instancier les schémas relatifs aux énoncés.

3.2. La composante syntaxico-sémantique

Cette composante délivre la c-structure et la f-structure conformément au modèle LF-Grammar (grammaire lexicale fonctionelle) de Bresnan/Kaplan [Kaplan and Bresnan, 82] mis en oeuvre dans un ATN par Reynier [Reynier et al, 89]. Elle contrôle la composante d'amorçage sémantique (C^2M^2) via le superviseur. Elle n'est pas détaillée dans cet article.

3.3. La composante C^2M^2, compréhension en constituants majeurs et mineurs

Cette composante peut fonctionner en vérification et en proposition comme les autres modules du système.
(a) en vérification, elle établit les relations, de type "forward" et "backward", entre les constituants de la phrase, et écrit dans la mémoire commune un drapeau qui indique le degré de compréhension atteint (pour le moment c'est une réponse binaire "vrai" ou "faux").
(b) en proposition, elle prédit les structures de sens qui peuvent convenir après une séquence de mots donnés.

Actuellement, seule la partie (a) a été implantée dans le système de la manière suivante :

Les relations entre les constituants majeurs sont classées dans des groupes de relations qui dépendent de la structure fonctionnelle de la phrase (ce sont des prédicats Prolog). Les relations entre les constituants mineurs portent sur les mots se trouvant à l'intérieur d'un même groupe syntaxique. Ceci est illustré ci-après pour une grammaire finalisée.

On peut distinguer plusieurs grands types de phrases dans l'application de robotique envisagée ici :

1. les phrases de type (GV) :

$$P \text{ ---> } GV \quad p = f$$

$$GV \text{ ---> } V \quad [Adv] \quad p = f$$

L'analyse fonctionnelle de la phrase est réduite dans ce cas à sa plus simple expression, puisqu'il n'y a qu'un seul constituant majeur. La compréhension de ce type de phrase est donc très simple et consiste à mesurer l'amorçage sémantique entre les mots lexicaux (V, Adv) (constituants mineurs) à l'intérieur du GV. Le schéma des verbes compatibles avec cette structure de phrase, est :

verbe : temps = infinitif
concept=<verbe, COD=vide, COI=vide>
modifieur = adverbe
avec COD=complément objet direct ; COI=complément objet indirect.

Les prédicats d'amorçage sémantique (AS) entre les constituants mineurs sont associés aux verbes (en réalité sous la forme d'une clause Prolog) de la manière suivante :

AS(verbe,adverbe) :
aller(manière, vitesse);
arrêter(nil);
etc.

Ex : La phrase "arrêter" est compréhensible en soi au sens défini dans le §2.2. Bien sûr la compréhension complète —au sens d'action possible— n'est assurée que si l'historique de navigation du robot contient une commande précédente de mouvement comme (aller, avancer, etc.) : ceci n'est pas du

ressort de C^2M^2 mais de l'analyseur pragmatique. Par contre la phrase "aller chaudement" est rejetée avant l'analyse pragmatique, puisque "chaudement" n'a pas de sème de manière ni de vitesse.

2. les phrases de type (GV GN) :

P ---> GV GN
p = f (p COD) = f

GN ---> [Art] N
p = f

GV ---> V [Adv]
p = f

L'analyse fonctionnelle de la phrase se ramène dans ce cas à celui du groupe verbal, tandis que la structure fonctionnelle du complément d'objet direct est celui du groupe nominal.

Le verbe est représenté par :

verbe : temps = infinitif
concept = <verbe, COD, COI=vide>
modifieur = adverbe

Pour comprendre ce type de phrase il faut évaluer un possible amorçage sémantique du type "forward" entre le verbe du GV et le nom du GN (constituants majeurs). Donc il faut établir une relation logique du type (verbe, nom), puis il faut procéder à une évaluation de l'amorçage à l'intérieur des deux groupes, ce qui se simplifie pour le deuxième (GN) qui ne contient qu'un mot lexical. Pour cela on dispose des prédicats suivants :

AS(verbe, nom) :
arrêter(instrument,procédure);
longer(batiment,partie_batiment,meuble);
etc.

et des prédicats précédents AS(verbe, adverbe) pour les constituants mineurs du GV.

3. les phrases de type (GV GP) :

P ---> GV GP
p = f (p COI) = f

GP ---> Prép GN
p = f

L'analyse fonctionnelle de la phrase se ramène dans ce cas à celui du groupe verbal, tandis que la structure fonctionnelle du complément d'objet indirect est celui du groupe prépositionnel, et celle du groupe prépositionnel est celui du groupe nominal.

Les verbes sont représentés par :

verbe : temps = infinitif
concept = <verbe, COD=vide, COI>
modifieur = adverbe

Pour comprendre ce type de phrase il faut mesurer l'amorçage sémantique du type forward entre le verbe du GV et la préposition du GP, puis l'amorçage de type forward entre le verbe du GV et le nom du GN et enfin l'amorçage entre la préposition du GP et le nom du GN. Pour cela il faur disposer des relations logiques de type (verbe, prép), (verbe, nom) et (prép, nom)

Cette séquence de vérification est importante pour comprendre la phrase de gauche à droite comme il est illustré dans la fig. 3.

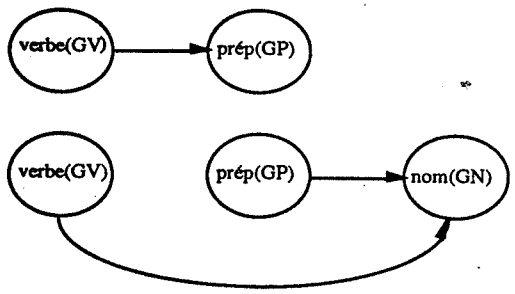


Fig. 3: L'amorçage sémantique dans la phrase de type (GV GP)

Ex. : la phrase "aller vers la porte" devient compréhensible si l'on arrive à établir les relations entre (aller, vers), et (aller, porte) puis entre (vers, porte).

4. etc. pour tous les autres types de phrase possibles

3.4. La stratégie de compréhension

La compréhension d'une phrase est soumise à deux conditions :
a. elle doit être syntaxiquement correcte et,
b. l'amorçage sémantique entre les mots lexicaux doit pouvoir s'établir.

La stratégie consiste donc à activer l'ASS (analyseur syntaxico-sémantique) en mode descendant de gauche à droite pour vérifier tout d'abord que les structures sont correctes. Dans un tel cas l'ASS envoie un message au superviseur pour lui indiquer d'activer le module C^2M^2 dont le rôle est alors d'établir l'amorçage sémantique entre les constituants déterminés par l'ASS et présents dans le blackboard. En cas de succès la phrase (ou porion de phrase) est dite "compréhensible" ce qui permet au superviseur de valider ses hypothèses.

4. RESULTATS

Une des mission importante du module de compréhension (en mode vérification) dans le système DIRA est de filtrer les hypothèses de mots au cours de l'exploration de gauche à droite des phrases. La Fig. 4 montre la comparaison entre deux courbes visualisant le branchement lexical pour la phrase "lâcher la clé", avec et sans ce module de compréhension pour la même chaîne phonétique d'entrée.On constate bien évidemment que le facteur de branchement dans le premier cas est constamment inférieur au second, avec un rapport significatif (réduction dans un rapport 4). A elle seule, cette amélioration justifie déjà pleinement l'existence d'un tel module.

Par ailleurs cette composante accélère le traitement sémantique dans la mesure où elle évite d'instancier systématiquement les schémas des objets et des tâches liés à l'application.

5. CONCLUSION

L'amorçage sémantique introduit dans ce système remplit bien le rôle attendu dans le filtrage des hypothèses lexicales en cours de production de gauche à droite :
- il diminue le facteur de branchement lexical à chaque instant dans un rapport intéressant,

- il accélère le filtrage dans la mesure où il élimine déjà quelques hypothèses avant l'utilisation des informations pragmatiques —en évitant notamment l'instanciation de certains schémas de la composante pragmatique,
- il permet d'obtenir une forme logique de l'énoncé qui peut éventuellement permettre d'éviter de reconnaître systématiquement tous les mots.

Un tel modèle semble compatible avec les hypothèses actuelles concernant les processus de compréhension humains à savoir le calcul d'une forme logique avant l'insertion des connaissances spécifiques au domaine auquel se rapporte l'énoncé.

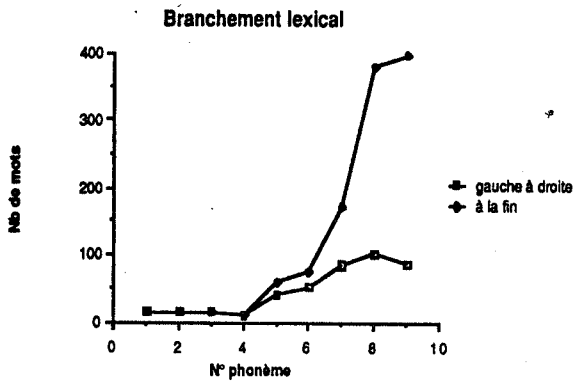


Fig. 4. comparaison entre le branchement lexical dans les cas où on a (1) un processus de compréhension progressif de gauche à droite et (2) un processus de compréhension à la fin de la phrase seulement.

6. REFERENCES

CAELEN, J. et TATTEGRAIN, H. (1988)
Le décodeur acoustico-phonétique DIRA-DAP
Actes 17° JEP, Nancy, pp. 115-121.

BEVER. T.G. (1970)
"The cognitive basis for linguistic structures."
In J.R. Hayes (edition), cognition and the development of language (New York), pp. 279-362.

BRANSFORD, J. & JOHNSON, M. (1972)
"Contextuelle prerequisites for understanding : Some investigations of comprehension and recall."
Jounal of Verbal Learning and Verbal Behavior , 11, pp. 717-726.

BRESNAN, J. and KAPLAN, R.M. (1982)
Introduction: Grammar as mental representations of language. The mental representations of grammatical relations. In Bresnan ed., Cambridge Mass. & London, MIT Press.

CHAFE, W. L. (1974)
"Language and consciousness."
Linguistic Journal, 50, pp. 111-133.

CHOMSKY, N. (1980)
"Rules and Representation"
NewYork : Colombia University Press.

DOWTY, D.,WALL,R., PETERS, S. (1981)
"Introduction to Montague semantics"
Dordrecht : Hollande, Reidel.

P. FALZON (1989)
"Ergonomie cognitive du dialogue"
Presses Universitaires de Grenoble

FODOR, J. D. (1980)
"Semantics : theories of meaning in generative grammar."
Cambridge : Harvard University Press.

FODOR, J.D. (1983)
The Modularity of Mind. Cambridge, MA : the MIT Press.

FORSTER, K. I. (1979)
"Levels of processing and the structure of the language processor."
In W.E. Cooper and E.C.T. Walker (eds.), Sentence Processing: Psycholinguistic Studies , pp. 216-225.

FRASER, C., BELLUGI, U., BROWN, R.W. (1963)
"Control of grammar in imitation, comprehension, and production."
Journal of Verbal Learning and Verbal Behavior, 2, 121-135.

HALLIDAY, M. A. K. (1967)
"Notes on transitivity."
Part II Journal linguistic , 3, pp 199-244.

KAPLAN, R.M. and BRESNAN, J. (1982)
Lexical Functional Grammar : a formal system for grammatical representations. The mental representations of grammatical relations. In Bresnan ed., Cambridge Mass. & London, MIT Press.

KEMPER, S., CATLIN, J. (1979)
"On the role of semantic constraints in sentence comprehension."
LANGUAGE AND SPEECH, Vol.22, Part 3.

KIGER, J. I., GLASS, A. L. (1983)
"The facilitation of lexical decisions by a prime ocurring after the target."
Memory and Cognition, 11, 356-365.

LE-NY, J. F. (1979)
"La sémantique psychologique."
Press Universitaire, Paris.

D. LUZZATI (1989)
"Recherches sur le dialogue Homme-Machine: modèles linguistiques et traitements automatiques"
Thèse de doctorat d'Etat, Paris III.

MOUSEL, P., PIERREL, J.M. et ROUSSALANY, A. (1989)
Coopération entre syntaxe, sémantique et pragmatique dans un système de dialogue homme-machine
Actes 7° congrès AFCET-RFIA, Paris, pp. 371-386.

NASRI, M.K., CAELEN-HAUMONT, G. et CAELEN, J. (1989)
Using Prosodic Rules in Continuous Speech Recognition
Proc. of ICASSP-IEEE, Glasgow, Vol. 1, pp. 671-674.

REYNIER, E. et CAELEN J. (1989)
ATN Compiler and Parser for an ASR system
Proc. of EUROSPEECH'89, Paris, pp. 398-401.

WILSON, D. and SPERBER, D. (1979)
Remarques sur l'interprétation des énoncés selon Paul Grice.
Communication , 30, pp. 80-94.

WINOGRAD, T. (1972)
Understanding Natural Language. New York Academic Press.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

LE RÔLE DU DIALOGUE POUR LA RECONNAISSANCE DE PAROLE. LE CAS DU SYSTÈME PAGES JAUNES. [1]

GUYOMARD M.[o] SIROUX J.[*]
COZANNET A.[+]

o ENSSAT-IRISA, BP 447, F-22305 LANNION CEDEX
* IUT-IRISA, BP 150, F-22301 LANNION CEDEX
+ CNET, Département LAA/TSS/RCP, BP 40, F-22301 LANNION CEDEX

RÉSUMÉ Le caractère prédictif de certains types de dialogues peut être utilisé pour contraindre et améliorer les processus linguistiques de la reconnaissance et de la compréhension de la parole. Après un bref tour d'horizon de la mise en œuvre des prédictions dans certains systèmes de dialogue oral, nous développons le cas du système Pages Jaunes. Enfin nous étudions l'influence des stratégies de dialogue sur le comportement langagier du locuteur. Nous montrons en particulier quelques avantages de certaines formes de coopération.

Mots clés : Dialogue oral, compréhension de parole, prédictions, coopération, inattendus, stratégies de dialogue.

1. INTRODUCTION

Historiquement, la recherche de meilleures performances pour les systèmes de reconnaissance et de compréhension de la parole (RCP) fondés sur l'identification d'unités linguistiques peut s'analyser comme la représentation et la mise en œuvre de sources de connaissances visant à contraindre et à mieux contrôler le traitement associé à un niveau linguistique donné, à partir d'informations de niveaux supérieurs. L'apparition du niveau dialogue confirme cette analyse mais ajoute, de par le caractère évolutif du dialogue, une dimension prédictive inconnue jusque là. Comment alors représenter, répercuter et utiliser les informations issues du dialogue? Durant la réalisation du système de dialogue oral Pages Jaunes (PJ) [BIGO,88] [GUYO,88], nous avons été conduits à nous intéresser au rôle du composant dialogue pour la RCP, et à mettre en œuvre des solutions pratiques. Sur un autre plan, les études de dialogue oral ont, depuis quelques années, acquis une certaine autonomie. Les avancées qui en résultent sont elles compatibles avec une meilleure reconnaissance de parole? C'est une question à laquelle nous tentons de proposer ici un début de réponse.

Le rôle du dialogue peut être considéré sous deux aspects. Le premier, interne, technique, concerne la manière dont on peut tirer le meilleur parti (informatiquement parlant) du caractère prédictif de l'interaction; les paragraphes 2, 4 et 5 lui sont consacrés. Le second se rapporte aux effets de la stratégie de dialogue sur le comportement linguistique des locuteurs. Ces effets sont analysés au paragraphe 6. Le paragraphe 3 décrit de manière rapide l'architecture du système.

2. LES PRÉDICTIONS

Un grand nombre d'activités humaines observables tant par leurs conséquences mentales que physiques, ne s'expliquent que par le fait que leurs auteurs ont une idée consciente ou non, de ce qui constituera le futur (immédiat ou lointain).

La classe d'activités qui nous intéresse ici plus particulièrement correspond aux processus perceptifs et cognitifs mis en œuvre dans la communication, comme dans les deux exemples suivants où B anticipe la fin de la prononciation du nom de la ville par A (phénomène d'"amorçage lexical" [LE N,80]).

(1) A : J'ai deux possibilités Lannion et Plou
B : Plouaret, d'accord

(2) A : J'ai deux possibilités, Lannion et Plou
B : Ploubezre
A : non, Plouaret

Ce type de réaction semble être fondé pour une bonne part, sur des connaissances de nature hétérogène soumises à des adaptations incessantes en fonction du déroulement de la conversation.

Malgré quelques risques de résultats erronés (dialogue 2), il est clair que de tels processus (gestion et utilisation des connaissances) doivent être utilisés dans les systèmes de RCP. Nous appellerons prédictions les informations linguistiques qui permettent de favoriser, sous une forme ou une autre, les traitements de RCP.

2.1 Utilités des prédictions en RCP et dialogue

Les prédictions peuvent être utilisées de manière efficace à deux moments du traitement d'un énoncé de l'utilisateur : durant le processus de RCP, et durant le processus d'interprétation pour le dialogue. Durant la RCP, elles permettent de diminuer le nombre de mots, de limiter le choix des constructions syntaxiques réduisant ainsi l'espace de recherche. Pour l'interprétation, le fait de n'avoir en entrée que des éléments cohérents par rapport à l'état courant du dialogue, élimine les traitements de vérification et les sous-dialogues qui pourraient en résulter.

2.2 Quelques mises en œuvre

Les prédictions ont été utilisées de manière plus ou moins explicite dans bon nombre de systèmes [LEVI,85] [BRIE,86] [CARB,89]. Nous

1. Ce travail a été mené au sein de département TSS/RCP du CNET LANNION en partie dans le cadre d'une convention CNET-IRISA (n° 87 7B 075 00 790 9245 LAA).

détaillons ici trois mises en œuvre explicites de prédictions.

2.2.1 KEAL+CADI Dans le système constitué par les modules KEAL (partie RCP) et CADI (module de gestion du dialogue (MGD)) [SIRO,84] [SIRO,85], les prédictions sont essentiellement syntaxico-sémantiques. Leur production est principalement fondée sur le modèle de dialogue. Une prédiction émise vers les modules syntaxique et sémantique comprend la grammaire (grammaire sémantique) et les règles à utiliser pour l'interprétation sémantique de l'arbre syntaxique. Fournir la grammaire revient aussi implicitement à fournir le vocabulaire terminal qui pourra être trouvé dans l'énoncé du locuteur.

Ce type de fonctionnement permet de réduire considérablement l'espace de recherche durant l'analyse syntaxique (par rapport à une analyse qui prendrait en compte une grammaire générale de l'application) et d'éviter que le MGD reçoive des informations pragmatiquement non valides. Cependant deux inconvénients majeurs existent : le grain de définition de la prédiction n'est pas suffisamment fin : le MGD dispose d'un catalogue de prédictions figées, prédire revient à choisir l'un des éléments du catalogue sans pouvoir l'adapter à la situation particulière de dialogue (notamment pour le vocabulaire). On peut ainsi trouver des séquences de dialogue du type :

(3) A : à Lannion.
B : à Toulon ?
A : non.
B : où ?
A : à Lannion.
B : à Toulon ?

Ensuite le mode de fonctionnement de KEAL ne permet pas de répercuter les prédictions induites du vocabulaire terminal sur l'analyse lexicale.

2.2.2 VODIS Le système VODIS [PROC,89] [YOUN,89a] fonctionne, en ce qui concerne le dialogue, sur des fondements assez semblables à ceux de CADI. Les prédictions, émises à partir d'une grammaire sémantique et du modèle de dialogue, sont utilisées d'abord de manière peu contraignante pour un premier filtrage syntaxique, puis de manière plus stricte, pour terminer l'analyse linguistique. Il nous semble que, fondées sur une représentation du dialogue (par frame) hiérarchique et non complète (du point de vue contrôle du dialogue), ces prédictions sont encore trop rigides.

2.2.3 MINDS Dans le système MINDS [YOUN,89b], les prédictions sont produites à partir d'un ensemble de connaissances (application, dialogue) modélisé en réseau. Deux aspects particuliers concernant les prédictions sont à noter. D'abord les contraintes déduites des prédictions sont utilisées dès le niveau acoustico-phonétique du système. Ensuite pour pallier les problèmes d'échec (de reconnaissance ou de prédiction mal adaptée), le système produit plusieurs ensembles de prédictions allant du plus spécifique au plus général. Pour les prédictions les plus spécifiques, le système utilise toutes les contraintes possibles (pragmatiques, sémantiques, syntaxiques). Les ensembles successifs de prédictions deviennent de plus en plus généraux en considérant des espaces plus vastes du réseau de connaissances. Éventuellement, le dernier niveau de prédictions atteint peut correspondre à l'ensemble de la grammaire (sémantique) avec tous les mots possibles de l'application. Avec ces ensembles de prédictions, il est ainsi possible en cas d'échec de l'analyse d'un énoncé de reprendre l'analyse en considérant un ensemble de contraintes plus générales. Cependant il semble que certains énoncés (gestion de communication) ne puissent être traités qu'à l'aide de prédictions de niveau assez général (donc après un échec).

3. ARCHITECTURE DU SYSTÈME PAGES JAUNES

La figure 1 fournit une vue schématique du système PJ.
Le module MGD est considéré comme un intermédiaire entre trois entités : l'utilisateur, la partie reconnaissance et analyse linguistique (RCP) et l'application. Il gère sous un même formalisme les différents flux de données nécessaires. Le lecteur trouvera à la figure 2 un exemple de dialogue effectivement réalisé avec le système PJ ainsi que des détails sur les constituants du système dans [BIGO,88] [GUYO,88].

4. PRÉDICTIONS DANS LE SYSTÈME PAGES JAUNES

Ce que nous appelons prédictions dans le système PJ regroupe en fait deux types d'informations : le premier concerne les éléments nécessaires aux traitements linguistiques des ellipses et des anaphores, le second les prédictions proprement dites (vocabulaire et syntaxe). Nous examinons en premier lieu les éléments du MGD qui permettent d'expliquer le processus de génération des prédictions, avant de fournir quelques détails sur celles-ci.

4.1 Les actes de dialogue et les énoncés

Du point de vue dialogue, nous considérons les énoncés échangés durant l'interaction comme des actions de dialogue qui possèdent une fonction et éventuellement un contenu propositionnel. En faisant référence à la théorie des actes de langage [SEAR,72], ces actions sont appelées actes de dialogue (AD) [BUNT,80].

Dans notre cadre applicatif, le langage utilisé est assez stéréotypé [GUYO,87]. Il est cependant possible d'observer plusieurs possibilités de verbalisation pour un même acte de dialogue. Ces possibilités dépendent du locuteur mais aussi de l'état d'avancement du dialogue, par

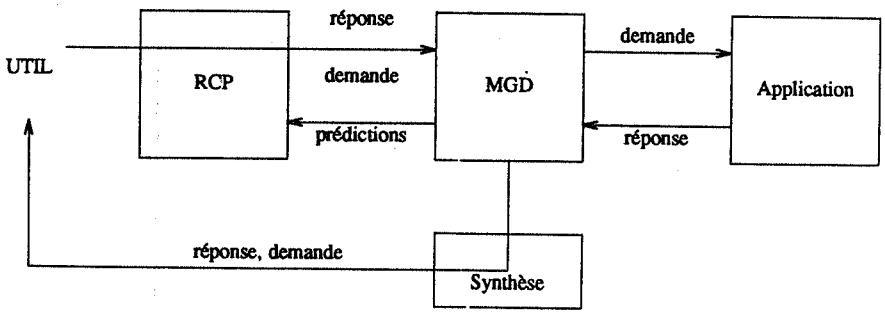


Figure 1. Structure du système PJ.

exemple l'énoncé de surface correspondant à une demande ne pourra pas être codé elliptiquement en début de dialogue.

S — Pages jaunes informatiques bonjour, veuillez donner votre demande simplement.
U — Je recherche un loueur de camionnettes.
S — Dans quelle localité souhaitez-vous la recherche?
U — À Plouaret.
S — Je n'ai pas d'information concernant cette rubrique mais je dispose de 2 rubriques proches pour tenter la recherche

— garage d'automobiles et

— location d'automobiles.

Souhaitez-vous que j'effectue la recherche dans l'une de ces rubriques?
U — Location d'automobiles.
S — Je recherche dans la localité de Plouaret dans la rubrique location d'automobiles.
Veuillez patienter. .
Il n'y a pas d'abonné dans la rubrique location d'automobiles dans la localité de Plouaret ni dans son voisinage,
mais il y a 2 abonnés à Plouaret à la rubrique garage d'automobiles.
Souhaitez-vous cette liste?
U — Non.
S — Voulez-vous un autre renseignement?
U — Non merci.
S — Au revoir.

Figure 2. Exemple de dialogue avec le système PJ.

