XVII[e] J.E.P.

17[e] JOURNEES D'ETUDE SUR LA PAROLE

NANCY

20-22 septembre 1988

S.F.A. Société Française d'Acoustique

G.C.P. Groupe Communication Parlée

Dédicace à
Mme le professeure de la
Sorbonne Jacqueline Vaissière Insedet !

C'est pour à moi
(et à Janine)
Comment vas tu
A bientôt
Dominique

XVII^e J.E.P.

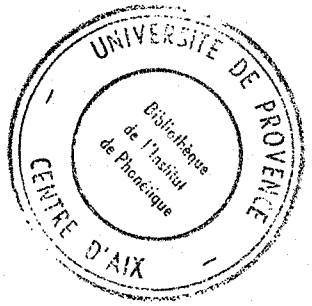
17^e JOURNEES D'ETUDE SUR LA PAROLE

NANCY, 20-22 septembre 1988

S.F.A.
Société Française d'Acoustique

G.C.P.
Groupe Communication Parlée

Société Française d'Acoustique S.F.A.
Groupe Communication Parlée G.C.P.



17e J.E.P.

Les 17e Journées d'Etude sur la parole du Groupe Communication Parlée de la Société Française d'Acoustique se sont tenues sur le campus Victor Grignard de l'Université de Nancy I (Nancy, France) du 20 au 22 septembre 1988

Elles ont été conjointement organisées par

Le Centre de Recherche en Informatique de Nancy
CRIN/CNRS

L'Institut de Phonétique de Nancy
Université de Nancy II

Comité de programme :

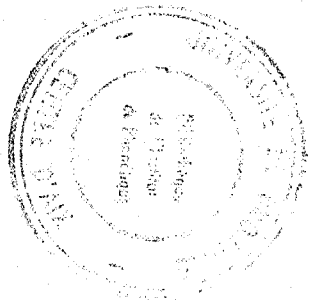
Président : J.-M. PIERREL, CRIN/INRIA-Lorraine
Vice-présidente : J. VAISSIERE, CNET-Lannion
Membres :
C. BENOIT, CNET-Lannion
L.-J. BOE, Institut phonétique Grenoble
A. BOTHOREL, Inst.phonétique Strasbourg
J. CAELEN, I. C. P. -Grenoble
J.-P. HATON, INRIA-Lorraine / CRIN
J.-M. HOMBERT, Université de Lyon
J.-P. LEFEVRE, OROS Grenoble
F. LONCHAMP, Institut phonétique Nancy
J.-S. LIENARD, LIMSI-Orsay
H. MELONI, GIA Aix-Marseille
G. PERENNOU, CERFIA Toulouse
M. ROSSI, Institut phonétique Aix
C. SORIN, CNET-Lannion
J.-P. TUBACH, ENST Paris

Comité d'organisation :

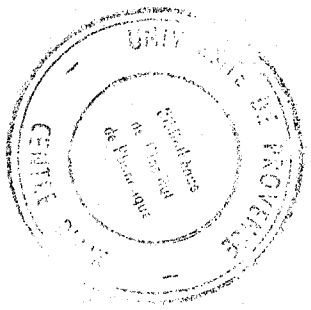
Président : J.-M. PIERREL
Membres :
C. BOURJOT
A. BOYER
N. CARBONELL
Y. LAPRIE
F. LONCHAMP
B. MANGEOL
L. ROMARY
A. ROUSSANALY

Secrétariat S.F.A. : CNET Lannion A, BP 40, 22301 LANNION
Secrétariat GCP : ENST Dept. SYC, 46 rue Barrault, 775634 PARIS Cedex 13

Sommaire



Décodage acoustico-phonétique
et
bases de données



Préliminaires méthodologiques pour une base de données acoustique phonétique

Guy Pérennou & Nadine Vigouroux

Laboratoire CERFIA, département CAPT, UA-CNRS N°824
118, Route de Narbonne - 31062 Toulouse.

ABSTRACT

— At the CERFIA laboratory we have been developing since 1984 an acoustic and phonetic data base, as well as an interface for labelling. In this paper we expose our methodological analysis and some specifications relative to problems with the labelling process that have been already remarked. We begin with two remarks concerning labelling problems: 1) current linguistic theories have very little chance of being in agreement on all the various points touching upon the manner of interpreting the speech signal, 2) the nature of the information that is researched can very according to the goals to be attained (i.e. research in phonetics, conception vocal interfaces, etc).