4.2 Structure du dialogue

La structure du dialogue géré par le système est fortement influencée par le type d'application (dialogue informatif pour base de données) et par la stratégie de dialogue mise en œuvre (coopération, relance du dialogue par réponses correctives et suggestives). Une des conséquences de ceci est que nous avons pu fonder la modélisation du dialogue sur une grammaire.

L'état du dialogue à un instant donné est représenté par l'arbre de dérivation ainsi que par des informations pertinentes rangées dans des structures de données (l'historique de dialogue). Il est donc possible en utilisant le caractère prédictif de la grammaire d'énumérer les actes qui sont susceptibles d'apparaître dans le prochain tour de parole du locuteur, et d'autre part, en considérant les valeurs d'informations déjà évoquées, de restreindre le vocabulaire.

4.3 Le problème du contexte parole

L'utilisation de l'oral comme moyen de communication peut provoquer à tout moment des interventions dont le but est de gérer le canal de communication. Nous distinguons deux niveaux de gestion : le niveau système dont les problèmes de communication correspondent à ceux de la reconnaissance (pas de reconnaissance, score médiocre) et le niveau utilisateur dont les problèmes se traduisent par des demandes de répétition ou d'épellation. Ces niveaux sont gérés de manière parallèle au dialogue normal à l'aide de règles qui prennent en compte le contexte d'apparition des problèmes. Ils permettent de prédire les actes nécessaires à la gestion des problèmes de communication.

4.4 Prédiction élémentaire, prédiction transmise à l'analyseur linguistique

Lorsque le MGD émet un acte de dialogue susceptible de provoquer une réaction de l'utilisateur, il produit un ensemble de prédictions correspondant aux différentes réponses (AD) possibles de l'utilisateur. Une prédiction pour un acte sera de la forme :

— nom de l'acte prédit;
— liste de noms de verbalisation. Ces noms de verbalisation dépendent du contexte de dialogue. Ils sont interprétés par l'analyseur;
— restrictions de vocabulaire sur certaines valeurs de paramètres. Elles permettent soit d'interdire certains mots, soit de restreindre le choix à une liste fermée de mots.

Une prédiction transmise à l'analyseur est un triplet. Le premier élément rappelle les paramètres de la dernière demande de l'usager; il contient principalement la rubrique et la localité (informations extraites de l'historique de dialogue). Le second élément est composé d'une liste d'items. Chaque item comprend les informations relatives à un professionnel transmises à l'utilisateur. Cette liste servira à résoudre les problèmes référentiels. Le dernier élément est une liste de prédictions élémentaires représentant les différentes alternatives d'occurrences d'actes possibles.

4.5 Problèmes et conclusions

Les prédictions telles que nous les présentons, possèdent des avantages autant du point de vue reconnaissance que du point de vue dialogue. Les principes sur lesquels s'appuie le processus de génération (structure du dialogue, contexte applicatif, règles implicites de la conversation) ne permettent pas cependant de faire face à toutes les situations générales de dialogue (changement de thème par exemple). Toutefois, l'état de l'art actuel en RCP interdit de viser, pour l'instant, la prise en compte de ces situations générales.

5. TRAITEMENT DES PRÉDICTIONS DANS L'ANALYSEUR

L'efficacité de l'analyse linguistique est grandement améliorée si l'on peut réduire dynamiquement, c'est-à-dire en fonction du contexte de l'énoncé à reconnaître l'extension des vocabulaires et grammaires sur lesquels portera l'analyse.

5.1 L'analyseur ALOEMDA

ALOEMDA est un analyseur linguistique conçu pour la reconnaissance d'énoncés oraux ou écrits [COZA,87]. Il accepte en entrée un treillis de détections lexicales. Nous appelons *détection lexicale* un quadruplet (**m, i, f, s**) où : **m** désigne une entrée du lexique (mot), **i** et **f** désignent les positions temporelles de début et de fin de la détection, **s** désigne un score, ou mesure de la ressemblance du mot **m** avec la tranche temporelle de signal située entre **i** et **f**. Cette description d'une détection présente une forte analogie avec celle d'un arc au sein d'un graphe étiqueté ou d'un réseau de transition. On peut alors décrire une détection lexicale D comme l'étiquette de la transition du noeud **i** au noeud **f**, et **s** est alors vu comme une mesure de la probabilité d'emprunter D à partir de **i**, ou un coût sur un chemin contenant D. Cette analogie a été exploitée dans ALOEMDA, où les grammaires sont écrites sous forme de réseaux récursifs, augmentés (description de type ATN). Cependant, contrairement aux ATNs, la stratégie n'est pas implicitement induite par cette représentation.

La méthode d'analyse choisie est la méthode du diagramme actif [WINO,83], en lecture gauche droite. Aloemda cherche à construire une

transition entre le nœud initial (instant de début de l'énoncé) et le nœud final (instant de fin d'énoncé), en combinant les détections trouvées au moyen de la grammaire. Chaque nouvelle transition est munie d'un score, calculé à partir des scores des transitions qu'elle recouvre. La progression de la construction du diagramme est obtenue soit par ajout de nouvelles transitions *détections lexicales* à la suite des prédictions sur les successeurs possibles d'une transition active existante, soit par combinaison d'une transition *active* avec une transition *complète*. Une transition active correspond à un constituant de phrase incomplet : le réseau qui décrit un tel constituant permet de prédire les transitions possibles à partir de l'état atteint. Une transition complète est formée par un constituant achevé.

Un énoncé est dit reconnu si Aloemda trouve une transition recouvrant la totalité du signal d'entrée.

5.2 Extensions nécessaires pour le dialogue

En situation de dialogue avec prédictions, l'analyse doit s'adapter, selon deux axes, au cours de l'avancement du dialogue : sur l'axe lexical, en restreignant les détections au vocabulaire pertinent, et sur l'axe grammatical, en construisant une grammaire particulière à partir des schémas de réseaux. Pour cela, on substitue à l'axiome de la grammaire un nouveau réseau ne contenant que deux états (état initial EI et état final EF) et autant d'arcs que la prédiction contient de verbalisations différentes. Chacun des arcs ainsi générés a pour étiquette le nom d'une verbalisation, et dans sa partie action, l'appel à une procédure d'interprétation de l'énoncé reconnu, qui génère la forme interne à transmettre au MGD.

Dans une version "classique" de l'analyseur, le vocabulaire global et l'extension des classes lexicales sont gérés de façon implicite. Pour tenir compte de l'aspect dynamique des prédictions fournies par le MGD et en utilisant encore les mêmes procédures d'appel et de choix lors de l'analyse, nous avons d'une part défini explicitement pour chaque réseau les extensions (lexique et classes), et créé des registres statiques (utilisés comme valeur par défaut) et dynamiques (dont le contenu précise l'extension pour la situation de dialogue en contexte) et, d'autre part, écrit des procédures de mise à jour des différents registres, lors du passage à un nouvel énoncé.

5.3 Résolution des anaphores et ellipses en contexte

Aloemda est capable de traiter des cas restreints d'ellipse et d'anaphore (ex : "ceux de Trébeurden", "l'adresse du garage Renault"). Il utilise pour ce traitement les noms des verbalisations présents dans les prédictions élémentaires (les réseaux correspondants aux structures elliptique et anaphorique ne sont pas toujours actifs) ainsi que les deux premiers éléments de la prédiction reçue. Le premier élément qui porte les paramètres de la dernière demande sert à compléter les demandes elliptiques (par exemple rajouter le paramètre manquant à la demande "ceux de Trébeurden"). Le second élément contient le contexte nécessaire pour résoudre les problèmes référentiels (par exemple "le garage Renault" est un élément du $n^{ème}$ professionnel transmis à l'utilisateur).

6. INFLUENCE DE LA STRATÉGIE DE DIALOGUE SUR LA RECONNAISSANCE

Si, pour l'amélioration des performances de la reconnaissance de la parole, le rôle prédictif du dialogue est incontestable, en revanche il est possible de s'interroger sur l'influence de la stratégie utilisée dans la gestion du dialogue. Certains, comme Young et Proctor [YOUN,89a], pensent que qualité ergonomique du dialogue et amélioration de la reconnaissance s'opposent, et recommandent de porter les efforts sur les progrès de la reconnaissance avant de se préoccuper de la qualité du dialogue.

Notre avis est différent. Outre le fait que la crédibilité des applications est aussi conditionnée par la qualité du dialogue, nous pensons qu'une stratégie de dialogue qui soit à la fois efficace et plaisante pour l'utilisateur est une source — indirecte — d'une reconnaissance améliorée.

6.1 L'expérimentation

Il existe des critères objectifs de mesure de la difficulté de la reconnaissance (la taille du lexique, le facteur de branchement de la grammaire, sa perplexité), et le rôle des prédictions est justement d'agir favorablement sur ces facteurs. Mais certains éléments soit sont plus difficiles à quantifier — c'est le cas de l'influence du nombre d'interventions à qualité informative constante —, soit posent des problèmes de modélisation — à l'instar des phénomènes d'hésitations (les inattendus). Pour évaluer leur influence nous avons choisi d'utiliser les transcriptions de l'expérimentation qui a précédé la mise en œuvre du système PJ [GUYO,87]. Cette expérimentation, de type "Wizard of Oz" est constituée de trois phases : une phase de dialogues dirigés, proche d'un dialogue par langage de commande (24 dialogues), une phase de dialogues libres (48 dialogues) et enfin une phase de dialogues "coopératifs" (70 dialogues). Cette dernière phase fut conçue pour valider la spécification du système PJ (la figure 2 fournit un exemple de dialogue coopératif réalisé avec le système PJ).

6.2 La présence d'inattendus

Si, pour l'écrit, beaucoup d'inattendus ont été identifiés, répertoriés et expliqués [FOUR,87] [SOND,80] [VÉRO,88], pour la parole, à l'exclusion d'études purement linguistiques [BOOM,65] [COOK,74] [HIEK,81] [LÉON,71] [GROS,75] [COST,86] peu de chercheurs ont abordé le problème sous l'angle de la reconnaissance de parole.

Table 1. Taux d'inattendus par dialogue dans l'expérimentation Pages Jaunes.

	dialogues dirigés	dialogues libres	dialogues coopératifs
remplissages	2,39	3,06	1,47
répétitions	0,22	0,92	0,40
auto-corrections	0,52	0,96	0,36
total	3,13	4,94	2,23

Dans le cadre de l'étude du corpus issu de l'expérimentation, R. Rigault [RIGA,87] a étudié l'occurrence de trois types de phénomènes d'hésitations :

— les remplissages ("Donc *euh* ce sont *euh*..."),
— les répétitions ("j'voudrais savoir *ma voit ma voi*ture..."),
— les auto-corrections ("le numéro *des de* téléphone...").

La table 1 montre que le plus faible taux d'inattendus se présente pour les dialogues coopératifs. Vis-à-vis des dialogues libres ce résultat était prévisible, liberté et hésitations allant souvent de paire. En ce qui concerne les dialogues dirigés, les chiffres sont plus surprenants mais, selon nous, ils s'expliquent en partie par une maîtrise imparfaite du langage de commande et en partie par les difficultés auxquelles se heurte l'utilisateur pour mettre en place une stratégie de dialogue satisfaisante dans un cadre aussi rigide.

Bien qu'encourageant, ce résultat ne doit pas occulter le problème de modélisation du phénomène, qui reste entier.

6.3 L'influence de la durée d'intervention de l'utilisateur

Il serait tentant d'affirmer qu'un énoncé est d'autant plus facile à reconnaître qu'il est court. La réalité est plus complexe et démontre l'influence de nombreux facteurs. En revanche, on peut, sans risquer d'être contredit, affirmer que, toutes choses étant égales par ailleurs, il est préférable de diminuer le nombre d'interventions de l'utilisateur. La table 2 montre l'influence bénéfique d'une stratégie coopérative, le nombre d'interventions par dialogue étant minimal pour ce type de dialogue.

Table 2. Nombre et longueur des énoncés utilisateurs dans l'expérimentation Pages Jaunes.

	dialogues dirigés	dialogues libres	dialogues coopératifs
nbre d'interventions utilisat./dialogue	14,35	6,73	6,17
nbre de mots utilisat./intervention	2,76	10,02	6,59
nbre de mots utilisat./dialogue	39,60	67,43	40,66

7. CONCLUSION

En nous appuyant sur un exemple de réalisation, nous avons illustré les divers avantages qu'il est possible de tirer du niveau dialogue dans un système de communication orale personne-machine.

Ces avantages proviennent du caractère prédictif du dialogue et de la stratégie utilisée pour l'interaction. Prédictions et stratégie soulèvent cependant des problèmes tant théoriques que pratiques. Parmi eux, citons celui de la finesse des prédictions qui doit être lié à celui des performances du système, celui de la pertinence et du contenu du modèle utilisateur ou enfin celui concernant le débat (éternel?) liberté de l'utilisateur versus robustesse du système. En résumé donc, le sujet n'est pas clos!

BIBLIOGRAPHIE

[BIGO,88] BIGORGNE D., COZANNET A., GUYOMARD M., MERCIER G., MICLET L., SIROUX J., A Versatile Speaker Dependant Continuous Speech Understanding System. *Proceedings of ICASSP-88*, New York, pp. 303-306, 1988.

[BOOM,65] BOOMER D.S., Hesitation and Grammatical Encoding. *Language and Speech*, Vol. 8, Pt. 3, pp. 148-158, Jul.-Sept. 1965.

[BRIE,86] BRIETZMAN A., EHRLICH U., The role of Semantic Processing in an Automatic Speech Understanding System. *Proceedings of COLING-86*, Bonn, pp. 596-598, 1986.

[BUNT,80] BUNT H., van KATWIJK F.F., MULLER L.F., van NESS F.L., Dialogue Control Acts. *IPO annual progress report*, Pays-Bas, pp. 95-99, 1980.

[CARB,89] CARBONELL N., PIERREL J.-M., Architecture and Knowledge Sources in a Human-Computer Oral Dialogue System. *The Structure of Multimodal Dialogue*, Taylor M.M., Néel F. and Bouwhuis D.G., Éditeurs, North-Holland, Pays-Bas, pp. 417-429, 1989.

[COOK,74] COOK M., SMITH J., LALLJEE M., Filled Pauses and Syntactic Complexity. *Language and Speech*, Vol. 13, Pt. 1, pp. 11-16, Jan.-Mar. 1974.

[COST,86] COSTE D., Auto-interruptions et reprises. *DRLAV*, n° 34-35, pp. 127-139, 1986.

[COZA,87] COZANNET A., ALOEMDA, un analyseur pour l'oral et l'écrit. *Actes du congrès AFCET-RFIA*, pp. 381-390, Antibes, 1987.

[FOUR,87] FOURNIER J.P., HERMAN P., SABAH G., VILNAT A., BURGAUD N., GILLOUX M., Traitement des mots inconnus dans un système de Questions-Réponses en langue naturelle. *Actes du congrès AFCET-RFIA*, pp. 653-667, Antibes, 1987.

[GROS,75] GROSJEAN F., DESCHAMPS A., Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica*, Vol. 31, n° 3-4, pp. 144-184, 1975.

[GUYO,87] GUYOMARD M., SIROUX J., Constitution Incrémentale d'un corpus de dialogues oraux coopératifs. 16*èmes JEP*, pp. 179-182, Hammamet (Tunisie), 3-10 Oct. 1987.

[GUYO,88] GUYOMARD M., SIROUX J., Une approche de la coopération dans le dialogue oral homme-machine. *Actes du colloque ERGO-IA 88*, pp. 287-301, Biarritz, 1988.

[HIEK,81] HIEKE A.E., A Content Processing View of Hesitation Phenomena. *Language and Speech*, Vol. 24, Pt. 2, pp. 147-160, Apr.-Jun. 1981.

[LE N,80] LE NY J.-F., DENIS M., Identification et compréhension du langage naturel : perspectives cognitives. *Actes du séminaire syntaxe et sémantique en compréhension de la parole*, AFCET-GALF, Paimpont, pp. 3-16, 1980.

[LÉON,71] LÉON P.R., Aspects phonostylistiques des niveaux de langue. *La grammaire du français parlé*, pp. 150-159, Hachette, 1971.

[LEVI,85] LEVINSON S.E., RABINER L.R., A Task Oriented Conversational Mode Speech Understanding System. *in Bibliotheca Phonetica*, n° 12, Karger, Basel, pp. 149-196, 1985.

[PROC,89] PROCTOR C.E., YOUNG S.J., Dialogue Control in Conversational Speech Interfaces. *The Structure of Multimodal Dialogue*, Taylor M.M., Néel F. and Bouwhuis D.G., Éditeurs, North-Holland, Pays-Bas, pp. 385-398, 1989.

[RIGA,87] RIGAULT R., *Modélisation de certains phénomènes de l'oral*. Stage de DEA, NT/LAA/TSS/345 CNET Lannion, Août 1987.

[SEAR,72] SEARLE J.R., *Les actes de langage*. Herman, Paris, 1972; traduction française de *Speech Acts*, Cambridge University Press, 1969.

[SIRO,84] SIROUX J., VIVÈS R., MERCIER G., Dix ans de communication orale et de dialogue entre l'homme et la machine avec le système KEAL. *Actes du séminaire GRECO-GALF Dialogue homme-machine à composante orale*, Nancy, 1984.

[SIRO,85] SIROUX J., GILLET D., A System for Man-Machine Communication Using Speech. *Speech Communication*, Vol. 4, n°4, pp. 289-315, 1985.

[SOND,80] SONDHEIMER N.K., A Rule-bases Approach to Ill-Formed Input. *Proceedings of COLING-80*, Tokyo, 1980.

[VÉRO,88] VÉRONIS J., FOURNIER J.-P., Traitement des erreurs dans la communication homme-machine en langage naturel. *Premières journées nationales du GRECO-PRC communication homme-machine*, pp. 129-149, Paris, 24-25 Nov. 1988.

[WINO,83] WINOGRAD T., *Language as a Cognitive Process*. Vol. 1 Syntax, Addison-Wesley, pp. 116-127, 1983.

[YOUN,89a] YOUNG S.J., PROCTOR C.E., The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems. *Computer Speech and Language*, 3, pp. 329-353, 1989.

[YOUN,89b] YOUNG S.R., HAUPTMAN A.G., WARD W.H., SMITH E.T., WERNER P., High Level Knowledge Sources in Usable Speech Recognition System. *Communication of ACM*, Vol. 32, pp. 183-194, 1989.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

DIAPASON

un système de DIAlogue Pour la commande orAle d'une console SONar

SOUVAY Gilles* PIERREL Jean-Marie* GALLAIS Evelyne° ALINAT Pierre°

*CRIN INRIA LORRAINE BP 239 F54506 VANDOEUVRE-LES-NANCY CEDEX
°THOMSON-SINTRA DASM BP 53 F06801 CAGNES-SUR-MER CEDEX

ABSTRACT

We don't present in this paper a speech recognition system but a real man-machine dialog system with a vocal input : DIAPASON. The historic of the dialog and the managed task are totally used as knowlegde sources. This allows to increase the overall performence of the system. This paper deals with the architecture of DIAPASON and its components. It also presents the results of an ergonomical study of the system.

1 INTRODUCTION

Dans le domaine de la reconnaissance de la parole, les systèmes existants sont en majorité des systèmes de reconnaissance de phrase. Mais ils s'avèrent insuffisants car il faut non seulement disposer d'un système de reconnaissance fiable mais aussi gérer le résultat de la reconnaissance dans le cadre de l'application. Cela consiste à valider le résultat de la reconnaissance (correction des erreurs, lever des ambiguïtés), à interpréter la phrase en fonction du contexte (détection des erreurs, résolution des références elliptiques ou anaphoriques), à engendrer des réponses aux demandes de l'utilisateur et à prévoir les enchaînements (gestion de l'évolution du dialogue).

Un tel système a été développé conjointement par THOMSON DASM (Division Activités Sous-Marines) pour la reconnaissance vocale, et par le CRIN pour la gestion du dialogue, grâce au soutient de la DRET (Direction des Recherches et des Etudes Techniques du Ministère de la Défense) [1]. Les motivations de cette association sont multiples. Elle permet un échange d'idées entre des industriels soucieux de proposer un produit et des théoriciens qui concoivent des modèles. Elle permet à THOMSON de tester la fiabilité d'un système de reconnaissance original et très performant en condition réelle, et au CRIN de valider des idées et des modèles développés depuis plusieurs années.

L'ambition de DIAPASON est de gérer un dialogue intelligent dans un domaine limité celui de la commande de processus. Nous étudierons dans cet article la partie gestion de dialogue développée au CRIN. Nous présenterons l'application choisie (la commande d'une console sonar), les spécificités du dialogue dans DIAPASON, les composantes du système et leur mise en œuvre, et nous terminerons par la présentation des résultats obtenus.

2 DESCRIPTION DE LA TACHE

Afin de pouvoir évaluer l'apport de la commande orale, nous avons voulu nous placer dans un cas réaliste celui de la commande de la console d'un sonar réel. Ce système est dirigé par des menus déroulants (profondeur maximum 3) à l'aide d'un clavier logiciel, d'une boule et de deux commutateurs ce qui en fait un système complexe à utiliser et demande un changement fréquent de l'opérateur, rapidement fatigué par sa tâche.

Il n'était pas possible de modifier le matériel existant pour y intégrer le système de dialogue. Nous avons donc simulé le sonar sur un micro-ordinateur de type PC connecté au système de reconnaissance. Nous avons choisi deux images représentatives de la tâche à effectuer l'image "initialisation" et l'image "veille" [FIGURE 1]. La première permet la gestion du contenu des mémoires de visualisation et la gestion des paramètres de base du sonar. La deuxième est composé d'une image (azimut, temps) sur laquelle sont représentées les pistes correspondant aux contacts détectés et d'une image annexe grâce à laquelle l'opérateur peut avoir des renseignements complémentaires sur les contacts. Sur cette image l'opérateur peut régler les échelles, décider des informations à visualiser, pointer des traitements particuliers dans des directions déterminées, régler les paramètres de certains traitements et envoyer des renseignements vers l'extérieur.

3 LA NATURE DU DIALOGUE DANS DIAPASON

Il est constitué d'une suite d'échanges entre l'utilisateur et le système. Il se déroule en trois phases. L'opérateur donne tout d'abord un ordre à l'aide d'une des entrées à sa disposition en respectant un vocabulaire et une syntaxe prédéfinie, vient ensuite sa mise au point (éventuelles corrections) et enfin son exécution. DIAPASON est alors prêt à recevoir un nouvel ordre.

3.1 Les moyens du dialogue

Pour donner un ordre, l'opérateur dispose soit du clavier + souris soit d'un micro + souris. La souris sert à désigner des objets à l'écran (mémoire, contacts...). Le mode oral est le mode privilégié. Le mode écrit est utilisé en cas de défaillance du système de reconnaissance (bruit d'ambiance ponctuellement trop important, mot récalcitrant pour un locuteur donné...).

3.2 Le vocabulaire, la grammaire

La syntaxe des ordres est identique en mode oral et écrit. Le langage utilisé est de type artificiel et décrit par une grammaire. Le vocabulaire comprend une centaine de mots.

3.3 Notion d'ordre dans DIAPASON

Un ordre correspond à une tâche bien définie que l'opérateur veut réaliser. Quel qu'il soit, on distingue tout d'abord deux éléments : la **commande** et les **paramètres**. Le premier détermine le type d'action que l'on veut effectuer (affectation ou libération d'une mémoire, affichage d'un bruiteur). Les seconds précisent les objets sur lesquels la commande agit. Exemple : afficher (commande) bruiteur alpha 1 (paramètre). Le nombre de paramètres varie de 0 à 2.

3.4 La phase de mise au point

Afin de ne pas alourdir le dialogue, elle devra être la plus courte possible. Dès que DIAPASON a analysé une demande, il envoie en réponse un message. Dans le cas général, il s'agit de l'interprétation de l'ordre. Si le locuteur est satisfait on passe alors à la phase d'exécution. Dans le cas contraire, mauvaise reconnaissance ou erreur de l'opérateur, le système offre alors la possibilité de corriger la demande. La **correction** porte sur un ou plusieurs des éléments composant l'ordre.