We therefore adopt a guideline which consists in opposing particular points of view of different users with a set of points of view as neutral as possible.

In this framework, we indicate how, in the particular point of view of the CERFIA (but open to allcomers) our acoustic and phonetic data base must be linked to a lexical data base and to an expert system in phonology (developed at the CERFIA within the framework of the GRECO "Communication Parlée").

1. INTRODUCTION

Le traitement automatique de la parole et les recherches fondamentales en phonétique font appel à d'importants corpus de parole numérisée. Chaque équipe travaillant dans le domaine peut être amenée à en créer, mais dès qu'il s'agit de faciliter les échanges, de comparer des résultats de recherche ou d'évaluer les performances de cartes vocales, des corpus communs deviennent nécessaires. Ils se justifient également par des considérations d'économie car des corpus numérisés, suffisamment étendus pour permettre d'asseoir des résultats sur des statistiques suffisantes, peuvent demander des efforts et des investissements importants qu'il vaut mieux envisager à l'échelle d'une communauté.

C'est dans cet esprit qu'est développée au sein du GRECO de la Communication Parlée une *base de données des sons du français BDSON* [Carré,84] —d'autres projets du même type existent aux Etats-Unis [Baker,83] au Japon [Itahashi,86], [Takeda,87] et en Europe [Carlson,86], ...—.

A échelle plus réduite, les équipes de recherche en disposent généralement sur leurs *postes de travail parole*. L'organisation de ces bases, la nature et la structure des informations acoustiques et phonétiques qui s'y trouvent, conditionnent pour une grande part les possibilités réelles d'échanges dont il vient d'être question. Il faut donc y porter une attention particulière si l'on veut éviter en ce domaine une Babel informatique.

Une bonne exploitation des corpus est liée aux possibilités d'accès aux segments de signal vocal, réalisations d'énoncés ou d'unités linguistiques données. Par exemple, s'il s'agit de décodage acoustico-phonétique, on voudra extraire toutes les occurrences d'un allophone apparaissant dans un contexte donné; dans d'autres applications les segments extraits devront correspondre à des mots —cas où l'on procède à l'évaluation de cartes de reconnaissance de mots isolés—, des groupes de mots, des phrases etc. C'est pourquoi BDSON prévoit des corpus étiquetés, autrement dit, que le signal numérisé soit complété par des tables assignant des unités linguistiques et/ou phonétiques à des instants, ou des intervalles, du signal numérisé.

Or l'étiquetage se révèle être une tâche délicate et coûteuse dès que l'on s'intéresse aux unités inférieures au mot: délicate, si l'on considère les désaccords entre phonéticiens quant au statut des unités phonologiques et phonétiques, à leur polymorphisme et aux difficultés qu'il y a parfois à les localiser dans le signal; coûteuse car ce type d'étiquetage demande beaucoup de temps à un spécialiste muni d'un poste de travail convenablement équipé.

Par exemple, d'après diverses expériences, on peut estimer que pour une seconde de signal l'étiquetage en phonèmes demande 3 heures, soit un rapport de $1/10800^e$ entre la durée de l'enregistrement et le temps nécessaire à son étiquetage.

En conséquence, le choix de la finesse d'étiquetage d'un corpus, et même le choix des corpus, ne peuvent être faits sans que les critères économiques soient examinés; une base de données conçue pour gérer des corpus de parole devra donc intégrer des informations d'étiquetage à *finesse variable* —en paraphrasant ici l'*analyse à profondeur variable* des spécialistes de la communication en langage naturel—.

Au laboratoire CERFIA nous avons développé depuis 1984 une *base de données acoustique phonétique (BDAP)* et un poste de travail parole permettant, entre autre, l'étiquetage.

Nous exposerons ici notre analyse méthodologique et quelques spécifications relatives aux problèmes d'étiquetage soulevés précédemment partant du constat 1) que les théories linguistiques en présence ont peu de chance de s'accorder sur divers points touchant à la manière d'interpréter le signal de parole, 2) que selon les buts poursuivis —recherches en phonétique, conception d'interfaces vocales, etc— on peut y rechercher des informations de nature très différentes.

Par ailleurs, il nous semble que diversité est plutôt une richesse qu'il n'y a pas lieu d'aborder avec des idées réductrices. Nous en déduisons un principe directeur, conforme à la méthodologie des bases de données, à savoir qu'une BDAP doit opposer:

- un *noyau aussi neutre que possible* par rapport aux particularités des démarches linguistiques,
- un ensemble de *vues particulières* .

Ces dernières sont à la charge des utilisateurs de cette base. Le gestionnaire de la base a la responsabilité du noyau dont la structuration doit permettre aux vues particulières de se développer sans difficultés.

Ce partage n'est pas toujours clair; il faut bien, par exemple, que des unités d'étiquetage soient adoptées, ce qui suppose un minimum de consensus linguistique et ne va jamais sans quelques discussions. Mais en ayant présent à l'esprit les principes précédents on évitera bien des faux problèmes. Une BDAP ne doit pas être le lieu où s'affrontent les idées concurrentes, parfois au sein d'un même laboratoire —les congrès et les revues sont faits pour cela— mais le réservoir où chacun peut puiser des matériaux nécessaires à sa recherche à travers les vues particulières qu'il peut contribuer à élaborer.

Dans cet esprit, nous indiquons comment, en tant que vue particulière du CERFIA, notre BDAP doit être reliée à la base de données lexicales BDLEX et au système expert en phonologie GEPH (développés au CERFIA dans le cadre du GRECO CP).

2. LES INFORMATIONS D'UNE BDAP

Nous allons passer en revue les trois types d'information d'une BDAP, à savoir: les enregistrements de signal vocal, les descripteurs de corpus et les tables de transcription et d'étiquetage.

2.1. Le signal vocal

Rappelons brièvement les formes sous lesquelles on peut accéder au signal —on pourra se reporter à [Vigouroux,87] pour plus de détails—.

• Le signal numérisé SN: Relation SN(t*,s) où t est le temps, clé de la relation (notation avec étoile).

• Les spectres et indices par *centisecondes* ou SIC:
Relation SIC(n*,e_1,...,e_{24},i_1,...,i_7) où n est le numéro du centiseconde, e_1,...,e_{24} les valeurs spectrales par rapport à un banc de filtres i_1,...,i_7 des indices déduits des valeurs précédentes.

• Les spectres et indices par *segments homogènes* ou SISH
Relation: SISH(m*,e_1,...,e_{24},i_1,...,i_7,j_1,...,j_K), où m est le n° du premier centiseconde, i_1,...,i_7 sont les moyennes des indices de centisecondes, j_1 la durée et j_2, ..., j_K les autres indices propres au segment.

2.2. Les informations globales sur les corpus

Chaque corpus considéré comme un tout renvoie à un descripteur dans un fichier de textes. Il se décompose en *unités d'énonciation (UE)* successives constituant pour le locuteur autant de tâches indépendantes, possédant chacune un descripteur. Une UE sera selon le cas un texte, une phrase, un mot isolé, un logatome, etc.

2.2.1. Descripteurs de corpus

Ils ont pour fonction d'identifer et de donner les caractéristiques générales des locuteurs (nom, âge, caractéristiques linguistiques, ...) et des corpus (nom et type de corpus, nombre d'entités parole, types d'entités parole: phrase, mot isolé, logatome...). Une BDAP doit comporter un fichier regroupant ces descripteurs et permettant diverses recherches sur le contenu de la base.

2.2.2. Descripteurs d'UE

Forme orthographique — Le plus simple est d'associer à une UE sa forme orthographique conventionnelle. Il ne faut cependant pas exclure la possibilité d'y inclure des informations prosodiques et/ou phonétiques; un corpus destiné à des statistiques sur le «e caduc» pourrait, par exemple, avoir un descripteur orthographique où chute et maintien de «e» sont indiqués —exemple: «la p'tit' fenêtr'»—;
dans le même esprit, pourraient-être indiquées les liaisons —exemple: «extrêm'ment_habil'»—, les accents —exemple: «c'est TERRIbl'»—, etc.