Exemple :
O : Afficher bruiteur alpha 2
D : Afficher bruiteur alpha 1
O : **Négatif alpha 2**
D : Afficher bruiteur alpha 2

O = Opérateur
D = DIAPASON

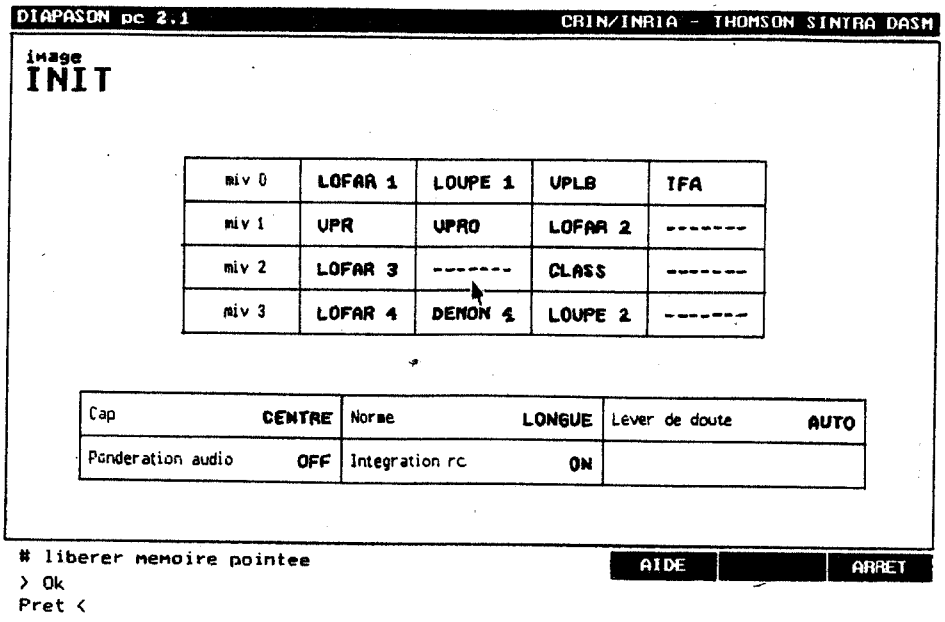


FIGURE 1a : IMAGE INIT

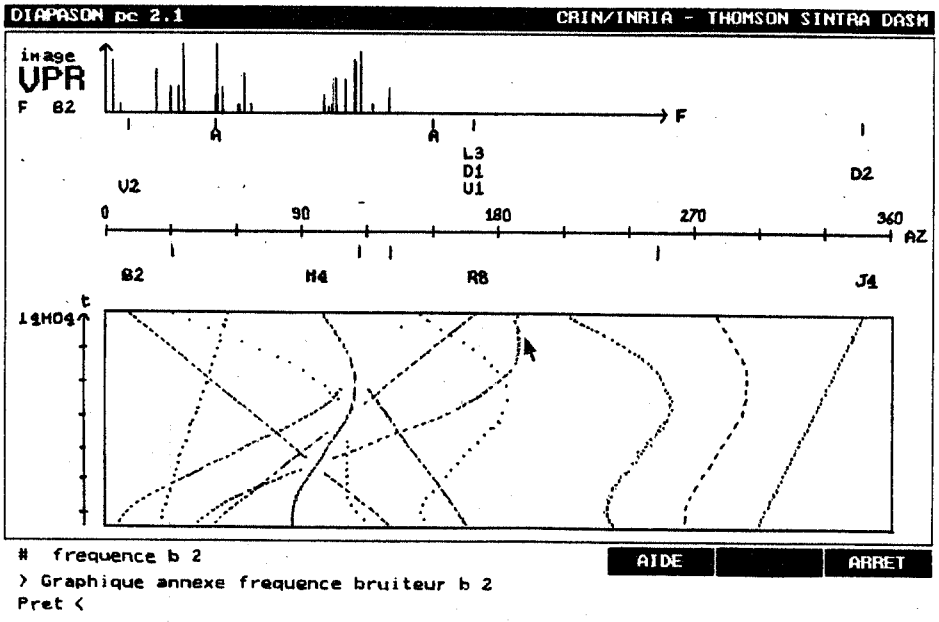


FIGURE 1b : IMAGE VEILLE

implicite : si le système n'observe pas de contestation, il exécute l'ordre au bout d'un temps t ou lorsque la phrase suivante ne contient pas de correction ; **inexistante** : l'ordre s'exécute immédiatement sans intervention de l'opérateur.

Selon l'importance de la commande, on va lui associer un type de confirmation. Cela se fait en fonction de la gravité de l'ordre. Une libération de fonction est une action qui fait perdre irrémédiablement des données. On lui associe donc une confirmation explicite. Un changement de paramètre (la cap par exemple) n'a pas de conséquence grave sur son fonctionnement, on lui associe une confirmation inexistente.

Exemple : libération d'une mémoire

	O :	Libérer mémoire LOFAR 1
confirmation explicite	D :	Confirmez : libérer mémoire LOFAR 1
	O :	Ok
	D :	Ok
		# exécution de libérer mémoire LOFAR 1 #
confirmation implicite	O :	Libérer mémoire LOFAR 1
	D :	Libérer mémoire LOFAR 1
	O :	Cap bâtiment
		# exécution de libérer mémoire LOFAR 1 #
confirmation inexistente	O :	Libérer mémoire LOFAR 1
	D :	Libérer mémoire LOFAR 1
		# exécution de libérer mémoire LOFAR 1 #

Il se peut que le système détecte une incohérence dans les propos du locuteur. Il envoie alors un message indiquant exactement la nature du problème.

Exemple : O : Afficher bruiteur alpha 2
D : Il n'y a pas de bruiteur alpha 2

Le système peut demander un complément d'information ou l'opérateur a la possibilité d'annuler un ordre.

Exemples :

	O :	Afficher ...
complément d'info	D :	Afficher quel bruiteur
	O :	**Alpha 2**
	D :	Afficher bruiteur alpha 2
	O :	Effacer bruiteurs
annulation	D :	Raz bruiteurs
	O :	**Négatif**
	D :	Que dois-je faire ?

3.5 L'exécution d'un ordre

Lorsque la commande et les paramètres répondent à l'attente du locuteur, il faut confirmer l'ordre pour l'exécuter. La **confirmation** valide le couple (commande , paramètres). Elle peut être **explicite** : on impose au locuteur la confirmation systématique de l'ordre, on ne pourra rien faire d'autre tant qu'il ne sera pas validé ou infirmé ;

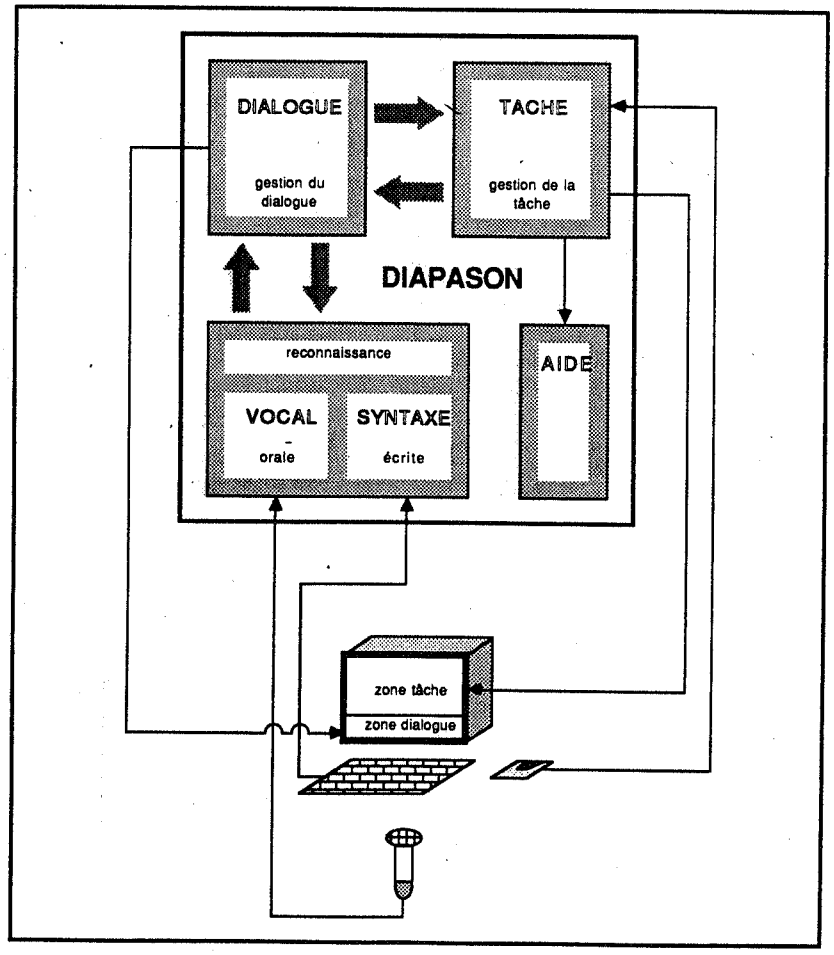


FIGURE 2 : ARCHITECTURE GENERALE DU SYSTEME

Un ordre est donc composé de trois éléments : commande, paramètres et confirmation

3.6 L'enchaînement des ordres dans DIAPASON

Un point fort des systèmes de dialogue homme-machine finalisé est de pouvoir omettre certains éléments d'une phrase en fonction du contexte. DIAPASON permet ainsi l'ellipse sur la commande ou un des paramètres. L'ellipse sur la commande est utilisée pour effectuer une même opération sur un objet différent. L'ellipse de paramètre est possible lorsque l'on effectue plusieurs actions sur un même objet.

Exemples : O : Libérer mémoire LOFAR 1
D : Libérer mémoire LOFAR 1
O : mémoire CLASS
D : Libérer mémoire CLASS

O : Afficher bruiteur alpha 1
D : Afficher bruiteur alpha 1
O : Lever de doute gauche
D : Lever de doute **bruiteur alpha 1** gauche

4 COMPOSANTES ET FONCTIONNEMENT DE DIAPASON

4.1 Architecture générale du système

On distingue dans DIAPASON trois niveaux : la reconnaissance de phrase prise en charge par RECONN, la gestion de la tâche avec TACHE et la gestion du dialogue avec DIALOGUE. Ces modules échangent des informations entre-eux comme indiqué sur la figure 2. Cette architecture est très proche du système PARTNER [2].

4.2 Le module de reconnaissance RECONN

Il y a deux modes de communication avec le système l'écrit ou la parole. Il existe pour chacun un module de reconnaissance. Ils reprennent tous les deux les principes développés pour MYRTILLE I [3][4]. Il s'agit de systèmes guidés par la syntaxe qui à chaque étape, émettent des hypothèses sur les mots à reconnaître et les valident par un sous-module de reconnaissance de mots. **SYNTAXE** effectue outre l'analyse syntaxique, une simple comparaison de mots. Le sous-module de reconnaissance orale **VOCAL** developpé par THOMSON est de type analytique, mots enchaînés, multi-locuteur [5], et utilise au niveau bas un modèle qui s'inspire du fonctionnement de la cochlée humaine [6].

SYNTAXE et VOCAL travaillent à partir de grammaires fournies par DIALOGUE et lui proposent la représentation sémantique de la phrase analysée. Le langage utilisé dans DIAPASON est de type artificiel et présente donc de forts aspects sémantiques. La construction de la structure sémantique se fait en recueillant les éléments pertinents dans les arbres syntaxiques résultats.

4.3 le module de gestion de la tâche : TACHE

Il dirige le fonctionnement de DIAPASON et contrôle le reste du système en formulant à DIALOGUE des requêtes. Son but est d'obtenir les éléments nécessaires à l'exécution d'un ordre : commande, paramètre et confirmation. Il s'efforce de gérer au mieux les incohérences qu'il pourrait détecter.

4.3.1 le fonctionnement de TACHE

TACHE est un automate d'états finis dont le but est de se mettre dans un état de satisfaction Les transitions sont constituées d'appels au module DIALOGUE, qui lui fournira successivement la commande, les paramètres et demandera s'il peut se mettre dans l'état de satisfaction en fonction du type de confirmation associé à la commande. Le fonctionnement ne se fait pas de manière linéaire. Il est influencé par les réponses fournies par DIALOGUE et peut revenir dans un état antérieur dans le cas d'une correction. Une incohérence peut provoquer un retour à l'état initial ou faire avancer dans un état de satisfaction avec envoie d'un message d'erreur. Enfin lorsqu'un ordre est accepté, on l'exécute et TACHE retourne dans son état initial.

4.3.2 le traitement de l'incohérence

Il existe des conditions de validité pour le couple (commande, paramètres). Lorsque l'une de ces règles est violée on dit qu'il y a **incohérence**. Elle a deux origines possibles : le niveau reconnaissance a proposé une hypothèse ne correspondant pas aux propos du locuteur ou l'opérateur ne tient pas lui-même des propos cohérents. Une manière approximative de distinguer ces deux cas est de se fier aux scores de reconnaissance. Si l'on est au dessus d'un certain seuil on considère que le locuteur s'est trompé et l'on envoie un message d'incohérence. Si l'on est en dessous on la refuse et on poursuit l'analyse. Dans ce cas de figure, on explore successivement toutes les hypothèses fournies par RECONN. On arrête la recherche lorsque l'on rencontre un ordre cohérent ou lorsqu'il n'y a plus d'hypothèse. Dans ce dernier cas on récupère la première hypothèse rejetée.

4.4 Le module de gestion du dialogue : DIALOGUE

Le module de gestion du dialogue joue un rôle tampon entre RECONN et TACHE. Il doit répondre aux requêtes formulées par TACHE. Pour cela il fait appel soit au module RECONN, soit à sa mémoire suivant une stratégie préétablie.

4.4.1 le fonctionnement de DIALOGUE

DIALOGUE est un automate d'états fini. Ses transitions sont plus complexes que dans le cas de TACHE car elles sont liées à des appels aux deux modules TACHE et RECONN et à des événements internes.

Les échanges DIALOGUE ->RECONN sont de trois natures. La demande d'hypothèse est formulée lorsque l'on détecte de la parole ou un accès au clavier. L'acceptation d'hypothèse indique au module de reconnaissance concerné qu'il a terminé son travail. Il est réinitialisé à partir de données fournies par DIALOGUE. Le rejet d'hypothèse demande à l'analyseur de fournir l'hypothèse suivante.

Les échanges DIALOGUE->TACHE correspondent tout d'abord aux réponses aux requêtes formulées par TACHE : l'envoi de valeur pour la demande de valeur, la réponse de confirmation pour la demande de confirmation. Il existe deux autres requêtes permettant la synchronisation : le signal d'activation qui met TACHE dans son état initial et le signal de restauration pour la gestion de l'incohérence.

4.4.2 la mémoire du dialogue

a) composition de la mémoire : la mémoire du dialogue a deux composantes. La première correspond à la mise au point de l'ordre courant. Le but de DIAPASON est d'arriver dans un des états de satisfaction de TACHE en recueillant commande, paramètres et satisfaction. Cette collecte peut se faire avec plusieurs interventions de l'opérateur (demande de complément d'information du système, corrections de la part de l'utilisateur). La **mémoire à court terme** conserve toutes les hypothèses proposées par le niveau reconnaissance au cours de cette mise au point. A chaque phrase on associe un **statut** qui est fonction de la réponse fournie par TACHE (refus de la commande ou du paramètre pour cause d'incohérence) ou fonction des corrections du locuteur (négation de commande, refus du paramètre). La deuxième correspond au dialogue global. On stocke dans la **mémoire à long terme** tous les ordres exécutés par le système. On y conserve aussi tous ceux qui étaient incohérents mais finalement acceptés ainsi.

b) stratégie de recherche de valeur - résolution des ellipses : pour répondre à la demande de valeur formulée par TACHE, DIALOGUE dispose de quatre sources : la mémoire à long terme, la mémoire à court terme, les valeurs par défaut et le niveau reconnaissance. Il consulte tout d'abord la mémoire à court terme, puis la mémoire à long terme. Il regarde ensuite si une valeur par défaut existe. Enfin il fait appel en dernier ressort au module de reconnaissance. Dans ce dernier cas, l'hypothèse résultat est rangée dans la mémoire à court terme et sera utilisée le cycle suivant.

Une des caractéristiques de DIAPASON et de ne pas proposer deux fois le même ordre. Cela est possible grâce au statut évoqué dans le paragraphe précédent, DIALOGUE ne propose plus de commande ou de paramètre qui ont été niés ou réfusés une première fois pour cause d'incohérence ou de correction.

c) incohérence et mémoire : lorsque nous avons traité l'incohérence 4.3.2, nous n'avions pas encore indiqué que les valeurs pouvaient avoir deux origines, mémoire à court terme ou à long terme. Le message envoyé à l'opérateur était un constat d'échec. DIALOGUE proposait des valeurs erronées et on n'y pouvait rien. En cas d'incohérence un traitement supplémentaire est effectué qui permet de revenir sur un choix. Le principe est le suivant : si une valeur provient de la mémoire à long terme et provoque une incohérence c'est qu'il ne fallait pas la proposer. On pose une question au lieu d'envoyer un message d'erreur.

Exemple : O : Afficher bruiteur pointé.
D : Afficher bruiteur pointé
O : Afficher
réponse sans traitement :
D : Le bruiteur pointé est déjà affiché
réponse avec traitement :
D : Afficher quel bruiteur ?

5 PRESENTATION DES RESULTATS

Afin d'évaluer les performances du système nous avons procédé à une comparaison des dialogues de commande du sonar réel et de la maquette implantée sur PC.

5.1 Conditions de l'expérimentation

La manipulation s'est déroulée sur la plate-forme d'intégration des sonars. Le **bruit d'ambiance** était relativement important, 6 bases d'électronique dans la pièce et des conversations fréquentes de 0,5 à 2 m du micro (omnidirectionnel pour l'expérimentation). Les opérateurs étaient entraînés soit sur le sonar, soit sur le PC, le temps ne permettait pas de les entraîner sur les deux systèmes. Le sonar fonctionnait en simulation, les tests n'ont donc pu être effectués en environnement opérationnel sur une tâche opérationnelle, mais sur des sous-tâches simulées.

5.2 Mesures ergonomiques

Vu l'impossibilité de disposer du sonar pendant un temps important et vu le fait que les tests ne pouvaient être faits sur une tâche réelle, nous avons concentré nos efforts de comparaison sur quelques sous-tâches exécutées par quelques opérateurs en gardant bien à l'esprit que seules les comparaisons sur les tâches réelles en ambiance réelle permettent des conclusions quasi définitives.

Trois sous-tâches ont été effectuées par six opérateurs sur le sonar, puis sur le PC. Pour chaque sous-tâche on a mesuré la durée totale et le nombre d'opérations effectuées.

En plus de ces essais, une vingtaine de locuteurs non entraînés ont pu essayer plus ou moins longuement le système de commande vocale montrant ainsi les qualités du système : multilocuteur sans apprentissage, résistance aux variations de prononciation (accents régionaux, rapidité d'élocution, silence ou liaisons entre mots, hésitation ...), résistance au bruit d'ambiance et aux conversations de proximité.

5.3 Résultats acquis

La comparaison entre la console du sonar et sa simulation sur PC nous a permis d'acquérir des résultats relatifs à l'emploi et aux avantages de la commande vocale pour les consoles.

5.3.1 Mode d'emploi de l'entrée vocale

Les **temps de réponse** cumulés des systèmes de reconnaissance de la parole et de compréhension de l'ordre ne doivent pas dépasser la seconde. Au-delà l'opérateur perd du temps à attendre.

Les **demandes de confirmation** d'ordre doivent le plus souvent possible être faites "en place" c'est-à-dire se matérialiser par une modification de l'image proche de l'endroit où l'opérateur est censé observer à cet instant. Nous avions réservé une zone de l'écran pour les dialogues où les phrases reconnues et leur interprétation s'affichait. A l'usage il s'est avéré qu'elle faisait perdre une partie des avantages principaux de la parole qui est que le regard reste fixé sur l'objet manipulé.

L'**utilisation d'un commutateur** ("pédale micro") permettant à l'opérateur d'indiquer quand il donne un ordre vocal s'avère un moyen simple, pratique à utiliser et très efficace pour se protéger de la réception d'ordres parasites dûs aux conversations se tenant près du microphone.

Pour les ordres vocaux l'opérateur n'est plus guidé dans son choix par des possibilités affichées sur l'écran. Dans ce cas les **moyens d'aide** deviennent très utiles, surtout pour les opérateurs débutants.

Dans les systèmes à commande par mots enchaînés, il faut limiter la taille du vocabulaire et la richesse de la syntaxe de façon que l'opérateur puisse l'apprendre. La taille maximale semble être de 150 à 200 mots, ce qui est suffisant pour les applications de type console.

5.3.2 Aspects ergonomiques de l'avantage à employer une entrée vocale

Le regard de l'opérateur peut rester fixé sur l'objet manipulé, autrement dit il n'est pas nécessaire, pour un ordre vocal, de regarder le clavier ou des menus affichés à l'écran. Malgré l'usage de la zone dialogue on a remarqué que le regard était quatre fois moins distrait ; on devrait arriver à quinze fois en effectuant la confirmation en place.

L'usage de la commande vocale permet de concentrer en une seule phrase de nombreuses manipulations élémentaires sur le sonar.

Les échanges sont plus rapides, dans le cas de système avec un temps de réponse de une seconde, on obtient pour l'ensemble des opérateurs et des sous-tâches une amélioration de la vitesse de travail de 1,66 en faveur de la commande vocale (entre 1,1 et 2 selon le cas). Cet avantage est d'autant plus net que les ordres correspondent à des phrases longues.

Les menus occupent de la place sur l'écran. Cette occupation peut être permanente ou temporaire. Dans tous les cas c'est de la surface perdue pour l'affichage des informations propement dites. La commande vocale permet de récupérer cette surface.

6 CONCLUSION

L'étude décrite ici a montré l'interêt du dialogue lors de l'intégration d'une entrée vocale. La parole apporte un certain naturel à la commande, et la possibilité de donner des ordres incomplets et de les enchaîner renforce cette impression. L'opérateur peut garder son regard fixé sur la zone où se produit l'événement plutôt que d'être distrait par des menus situés dans d'autres régions de l'écran. Ce dernier n'est plus systématiquement surchargé au détriment de l'image. La place ainsi gagnée permettra éventuellement d'afficher des informations pertinentes qui ne pouvaient l'être auparavant. Globalement on observe un gain de temps appréciable et une charge de travail réduite pour l'opérateur .

7 BIBLIOGRAPHIE

[1] GALLAIS E. SOUVAY G. "Commande vocale de console". Rapport de synthèse final THOMSON DASM août 1989.

[2] MORIN P. PIERREL J.M. "PARTNER : un système de dialogue oral homme-machine". Actes COGNITIVA 87, Paris, 1987.

[3] PIERREL J.M. "Un système de compréhension automatique du discours continu utilisant des contraintes morphologiques, syntaxiques et sémantiques", RAIRO informatique, Vol. 12-2, p83-105, 1978.

[4] PIERREL J.M. "Dialogue Oral Homme-Machine". Hermès, Paris, 1987.

[5] SOUVAY G. PIERREL J.M. GALLAIS E. ALINAT P. "Intégration d'un système de reconnaissance analytique de la parole dans une console sonar : vers un dialogue naturel". Rapport interne CRIN Septembre 1989.

[6] ALINAT P. "Etude des phonèmes de la langue française au moyen d'une cochlée artificielle - application à la reconnaissance de la parole". Revue Technique Thomson CSF Vol 7 n°1 mars 1975.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

RECONNAISSANCE DE PAROLE EN ENTREE D'UN SYSTEME DE TRADUCTION AUTOMATIQUE, EN FRANCAIS ET EN ANGLAIS.

Jean-Pierre TUBACH *, Raymond DESCOUT **, Pierre ISABELLE **

* Télécom Paris, Département Signal (CNRS, URA 820) et CCRIT
** CCRIT (Centre Canadien de Recherche sur l'Informatisation du Travail), Montréal.

RESUME

Cet article rend compte d'un travail effectué au CCRIT (Centre Canadien de Recherche sur l'Informatisation du Travail) (Groupe des Technologies Vocales (GTV), en collaboration avec le Groupe de Traduction Assistée (GTA)). Il ne s'agit pas, ou fort peu, d'une recherche sur les algorithmes de reconnaissance de parole, mais principalement d'une investigation dans le domaine de la mise en oeuvre des technologies vocales disponibles sur le marché, pour la réalisation d'applications avancées. Ce type d'étude nous semble actuellement absolument nécessaire, et prometteur.

Le GTA du CCRIT a développé un système de traduction automatique (CRITTER), bi-directionnel entre le français et l'anglais, pour les rapports d' Agriculture Canada sur le marché de la viande et du bétail. Une première interface d'entréee sortie vocale (IRMA) a été donnée à ce système par le GTV, qui a été démontrée avec succès à l'exposition Expotec de Montréal, pendant l'été 1989.

Le but de ce travail est de fournir à CRITTER une interface d'entrée vocale plus avancée que celle mise en oeuvre dans IRMA (parole continue (mots connectés, pour être précis), multilocuteur ou bien haute qualité). Les cartes de reconnaisance utilisées sont VECSYS Datavox et X-COM Media50.

On aboutit à un système de traduction avec entrée vocale, bidirectionnel français <--> anglais, ce qui constitue, à notre connaissance une "première".

Il en résulte des perspectives ergonomiques intéressantes dans la cadre du projet de poste de travail pour traducteur du CCRIT, puisque l'utilisateur pourrait, à son choix, entrer vocalement le texte source pour traduction par CRITTER, ou le résultat de sa traduction "humaine" dans la langue cible.

I PRESENTATION

I.1 Brève présentation de CRITTER

On sait que, si la traduction automatique d'une langue dans une autre pour n'importe quel texte demeure actuellement un difficile objectif à très long terme, elle est possible dans le contexte restrictif d'un langage et d'un champ sémantique relatifs à certaines applications.

C'est ce que réalise CRITTER , pour les rapports hebdomadaires publiés par le Ministère Canadien de l'Agriculture, sur les marchés de la viande et du bétail dans le différentes provinces du Canada.

Ce logiciel, décrit dans Isabelle et al., 1988, repose sur un système de transfert, et a pour caractéristique particulièrement intéressante d'être réversible entre le français et l'anglais. Il est mis en oeuvre sur une station de travail SUN.

I.2 Brève présentation d'IRMA

Le système IRMA, dans sa première version, repose, en ce qui concerne la reconnaissance, sur l'utilisation de la carte DRAGON VoiceScribe 1000. L'élocution est faite nécessairement par mots isolés. La taille du vocabulaire est de l'ordre de 150 mots pour le français et autant pour l'anglais. Aucune syntaxe n'est utilisée au niveau de la carte de reconnaissance.

On fournit à CRITTER un treillis de mots, comportant pour le moins tous les homophones ou quasi-homophones d'un mot prononcé (exemples: faible, faibles; était, étaient, été ; ...). Le système de traduction assure dans ce treillis la désambiguïsation, par recherche d'une analyse parvenant avec succès au terme de la phrase.

Toutes les fonctions de contrôle (changement de langue, mise en service et hors service du micro, édition des erreurs) sont contrôlées depuis la SUN hôte de CRITTER, et bien intégrées dans ses excellents écrans de démonstration.

Les erreurs de reconnaissance sont nombreuses, malgré le fonctionnement en mode monolocuteur, ce qui rend le système pénible à utiliser. Les raisons en sont les suivantes :
- la carte utilisée ne représente plus l'état de l'art
- il n'y a prise en compte d'aucune contrainte syntaxique pour la reconnaissance

I.3 Les options retenues pour ce travail

Nous avons choisi d'exploiter au mieux les nouvelles cartes de reconnaissance, Vecsys Datavox et X-Com Media50 (présentées en détail dans Tubach et al., 1989), de façon à fournir une bonne mise en valeur des possibilités actuelles des technologies vocales, en supplément à celles de la traduction automatique:
- l'élocution sera continue et non par mots isolés, grâce à l'utilisation de cartes permettant la reconnaissance de mots connectés
- une syntaxe sera utilisée à la reconnaissance, et permettra de fournir à CRITTER des chaînes orthographiquement correctes, directement assimilables
- le changement de langue et le fonctionnement du micro pourront être commandés vocalement, ainsi que des corrections en cas d'erreurs ("édition vocale").

I.4 La définition du langage

Le domaine pris en compte par CRITTER est celui des rapports sur le marché des bestiaux et de la viande au Canada (Canada livestock and meat

trade reports). Une étude du langage spécifique de cette application a d'abord été effectuée, grâce à des discussions avec le Groupe de traduction assistée, à la lecture de rapports d'Agriculture Canada, et à des essais du système CRITTER.

Un sous ensemble de ce langage, de complexité adaptée aux possibilités actuelles de la reconnaissance vocale, a été défini.

Le langage considéré comporte des phrases fréquemment très longues, tout particulièrement en Français (exemple: "à Vancouver, les prix de gros des bouvillons de court engraissement ont gagné 3 dollars le kilo en raison d'arrivages peu abondants au début de la semaine").

Les premières expérimentations ont montré que, même en permettant des pauses et-ou hésitations à certains endroits spécifiques, la dictée de telles phrases en tant qu'entités uniques n'est pas réaliste, mais hautement artificielle, ce qui va à l'encontre de nos objectifs. Des inconvénients techniques secondaires sont également liés à ce mode de fonctionnement: par exemple, le temps de réponse apparent en fin de phrase est systématiquement majoré de la durée maximale admise pour une pause (1 seconde au moins).

Nous avons donc défini un autre mode d'élocution, beaucoup plus satisfaisant ergonomiquement dans le cadre de cette application: l'élocution par membres de phrases, qui permet d'énoncer comme entités indépendantes les constituants syntaxiques majeurs des phrases: lieu,temps, sujet, verbe-complément, cause (sujet et verbe-complément pouvant optionnellement être coupés en deux parties). Exemple: "Les prix de gros / des vaches de boucherie // ont augmenté / de trois dollars le kilo // à la fin de la semaine // à Montréal // en raison d'arrivages très faibles // point final " (/ désignant une pause facultative et // une pause obligatoire).

On remarque qu'une indication spécifique ("point final", ou "fin de phrase") devient alors nécessaire pour indiquer qu'une phrase est terminée. On fournit aussi la possibilité d'énoncer d'un trait des portions de phrases plus vastes (sujet verbe complément) si elles sont simples. (On trouvera la définition complète du langage dans Tubach, 1989).

Le problème des hésitations (euh, mmm, aah) du locuteur, et des pauses, devient beaucoup moins crucial avec ce type d'élocution, puisqu'elles se produisent principalement aux limites que nous avons choisies pour les membres de phrases. On prend néanmoins en compte le phénomène relevé par Hauptmann et Rudincky, 1988, d'hésitation survenant fréquemment avant la fourniture d'une information particulièrement précise (les prix ont augmenté // de euh 3 dollars le kilo)

Les contraintes syntaxiques sont moins fortes dans le cas du langage par membres de phrase (par exemple, on ne "sait" plus si le sujet était singulier ou pluriel au moment de distinguer "ont augmenté" de "a augmenté"). Mais la qualité de reconnaissance obtenue au terme de notre travail permet un fonctionnement correct avec ces contraintes réduites (en multilocuteur, ce n'est vrai que pour la langue française, voir plus loin) .

Pour ce qui est de l'anglais, les raffinements de la syntaxe utilisés dans le reconnaisseur pour prendre en compte certaines spécificités de la langue française disparaissent, et n'ont pas de contrepartie. Il s'agit des liaisons (des vaches, des (z) agneaux), des articles élidés (de veaux, d'agneaux), de l'accord de l'adjectif et du nom (demande faible, arrivages faibles). Au niveau du vocabulaire, les mots anglais sont beaucoup plus courts, en particulier dans ce domaine ("yearlings" pour "sujets d'un an", "stockers" pour "bouvillons de long engraissement"), ce qui n'est a priori pas favorable pour la reconnaissance, mais là aussi ne pose pas de problème majeur grâce à la bonne qualité de reconnaissance.

II UTILISATION DE LA CARTE XCOM MEDIA50

II.1 Présentation

Cette carte, commercialisée, entre autres, par XCOM (Grenoble, France), sous le nom de Media50, est issue de travaux menés au Centre National d'Etudes des Télécommunications de France Télécom par C. Gagnoulet, D. Jouvet, J. Monné. Elle trouve place dans un compatible PC, où elle ocupe une seule position. Elle est utilisée à France Télécom dans le serveur téléphonique vocal Mairievox, et les mêmes algorithmes le sont dans la cabine téléphonique à numérotation vocale Publivox (Gagnoulet et Jouvet, 1989) .

Elle met en oeuvre des algorithmes d'apprentissage et de reconnaissance utilisant la modélisation par chaînes de Markov. Prévue initialement pour la reconnaissance de mots connectés en nombre limité, avec un automate de Markov par mot, elle peut également être utilisée en reconnaissance phonétique (un automate élémentaire par unité phonétique), ce qui permet des tailles de vocabulaire supérieures (de l'ordre de 120 mots) et une syntaxe plus fine. Fait très important, un apprentissage à partir d'enregistrements de plusieurs locuteurs permet une reconnaisance multilocuteur sans augmentation de la taille mémoire nécessaire.

L'environnement de développement utilisé est le logiciel PHIL 86 du CNET. Il permet la description de la syntaxe et du vocabulaire de l'application, et l'apprentissage du réseau de Markov correspondant, sur une base de données multilocuteur, puis la reconnaissance et l'évaluation. On trouvera dans Tubach, 1989 des informations très détaillées sur l'utilisation pratique de ce logiciel.

II.2 Les modèles de Markov utilisés

Dans sa thèse de doctorat, D. Jouvet (1988) a exploré un certain nombre de modèles de Markov pour la reconnaissance phonétique. Les meilleurs résultats ont été obtenus pour des modèles comportant un nombre important d'états, de transitions et de fonctions de probabilités (typiquement 5 états et 12 transitions pour les voyelles et les consonnes, 7 états et 18 transitions pour les successions de deux phonèmes qu'il est parfois nécesaire de considérer).

Leur utilisation pour une application aussi "grosse" que celle considérée ici entraîne malheureusement un encombrement en mémoire de données incompatible avec les 128Ko de la carte Media50.

Confronté plus tôt à cette difficulté, C. Gagnoulet (communication personnelle) a proposé des modèles beaucoup plus compacts (typiquement 4 états et 5 transitions pour les voyelles, 3 états et 3 transitions pour les consonnes, 5 états et 7 transitions pour les groupes de 2 phonèmes). Nous les avons testés, et avons constaté que, si la représentation des voyelles est suffisamment bonne, celle des consonnes laisse trop à désirer pour notre application.

Nous avons donc finalement adopté des modèles "intermédiaires", qui représentent un compromis entre taille mémoire et performance en reconnaissance. Il s'agit essentiellement de renoncer à utiliser le modèle le plus court (3 états) du paragraphe précédent, et à choisir le modèle à 4 états et 5 transitions pour les voyelles et les consonnes, et le modèle à 5 états et 7 transitions pour les groupes de phonèmes.

II.3 Données sur le l'application "membres de phrase" en français

Le vocabulaire compte 120 mots, choisis uniquement en fonction de l'application (c'est à dire sans artifices destinés à faciliter la reconnaissance).

48 unités phonétiques sont utiliséees:
13 voyelles (a, i, o, oe, eu, e_muet, e, ai, y, u, an, in, on)
18 consonnes (p, t, k, b, d, g, m, n, v, f, s, ch, z, jj, l, r, l_final, r_final)
15 diphones (jai, je, jo, jon, jin, jeu, wa, wi, ye, yi, tr, gr, pr, bl, gl)
2 unités supplémentaires (euh et mmm, pour les hésitations)

Le facteur de branchement statique varie de 1 à 50, avec une moyenne de 5,1. Le facteur de branchement dynamique est 3,0. La longueur des

"phrases" varie de 3 à 11 "mots", avec une moyenne de 8,3. Le langage compte environ 10000 phrases différentes. (Ces informations sont fournies par l'utilitaire REBUS2 de Datavox/Vecsys).

L'apprentissage a été fait pour 10 locuteurs (6 québecois, 4 français) (il serait en fait plus correct de parler de "Montréalais" et de "Parisiens", aucun accent régional marqué n'ayant été pris en compte). Cet apprentissage requiert environ 6 heures de l'ensemble PC + carte Media50. Chaque locuteur d'apprentissage a prononcé 105 membres de phrases ou phrases courtes. La plupart l'ont fait une seconde fois, pour fournir des données d'évaluation.

D'autres apprentissages ont été faits pour les québecois seulement, et pour les français seulement (voir paragraphe Evaluation).

II.4 Données sur l'application "membres de phrase" en anglais

Le vocabulaire compte 106 mots, là aussi choisis uniquement en fonction de l'application.

40 unités phonétiques sont utiliséees:
- 10 voyelles non diphtonguées
- 20 consonnes
- 9 voyelles diphtonguées et diphones
- 1 unité supplémentaire (aah, pour les hésitations)

Les transcriptions phonétiques du vocabulaire ont été faites selon les indications du "Gage Canadian Dictionary"

Le facteur de branchement statique varie de 1 à 47, avec une moyenne de 8,4. Le facteur de branchement dynamique est égal à 3,2. La longueur des "phrases" varie de 3 à 11 "mots", avec une moyenne de 7,4. Le langage compte environ 6000 phrases différentes.

L'apprentissage a été fait pour 6 locuteurs anglophones ou bilingues. (durée quatre heures environ).

Chaque locuteur d'apprentissage a prononcé 100 membres de phrases ou phrases courtes. La plupart l'ont fait une seconde fois, pour fournir des données d'évaluation. Des données de test ont de plus été fournies par deux francophones (1 québecois, 1 français)

III UTILISATION DU SYSTEME DE RECONNAISSANCE VECSYS DATAVOX

Ce système de reconnaissance, commercialisé par VECSYS (Bièvres, France), est issu de travaux menés au LIMSI-CNRS d'Orsay, par J.L. Gauvain et J.C. Gangolf. Il s'agit d'un ensemble de deux cartes pour PC et compatibles, et d'un boitier externe, chargé des fonctions analogiques et de conversion analogique - numérique.

Les cartes mettent en oeuvre des algorithmes d'apprentissage et de reconnaissance utilisant la Programmation Dynamique. Une des cartes comporte d'ailleurs un circuit intégré spécialisé, appelé uPCD (micro Processeur de Comparaison Dynamique), responsable principal des hautes performances de l'ensemble (Quénot et al., 1989).

Ce produit est orienté vers une reconnaissance en mode monolocuteur, avec une excellente qualité de reconnaissance.

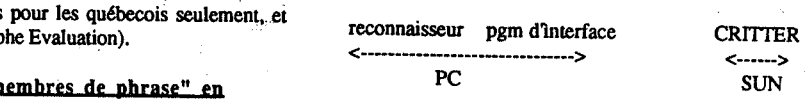
L'environnement de développement utilisé est le logiciel Datavox de Vecsys, qui contrôle, de façon ergonomique, la définition du langage, l'apprentissage et la reconnaissance, en mots connectés, monolocuteur, pour des vocabulaires de plus de 200 mots. (D'autres logiciels sont en développement au LIMSI pour des applications plus ambitieuses).

Le vocabulaire est, dans les deux langues le même que précédemment; (le langage est en fait un peu plus vaste, par exemple dictée des informations numériques sous la forme "chiffre point chiffre" ou "chiffre virgule chiffre" au lieu de "chiffre". La description est faite au niveau mot, et non phonème, ce qui rend la mise en oeuvre plus aisée.

On peut penser d'après les spécifications des cartes et de leur logiciel que, pour cette application, le système est encore loin de ses limites, et qu'un doublement de la taille du vocabulaire pourrait être envisagé.

IV PROGRAMMES D'INTERFACE

Un programme d'interface, résidant dans le PC est nécesaire entre les cartes de reconnaissance dont il recevra les résultats, et CRITTER auquel il les transmettra après mise en forme.



Lors du choix du protocole de communication entre les deux systèmes, et de leurs rôles respectifs sont apparues deux conceptions possibles, que l'on peut, un peu carricaturalement, décrire comme : système de traduction avec un serveur vocal, ou bien système vocal avec un serveur de traduction. Il serait en fait souhaitable de cumuler les avantages des deux formules, et de permettre l'initiaive à l'un ou l'autre des deux systèmes, en matière de contrôle: changement de langage, ouverture/fermeture du micro, correction d'erreurs.

Il a été décidé, pour parvenir en temps limité à un ensemble "qui tourne", de conserver exactement les protocoles de la première version d'IRMA. C'est dans le cadre des protocoles de communication développés au GTV, que pourra prendre place une coopération plus évoluée entre les deux systèmes (ces protocoles fournissent en effet un cadre général et des outils pour les échanges entre un système Unix et un PC muni de dispositifs vocaux).

V EVALUATION

Les études d'évaluation qui sont décrites ci-après portent sur le cas du multilocuteur sur la carte Media50. C'est dans ce contexte que se pose le problème de l'influence des diverses variantes du français, alors que ce n'est pas le cas pour un système monolocuteur.

Indiquons cependant auparavant que l'utilisation "en ligne" du reconnaisseur Vecsys est plus "impressionnante", tant par la qualité de reconnaissance (monolocuteur) que par le temps de réponse. Nous n'avons pas disposé d'un outil d'évaluation pour ce système (c'est dans le cadre du projet Esprit SAM que cet outil deviendra disponible).

V.1 En français

Afin d'étudier l'influence de deux variantes (québecoise et française, ou plutôt montréalaise et parisienne) de la prononciation de la langue française, nous avons effectué, pour Media50 :
- un apprentissage comportant tous les 10 locuteurs enregistrés, que nous nommons T (pour Tous)
- un apprentissage avec les 6 locuteurs québecois seulement, ou Q (pour Québec)
- un apprentissage avec les 4 locuteurs français seulement, ou F (pour France)

Une seule femme, québecoise, figurait parmi les locuteurs d'apprentissage.

On présente dans les tables qui suivent les pourcentages d'erreurs au niveau des membres de phrases (unités de dictée) dans la colone "phrases", et au niveau des mots (sans distinguer substitutions, insertions, omissions). On distingue les résultats sur l'échantillon d'apprentissage, et sur des données de test.

Apprentissage T

Résultats sur tous locuteurs

	phrases	mots
apprentissage	7.7	2.9
test	9.2	3.9

Apprentissage Q

Résultats sur les locuteurs québecois

	phrases	mots
apprentissage	6.8	2.5
test	7.2	2.7

Résultats sur les locuteurs français

test	13.2	6.6

Apprentissage F

Résultats sur les locuteurs français

	phrases	mots
apprentissage	4.2	1.5
test	6.3	2.7

Résultats sur les locuteurs québecois

test	15.6	9.0

On constate tout d'abord des performances assez satisfaisantes pour le cas T (tous les locuteurs, en apprentissage comme en reconnaissance). Et également des résultats un peu meilleurs pour les deux cas de séparation des deux ensembles de locuteurs.

Les différences entre le cas F et le cas Q doivent être interprétées avec prudence, elles ne traduisent probablement que l'inégalité des nombres de locuteurs (6 pour Q, 4 pour F)

La dégradation dans le cas des tests croisés F / Q est sensible; mais il faut tempérer cette première impression par le fait que c'est le seul cas où sont testés des locuteurs n'ayant pas participé à l'apprentissage correspondant. Il serait donc plus approprié de comparer ces chiffres aux résultats d'autres locuteurs québecois sur Q, et français sur F. Ceci n'a pu être fait, faute de temps.

On retiendra principalement donc les taux d'erreur de 3.9% sur les mots, et 9.2% sur les membres de phrases pour le cas test avec tous locuteurs. Ils permettent un fonctionnement acceptable du système, grâce à la fonction d'édition vocale "correction" (ou de l'édition sur IRMA/CRITTER).

V.2 En anglais

L'apprentissage a été fait pour 6 locuteurs anglophones ou bilingues (3 hommes et 3 femmes). Les tests ont été faits sur ces mêmes locuteurs, et séparément sur deux "non anglophones" (un québecois et un français).

Les résultats sont les suivants :

Résultats sur locuteurs anglophones ou bilingues

	phrases	mots
apprentissage	18.3	11.1
test	23.5	13.1

Résultats sur locuteurs non anglophones

test	30.2	19.5

Ces résultats ne sont pas excellents, et ne permettent pas un fonctionnement réaliste du système. Nous proposons les explications suivantes (par ordre d'importance décroissante):
- le vocabulaire anglais s'avère contenir des mots très proches phonétiquement: (calves /cows), (was / were) (increased / decreased). Si l'on laisse de côté ces confusions, les taux d'erreur reviennent près de ceux constatés en français
- dans la description phonétique des mots, il n'est pas fait de distinction entre voyelles accentuées et non accentuées, ce qui constitue incontestablement une erreur ... bien française
- l'initialisation de l'apprentissage est faite à partir d'un réseau de mots isolés appris pour un seul locuteur, français et non canadien anglophone.

Cependant, des tests informels au micro avec locuteurs anglophones donnent une impression plus favorable que ne le laissent prévoir les chiffres ci-dessus.

VI CONCLUSION

Pour un premier développement d'application complexe sur Media50, avec reconnaissance phonétique, une durée d'un mois parait un minimum, et une connaissance des principes des algorithmes utilisés est utile. En contrepartie, le fonctionnement multilocuteur présente un intérêt certain.

Sur Vecsys Datavox, la prise en main du reconnaisseur et de son environnement de développement est assez aisée, et une semaine parait un délai raisonnable pour parvenir à une première réalisation non élémentaire, même pour un informaticien peu familier des algorithmes de traitement de la parole. La rançon en est le fonctionnement monolocuteur, néanmoins acceptable dans des applications de type bureautique comme celle ci. Temps de réponse et qualité de reconnaissance sont "impressionnants"

Notre travail a conduit à la conception et à la mise en oeuvre d'une entrée vocale en parole continue pour le système de traduction automatique CRITTER, en français et en anglais, la dictée étant faite par membres de phrase, entités bien adaptées ergonomiquement à cette application. Il démontre bien l'intérêt et la faisabilité de l'entrée vocale dans le cadre d'un poste de travail pour traducteur.

VII PERSPECTIVES

Pour orienter les suites à donner à cette étude, on peut faire les remarques suivantes:

On a utilisé ici dans les cartes de reconnaissance une syntaxe assez évoluée pour représenter un langage assez souple, de façon orthographiquement correcte. Mais on n'est pas du tout au niveau des connaissances syntaxiques mises en oeuvre dans CRITTER lui même, il s'en faut de beaucoup.

Il pourrait donc sembler beaucoup plus naturel et efficace de ne demander à un serveur de reconnaissance vocale qu'une reconnaissance d'entités phonétiques, disons pour simplifier de phonèmes, qui seraient ensuite traitées par l'analyseur syntaxique de CRITTER, adapté au cas phonétique (l'intermédiaire d'une représentation orthographique n'étant nullement indispensable). Il ne faut cependant pas sous-estimer l'effort que représente l'adaptation de cet analyseur syntaxique au cas phonétique.

Cette démarche n'a pas été possible avec les produits considérés, qui, même lorsqu'ils effectuent une reconnaissance phonétique (Media50) gèrent également le niveau mots, et exigent une syntaxe pour que le fonctionnement soit satisfaisant; le formalisme fourni pour exprimer cette syntaxe (équivalent à des grammaires régulières dans les deux cas) est insuffisant pour y implanter les connaissances de CRITTER.

A notre connaissance, le seul système commercialisé voisin d'un reconnaisseur phonétique est, pour l'anglais, le "Phonetic Engine" de Speech Systems Inc. (SSI), destiné précisément à être connecté à une station de travail comme celle de CRITTER.

Il faut aussi considérer pour le français la "machine phonétique" qui devrait résulter, prochainement, des travaux du CNET sur le système KEAL.

On devra suivre de près les travaux susceptibles de déboucher sur un reconnaisseur fournissant un jeu de phonèmes multilocuteur, indépendant de l'application (mais pas de la langue !). On peut penser actuellement que ce sont les équipes utilisant la modélisation de Markov qui y parviendront d'abord.

On pense donc au travail du CNET de Lannion (C Gagnoulet et al) pour le français, de l'Université Carnegie Mellon de Pitsburg (KF Lee et al), de BBN à Cambridge, MA (Schwarz et al.), de Bell Labs. (S Levinson et al)... pour l'anglais.

REMERCIEMENTS

J.P. Tubach souhaite exprimer ses remerciements :

- à la Direction du CCRIT de Montréal, et à la Direction de Télécom Paris pour avoir rendu possible le séjour sabbatique au cours duquel a été réalisé ce travail.
- à Pierre Hamel, Sylvain Faucher, Elliott Macklovitch, Michel Simard, du GTV et du GTA du CCRIT, qui l'ont grandement aidé.
- à Denis Jouvet, Jean Monné du CNET de Lannion, et Bernard Prouts de VECSYS, qui lui ont apporté à distance leur assistance.

BIBLIOGRAPHIE

Gage Canadian Dictionary (Avis S.W. et al.) Gage publishing limited, Toronto, Canada

Gagnoulet C., Jouvet D. "Développements récents en reconnaissance de la parole". L'Echo des recherches (CNET - ENST), No 135, 1989, pp 27-36

Hauptmann G.A., Rudincky, A.I. "Talking to computers : an empirical investigation". International Journal of Man-Machine studies, vol 28, pp 583-604 (1988)

Isabelle P., Dymetman M., Macklovitch E.: "CRITTER : un système de traduction pour les rapports sur les marchés agricoles". Document CCRIT, Montréal, Canada, No Co 28-1/25-1988F. (Version française de : "CRITTER, a translation system for agricultural market reports", 12ème conférence internationale sur la linguistique computationnelle (COLING), Budapest, Hongrie, Août 1988; et Document CCRIT, Montréal, Co 28-1/25-1988E.