Une telle démarche présenterait un danger de confusion entre les tâches de gestion de la base de données et celles de la transcription qui requièrent le jugement d'un expert phonéticien. Aussi, si elle était jugée nécessaire, elle demanderait qu'il y ait deux champs orthographiques. L'un, en orthographe conventionnelle, pourrait alors être placé sous la responsabilité du gestionnaire de la base et y être enregistré en même temps que l'enregistrement correspondant; l'autre, sous la responsabilité d'un expert phonéticien, pourra être créé à sa discrétion et inclure des indications prosodiques et/ou phonétiques.
N.B. — Lorsque le corpus pris en charge contient des logatomes, le descripteur est l'énoncé dans la forme communiquée au locuteur.

Représentation phonologique — A ce stade, un énoncé est représenté au moyen des trois catégories d'éléments phonologiques —nous nous inspirons ici de [Dell,86a]—:

• Les frontières
- de morphèmes +
- de mots #
- de syntagmes ##
- de groupes phonologiques §

• Les phonèmes pertinents que l'on peut dégager par l'analyse en traits phonologiques non redondants au plan lexical. Chaque phonème est désigné par un symbole de l'API.

• Les diacritiques phonologiques comprenant essentiellement:
- la marque " de latence destinée aux consonnes qui ne se prononcent qu'en contexte de liaison ou de soudure morphologique,
- la marque ' de consonne semi-fixe (ex.: s final de «six, *adj card* » transcrit /si s'/),
- la marque ˙ de consonne facultative en finale (ex.: /ananas˙/).

Exemple:«prendre son temps» ---> § prɑ̃drə # sɔ̃n" # tɑ̃ § (1)

Comme nous l'avons déjà indiqué il n'y a pas lieu ici d'exacerber les controverses opposant les différentes théories linguistiques. Ainsi par rapport à celles qui soutiennent qu'en français les représentations phonologiques ne doivent pas contenir de voyelles nasales, voir par exemple [Schane,68], précisons que la représentation choisie se place après qu'un premier cycle de règles abstraites préflexionnelles ait pris effet; ainsi la représentation |prɛn+d+r+ə| (ib.pp. 115,119) a pu se transformer en /prɑ̃drə/, le groupe |ɛn| se transformant en /ɑ̃/ (ib. p.50) dans un tel contexte.

Ce qu'il faut bien avoir présent à l'esprit, c'est que la représentation phonologique est transformée successivement par les règles phonologiques partiellement ordonnées jusqu'à obtenir la forme phonétique. La forme sous-jacente retenue se situe à un stade de transformation qui nous semble un compromis acceptable entre l'abstraction, de toute manière nécessaire, et le naturel. Malgré tout, elle est marquée par le modèle linguistique adopté, de sorte qu'il faudrait éviter d'y référer dans le noyau de la base.

Représentations phonotypiques — Ce sont celles que l'on obtient en débarassant les énoncés du type précédent des éléments non pertinents pour la prononciation: symboles de frontières ou phonèmes normalement appelés à être élidés etc. L'énoncé (1) pourrait alors devenir /prɑ̃drəsɔ̃tɑ̃/. Ce niveau de représentation reste abstrait en ce sens qu'il ne requiert pas d'information sur les particularités de la prononciation effective de l'énoncé. Il exige des conventions précises relatives aux règles phonologiques que l'on doit faire intervenir pour dériver les représentations phonotypiques des représentations phonologiques. A titre d'exemple on pourra conserver les /ə/ sauf devant voyelle à l'intérieur d'un mot ou en finale après voyelle, appliquer les règles de liaison obligatoires, etc. —on pourra se reporter à [Autesserre,88] pour une discussion sur ces questions—.

«les petites amies» —> #lEz"#pətit"+ə+z"#ami+ə+z"##
—> /lEpətitəzami/
«perdre son temps »—> #pɛrdrə#sɔ̃#tɑ̃## (2)
—> /pɛrdrəsɔ̃tɑ̃/ (2a)

Lorsque les applications visées ne le nécessitent pas, il est préférable d'utiliser le symbole d'archiphonème pour les voyelles à double timbre. C'est ici le cas du E dans «les». Il est cependant possible de transcrire le timbre précis dans les cas où il existe une règle systématique pour le déterminer. On pourra ainsi prendre /mɛrkrədI/ pour représentation phonotypique de «mercredi» compte tenu du fait qu'en syllabe fermée on a toujours /ɛ/.