Jouvet D. "Reconnaissance de mots connectés indépendemment du locuteur par des méthodes statistiques". Thèse de Doctorat de Télécom Paris (ENST), Juin 1988.

Quénot G., Gauvain J.L., Gangolf J.J., Mariani J.J. : "A dynamic programming processor for speech recognition". IEEE Transactions on Integrated Circuits. 1989

Tubach J.P., Gagnoulet C., Gauvain J.L. : "Advances in speech recognition products from France", Speech Tech'89 Conference, New York, Mai 1989

Tubach J.P. : "Reconnaissance de parole continue en entrée d'un système de traduction automatique, en français et en anglais", Document technique Télécom Paris, 89D007, 1989, et Document CCRIT, Montréal, 1990

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Idées et concepts de réalisation d'une machine à dicter destinée aux grands vocabulaires

K.Smaili, F.Charpillet, JM.Pierrel, JP.Haton

CRIN BP 239
54506 Vandoeuvre les-Nancy France

Résumé

Avec comme objectif de permettre la saisie de texte, la machine à dicter est sans doute , à terme, l'un des domaines les plus prometteurs du traitement de la parole. Cependant, compte tenu de l'état actuel des recherches, des avancés techniques importants restent à faire pour qu'une telle machine soit réellement opérationnelle. Nous présentons dans ce papier les efforts de recherches que nous menons actuellement à Nancy dans ce domaine.
Nous aborderons successivement dans ce papier l'architecture générale du système, la composante syntaxique, le niveau lexical et l'éditeur associé.

I/ Introduction

L'objectif essentiel des études sur la reconnaissance automatique de la parole est de permettre à terme une interaction, la plus naturelle possible entre l'homme et la machine dans le cadre d'une application spécifique [Pierrel 1989]. En ce sens, ces recherches sont nettement finalisées, même si elles nécessitent d'aborder des problèmes fondamentaux liés à la langue, support de la communication. Les applications potentielles de la reconnaissance et compréhension automatique de la parole font apparaître divers types de communication orale possibles entre l'homme et la machine, qui se différencient par l'utilisation faite de la parole, le niveau de langue traité et par conséquence, les traitements mis en œuvre .
Avec comme objectif de permettre la saisie orale de textes en machine, la machine à dicter est sans doute , à long terme, l'un des domaines les plus prometteurs du traitement de la parole. Certes, compte tenu de l'avancée des techniques, il n'est pas possible aujourd'hui de proposer des systèmes de saisies de texte efficace et facile à utiliser.. Pour parvenir à cet objectif divers problèmes restent ouverts :
- la définition de modèle syntaxique conduisant à une couverture maximale de la langue française,
- l'utilisation de très grands vocabulaires (plus de 5000 mots ou 15 000 formes),
- la possibilité d'utilisation de la parole continue et donc la nécessité d'un décodage acoustico-phonétique robuste et fiable,
- l'intégration dans un poste de travail d'un tel système de saisie.

II/ Etat de l'art dans ce domaine

Parmi les recherches entreprises dans ce domaine, citons par exemple IBM Yorktown aux Etats unis [Jelinek 82], le CERFIA à Toulouse [Pérennou 80], IBM Paris [Derouault, Merialdo 86], le LIMSI à Paris [Mariani 87], le CRIN à Nancy [Charpillet 85] et bien d'autres. Les Japonais ont été les premiers à donner le coup d'envoi avec l'annonce par leurs principaux constructeurs électroniques (Hitachi, Toshiba, Matsuchita) de l'arrivée imminente de tels systèmes. Cependant les particularités de la langue japonaise ne permettent pas d'étendre leur approche aux autres langues. En effet, une centaine de syllabes principales suffisent à caractériser cette langue. Ce chiffre est à comparer aux quelques milliers de syllabes des langues latines et anglo-saxonnes, pour lesquelles des unités plus fines et plus instables tel que le phonème doivent être utilisées.
Dans la littérature, on confond souvent système de reconnaissance de la parole (SRP) et machine à dicter (MAD). Il est vrai que la MAD est fortement dépendante du système de reconnaissance de la parole mais elle a besoin d'un certain nombre d'outils (que nous verrons en VI) pour qu'elle puisse fonctionner en tant que telle.
Dans le tableau ci-dessous (fig 1) nous citons différents SRP et MAD utilisant de grands vocabulaires (plus de 10 000 entrées) .

	Système	Elocution	Mode d'utilisation	Nombre de mots	Performances
IBM France	Tangora	Syllabique	Mono locuteur	200 000	87,3 % sur 79 phrases
INRS / BELL Northern	INRS	Mots isolés	Mono locuteur	75 000	90 %
AT et T	AT et T	Mots isolés	Mono locuteur	52 000	60 %
SSI Tarsans	PE 200	parole continue	Multi-locuteur	36 000	?
Olivetti	Olivetti	Mots isolés	Mono locuteur	30 000	90 % sur 361 mots
Dragon	Dragon	Mots isolés	Mono locuteur	20 000	90 %
IBM Yorktown	TangoraMots isolés		Mono locuteur	20 000	93 % sur 50 phrases
LIMSI	Hamlet	Mots isolés	Mono locuteur	quelques milliers	95 % sur 100 phrases

SRP et MAD utilisant de grands vocabulaires
Fig 1

Parmi ceux-ci, le seul système fonctionnant en parole continue est le PE200 [Meisel 89] qui ne nécessite aucune phase d'apprentissage. Les paramètres acoustiques correspondant au signal d'entrée sont convertis en phonèmes grâce à un réseau neuronal très particulier; le passage des phonèmes aux mots se fait à l'aide d'une description syntaxique des suites de mots et d'un lexique d'une grande taille. Malheureusement, nous ne connaissons pas les performances de ce système pour pouvoir les comparer aux autres. Un autre système se démarque dans ce tableau : Tangora d'IBM France. Il utilise un très grand vocabulaire, mais son inconvénient majeur est le mode d'élocution syllabique très contraignant. La caractéristique principale de la plupart des systèmes cités ci-dessus est l'apprentissage plus ou moins fastidieux nécessaire pour chaque nouveau locuteur. C'est le cas de Tangora 200 000 qui nécessite la prononciation de pas moins de 400 phrases pour tout locuteur. Le système d'Olivetti est le moins exigeant en apprentissage. L'élocution de 40 mots pour chaque locuteur est suffisante.
Quant aux performances, il est très difficile de les comparer parce que les conditions de tests ne sont jamais identiques, la taille du vocabulaire diffère d'un système à l'autre, la langue utilisée présente plus ou moins des difficultés phonétiques et syntaxiques et surtout les tests effectués manquent toujours de précisions. On ne sait pas, par exemple, si le résultat cité est obtenu en première réponse du système ou parmi les n premières propositions.

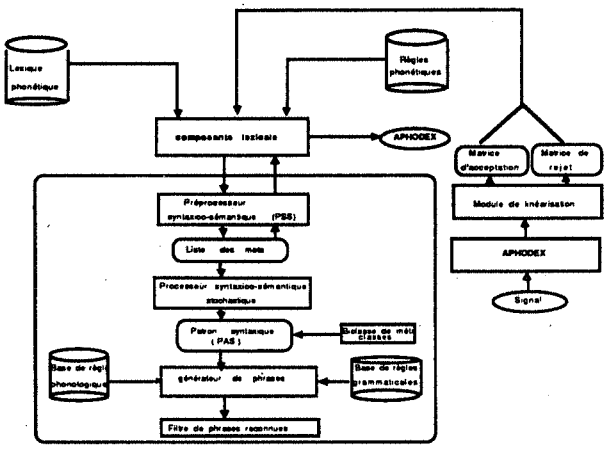
A titre d'exemple, le système d'AT et T donne un score relativement faible (60%) [Levinson 89] du fait qu'il prend en compte seulement les 10 premières propositions.

III/ Architecture générale du système de reconnaissance

III.1/ Présentation de l'architecture

Le système de reconnaissance que nous proposons est fondé sur trois composantes principales :
- la composante acoustico-phonétique qui à partir d'un signal acoustique délivre en sortie un treillis phonétique d'acceptation (TPA) [Smaili 89] utilisé pour identifier les entités lexicales et un treillis phonétique de rejet qui permet de réfuter un certain nombre d'hypothèses émises par le niveau lexical,
- la composante lexicale qui est le noyau de ce système de reconnaissance permet d'extraire du TPA les mots susceptibles d'avoir été prononcés,
- la composante syntaxico-sémantique a un double rôle dont le premier est surtout prédictif et le second sélectif.

Le schéma général du système de reconnaissance est illustré par la figure 2 et les différentes composantes sont explicitées dans les points qui suivent.



Architecture générale du SRP
Fig 2

III.2/ Le décodage acoustico-phonétique (DAP)

Etant donné les objectifs de reconnaissance de la parole continue que nous nous sommes fixés, nous avons opté pour un décodage analytique phonémique. Le décodage acoustico-phonétique que nous utilisons correspond au système APHODEX [Fohr 87] : il s'appuie sur une approche à base de connaissances explicites qui présente l'avantage intrinsèque d'offrir de nombreuses facilités pour mettre en œuvre les connaissances acquises par des phonéticiens sur le propriétés acoustico-phonétiques de la parole continue et pour les intégrer au processus de reconnaissance. Dans ce cadre, nous avons axé tous nos efforts sur le DAP multi-locuteur. Actuellement la qualité des performances d'APHODEX ainsi que sa robustesse à la variabilité inter-locuteur sont démontrées. Néanmoins il est souhaitable d'améliorer sensiblement notre DAP qui ne fournit en fait que 70 % de bonne identification phonétique en parole continue multi-locuteur sans apprentissage. Nous travaillons en particulier à une version incluant une adaptation au locuteur.

IV/ Le modèle syntaxique

Le rôle de la syntaxe en reconnaissance de la parole est de participer au choix du prochain mot à reconnaître et à l'élimination d'un certain nombre de solutions incorrectes. Deux approches sont envisageables pour mettre en oeuvre un tel modèle, selon que le domaine d'application est restreint ou non. Dans le premier cas, les phénomènes liés à l'application sont modélisables, par conséquent la solution préconisée est l'écriture d'une grammaire [Pierrel 87]. Par contre, lorsque l'on s'intéresse au langage naturel, le problème est beaucoup plus difficile. En effet, la linguistique n'a pas progressé au point de nous fournir une grammaire de taille raisonnable tenant compte de tous les phénomènes observés. C'est pourquoi, les linguistes préfèrent en général l'écriture de grammaires se limitant à quelques phénomènes de la langue, c'est ainsi que, Lacouture et Lapalme [Lacouture, Lapalme 88] se sont inspirés d'une étude ancienne pour implanter les règles du français fondamental. Ce système ne pourra pas être utilisé seul comme l'ont précisé les auteurs eux mêmes dans une application du type MAD, mais pourrait faire l'objet d'un noyau grammatical autour duquel viendraient s'agréger des règles de constructions grammaticales plus élaborées. Une autre approche utilisée dans la plupart des SRP existants est l'approche probabiliste.
Nous décrivons dans le paragraphe suivant le principe de cette méthode pour laquelle nous avons opté.

IV.1/ Présentation du modèle choisi

La séquence de mots d'une phrase prononcée ou écrite obéit à des contraintes syntaxico-sémantiques et comme la probabilité de production d'un mot dépend conditionnellement de toute la première partie de la phrase prononcée [Jelinek 82], il est donc tout à fait naturel de penser à l'utilisation d'un modèle probabiliste. Mais en pratique, il est très difficile d'estimer cette probabilité, c'est pourquoi Jelinek a proposé de réduire le calcul aux deux derniers mots de l'entité en cours d'interprétation. Cependant, pour des problèmes d'apprentissage du modèle, l'unité syntaxique ne peut être le mot. En effet, le calcul de probabilité d'occurrence sur les séquences de trois mots demanderait des corpus d'apprentissage de taille prohibitive. Pour remédier à ce problème, l'idée utilisée est de revenir aux matrices de précédences fréquentielles [Debili 77] qui est une idée semblable à celle des suites de mots en substituant à ces derniers leurs classes syntaxico-sémantiques.

V.2/ Mise en œuvre

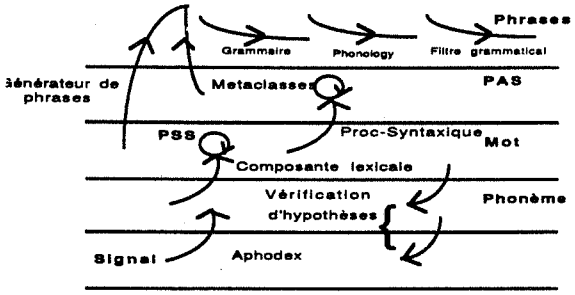
Le modèle linguistique que nous avons choisi est fortement dépendant des fréquences d'apparition de bi et triclasses mais utilise d'autres concepts plus spécifiques à la grammaire française. Ce modèle est composé de sept modules syntaxico-sémantiques (l'interaction entre les différents modules est schématisée par la figure 3) :
- Un préprocesseur syntaxico-sémantique (PSS) qui collabore avec la composante lexicale pour filtrer les premiers flux de mots. Le PSS est basé exclusivement sur le modèle biclasse, ce qui permet d'avoir rapidement une première interprétation du TPA.
- Un processeur syntaxico-sémantique stochastique qui à partir de la liste de mots fournit par le PSS génère la liste des patrons syntaxiques (PAS) (une suite de classes syntaxico-sémantiques représentative de la phrase). A chaque patron syntaxique, un coefficient de plausibilité (CP) est affectée. Ce coefficient est calculé comme suit :

$$CP = \prod_{i=2}^{n} (F^{ci}_{ci-2,ci-1} + k*F^{ci}_{ci-1} + h)$$

où :
$F^{ci}_{ci-2,ci-1}$ représente la fréquence d'apparition de la triclasse (ci-2,ci-1,ci)

et
F^{ci}_{ci-1} la fréquence d'apparition de la biclasse (ci-1,ci)

k et h sont deux constantes qui varient en fonction du nombre de bi et triclasses.



Interaction entre les différentes sources de connaissances
Fig 3

- Un filtre éliminant les PAS qui n'ont pas de sens grammatical (suite de classes syntaxiques composée uniquement d'adjectifs, par exemple). Ce filtre est basé sur la fréquence d'apparition d'une suite d'ordre 2 de métaclasses où la métaclasse n'est en fait qu'une triclasse existante.
- Le générateur de phrases utilise les PAS et la liste des mots générées par le PSS pour afficher les phrases les plus probables.
- La base de règles grammaticales qui permet de prendre en compte les phénomènes syntaxico-sémantiques qui ne peuvent l'être uniquement avec les modèles fréquentiels. Ces règles contribuent à l'élimination d'un certain nombre de phrases parasites.
- La base de règles phonologiques permet de prendre en compte les phénomènes phonologiques de la langue française. On trouvera dans cette base des règles du type :

```
Statut = Si Mot1 = "mon"
            Alors  Si  voyelle (Mot2[1]) && féminin (Mot2)
                       Alors   "Accepté"
                   Fsi
         Fsi
```

- Enfin un filtre de phrases reconnues qui n'est qu'un ensemble de règles grammaticales plus générales ayant une portée plus longue que celle des règles grammaticales utilisées précédemment.

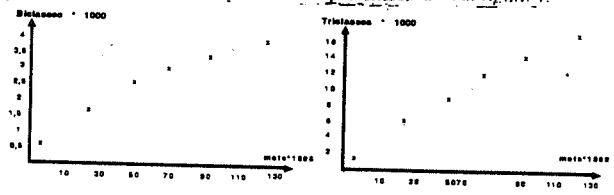
Il est tout à fait évident que malgré le nombre élevé de modules syntaxiques de ce système, il restera toujours des phrases qui ne pourront pas être éliminées. La solution dans ce cas est de laisser l'utilisateur intervenir à l'aide d'une interface ergonomique qui lui permettra de sélectionner la bonne phrase.

V.2.1/ Les classes syntaxico-sémantiques

Elles sont au nombre de 200. A partir des 8 classes grammaticales, nous avons affiné chacune d'elles pour en construire 200. Ce nombre élevé s'explique par le fait qu'on a voulu introduire d'une façon indirecte la sémantique dans les classes. C'est ainsi que l' on trouve beaucoup de classes ne possédant qu'un seul élément. Ceci qui permet de connaître d'une façon exacte le contexte de la classe donc du mot. La règle d'appartenance d'un mot à une classe obéit à une contrainte théorique très stricte : tout mot d'une classe peut être substitué à un autre de la même classe sans que la structure syntaxique de la phrase ne soit modifiée.

V.2.2 / Le module d'étiquetage et ses outils

Le modèle probabiliste est construit par apprentissage semi-automatique sur des corpus de textes très diversifiés. En effet, l'apprentissage a d'abord été fait de façon manuelle sur un ensemble de lettres internes du GRECO communication parlée qui est composé de 3500 mots puis sur un livre très technique [Pierrel 87] de 23 000 mots mais cette fois-ci d'une façon semi-automatique [Smaili 89]. Actuellement l'étiquetage se fait sur un corpus de taille beaucoup plus importante (340 000 mots) que nous avons extraits des nouvelles d'un journal local (L'Est républicain). Les statistiques dont nous disposons à l'heure actuelle sont représentées dans la figure 4.



Variation du nombre de bi et triclasses en fonction du nombre de mots

Fig 4

Nous pouvons constater qu'avec 130 000 mots la courbe des triclasses ne converge pas. Il faut encore davantage de textes pour espérer arriver à la convergence. Quant à la courbe des biclasses, nous pensons qu'elle convergera très vite et que le texte dont nous disposons suffira amplement.

Tous les textes étiquetés automatiquement ont été vérifiés par un linguiste qui a corrigé les éventuelles erreurs de l'étiqueteur automatique. Pour ce faire, nous avons développé un outil d'aide à l'étiquetage qui est un éditeur multi-fenêtre permettant l'étiquetage et la correction d'une façon très ergonomique.

V/ Le niveau lexical

V.1/ Revue critique des systèmes à très grand vocabulaire (TGV)

Lors de la conception d'un SRP, le problème du choix de l'unité de décision se pose toujours. Ce problème est clairement posé dans [Lea 80] en termes d'avantages et d'inconvénients. Jusqu'à ces dernières années, la prédominance des systèmes de reconnaissance globale a fait qu'on confondait souvent la reconnaissance proprement dite et l'accès lexical. Les systèmes de reconnaissance analytique ont, de par leur nature, une étape d'accès lexical. La conception de composante lexicale pour les systèmes à TGV repose [Adda 87] sur une stratégie d'analyse ascendante et sur l'utilisation de nombreuses sources de connaissance :

- Entre une analyse de type ascendant ou descendant, seul le premier ou une combinaison des deux est réaliste.

En effet, dans le cadre d'une MAD, le niveau syntaxique n'est pas suffisamment déterministe pour baser la reconnaissance sur l'ensemble des prédictions syntaxiques.

- L'utilisation de nombreuses sources de connaissances est une technique très prometteuse parce qu'elle tente d'exploiter plusieurs informations qui proviennent de sources indépendantes [Haton 87], ce qui permet d'augmenter les performances du système. Mais le problème qui se pose c'est l'attribution des poids aux différentes méthodes qui pourraient être utilisées pour interpréter le treillis phonétique.

L'utilisation de traits robustes est un autre principe des SRP à très grand vocabulaire en mots isolés, mais son utilisation en parole continue est peu discriminante. En effet, la plupart des mots ne sont pas détectés avec les classes phonétiques majeures, la quasi-totalité des erreurs (fausse détection) s'effectuant à la jonction entre plusieurs mots.

V.2/ Les choix lexicaux de notre système

a / Vocabulaire

Le dictionnaire est composé de 37 000 entrées lexicales (formes fléchies incluses) extraites du dictionnaire de la base de données lexicale BDLEX [Pérennou 87] qui en compte 230 000. Chaque entrée du vocabulaire sélectionné est composée des champs suivants :

forme orthographique du mot, forme phonétique, finale phonologique qui permet de traiter les phénomènes de frontières de mots, classe syntaxique et quelques traits syntaxiques comme le genre et le nombre pour les noms et les adjectifs, le temps de conjugaison pour les verbes ...etc.

Pour des raisons de rapidité d'accès aux différents mots d'une même classe, nous avons préféré les regrouper dans un même sous vocabulaire. Ce principe a été généralisé à toutes les classes grammaticales

b / Stratégie

Etant donné l'objectif final que nous nous sommes fixés à savoir la reconnaissance de la parole continue, nous avons écarté l'idée d'utilisation des traits robustes pour les raisons citées plus haut. Quant à la stratégie nous avons voulu faire jouer au modèle syntaxique un rôle sélectif et prédictif vu l'importance de la tâche. Par conséquent la stratégie est de type mixte ayant comme point d'interaction le niveau lexical.

c / Règles et données phonétiques

L'un des facteurs rendant la recherche lexicale ardue est sans doute la non fiabilité du treillis phonétique. De ce fait, la composante lexicale n'a pas d'autre alternative que d'opérer sur une information incertaine. Pour pallier ces problèmes le système utilise des règles phonétiques et des données phonétiques calculées sur un corpus de 14 000 phonèmes qui sont représentées dans les 4 matrices suivantes :

- une matrice de confusion qui donne la probabilité de similitude entre deux phonèmes. Cette matrice de confusion est indépendante du DAP utilisé,
- une matrice de confusion du système qui modélise globalement le comportement du DAP. Les pondérations données par les deux matrices de confusion sont combinées à l'aide de la méthode de Dempster-Shafer [Zafeh 86] pour n'en former qu'une,
- deux autres matrices sont utilisées pour éviter au modèle lexical d'autoriser n'importe quel type d'élision et/ou insertion.

d / Le retour arrière vers le DAP

La matrice d'élision réduit considérablement le nombre d'opérations d'élisions mais certaines ne peuvent être évitées. En effet, les liquides par exemple sont des phonèmes qui sont très souvent omis par APHODEX ce qui se traduit par une probabilité d'élision importante. Le retour arrière relancera le décodage sur un segment donné mais cette fois-ci avec des informations morphologiques précises. La réponse du DAP sera de confirmer ou d'infirmer l'hypothèse d'élision.

VI/ L'éditeur associé à la MAD

V.1/ De la nécessite d'un éditeur associé

Lorsqu'on dicte une lettre administrative, par exemple, il arrive très souvent qu'on ait envie de remplacer un mot par un autre ou tout simplement de supprimer ce mot; un simple système de reconnaissance ne suffit plus, il faut y ajouter un éditeur performant permettant d'effectuer facilement les modifications ou corrections sur le texte reconnu. Deux approches sont alors possibles suivant que l'on choisit un éditeur vocal ou non.

La première est plutôt difficile à mettre en œuvre car il faut pouvoir distinguer du flot des données dictées les mots du langage de commande de l'éditeur.

Quant à la deuxième approche, elle est beaucoup plus abordable surtout dans un poste de travail multi-média. Il suffit de permettre au locuteur de pouvoir interrompre le système de reconnaissance avec par exemple la souris et de donner la main à un éditeur de texte qui permettra de cliquer sur le mot à modifier. Un système de reconnaissance de mots isolés pourra être déclenché pour remplacer le mot sélectionné. Une solution plus facile consisterait à modifier le mot désigné par le clavier ou à sélectionner, par la souris, le mot correct parmi les autres hypothèses fournies par le système [Baker 89].

VI.2/ Solutions retenues dans notre système

Etant donnée la complexité de la mise en œuvre de la solution éditeur vocal, nous avons opté pour la deuxième approche. L'éditeur actuel est plutôt un système de désignation par souris orienté concepteur de MAD.

En effet, plusieurs informations importantes comme le TPA, les différents scores des mots reconnus sont affichés à l'écran dans des fenêtres particulières. Cependant nous sommes conscients que ces informations peuvent ne pas intéresser l'utilisateur final. C'est pourquoi nous envisageons à terme de développer une interface plus orienté vers l'utilisateur final. Les mots reconnus forment un treillis dont chaque élément est représenté par une fenêtre. Chaque fenêtre contient l'ensemble des mots reconnus correspondant à une même portion du signal. A tout moment le SRP peut être interrompu pour interagir avec l'utilisateur final par l'intermédiaire d'un éditeur qui propose les fonctions suivantes :

poursuite de la reconnaissance sur le reste du TPA, choix d'un mot parmi ceux qui sont reconnus, proposition d'un mot non reconnu par le système, consultation des autres zones et affichage des scores des mots reconnus.

Une fenêtre de dialogue est constamment présente à l'écran. C'est par l'intermédiaire de celle-ci que s'effectue certaines tâches décrites ci-dessus. Elle affiche également une trace du fonctionnement du système.

Les phrases reconnues et validées par l'utilisateur sont éditées dans un fichier dont une image apparaît à l'écran.

VII/ Implantation et exemples de résultats

Le système de transcription comprenant le modèle linguistique, les modules de décodage, l'interface graphique et les différentes sources de connaissances sont dévelopés sur une station masscomp. Le filtre méta-classes est le seul module à ne pas avoir été implanté à cause de la non convergence du modèle triclasses. Nous donnons ci-dessous les premiers résultats obtenus en mots isolés sur un vocabulaire de 6700 entrées lexicales:

Phrase prononcée : *La bise et le soleil se diputaient*

TPA généré:

a	p	i	s	ai	l	u	s	i	l	e	j	s	ai	l	i	s	p	y	p	e
œ	b	e	z	e	m	o	z	e	m	y	l		y	j	e	z	b		b	ai
			l	y	j			ai	j	œ	m			r		l				œ
					&						n									
					œ															

Mots générés par le PSS:

1/ (**à la** le as deux de ne âme(s) âne(s) ail(s) oeil me eux là)
2/ (pile **bel** belle(s) pire(s) baisse(s) bile(s) **bise** bol(s) bosse(s) caisse(s) père(s) ville(s) tel(le)(s) dise(s)(nt) quel(s))
3/ (dû du(e) gai(e)(s) laid(s) nu(e)(s) mais le la les une(s) ai deux de du des es est **et** ne aide(s) aile(s) nez oeil me elle(s) eux né(e)(s) aide(s)(nt) aime(s)(nt) naît nais dès mai)
4/ (**le** mou dieu(x) lion(s) me mien)
5/ (civil(s) civile(s) **soleil** sommeil(s))
6/ (ces sale sa ses six chez celle ceux selle salle(s) scie selle(s) sel(s) sol(s) somme(s) sot(s) **se** sa sue si sais sait)
7/ (**disput**(ais ait **aient** er ez é(e)(s)) multipli(ais ait aient er ez é(e)(s)))

Chaque liste de mots est représentée par une fenêtre au niveau de l'éditeur graphique.

Phrases générées:

à baisse(s) / bise mou soleil se disput [aient , ait]
la baisse / bise est mou soleil sain disputé
la baisse / bise est mou soleil se disput [aient , ait]
la baisse / bise et mou soleil se disput [aient , ait]
la baisse / bise est le soleil sain disputé
la baisse / bise et le soleil se disput[aient , ait]
la baisse / bise est le soleil se disput[aient , ait]
la baisse / bise est mou soleil sot disputé
la baisse / bise est le soleil sot disputé

D'aprés ces résultats nous pouvons constater que le PSS a reconnu tous les mots de cette phrase. Aucune intervention de la part de l'utilisateur n'a été nécessaire. D'autre part la phrase prononcée a été reconnue mais malheureusement d'autres l'ont été également. Aucun modèle linguistique ne pourra éliminer une phrase comme : *La baisse est le soleil sain disputé* sauf au détriment d'une sémantique très stricte et appliquée à un domaine d'application restreint.

D'autres phrases ont un score global meilleur que celui de la phrase prononcée. C' est le cas de la phrase *à bise le soleil se disputait*. En effet le mot "à" a été reconnu avec un score de 100% puisque le /l/ du mot "la" n'a pas été reconnu par Aphodex.

VIII/ Conclusion

La machine à dicter est une idée attractive pour les postes de travail du futur mais malheureusement beaucoup de problèmes liés à la reconnaissance de la parole continue restent sans solutions définitives. Dans ce papier, nous avons présenté un système qui commence à donner ses premiers résultats en mots isolés. Ces résultats sont encourageants et nous pensons encore les améliorer. En effet, l'apprentissage du modèle stochastique est entrain de se faire sur un corpus d'apprentissage de taille beaucoup plus importante. Nous pensons également introduire la fréquence d'apparition des mots (de quelques mots qui présentent des ambiguïtés syntaxiques) ce qui permettra de rendre le modèle plus fin et plus précis. La plausibilité du patron syntaxique n'est pas introduite dans le calcul du score global de la phrase [Smaili 89]. Une fois que le modèle convergera, il sera trés intéressant d'introduire ce coefficient de plausibilité, ce qui permettra probablement d'éliminer un grand nombre de phrases engendrées.

Nous estimons qu'avec notre savoir faire actuel en linguistique, nous ne pouvons pas faire mieux au niveau syntaxique. En revanche beaucoup d'efforts restent à faire au niveau du DAP, notamment l'ajout de règles phonétiques permettant de prendre en compte tous les phénomènes de la langue. Nous pensons également que les résultats du DAP ne peuvent être améliorés qu'avec l'utilisation de plusieurs sources de connaissances et de plusieurs méthodes de décodage acoustico-phonétique agissant sur un même signal.

Références

[Adda 87] : G.Adda " Reconnaissance de grands vocabulaires: une étude syntaxique et lexicale", *Thèse de Docteur-ingénieur en informatique, Paris XI DEC 1987.*

[Baker 89] : JK.Baker "A second-generation large vocabulary system" Speech Technology
man / machine voice communications, pp 20-24, May 1989.

[Charpillet 85] : F.Charpillet " Un système de reconnaissance de parole continue pour la saisie de textes lus", *Thèse de Docteur d'université en informatique, NANCY I 1985.*

[Debili 77] : F.Debili " Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage" *Thèse de docteur ingénieur, spécialité électronique, option traitement de l'information Sep 1977*

[Derouault, Mérialdo 86]: AM.Derouault, B.Mérialdo " Natural language modeling for phoneme to text transcription", *IEEE Transactions on pattern analysis and machine intelligence VOL PAMI-8 N°6 NOV 1986*

[Fohr 87] : N.Carbonell, D.Fohr, JP.Haton " Aphodex, an acoustic-phonetic decoding expert system" *International Journal of Pattern Recognition and Artificial Intelligence VOL 1, N°2 1987*

[Haton 87] : JP.Haton " Interaction between stochastic modeling and knowledge-based techniques in acoustic-phonetic decoding of speech", *Proc IEEE ICASSP 1987*

[Jelinek 82] : F.Jelinek, R.L Mercer, LR.Bahl " Continuous speech recognition: statistical methods", *CSR group, IBM TJ Watson research center, Yorktown Heigths NY 10598*

[Lacouture, Lapalme 88] : R.Lacouture, G.Lapalme " Une implantation informatique du français fondamental", *Technique et Science Informatiques Mai 1988*

[Lea 80] : WA.Lea, JE.Shoup " Contributions of the ARPA-SUR project" *in Trends in speech recognition, prenctice hall.*

[Levinson 89] : SE.Levinson, A.Ljolje, LG.Miller " Large vocabulary speech recognition using a hidden Markov model for acoustic/phonetic classification", *Speech Technology Apr / May 1989*

[Meisel 89] : WS.Meisel, MP.Fortunato, WD.Michalek " A phonetically based speech recognition system", *Speech Technology Apr / May 1989*

[Pierrel 87] : JM.Pierrel " Dialogue oral homme machine", *edition HERMES , 1987*

[Pierrel 90] : JM.Pierrel, N.Carbonell, JP.Haton, K.Smaili " Vers une meilleure intégration de la parole dans des systèmes de communication homme-machine", *soumis à la revue traitement de signal 1989.*

[Pérennou 80] : G.Pérennou, A.Fatholahzadeh, M.de Calmes " le filtrage syntaxique pour l'entrée vocale de texte" *Syntax and semantics in speech understanding, Paimpont sept 1980.*

[Pérennou 87] : G.Pérennou "Bdlex : a data and cognition base of spoken French" Research and development in language processing INRIA , PARIS 1987.

[Smaili 89] : JP.Haton, K.Smaili " A dictation machine based upon 35 000 french words", *Rapport interne CRIN SEP 1989*

[Zadeh 86] : L.Zadeh " A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination", *AI Magazine 1986*

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Le triplet Phonétique en décodage acoustico-phonétique

Yves Laprie

CRIN -INRIA Lorraine
BP 239
54506 Vandœuvre-lès-Nancy

Résumé

Cet article est consacré à la présentation du triplet phonétique en décodage acoustico-phonétique. Un triplet modélise en termes d'événements acoustiques et d'indices articulatoires un phonème en contexte. Nous étudions d'abord le processus d'acquisition et la structuration des connaissances. Ensuite nous présentons le mécanisme de décodage qui opère en deux étapes. La première est destinée à apparier aux instances de triplets de la phrase à décoder les triplets de références les plus proches en termes d'indices articulatoires ou éventuellement d'événements acoustiques. La seconde étape opère par relaxation sur les contraintes existant entre les triplets et permet d'éliminer, parmi les triplets candidats, ceux qui sont incohérents par rapport au reste de la phrase.

1 Introduction

Les modèles de Markov cachés (HMM) ont remporté ces dernières années un vif succès dans le domaine de la reconnaissance automatique de la parole continue multi-locuteurs. Ils permettent en effet de modéliser le signal de parole sans nécessiter aucune connaissance des phénomènes liés à la production de la parole. Cette absence de connaissance permet certes d'obtenir de bons taux de reconnaissance globaux pour une application limitée (environ un millier de mots pour SPHINX [16]), mais rend fort difficile une amélioration du système de décodage.

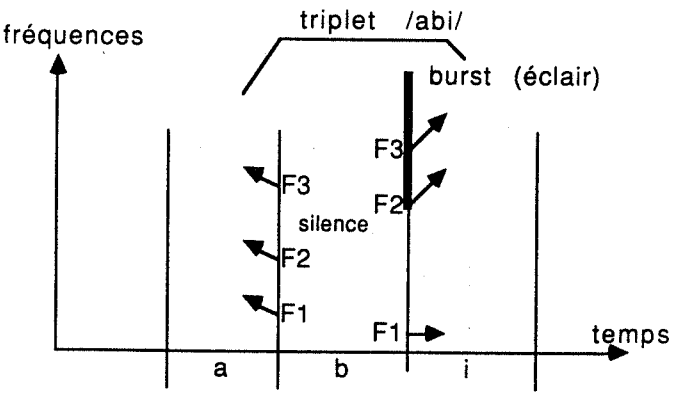
Plusieurs expériences en lecture de spectrogrammes [1] [2] ont montré qu'un expert phonéticien reconnaît plus de 80% des phonèmes d'une phrase, ce qui est mieux que le niveau acoustique d'un système à base de HMM. En revanche les sytèmes experts en lecture de spectrogrammes développés depuis le début des années 80 n'égalent pas ces performances [3]. L'équipe dans laquelle nous travaillons a conçu et implanté un système de ce genre -Aphodex- [3] grâce à l'aide de F. Lonchamp de l'Institut de Phonétique de Nancy. Le formalisme retenu pour modéliser la connaissance, en l'occurence un ensemble de règles de production s'est révélé assez efficace pour simuler le raisonnement de l'expert ; il est par contre assez difficile de connaître l'état de l'expertise acoustico-phonétique que le système a reçue de la part du phonéticien. La difficulté à évaluer les connaissances du système est d'ailleurs accrue par le grand nombre de règles qu'il faut ajouter pour modéliser correctement les phénomènes contextuels de la parole continue.

2 Définition et intérêt du triplet.

Comme l'expertise acoustico-phonétique est destinée à la classification de phones, il est naturel de construire des prototypes représentant les phones à reconnaître. Pour que ces prototypes soient utilisables, ils doivent modéliser la réalisation acoustique en prenant en compte les phénomènes de coarticulation liés à la présence des phones voisins. Cela nous a donc conduit à proposer un système dont le grain de connaissance est le triplet :

un phone avec son contexte phonétique.

Il est important de noter qu'un triplet ne doit pas inclure les phonèmes voisins en entier sinon il devient à son tour soumis aux phénomènes contextuels auxquels on veut précisément échapper.



Exemple de description d'un triplet

Exemple de description d'un triplet

Figure 1

Un triplet s'articule donc autour des deux frontières du phone central comme le montre la figure 1. Il faut plutôt considérer une frontière comme une zone dans laquelle apparaissent l'essentiel des phénomènes de coarticulation, que comme une limite précise impossible à trouver dans de nombreux cas. A chacune des frontières sont attachés les événements acoustiques suivants s'ils sont présents :

- transitions formantiques,
- barre d'explosion décrit par sa durée, son intensité bruit de friction d'un triplet, centré sur un son fricatif, est scindé en deux parties, l'une accrochée à la frontière gauche, l'autre à la frontière droite. Cette séparation artificielle a l'avantage de pouvoir prendre en compte facilement des limites inférieures de bruit montantes (resp. descendantes) puis descendantes (resp. montantes), ainsi que le bruit de friction apparaissant après la barre d'explosion de certaines occlusives.
- points d'échange des cavités formantiques,
- profil de la micromélodie.

Pour que les références fréquentielles soient suffisamment précises nous avons ajouté un "centre de triplet" qui indique les fréquences des formants au centre du triplet.

Tel que nous venons de le définir, un triplet n'est que la description acoustique d'un phonème en contexte et ne porte donc aucune information sur l'expertise qu'un lecteur de spectrogramme a pu accumuler. La description acoustique du triplet est donc complétée par les indices qu'utilise un expert phonéticien. Un indice est un corrélat acoustique que l'expert peut justifier sur le plan articulatoire ; il doit être indépendant du locuteur, résistant et reconnu par l'expert comme significatif. Il s'agit par exemple :

- de la pince vélaire (rapprochement de F2 et F3 à la frontière d'une occlusive vélaire),
- de la position relative de la concentration d'énergie d'un burst par rapport à celle des formants.

Les deux facettes de cette représentation permettent d'orienter le décodage soit vers la recherche des triplets en fonction d'indices, quand les indices apparaissent clairement dans le signal, soit vers la recherche des triplets en fonction de leur réalisation acoustique quand peu d'indices décisifs ont été découverts.

La facette indicelle d'un triplet n'est d'ailleurs pas toujours présente, soit qu'elle n'ait pas été construite par l'expert, soit que l'état des connaissances acoustico-phonétiques ne permette pas de définir des indices pertinents. Cette facette est cependant très intéressante car elle permet de reconnaître avec plus de certitude les triplets qui contiennent des indices très caractéristiques, ceux que l'on commence maintenant à bien connaître.

A l'intérêt que représente le triplet pour exprimer les connaissances acoustico-phonétiques et assurer la cohérence de l'expertise, s'ajoute la possibilité d'adapter le décodage en fonction du locuteur. Il est évidemment possible de normaliser, notamment en fréquence, les triplets de la base de connaissances ; cette adaptation ne peut être que grossière puisque l'on veut qu'elle soit très rapide. Mais on peut en fait aller beaucoup plus loin. Il est en effet possible de vérifier que les relations fréquentielles existant entre deux triplets de la base de connaissances, proposés comme solutions du décodage pour deux segments à décoder, existent aussi entre leurs instances dans la phrase à décoder . Nous reviendrons en détails sur ce point au § 4.2 qui permet de tirer profit de l'unicité du locuteur pour construire un décodage global et cohérent.

3 Organisation des connaissances.

Le volume de connaissances à acquérir est évidemment considérable même si de nombreux triplets ne se rencontrent pas en Français. L'étude de Tubach et Boé [4] permet de constater qu'environ 1750 triplets suffisent à représenter 75% des triplets d'un très vaste corpus (300000 phonèmes). Il reste que pour valider l'intérêt du triplet en décodage, nous nous limiterons à un sous-ensemble de l'ordre de 500 triplets représentant essentiellement des occlusives et des sonantes qui sont des phonèmes assez difficiles à reconnaître avec Aphodex. Notons que ce nombre de triplets ne représente que la moitié des triplets construits à partir des occlusives en contexte vocalique.

Les triplets sont organisés suivant leur profil phonétique (le triplet de la figure 1 appartient donc à la classe [Voyelle Occlusive voisée Voyelle]). Cela représente encore un grand nombre de triplets par classe ([Voyelle Occlusive voisée Voyelle] conduit à 13 * 3 * 13 = 507 triplets). Nous verrons au § 4 comment affiner cette organisation.

La réduction du nombre des triplets (et donc de l'effort d'acquisition des connaissances acoustico-phonétiques) destinée à faciliter l'évaluation de cette approche ne doit pas occulter le fait qu'une très vaste connaissance des phénomènes de parole doit être disponible pour améliorer les performances de l'approche experte, cela quelle que soit la manière dont l'expertise est formalisée.

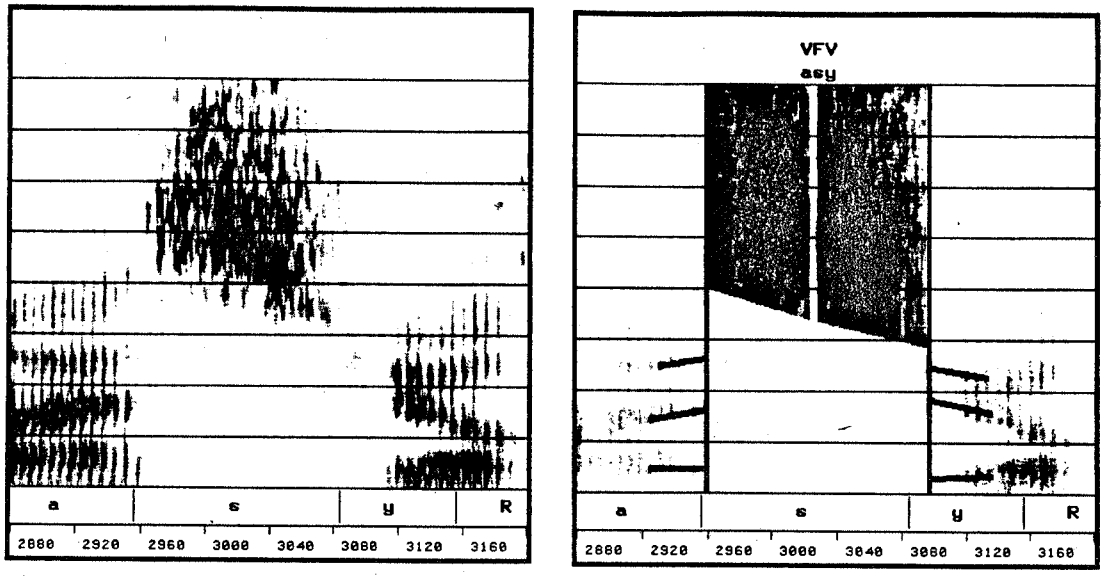
4 Acquisition des connaissances

L'acquisition des connaissances est scindée en deux étapes. La première consiste à collecter les triplets d'un corpus. La description est effectuée grâce à un éditeur de triplets convivial que nous avons développé sur le noyau du système Snorri [5]. La description a lieu directement sur le spectrogramme, l'utilisateur "dessinant" (fig. 2) avec la souris les différents événements acoustiques. Il est possible de procéder de manière semi-automatique pour les formants en utilisant le suivi que construit Snorri et en le corrigeant éventuellement ensuite. Comme il est essentiel que l'apprentissage se fasse sur des données correctes, nous préférons, pour l'instant du moins, conserver une partie manuelle qui permet de s'assurer de la pertinence des triplets collectés. Néanmoins, le fait que la description acoustique puisse être construite semi-automatiquement, avec les détecteurs d'événements acoustiques dont nous disposons, assure qu'il est possible de retrouver cette description durant la phase de reconnaissance.

Snorri permettant d'extraire une séquence de phonèmes donnée d'un corpus étiqueté, l'utilisateur peut collecter tous les triplets ayant le même profil phonétique et donc concentrer son attention sur une classe de triplets bien précise.

La seconde partie de l'apprentissage consiste à élaborer les triplets de référence à partir des triplets qui viennent d'être collectés. C'est là qu'intervient l'expertise du phonéticien qui permet de retenir les événements acoustiques pertinents et de définir parallèlement les corrélats acoustiques qui serviront comme indices de décodage.

Les corpus dont nous disposons ("La bise et le soleil" extrait du corpus BDSONS ainsi que le corpus d'O. Mella [6]) et grâce auxquels nous allons construire les triplets de test ne contiennent qu'environ un millier de triplets. Pour étendre cette base de connaissances il faudra donc, soit de nombreux autres corpus, soit pouvoir construire directement les triplets de référence.



Exemple de construction d'un triplet (formants et bruit de friction)

Figure 2

Comme l'apprentissage, en dehors de celui que nous allons réaliser pour le test, représente un volume de travail gigantesque, nous étudions la possibilité de le réaliser automatiquement. La première partie de l'apprentissage est tout à fait faisable en remplaçant le phonéticien ou le chercheur en parole par les détecteurs automatiques d'indices dont nous disposons déja pour le système Aphodex. La seconde partie, consistant à regrouper les descriptions disponibles pour en extraire des modèles représentatifs pourrait être réalisée en utilisant la classification conceptuelle [7]. Etant donné un certain nombre d'exemples et de classes, cette technique permet en effet de construire une représentation symbolique de chacune de ces classes.

5 Problèmes de reconnaissance

Le processus de décodage que nous allons maintenant décrire intervient après la segmentation et la détection des événements acoustiques. Cette étape préliminaire est destinée à construire à partir de la phrase une suite d'instances de triplets qui vont être identifiés grâce aux triplets de référence (§5.1).

Le décodage proprement dit se décompose en deux étapes successives. La première conduit à proposer, pour chaque instance de triplet, la liste des triplets de référence qui s'apparient le mieux avec le segment inconnu. La seconde étape est destinée à augmenter la cohérence de la solution globale pour la phrase en assurant que les contraintes qui existent entre les triplets de la base de connaissances qui ont été proposés comme solutions, sont aussi vérifiées par les instances de triplets de la phrase à décoder.

5.1 Décodage grossier

Ce décodage est essentiellement un décodage local et repose sur l'appariement du triplet inconnu aux triplets de la base de connaissances qui lui sont le plus proches. Précisons néanmoins que, connaissant la classe phonétique du triplet, seuls les triplets de même profil phonétique sont concernés par cet appariement. Pour que la comparaison soit significative il est nécessaire de prendre en compte la variabilité interlocuteurs et donc d'effectuer une normalisation des triplets par rapport au locuteur qui a énoncé la phrase à décoder. Cette normalisation doit être rapide (tout au plus quelques mots) et nous avons choisi la méthode de modélisation implicite de la longueur du conduit vocal [8]. Le fait de représenter la connaissance sous la forme de prototypes de sons en contexte permet d'envisager de prendre en compte d'autres déformations que peut subir un triplet, par exemple sous l'effet de l'accentuation ou encore d'un changement de vitesse d'élocution [15].

L'un des points clé d'un système à base de prototypes [9] est l'organisation des connaissances afin que la recherche des triplets les plus proches du triplet inconnu ne nécessite de parcourir qu'une petite partie de la base de connaissances. La solution couramment adoptée en image [10], comme en parole [11] est de hiérarchiser les prototypes suivant leurs indices caractéristiques. S'il est possible de structurer de cette manière une partie des triplets -c'est-à-dire ceux pour lesquels nos connaissances actuelles des phénomènes de parole permettent de définir des indices acoustiques clairs- il reste que pour une partie importante des triplets peu d'indices décisifs en reconnaissance sont connus. Il est cependant très important de structurer de cette manière une partie des connaissances, car l'expert ne place pas toutes ses connaissances au même niveau. Prenons l'exemple du triplet /syR/ ; la forte descente de F2 vers F1 qui est un indice très caractéristique, et justifiable sur le plan articulatoire,

permet à l'expert de conclure rapidement qu'il s'agit d'un triplet se terminant par /R/ et centré sur une voyelle proche de /y/. Quand ce "raccourci" de raisonnement n'est pas possible, ce qui est fréquemment le cas, il faut recourir au second mode d'organisation des connaissances qui repose cette fois sur les profils formantiques (et donc sur les références fréquentielles des triplets). Les triplets sont classés suivant les fréquences des formants F1, F2, F3 aux frontières du phone central. Le classement est effectué par classe d'amplitude relativement large pour tenir compte de l'imprécision des fréquences due aux fortes transitions formantiques. Le mode de structuration donne lieu à un mode de raisonnement tout à fait différent du précédent: les triplets sont cette fois comparés grâce à leur réalisation acoustique. Les indices acoustiques, définis par le phonéticien, ne sont utilisés dans ce cas que pour moduler le résultat de la comparaison.