Représentations phonétiques — Lorsque l'on précise la réalisation des traits sur des échelles non binaires (ex.: dévoisement partiel) et/ou que l'on assigne des valeurs aux traits redondants, les représentations précédentes deviennent à proprement parlé des représentations phonétiques. Les allophones et les diacritiques nécessaires pour ces représentations peuvent être transcrits au moyen de symboles de l'API, ce qui dans l'exemple précédent donnerait diverses formes comme (l'indice 0 signifiant dévoisement partiel):

[pɛʁdRəsɔ̃tɑ̃] [pɛʁ$_0$tsɔ̃tɑ̃] [pɛʁd$_0$sɔ̃tɑ̃] ... (2b)

si l'on ne prend en compte que certaines variations allophoniques du /r/ et le dévoisement possible de /d/.

L'assimilation de labialité des consonnes conduit de même à introduire de nouvelles transcriptions: ici il y aurait lieu de noter l'arrondissement possible du [s] --noté par l'indice ω-- au sein de la syllabe de noyau [ɔ̃] [Morin,71]; [Plénat,86] ce qui conduit à

[pɛʁdRəS$_ω$ɔ̃tɑ̃] [pɛʁ$_0$tS$_ω$ɔ̃tɑ̃]... (2c)

Enfin, si l'on envisage les diverses possibilités de réalisation de la nasale et des occlusives il existe encore bien d'autres variantes. Il est donc clair que différents types de représentations phonétiques doivent être envisagés en fonction des précisions phonétiques que l'on veut décrire.

Revenons aux questions de transcription qui ont pour objectifs de donner les représentations des énoncés.

Transcriptions phonologiques et phonotypiques— Les transcriptions phonologiques résultent de la structure syntaxique de l'énoncé et des formes phonologiques associées aux entrées lexicales référencées. Par conséquent, elles sont redondantes dans un corpus étiqueté possédant des descripteurs orthographiques. Comme nous l'avons déjà indiqué, il est préférable de ne pas l'intégrer au noyau de la base. Un utilisateur pourra toujours, pour sa vue particulière de la base, construire les transcriptions phonologiques conformes à ses options linguistiques.

Cependant le maintien d'une transcription phonologique se révèle souvent utile pour clarifer les transcriptions. En effet, l'absence d'une telle transcription conduit souvent à insérer dans les autres transcriptions des informations de type phonologique.

Les transcriptions phonotypiques doivent assigner aux énoncés leur représentation phonotypique. Elles sont plus simples que les précédentes. Comme pour les transcriptions phonologiques, ni l'écoute des enregistrements, ni la connaissance du locuteur ne sont nécessaires pour les obtenir.

Transcriptions phonétiques — Elles doivent préciser la prononciation effective —c.-à-d. donner les représentations phonétiques— en jugeant à l'oreille les énoncés produits par le locuteur. Plusieurs niveaux de transcription peuvent donc être envisagés selon le degré de précision phonétique souhaitée.

La trancription phonétique large est la moins précise en ce sens que l'on ne spécifie ni les variantes allophoniques, ni les diacritiques phonétiques. Beaucoup de personnes, après un minimum d'apprentissage, sont aptes à une telle tâche —ceci est confirmé par les expérience conduites dans divers projets, voir par exemple [Takeda,87]—.

Il existe cependant quelques points délicats, particulièrement en ce qui concerne les assimilations, les paires de phonèmes en opposition faible et les segments épenthétiques.

Considérons par exemple la nasalité progressive qui peut rendre la prononciation de

«on t' parl' pas» (réalisation [ɔ̃npArlpA]) (3)

homophone de

«on n' parl' pas» (3a)

—exemple emprunté à [Dell,86b], voir aussi [Morin,71], [Malécot,72], [Walter,76]—. Un transcripteur non averti peut ne pas avoir conscience du [n] lors de la transcription du premier énoncé alors que ce [n] lui semblera naturel dans le deuxième énoncé.