Ces deux modes de raisonnement, l'un guidé par les indices acoustiques, l'autre guidé par les réalisations acoustiques ne sont, tels que nous venons de les décrire, utilisables que si une segmentation de la phrase est disponible. Si aucune segmentation préexiste au décodage, ou s'il est indispensable de la remettre en cause, la structuration suivant les indices peut permettre de construire une solution. Les indices sont d'abord systématiquement recherchés dans la partie à décoder ; les indices découverts permettent alors de segmenter la parole et de proposer une solution composée de triplets caractérisés par ces indices. Cette démarche est comparable à celle qu'utilisent P.D. Green et al. dont le système tente de reconstruire des "Acoustic Sketchs" à partir des trajectoires formantiques [12].

Le résultat final du décodage grossier est donc pour chaque segment la liste des triplets les plus proches, soit en termes d'indices acoustiques, soit en termes de réalisations acoustiques.

5.2 Cohérence globale de la solution

Le résultat du décodage précédent n'est que la juxtaposition des solutions locales pour chacun des segments de la phrase. Il est donc possible que les contraintes liant deux triplets de références (une contrainte portant sur les positions relatives des formants des deux triplets) ne se retrouvent pas sur les instances des triplets de la phrase à décoder. Eliminer de telles incohérences, permet donc de proposer une solution consistante sur toute la phrase et donc en somme de modéliser l'unicité du locuteur dans le processus de décodage.

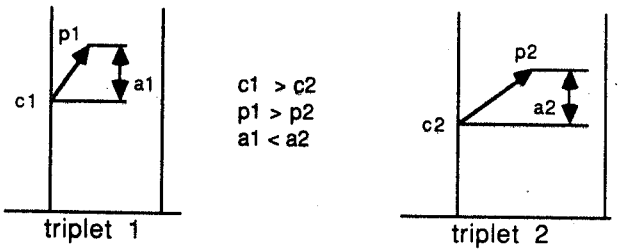
Le problème à résoudre est le suivant. Etant donnés n nœuds (les segments de parole) notés i et les ensembles d'étiquettes Li proposés pour chaque nœud (l'ensemble des étiquettes étant l'ensemble des triplets proposés pour le segment i) comment éliminer les étiquettes incompatibles avec celles des autres nœuds de la phrase? Il s'agit là du schéma classique de la relaxation discrète, connue aussi sous le nom de filtrage de Waltz [13].

Il existe deux types de consistance que l'on peut atteindre:

-consistance globale: les contraintes sont satisfaites simultanément pour l'ensemble des étiquettes de tous les nœuds.

-consistance sur les arcs: cette consistance est plus faible que la précédente et assure que les étiquettes de deux nœuds contraints l'un par l'autre sont compatibles entre elles.

C'est le second type de consistance qui nous intéresse et de nombreux algorithmes existent pour résoudre ce problème.



(seuls les frontières des triplets et F2 sont représentés)

Sous-contrainte liant les formants F2 de deux triplets

Figure 3

Les contraintes sont construites sur les profils formantiques et font intervenir trois critères (fig. 3):

-positions relatives des trajectoires formantiques aux frontières du triplet (c1 > c2),

-relation entre les pentes des transitions formantiques (p1>p2),

-relation entre les amplitudes fréquentielles des transitions (a2>a1).

Le fait que certains segments de parole soient mal placés ou bruités risque de conduire à un étiquetage global vide et il est donc nécessaire de disposer d'un algorithme de relaxation flou. Nous avons choisi GAC4 (développé dans notre équipe pour la vision par ordinateur) et sa version floue [14] qui permet de pallier le risque d'éliminer toutes les étiquettes.

Cette étape de relaxation, après le décodage grossier a deux avantages importants:

-elle permet d'approcher le comportement d'un expert qui propose une solution cohérente pour toute la phrase,

-grâce aux relations de fréquences qu'imposent les contraintes, il n'est pas nécessaire pour le décodage grossier d'effectuer une comparaison fréquentielle très précise.

6 Conclusion

De nombreuses applications de systèmes de reconnaissance de parole portent sur un vocabulaire limité ; il est donc envisageable de construire un système de décodage acoustico-phonétique à base de triplets. L'étape d'apprentissage même si elle reste importante devient alors techniquement réalisable ; ainsi Lee pour un système de dialogue limité à 997 mots a recensé 2380 triplets [16]. Une autre solution est de faire coopérer APHODEX et un système à base de triplets pour une classe phonétique particulière.

A l'heure actuelle la version du système réalisant le décodage acoustico-phonétique grossier et que nous sommes en train de tester utilise de manière simpliste la segmentation du système APHODEX car nous n'abordons dans un premier temps que des phonèmes pour lesquels la segmentation est relativement simple (les occlusives). Il faudra cependant compléter le système de décodage par des stratégies

permettant de remettre en cause une segmentation qui semble erronnée.

Le choix du triplet comme unité de décodage repose essentiellement sur sa bonne résistance aux effets contextuels. Nous estimons que les variations dues notamment à la position des triplets par rapport aux frontières de mots sont suffisamment faibles pour être négligées, du moins en première approche. Nous avons jusque là considéré uniquement l'utilisation des triplets dans le cadre du décodage acoustico-phonétique. L'utilisation de modèles représentant les unités de parole offrent d'autres avantages pour un système de reconnaissance complet. Comme il est relativement facile de déformer un triplet, cela permet d'envisager que les niveaux supérieurs émettent des hypothèses : d'accentuation d'une voyelle (se traduisant par un renforcement de la couleur d'une voyelle et une élévation de la fréquence fondamentale), de diminution de la durée d'une voyelle (réduisant l'amplitude des transitions formantiques), de nature lexicale proposant un triplet pour un segment de parole.

Remerciements

Je tiens à remercier François Lonchamp qui est à l'origine de l'idée des triplets phonétiques, Noëlle Carbonell et Dominique Fohr pour les fructueuses discussions.

Bibliographie

[1] R.A. Cole, A.I. Rudnicky et V.W. Zue "Performance of an Expert Spectrogram Reader", JASA, vol. 65, Supp. 1 p. S81. (Paper presented at the 97th meeting of the ASA)Boston, MA 1979

[2] F. Lonchamp "Reading Spectrograms: The View from the Expert" in J.P. Haton (ed.), Fundamentals in Computer Understanding : Speech and Vision, Cambridge University Press, 1987

[3] D. Fohr, N. Carbonell et J.P. Haton "Phonetic Decoding of Continuous Speech with the APHODEX Expert System", Proceedings of EuroSpeech 89, Vol. 2. p. 609-612, Paris, September 1989

[4] J.P. Tubach et L.J. Boé "Un corpus de transcriptions phonétiques: constitution et exploitation statistique" Rapport ENST - 85D001, Avril 1985.

[5] D. Fohr et Y. Laprie "Snorri : An interactive system for speech analysis" Proceedings of EuroSpeech 89, Vol. 1. p. 613-616, Paris, September 1989

[6] O. Mella "Méthodologie d'étude de la pertinenece de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur" Actes du Séminaire variabilité et spécificité du locuteur, Luminy, Juin 1989.

[7] R.S. Michalski et R.E. Stepp "Learning from observation: Conceptual Clustering", chapitre 11, "Machine Learning: An Artificial Intelligence Approach" R.S. Michalski, J.G. Carbonnell and T.M. Mitchell Ed., Springer Verlag, 1984.

[8] A. Bonneau et D. Fohr "Normalisation du locuteur par modélisation implicite de la longueur du conduit vocal" Actes du Séminaire variabilité et spécificité du locuteur, Luminy, Juin 1989.

[9] M. Minsky "A Framework for representing knowledge" extrait de "The Psychology of Computer Vision" P.H. Winston Ed. Mc Graw-Hill Book Company, 1975.

[10] C. Granger "Reconnaissance d'objets par mise en correspondance en vision par ordinateur", Thèse de doctorat, Nice, 1985.

[11] J.P. Damestoy "Réalisation d'un système à base de prototypes pour le contrôle du décodage acoustico-phonétique de la parole", Thèse de l'Université de Nancy 1, 1986.

[12] P.D. Green, M.P. Cooke, H.H. Lafferty et A.J.H. Simons "A Speech Recognition Strategy Based on Making Acoustic Evidence Knowledge Explicit" in "Recent Advances in Speech Understanding and Dialog Systems", H. Niemann, M. Lang and G.Sagerer Ed., NATO ASI Series, Springer Verlag, 1988.

[13] D. Waltz. "Understanding Line Drawings of Scenes with Shadows. in "The Psychology of Computer Vision", P.H. Winston Ed., McGraw-Hill, New York, 1975.

[14] R. Mohr et G. Masini "Good Old Discrete Relaxation", Proceedings of ECAI, pages 651-656, Munich, 1988

[15] O. Engstrand, "Articulatory correlates of stress and speaking rate in swedish VCV utterances",JASA, Vol. 85 (5), pp 1863-1875:May 1988.

[16] K.F. Lee, Hsiao-Wuen Hon, Mei-Yyuh Hwang, S. Mahajan et R. Reddy, "The SPHINX Speech Recognition System", Proc. IEEE ICASSP-89, Glasgow, Scotland, 1989

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

DECOMPOSITION TEMPORELLE : UNE TECHNIQUE CINEMATIQUE DE SEGMENTATION ET DE DECODAGE ACOUSTICO-PHONETIQUE ; EVALUATIONS.

P. Deléglise, C. Montacié, F. Bimbot.

Télécom Paris - Dépt SIG, C.N.R.S. - URA 820,
46 rue Barrault, 75634 PARIS cedex 13, FRANCE.

RESUME

Cet article présente un algorithme robuste de décomposition temporelle et décrit les grandes lignes de son application au décodage acoustico-phonétique.

L'algorithme de codage introduit par Atal [Atal 83] a été transformé dans le sens d'une robustesse accrue [Bimbot 88]. Nous en donnons ici une interprétation originale, orientée vers la cinématique, qui rejoint les idées de Caelen [Caelen 85]. L'algorithme peut être vu comme un détecteur des influences successives des phones dans le continuum acoustique, en ce sens qu'il recherche des déviations provoquées par ces phones sur la trajectoire spectrale, dans l'espace de paramétrisation.

Puis nous présentons des séries d'expériences visant à évaluer la technique de décomposition temporelle comme pré-traitement pour le décodage acoustico-phonétique. Un protocole complet de décodage a donc été mis en oeuvre afin de mesurer les performances de la technique.

On s'est intéressé aux rôles des facteurs suivants : choix de la paramétrisation spectrale, influence des cibles de transition, choix de la mesure de proximité spectrale, importance du classificateur, apport du fenêtrage adaptatif.

Ces expériences ont été conduites avec plusieurs locuteurs et plusieurs ensembles d'apprentissage. Le corpus test est constitué des noms épelés en éloçution continue.

1 PRESENTATION ET OBJECTIFS DE LA DECOMPOSITION TEMPORELLE.

La décomposition temporelle est une technique qui cherche à décrire l'évolution du contenu spectral du signal de parole par un modèle d'interpolation linéaire. Le vecteur spectral à l'instant t, noté $y(t)$, est estimé comme combinaison linéaire de n vecteurs spectraux $(g_k)_{1\leq k\leq n}$ indépendants de t. La contribution du $k^{ième}$ spectre à l'instant t est exprimée par une fonction d'interpolation $\phi_k(t)$.

Ceci se formule :

$$y^*(t) = \sum_{k=1}^{k=n} g_k \phi_k(t) \quad \text{pour } t \in [0,T],$$

où $y^*(t)$ est l'estimation de $y(t)$ par la décomposition temporelle.

Le but de l'algorithme de décomposition temporelle est de réaliser un choix optimal pour la valeur de n, pour les fonctions d'interpolation $\phi_k(t)$ et pour les vecteurs spectraux $(g_k)_{1\leq k\leq n}$ appelés cibles. Les critères de ce choix dépendent en partie de l'utilisation de la décomposition temporelle.

En segmentation, l'idéal serait que chaque $\phi_k(t)$ représente l'influence du $k^{ième}$ phone dans la trajectoire spectrale et que le vecteur g_k associé soit le représentant de ce phone dans l'espace de paramétrisation choisi. Ceci implique un synchronisme de tous les phénomènes concordant à la réalisation d'un phone, et un espace de paramétrisation où les transitions puissent se représenter comme interpolations linéaires entre les cibles.

Mais un tel synchronisme n'existe pas et un tel espace de représentation ne nous est pas accessible à partir du signal de parole. On se contente alors que la décomposition obtenue dépende essentiellement du contenu phonétique et soit indépendante de facteurs extérieurs tels que la vitesse d'élocution, la force articulatoire... En particulier, la décomposition temporelle traduisant les "non linéarités" de la trajectoire spectrales par l'ajout de cibles supplémentaires contextuelles, celles-ci devront être peu nombreuses, facilement identifiables en tant que telles, et si possible prédictibles d'après leur contexte.

Nous avons utilisé la décomposition temporelle comme pré-traitement du signal, dans le cadre d'expériences de décodage acoustico-phonétique. En effet, nous pensons que c'est là un des moyens d'évaluer quantitativement les performances de la technique en tant qu'outil de segmentation. A l'inverse, une reconnaissance au moins partielle des unités produites constitue un complément précieux de tout algorithme de segmentation. Les expériences sont conduites sur 9 locuteurs, dans un cadre de reconnaissance mono-locuteur.

Dans cet article nous donnons une interprétation de l'algorithme en termes de cinématique, puis nous décrivons les aspects généraux du processus de décodage que nous avons mis en oeuvre ; après la définition des corpus de test et d'apprentissage utilisés dans le cadre de nos expériences, on justifie la paramétrisation spectrale retenue. Ensuite, les stratégies de décodage proprement dites sont définies, notamment le choix du classificateur et de la distance spectrale. Une section est consacrée aux types d'erreurs rencontrées et aux moyens que nous avons utilisés pour y remédier (partiellement). Des variantes de la décomposition temporelle sont également comparées. Enfin, un ensemble de résultats détaillés de reconnaissance est donné comme évaluation de la décomposition temporelle.

2 DESCRIPTION DE L'ALGORITHME.

L'algorithme de décomposition temporelle se décompose en plusieurs parties.

2.1 PREMIERE PARTIE : CALCUL DES FONCTIONS D'INTERPOLATION.

Dans un premier temps on cherche à estimer le nombre n d'événements et les fonctions d'interpolations $\phi_k(t)$ associées à ceux-ci.

Dans ce but, on recherche les événements "importants", et on leur associe une fonction ϕ. Ces événements se détectent par la déviation qu'ils font subir à la trajectoire $y(t)$. Si l'on fait l'hypothèse que cette déviation se traduit par une attraction locale de la trajectoire, la projection de cette trajectoire sur l'axe de déviation maximum est une fonction de la variable t qui a l'allure suivante (fig. 1).

Pour rechercher ces déviations, on se donne une famille W de fonctions ressemblant à la forme ci-dessus (fig. 1), et l'on cherche pour chaque fonction de cette famille, la projection de la trajectoire la plus ressemblante. De cette projection, on déduit une fonction ϕ. Après cette opération, il est nécessaire d'effectuer un tri parmi les fonctions ϕ. Ce tri a pour objet de garder pour chaque événement détecté la fonction ϕ qui le représente le mieux.

Le premier traitement correspond à l'estimation d'une fonction d'interpolation à partir d'un contexte, le deuxième à la sélection des fonctions d'interpolation.

2.1.1 Estimation des fonctions d'interpolation.

Soit une fonction $w(t)$ élément de la famille W, et σ un seuil d'inertie minimal.

La fonction $w(t)$, définie sur un intervalle $[t_1,t_2]$, est généralement la fonction caractéristique d'un sous-intervalle de $[t_1,t_2]$.

σ est un seuil qui permet de ne considérer que les événements portant une valeur d'inertie supérieure à σ dans l'intervalle $[t_1,t_2]$.

L'axe de projection recherché est la droite λ, élément de $I(\sigma)$, telle que $P_\lambda(y(t))$ ressemble le plus à $w(t)$ sur l'intervalle $[t_1,t_2]$, où :

- $I(\sigma)$ désigne l'ensemble des droites, sur lesquelles la projection de la trajectoire $y(t)$, restreinte à $[t_1,t_2]$, a une inertie supérieure au seuil σ. C'est l'ensemble des droites du sous-espace engendré par les axes principaux d'inertie portant chacun une valeur d'inertie supérieure à σ.
- P_λ désigne la projection orthogonale sur l'axe λ.
- la ressemblance entre une fonction f et la fenêtre w est mesurée par la quantité suivante :

$$R(f,w) = \min_{c \in \Re} \left(\frac{\sum_{t=t_1}^{t=t_2} (f(t)-c)^2 w(t)^2}{\sum_{t=t_1}^{t=t_2} (f(t)-c)^2 \sum_{t=t_1}^{t=t_2} w(t)^2} \right)$$

La constante c est introduite pour neutraliser le problème du choix de l'origine dans la ressemblance de deux fonctions.

Le critère quadratique $\sum_{t=t_1}^{t=t_2} (f(t)-c)^2 w(t)^2$, a été préféré à $\sum_{t=t_1}^{t=t_2} (f(t)-c) w(t)$ (corrélation), afin de privilégier la coïncidence de valeurs importantes, plutôt que la ressemblance exacte des formes des deux fonctions.

L'algorithme comporte tout d'abord une décomposition en valeurs singulières pour trouver les axes principaux d'inertie et la projection de la trajectoire sur ces axes. Ensuite, l'axe de projection et la fonction ϕ associée se déterminent en recherchant le vecteur propre associé à la valeur propre maximale d'une matrice symétrique [Atal 83]. La fonction ϕ est calculée par $P_\lambda(y(t)) - c_0$, où λ est l'axe choisi et c_0 la constante minimisant l'expression de la ressemblance.

D'un point de vue pratique, cette méthode donne sensiblement les mêmes résultats que la technique décrite dans [Bimbot 88].

2.1.2 La sélection des fonctions d'interpolation.

Il s'agit de choisir une famille W de fonctions $w(t)$ et d'examiner l'ensemble des couples (fonction $w(t)$, fonction $\phi(t)$) pour répertorier les événements détectés et pour leur associer la fonction ϕ la plus adéquate. L'algorithme de décomposition temporelle a été utilisé avec deux types de familles de fonctions. Dans le premier cas les fonctions sont obtenues par translation d'une fonction de référence $w_0(t)$:

$$W_1 = \{w(t) / \exists\, p \in Z \text{ tel que } w(t) = w_0(t-p)\}$$

Dans le deuxième cas les fonctions sont obtenues par translation et dilatation de cette même fonction :

$$W_2 = \{w(t) / \exists\, p \in Z, \exists\, a \in Z \text{ tels que } w(t) = w_0\left(\frac{t-p}{a}\right)\}$$

Nous allons examiner les critères de sélections des fonctions et le choix de σ dans ces deux cas.

a. Etude de la famille W_1 et des algorithmes de sélection associés.

Différentes fonctions et différents algorithmes de sélection ont été proposés. Atal a choisi comme famille de fonctions l'ensemble des fonctions w_p définies par $w_p(t) = (t-p)^2$. L'intervalle de définition $[t_1,t_2]$ de w est fixe pour toutes les valeurs de p. Le seuil σ est un seuil absolu qui caractérise l'erreur tolérée pour la reconstruction des paramètres spectraux. Pour chaque valeur de p, Atal calcule le paramètre $v(p)$, mesure de la distance entre l'événement détecté par ϕ_p et l'instant $\tau=p$. Un événement est détecté à chaque fois que $v(p)$ passe des valeurs positives aux valeurs négatives.

Marcus [Marcus 84] choisit pour $w(t)$ des fonctions caractéristiques d'un sous-intervalle centré de $[t_1,t_2]$. Les durées sont de l'ordre de

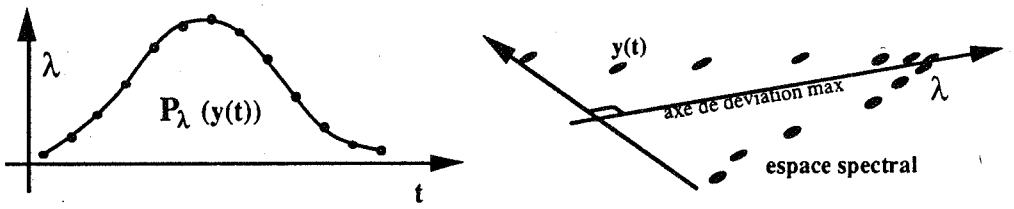


Figure 1.
à droite : trajectoire spectrale.
à gauche : évolution de la trajectoire projetée sur l'axe λ.
Le front montant de la projection traduit le rapprochement de la trajectoire vers le point d'agglomération, la partie quasi-constante correspond aux instants où les vecteurs spectraux sont quasiment identiques (stabilité spectrale), la portion descendante est associée à l'éloignement du point d'agglomération.

250 ms pour l'intervalle et de 80 ms pour le sous-intervalle. Le seuil σ est choisi comme un pourcentage fixe de l'inertie du nuage par rapport à l'origine. La sélection est obtenue en faisant une classification hiérarchique des fonctions ϕ trouvées. La coupure de cette hiérarchie à une hauteur donnée fournit le nombre d'événements et les fonctions ϕ associées à ces événements. Marcus a proposé cette technique pour corriger certains défauts de l'algorithme proposé par Atal et mis en évidence par les travaux de Zuk [Zuk 84].

b. Etude de la famille W_2 et des algorithmes de sélection associés.

Dans cette famille de fonctions, on a le plus souvent utilisé des fonctions caractéristiques d'intervalles. Une fonction de W_2 est déterminée par son intervalle de définition et le sous-intervalle sur lequel elle est non nulle. Comme il est très coûteux de calculer les fonctions ϕ pour tous les contextes, ces méthodes reposent sur le principe du fenêtrage adaptatif [Bimbot 88] [Dijk 89] .

2.1.3 Traitement annexes.

Divers traitements sont effectués sur les fonctions ϕ pour améliorer le résultat des étapes ultérieures :

- même si l'intervalle $[t_1,t_2]$ est correctement choisi, un seuillage est nécessaire pour éliminer les lobes secondaires de la fonction.
- la fonction ϕ est lissée pour faciliter le seuillage et améliorer la représentation des décompositions.

2.2 Remarque.

Toute la puissance de la décomposition temporelle réside dans la détection et la localisation des événements. Les traitements ultérieurs ne servent qu'à améliorer la cohérence de cette décomposition.

2.3 Partie 2 : calcul des cibles.

Deux méthodes peuvent être utilisées pour calculer les cibles. La première utilise la valeur exacte des fonctions ϕ : c'est une minimisation de l'erreur de reconstruction. La deuxième n'utilise que la segmentation induite par les fonction ϕ : c'est une interprétation géométrique de la trajectoire. Elle peut être utilisée à partir de toute segmentation.

2.3.1 Minimisation de l'erreur.

Il s'agit de trouver $(g_k)_{1\le k\le n}$ telle que

$\sum_{t=t_1}^{t=t_2} (y(t)-y^*(t))^2$ soit minimum, où $y^*(t)$ est l'estimation de $y(t)$ par la décomposition.

Les fonctions $(\phi_k)_{1\le k\le n}$ étant connues et peu corrélées entre elles, la solution est obtenue par la résolution d'un système linéaire bien conditionné. Une normalisation des fonctions ϕ peut être effectuée pour homogénéiser les cibles.

2.3.2 Recherche géométrique.

Cette méthode permet de s'affranchir de la forme exacte des fonctions d'interpolation pour le calcul des cibles : elle ne retient que leur structure générale.

Pour effectuer cette recherche géométrique, on fait l'hypothèse que, pour un instant donné, trois fonctions au plus interviennent dans la reconstruction. Cette technique de décomposition temporelle cherche donc à approximer la trajectoire par une suite de courbes planes. Comme la meilleure approximation plane d'une portion de courbe est sa projection sur son plan principal d'inertie, on a choisi de rechercher les cibles dans celui-ci. L'étude des fonctions ϕ permet de trouver les tronçons de trajectoire à considérer.

2.4 Reestimation des fonctions d'interpolations.

Si l'on veut obtenir les fonctions ϕ comme minimisation de l'expression $\sum_{t=t_1}^{t=t_2} (y(t)-y^*(t))^2$ (où les cibles sont connues), le problème de l'unicité des fonctions ϕ trouvées se pose.

En effet, l'ensemble des cibles associées à un segment de parole ne forme pas forcément un système de vecteurs indépendants. Pour résoudre cette difficulté, il existe des méthodes globales ou des méthodes locales.

2.4.1 Méthode globale.

Cette méthode, employée par Atal, consiste à se servir des anciennes valeurs des fonctions ϕ. On réestime donc successivement chaque fonction ϕ par recherche de celle qui minimise l'erreur de reconstruction en fixant toutes les cibles et les autres fonctions. Chaque valeur de la nouvelle estimation est obtenue algébriquement. La nouvelle fonction est lissée et tronquée pour éliminer ses lobes secondaires.