Le problème se complique encore du fait que l'assimilation peut être partielle —les deux énoncés donnés en exemple ne sont plus homophones— de sorte que le /t/ se réalise par l'allophone [~t] (c.-à-d.: t nasalisé) qui ne doit pas figurer à ce stade de transcription; il convient alors de transcrire simplement [t]. De tels cas se rencontrent fréquemment dans les corpus de nombres —par exemple: «vingt-six», «trente-quatre», etc—.

On notera, le problème étant convenablement posé au départ, que des erreurs d'appréciation à ce stade ne sont pas toujours très importantes: après tout, si un transcripteur a noté [t] dans le premier énoncé alors que la transcription [n] était justifiée, l'erreur pourra être corrigée lorsque l'expert-phonéticien effectuera *une transcription phonétique plus précise* (voir ci-après).

Une autre question est celle des paires de phonèmes dont le contraste est souvent atténué, voire annulé, surtout en syllabe non accentuée; ce sont (avec leur notation dans ce texte):

Œ= (ø œ)	O = (o ɔ)	A = (a ɑ)	E= (e ɛ)
IN= (ɛ̃ œ̃)	ñ = (n ɲ)	Ñ = (ɲ ŋ)	(4)

Même si la distinction est possible dans certain cas il paraît plus radical de ne pas demander un tel jugement lors de l'étiquetage phonétique large afin que celui-ci puisse se faire par des spécialistes du traitement automatique de la parole, non nécessairement experts en transcription phonétique précise.

Le «e caduc» /ə/, lorsqu'il est réalisé, a un timbre que l'on pourrait intégrer à l'ensemble E. Les éventuelles nuances que l'on peut parfois trouver —cf.[Walter,76] pour des données statistiques à ce sujet— ne justifient pas à elles seules une étiquette spécifique. C'est pourtant ce que nous proposons de faire dans la mesure où cela ménage des facilités pour effectuer des statistiques sur l'un des problèmes les plus embarassants de la phonologie du français.

Mais cette disposition à elle seule ne suffirait pas puisque les /ə/ qui tombent ne sont pas matérialisés. Dans sa vue particulière, un utilisateur s'intéressant spécifiquement à ce problème devrait donc disposer de descripteurs et/ou de tables d'étiquetage appropriés consignant cette information. Différentes solutions automatiques ou semi-automatiques existent, notamment si l'on peut disposer d'un système expert en phonologie comme GEPH. On notera dans l'exemple donné en annexe que la chute de /ə/ est accompagnée de celle de /j/ et qu'il y a bien d'autres cas d'élision de phonèmes dont il faudrait tenir compte. C'est pourquoi, placer ce type d'information dans les vues particulières, nous paraît une solution plus conforme à nos principes.

En ce qui concerne la prononciation de (n ɲ) et de (ɲ ŋ) dans divers contextes nous renvoyons à [Simon,70], [Morin,71], [Walter,76] où l'on trouvera bien d'autres références. Nous noterons simplement ici deux cas typiques. La bonne transcription d'une finale en «-ing» —c.-à-d. le choix entre [iɲ] et [iŋ]— serait une tâche délicate pour un transcripteur non phonéticien, de sorte qu'il vaut mieux en effet noter dans tous les cas [iÑ]. Ce problème se rencontre pour mots «camping» et «parking» des corpus ACC01 et ACC02 de BDSON.

La paire (n ɲ) pose des difficultés semblables en contexte palatal; de plus, une bonne transcription phonétique devrait distinguer entre les réalisations en un segment et celles qui en comportent deux; «tu grognes» sera souvent réalisé par [tygrOɲ(ə)] —voir aussi la réalisation de «peigne» dans l'exemple donné en annexe— et «vous grognez» par [vugrOñjE] avec apparition de [j] épenthétique. Si l'épenthèse n'a pas valeur d'allophone —réalisations de type ñ]— alors il conviendrait à ce stade de transcrire simplement [vugrOɲE]. Dans de telles situations, impliquant des appréciations relativement délicates, il faudra s'attendre de toute manière à quelques divergences entre transcripteurs. Les regroupements par paires donnés en (4) permettent au moins d'éviter quelques hésitations de transcription, bien inutiles à ce stade.