2.4.2 Les méthodes locales.

Il s'agit de calculer les fonctions d'interpolation sur des intervalles de temps où les cibles sont (linéairement) indépendantes. En particulier, si l'on prend tout ensemble de trois cibles (g_{k-1},g_k,g_{k+1}), la cible g_k est indépendante de ses deux voisines. On peut donc lui associer la fonction ϕ_k solution du problème local suivant :

$$\sum_{t=\tau_{k-1}}^{t=\tau_{k+1}} \left(y(t) - \sum_{j=k-1}^{j=k+1} g_j \phi_j(t) \right)^2 \text{ minimum,}$$

où τ_{k-1} et τ_{k+1} sont les instants associés aux cibles g_{k-1} et g_{k+1}.

Ce problème se résoud soit directement, soit par une méthode de gradient, si l'on rajoute des contraintes provenant de connaissances extérieures (modèle articulatoire et classification des phones associés aux fonctions) [Bailly 89]. La fonction ϕ_{k-1} (respectivement ϕ_{k+1}) est obtenue en minimisant le même problème pour les cibles (g_{k-2},g_{k-1},g_k) (respectivement (g_k,g_{k+1},g_{k+2})). Toutes les fonctions obtenues sont lissées et tronquées.

3 ALGORITHME GLOBAL.

Nous avons, pour la suite de nos expériences, enchaînés les traitements précédents de deux façons différentes.

La première méthode consiste en :

- Une estimation des fonctions d'interpolation,
- Un calcul de cibles,
- Deux itérations constituées chacune d'une réestimation des fonctions d'interpolations et d'un calcul de ces cibles.

Le calcul de cibles utilise la minimisation de l'erreur. La réestimation des fonctions utilise la méthode locale. Les itérations permettent de diminuer l'erreur de reconstruction et d'optimiser en partie les chevauchements des fonctions .

Cette méthode peut être utilisée avec ou sans fenêtrage adaptatif pour la première estimation des fonctions d'interpolations.

La seconde méthode fait appel à la recherche géométrique pour le calcul des cibles. Elle consiste, à partir des fonctions d'interpolation obtenues à la dernière étape de la méthode précédente, à effectuer un calcul des cibles par interprétation géométrique, puis une réestimation des fonctions par la méthode locale, et enfin une simple itération.

4 REPRESENTATION SPECTRALE POUR LA DECOMPOSITION TEMPORELLE.

Le choix de la représentation des vecteurs spectraux est un problème important dans l'utilisation de la décomposition temporelle. C'est un compromis entre les critères suivants :

- L'interpolation linéaire des paramètres doit avoir un sens physique. Il est souhaitable que la transition entre deux phones puisse être représentée comme interpolation linéaire entre les représentants des deux phones.

- La distance euclidienne sur ces paramètres doit être une mesure correcte de la différence perçue entre deux sons.

Comme premier jeu de paramètres, nous avons retenu les coefficients cepstraux à échelle Mel pondérés (M.F.C.C.). L'évaluation des systèmes de reconnaissance de mots connectés [Chollet 88] montre que la distance euclidienne sur ces coefficients présente de bonnes propriétés discriminantes [Cordeau 88].

Nous avons aussi choisi les L.A.R., car ils vérifient plusieurs propriétés intéressantes :

- La distance euclidienne sur ceux-ci correspond à la moyenne des valeurs absolues des différences entre les deux spectres [Viswanathan 75].
- L'interpolation linéaire entre deux sons proches donne des schémas de transitions formantiques réalistes, du moins pour les voyelles [Bimbot 88].
- Il est facile d'obtenir un spectre et un signal synthétique à partir des L.A.R.

Nous avons essayé d'améliorer les performances des L.A.R. en leur donnant une sensibilité plus proche de celle de l'oreille humaine. Pour effectuer cette tâche, nous avons eu recours à la technique de dilatation spectrale introduite par Oppenheim [Oppenheim 72] et appliquée au traitement de la parole par Chouzenoux [Chouzenoux 82] et Deléglise [Deléglise 83]. Cette méthode consiste, à partir d'une suite d'échantillons de parole (x_n) de spectre $\gamma_x(\omega)$, à construire une suite d'échantillons $(\tilde{x}_n)$ de spectre $\gamma_{\tilde{x}}(\tilde{\omega})$ telle que :

$$\gamma_{\tilde{x}}(\tilde{\omega}) = \gamma_x(\omega) \text{ pour } \tilde{\omega} = \theta(\omega).$$

Le modèle de prédiction linéaire calculé à partir de $(\tilde{x}_n)$ a une résolution variable $\theta(\omega)$ sur l'axe des fréquences et une fonction de poids égal à $\theta'(\omega)$ sur ce même axe. Au moyen de cette technique de dilatation spectrale, nous avons renforcé l'influence des composantes du spectre situées autour de 1500 Hz. Ce renforcement est un bon compromis entre la courbe d'isotonie et la courbe d'isosonie [Zwicker 81]. L'amélioration apportée par cette modélisation à résolution variable est évaluée dans les expériences de décodage présentées à la fin de cet article.

5 ASPECTS GENERAUX DU PROTOCOLE DE DECODAGE.

5.1 LES CORPUS.

Corpus de test : le corpus de test est constitué de 50 noms de famille épelés en élocution continue, c'est-à-dire sans pause entre lettres consécutives. Ils correspondent aux corpus REC01 et REC02 de la base de données de parole BD-SONS du Greco. C'est un corpus difficile, car les unités lexicales qui le constituent forment plusieurs séries minimales, sources de confusions : {A, E, I, O, U, é}, {B, C, D, G, P, T, V, é}, {F, L, M, N, R, S}, {A, H}, {G, J}, {I, J}, {A, K}, {K, Q}, {Q, U}. Ce corpus contient 513 phonèmes (pour environ 300 lettres). 9 locuteurs ont été testés : SL, BG, GM, RM, CF, DP, JF, FB et MJC. Parmi ceux-ci, SL, BG, GM, RM et FB sont du sexe masculin, CF, DP, JF et MJC du sexe féminin. Tous appartiennent à BD-SONS sauf FB et MJC qui ont été enregistrés à Télécom-Paris. Une évaluation comparative [Montacié 90] utilisant une carte de reconnaissance de parole à base de machines de Markov connectées a donné sur le même corpus de test (et le même corpus d'apprentissage) un score de reconnaissance inférieur de 15% en moyenne. L'enchaînement de lettres épelées provoque de nombreux effets de coarticulation, que les méthodes globales ne parviennent pas à résoudre de façon satisfaisante.

Corpus d'apprentissage : deux corpus d'apprentissage différents ont été utilisés.

- Pour les locuteurs provenant de BD-SONS, l'apprentissage a eu lieu sur 4 prononciations de l'alphabet en lettres isolées (ALA01). Ce corpus a été prononcé par chaque locuteur et représente environ 230 phonèmes (et 4 x (26 lettres + é) = 108 lettres).
- Pour FB et MJC, l'apprentissage a été conduit sur un ensemble de 270 logatomes porteurs de tous les diphones possibles dans une suite de lettres prononcées continument. Outre les contextes (non utilisés pour l'apprentissage), ce corpus contient 513 phonèmes. Ce corpus a été prononcé par chacun des 2 locuteurs.

5.2 DONNEES D'APPRENTISSAGE.

Le corpus d'apprentissage subit une décomposition temporelle. On étiquette ensuite les cibles obtenues par un symbole phonémique, pour celles associées sans ambiguïté à la réalisation d'un phonème. Un dictionnaire est ainsi constitué avec l'ensemble des cibles étiquetées. La procédure de création du dictionnaire est semi-automatique.

Le corpus des mots épelés ne fait intervenir que 24 phonèmes : a, i, e, ɛ, y, œ, u, o, p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, l, ʀ. En tout, on obtient 24 étiquettes possibles.

Chaque phonème est représenté au moins 4 fois dans le corpus d'apprentissage des lettres isolées (à cette restriction près que la décomposition temporelle peut avoir raté certains événements (≈ 10 %)).

Pour l'apprentissage par diphones, chaque phonème est représenté au moins 15 fois (environ), mais certains le sont deux fois plus, selon qu'ils apparaissent seulement en début ou en fin de certaines lettres, ou bien dans les deux positions.

6 IDENTIFICATION DES EVENEMENTS.

6.1 DISTANCE SPECTRALE.

La distance spectrale sert à sélectionner et à ordonner, pour une cible du corpus test, la liste des cibles candidates du dictionnaire.

L'essentiel des expériences a été mené avec une distance euclidienne entre les paramètres des cibles test et des cibles de référence.

Pour s'affranchir des problèmes de normalisation des fonctions d'interpolation ou de cibles non atteintes par la trajectoire lors de fortes coarticulations, on peut introduire une distance d'inférence [Deléglise 88], entre une cible g_k et un élément d du dictionnaire. Nous avons choisi comme distance d'inférence une mesure de l'aptitude de l'élément d du dictionnaire à remplacer g_k dans la décomposition temporelle $(\phi_k, g_k)_{1 \le k \le N}$ de l'élément du corpus de test. Pour celà, on considère les trois cibles g_{k-1}, g_k, g_{k+1} et les trois instants $\tau_{k-1}, \tau_k, \tau_{k+1}$ associés au milieu des fonctions d'interpolation $\phi_{k-1}, \phi_k, \phi_{k+1}$. Puis on cherche les trois fonctions $\phi^d_{k-1}, \phi^d_k, \phi^d_{k+1}$ telles que :

$$\sum_{t=\tau_{k-1}}^{t=\tau_{k+1}} \left(y(t) - \sum_{j=k-1}^{j=k+1} G_j \phi^d_j(t) \right)^2 \quad \text{minimum,}$$

où $y(t)$ est la trajectoire spectrale de l'élément du corpus traité,

et $\begin{cases} G_j = g_j \text{ pour } j \neq k, \\ G_k = d. \end{cases}$

La valeur du minimum définit la distance d'inférence entre la cible g_k et l'élément d du dictionnaire.

6.2 CIBLES DE TRANSITION.

Certaines cibles obtenues par décomposition temporelle n'ont pas d'interprétation phonétique immédiate [Bimbot 88]. En effet, on observe des cibles supplémentaires au niveau des transitions entre certains phonèmes ; celles-ci peuvent résulter d'asynchronismes articulatoires ou de non-linéarités de la trajectoire acoustique.

Ces phénomènes ont pour conséquence l'existence d'un nombre important d'insertions lors du décodage direct des événements acoustiques produits par la décomposition temporelle. Il est donc nécessaire de prendre en charge ces cibles de transition, c'est-à-dire de savoir les identifier en tant que telles dans la succession des événements.

6.3 INFORMATIONS COMPLEMENTAIRES SUR LES CIBLES.

Les cibles de transition ont un contenu spectral beaucoup moins prédictible que les cibles associées à des phones. Leur présence n'est pas systématique, et leurs caractéristiques spectrales apparaissent très variables.

L'adjonction d'information complémentaire sur les cibles du dictionnaire et les cibles de test permet de mieux discriminer les événements phonétiques des événements contextuels (transitions) lors de la reconnaissance. Ces paramètres sont les suivants :

- un paramètre d'énergie, qui caractérise l'énergie du signal au voisinage du centre de la fonction d'interpolation associée à l'événement,
- un paramètre de stabilité spectrale, calculé à partir de l'inertie d'une portion de la trajectoire par rapport à son centre de gravité,
- un paramètre de durée, qui représente la durée pendant laquelle la fonction d'interpolation associée à un événement est supérieure à 50 % de sa valeur maximale.

Chaque cible spectrale à reconnaître se voit donc adjoindre ces 3 paramètres. De même, chaque symbole phonétique (et non chaque cible) du dictionnaire est muni de caractéristiques complémentaires typiques (valeur moyenne et écart-type).

Au moment de la classification, on s'assure que les étiquettes phonétiques produites ne présentent pas d'incohérence avec l'un des paramètres complémentaires. Dans le cas contraire, on substitue une étiquette "+" (transition) à l'étiquette d'origine. Les cibles associées à une zone de silence reçoivent une étiquette "-".

Les cibles consécutives ayant la même étiquette phonétique sont regroupées, tant que la durée cumulée n'excède pas 200 ms.

6.4 CHOIX DU CLASSIFICATEUR.

On a utilisé la technique de classification non-paramétrique des k plus proches voisins (k-PPV) [Cover 67], qui consiste à retenir comme liste d'étiquettes candidates celles dont le représentant spectral est parmi les k plus proches de l'événement à identifier, au sens de la distance spectrale. La classification finale est obtenue en effectuant à effectuer un vote majoritaire sur l'ensemble de ces étiquettes. Quand l'ensemble d'apprentissage est suffisamment grand, cette technique tend vers l'estimation Bayésienne. Son inconvénient est l'important volume de calcul qu'elle entraine, par rapport aux estimations paramétriques (estimateur gaussien, par exemple).

Dans [Niles 89], il est conseillé de prendre k égal à $\sqrt{N}$, où N est le nombre moyen d'éléments par classe. Nous avons choisi k = 1 pour les apprentissages par lettres isolées, et k = 3 pour les apprentissages par diphones.

La figure 2 donne un exemple d'étiquetage du segment "ROUSSEL" [ʁoyɛsɛsœɛl] (apprentissage par lettres isolées, 1 PPV).

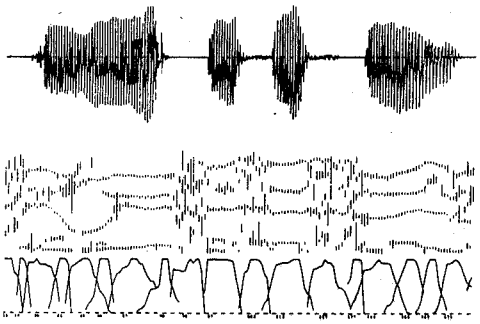


Figure 2 : Exemple de décodage.
ROUSSEL ([ʁoyɛsɛsœɛl]).

De haut en bas : signal, spectre LPC schématique, fonctions d'interpolation et chaîne phonétique trouvée.

6.5 EVALUATION DU DECODAGE.

La chaîne phonétique proposée par l'algorithme pour chaque élément du corpus est comparée à la transcription phonétique normative associée à cet élément. L'ensemble des comparaisons est effectué par un logiciel utilisant la distance d'édition de Wagner et Fischer [Wagner 74].

7 VARIANTES DE LA DECOMPOSITION TEMPORELLE.

Différentes variantes de la décomposition temporelle ont été testées dans le cadre du protocole décrit dans le paragraphe précédent. On donne ici un aperçu des différents résultats.

7.1 SENSIBILITE AU SEUIL D'INERTIE.

La pratique montre que les résultats sont insensibles à une variation de 20 % du seuil d'inertie autour de la valeur finalement retenue. Un seuil trop faible entraîne surtout un nombre accru d'insertions. Un seuil trop élevé augmente à la fois le nombre de substitutions et celui des omissions.

7.2 PARAMETRISATION SPECTRALE.

Bien que les LAR (Log Area Ratios) aient un comportement plus réaliste par interpolations linéaires et permettent une re-synthèse du signal [Bimbot 88], les WMFCC (Weighted Mel Frequency Cepstral Coefficients) [Tohkura 86] font preuve d'une nette supériorité en reconnaissance globale de mots isolés [Cordeau 88]. L'amélioration est d'environ 14 % (pour 8 WMFCC vs 12 LAR, et une métrique euclidienne). L'introduction de non-linéarités dans l'échelle fréquentielle pour le calcul des LAR (LAR - non linéaires) permet à ces derniers de combler la moitié de leur retard.

Ces résultats nous ont amenés à choisir les paramètres WMFCC pour la suite de nos expériences.

7.3 CHOIX DE LA DISTANCE.

Les expériences réalisées montrent que la distance d'inférence n'apporte globalement pas d'amélioration aux scores de reconnaissance, par rapport à la distance euclidienne. Néanmoins, un examen détaillé des résultats montre que la distance d'inférence a tendance à améliorer les scores de reconnaissance des énoncés dont les éléments sont bien reconnus avec la distance euclidienne. A l'inverse, la distance d'inférence dégrade les performances sur les énoncés déjà difficiles à identifier par des méthodes plus classiques ; en d'autres termes, la distance d'inférence apporte un plus sur les bonnes décompositions temporelles, mais manque encore de robustesse.

7.4 FENETRAGE FIXE VS FENETRAGE ADAPTATIF.

Les résultats montrent que la technique de fenêtrage adaptatif réduit d'environ 2/3 la sur-segmentation du corpus testé. Ceci se traduit par un taux d'insertion 2 fois moindre, à taux de reconnaissance égal.

8 RESULTATS.

Le décodage acoustico-phonétique par décomposition temporelle a été évalué sur 9 locuteurs, dans un cadre mono-locuteur. Le protocole retenu s'articule sur les points suivants :

- paramétrisation MFCC du signal.
- apprentissage sur lettres isolées ou diphones.
- décomposition temporelle par fenêtrage adaptatif.
- calcul des cibles par minimisation de l'erreur.
- distance spectrale euclidienne.
- utilisation d'informations complémentaires sur les cibles.
- décision de type k-PPV sur les cibles, pour la classification.
- évaluation du décodage par l'algorithme de Wagner et Fisher.

La Table 1 donne les scores de reconnaissance, de substitutions, d'omissions et d'insertions pour tous les locuteurs, ainsi que des scores moyens (hommes, femmes, ensemble). Ont été comptées comme confusions, toutes les différences entre les étiquettes phonétiques attendues et produites, sans aucun regroupement des phonèmes en macro-classes. L'intervalle de confiance de ces valeurs est de ± 3 % [Montacié 87].

Notons également que les différents paramètres de l'algorithme de décomposition temporelle et du décodage ont été mis au point sur le

locuteur SL et qu'ils n'ont été retouchés en aucune façon pour les autres locuteurs.

Ces scores montrent globalement de meilleurs résultats sur les voix masculines que sur les voix féminines, ce qui paraît être plutôt un artéfact de la paramétrisation spectrale qu'un biais introduit par le décomposition temporelle proprement dite.

Les confusions les plus fréquentes sont : m / n, œ / l, e / ɛ, b / d, f / s,... [l] et [ʁ] sont plutôt bien reconnus au prix d'un taux d'insertion sensiblement plus élevé de ces phonèmes. Les explosions de [p] sont extrèmement mal détectées par la décomposition temporelle, ce qui a pour conséquence un score très bas pour ce phonème.

7 CONCLUSIONS.

La décomposition temporelle code de manière efficace les évolutions du signal de la parole. Sa présentation en termes de modélisation de trajectoires acoustiques complète la description algébrique classique de l'algorithme.

Cette technique de décomposition temporelle apparaît donc comme un complément indispensable des techniques de segmentation classiques, en tant que nouvel outil de représentation de la parole. En effet, elle se montre particulièrement robuste pour segmenter les groupes vocaliques et liquides ; les portions de signal qui varient lentement sont décrites d'une façon globale ce qui permet d'éviter les décisions de type tout-ou-rien des algorithmes de segmentation classiques. En revanche, le modèle d'interpolation linéaire s'avère moins opérant pour représenter les transitions brutales, notamment aux endroits de changement de mode de production au niveau de la source. A ces endroits, les cibles perdent de leur pertinence. Ces défauts sont à imputer, au moins en partie, aux insuffisances des paramétrisations.

Les expériences de décodage acoustico-phonétique ont montré la pertinence des cibles comme représentants des événements acoustiques, et font de la décomposition temporelle une alternative possible à l'étiquetage centi-seconde.

Les faiblesses de la technique à modéliser les transitions brutales vont faire l'objet de recherches supplémentaires pour adjoindre un modèle de rupture à la décomposition temporelle [André-Obrecht 88], [Basseville 89].

9 BIBLIOGRAPHIE.

[André-Obrecht 88] R. ANDRE-OBRECHT : **A new statistical approach for the automatic segmentation of continuous speech.** IEEE Trans ASSP. Vol 36, pp 29-40. 1988.

[Atal 83] B.S. ATAL : **Efficient coding of LPC parameters by temporal decomposition.** Proc. ICASSP-83. 2.6. pp. 81

[Bailly 89] G. BAILLY, P.F. MARTEAU, C. ABRY : **A new algorithm for temporal decompostion of speech. Application to a numerical model of coarticulation.** Proc ICASSP-89. pp. 508-511.

[Basseville 89] M.BASSEVILLE, A. BENVENISTE : **Detection of abrupt changes in signals and dynamical systems.** Springer-Verlag. 1989.

[Bimbot 88] F. BIMBOT, G. CHOLLET, P. DELEGLISE, C. MONTACIE : **Temporal decomposition and acoustic-phonetic decoding of speech.** Proc ICASSP-88. pp. 425-428.

[Bimbot 88b] F. BIMBOT : **Synthèse de la parole : des segments aux règles, avec utilisation de la décomposition temporelle.** Thèse E.N.S.T. Décembre 1988.

[Caelen 85] J. CAELEN : **Introduction à une segmentation cinématique.** JEP 85. pp. 129-131.

[Cordeau 88] J.P. CORDEAU : **Analyse de quelques métriques pour la reconnaissance automatique des mots connectés.** Technical report (88D017) ENST Paris, France.

Cover 67] T. COVER, P.E. HART : **Nearest neighbour pattern classification** IEEE Trans. Inf. Theory. 1967.

[Deléglise 83] P. DELEGLISE : **Paramétrisation et détermination des noyaux stationnaires en vue de la rconnaissance de la parole continue.** thèse de 3ème cycle. Paris VI 1983.

[Deléglise 88] P. DELEGLISE, F. BIMBOT, C. MONTACIE, G. CHOLLET : **Temporal decomposition and acoustico-phonetic decoding for the automatic recognition of continuous speech.** ICPR 88. Rome, pp 839-841.

[Dijk 89] A.M.L. Van DIJK-KAPPERS, S.M. MARCUS : **Temporal decomposition of speech.** Speech communication. Vol. 8, No 2. 1989.

[Marcus 84] S.M. MARCUS, R.A.J.M. Van LIESHOUT : **Temporal decomposition of speech.** IPO annual progress report 19. pp. 25-31. 1984.

[Montacié 87] C.MONTACIE, G. CHOLLET : **Systémes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole.** JEP 87. pp. 323-326. Hammamet.

[Montacié 90] C.MONTACIE : Thèse de doctorat, ENST. A paraître.

[Niles 89] L.NILES, H.F.SILVERMAN : **How Limited Training Data can allow a Neural Network to Out-Perform an 'Optimal' Statistical Classifier.** Proc ICASSP. pp. 17-20. 1989.

[Oppenheim 72] A.V. OPPENHEIM, D.H. JOHNSON : **Discrete representation of signal.** IEEE-Trans ASSP. Vol. 60, pp. 681-691. June 1972.

Tohkura 86] Y. TOHKURA : **A weighted cepstral measure for speech recognition** Proc ICASSP-86. pp 761-764.

[Wagner 74] R.A WAGNER, M.J. FISHER : **The string to string correction problem.** JACM. 1974.

[Wiswanathan 75] R. WISWANATHAN, J. MAKHOUL : **Quantization properties of transmission parameters in linear predictive systems.** IEEE Trans ASSP (1975). Vol 23, . pp 309-321.

[Zwicker 81] E. ZWICKER, R. FELDTKELLER : **Psychoacoustique, l'oreille récepteur d'information.** Masson (1981)

[Zuk 84] E.A. ZUK : **An investigation of temporal decomposition of speech parameters for automatic segmentation of speech.** IPO annual progress report. No 459. 1984.

Locuteur	Identifiés	Substitués	Omis	Insérés
SL*(M)	82 %	17 %	1 %	15 %
BG(1)(M)	78 %	14 %	8 %	11 %
GM(1)(M)	72 %	22 %	6 %	11 %
RM(1)(M)	74 %	18 %	8 %	11 %
CF(1)(F)	69 %	25 %	6 %	11 %
DP(1)(F)	56 %	30 %	14 %	7 %
JF(1)(F)	65 %	28 %	7 %	15 %
FB(2)(M)	73 %	19 %	8 %	16 %
MJC(2)(F)	64 %	29 %	7 %	27 %
Moyenne H	76 %	18 %	6 %	13 %
Moyenne F	64 %	28 %	8 %	16 %
Ensemble	70 %	23 %	7 %	14 %

* -> locuteur sur lequel a été mis au point l'ensemble du processus de décodage.
(1) -> apprentissage par lettres isolées (* également).
(2) -> apprentissage par diphones.
(M) -> locuteur de sexe masculin.
(F) -> locuteur de sexe féminin.

Tableau 1 : reconnaissance phonétique, 1er choix, 513 phonèmes, REC01 et 02.

INDEX DES AUTEURS