Une BDAP doit prévoir la possibilité d'incorporer des transcriptions phonétiques simplifiées et/ou raffinées selon le cas. En particulier pour évaluer convenablement les systèmes de reconnaissance phonétique —phonétique pris au sens large—, ce ne peut être que par référence à une transcription phonétique.

Ainsi, par exemple, il est clair que la (ou les) décision prise pour le segment en correspondance avec la graphie «t'» de l'énoncé (3) dépendra beaucoup de l'allophone effectivement prononcé. Une méthode naïve de décompte des erreurs, telle que seule la réponse [t] soit considérée comme juste, est manifestement insuffisante. Même si on améliore l'évaluation en prenant pour référence une bonne transcription phonétique large, dans les cas d'assimilation partielle avec production de [~t] transcrit [t] par exemple, on retombe sur les inconvénients de la méthode naïve.

Une BDAP doit aussi contenir des tables permettant d'établir des liens entre les différents niveaux de transcription facilitant ainsi les accès en vue de statistiques phonologiques. A titre d'exemple considérons la phrase: «La jeune fille se peigne devant sa glace.» dont les transcriptions successives sont:

phonologique: § lA#ʒŒnə#fijə#sə#pEñə##dəvãt"#sA#glAsə§
phonotypique: /lAʒŒnəfijəsəpEñədəvãsAglAsə/
phonétique large: [lAʒŒnəfisəpEñədəvãsAglAs]

TPL	l	A	ʒ	Œ	n	ə	f	i	s	ə	p	E	ñ	ə	d	ə	v	ã	s	A	g	l	A	s
m°	1		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
								-jə																-ə

m°	8	24	Notations TPL: transcription phonétique large
TPH	-jə	-ə	TPH: transformations phonologiques
			m° : n° de l'unité phonétique

Les deux dernières peuvent être mise en correspondance entre elles par le premier des deux tableaux précédents. En pratique seul le second, que nous appellerons *tableau phonologique*, qui se réduit aux transformations phonologiques (TPH) effectives, est à considérer dans une base de données phonétiques. A chaque chaque prononciation d'un même énoncé et/ou à chaque transcription phonétique peut être associé un tel tableau. Notons que des correspondances de même type peuvent être établies avec les étiquetages phonétiques.

2.3. Les informations temporelles sur les corpus

La mise en correspondance temporelle de la transcription et du signal vocal d'une UE permet d'accéder, dans une BDAP, aux réalisations sonores répondant à des requêtes spécifiant des entités abstraites telles que: énoncés complexes, mots ou segments internes aux mots. S'il s'agit de logatomes, on pourra de même accéder aux segments phonétiques —au sens large— qu'il contient.

La tendance actuelle est de désigner cette opération de mise en correspondance, et son résultat, par *étiquetage —labelling* dans le rapport SAM [SAM rpt,87]—. Cela dit, on peut concevoir autant d'étiquetages que de variétés de transcriptions ou de représentations des énoncés. Cependant, dans une base de données il faut bien se restreindre, ne serait-ce que pour des raisons d'économie.

On remarquera que le descripteur d'UE est déjà une sorte de macro-étiquette attachée à l'intervalle de parole de cet énoncé. Elle peut être complétée par divers étiquetages (au sens strict) et des tables phonologiques associées. Nous retiendrons plus particulièrement les étiquetages suivants:

- étiquetage en mots: ETM(n*, mot, localisation),
- étiquetages phonétiques: ETPH1(m*,unité-phon,localisation),
  ETPH2(m*, allophone, localisation),
  ETPH3(m*, allophone, localisation),
  ...
- étiquetages acoustiques: ETAC1(l*, événement, localisation),
  ETAC2(l*, événement, localisation),
  ...

Le sens des quatre premiers est clair. Ils sont explicitement prévus dans le rapport SAM. Les derniers, aussi appelés *étiquetage fin*, visent à permettre l'accès à des sous-segments significatifs de la production d'un allophone tels que tenue occlusive, tenue vocalique, explosion, aspiration, implosion, transition, vibration, etc.

Plusieurs étiquetages phonétiques ETPHi, i=2,3,..., apportant plus ou moins de détails sur la prononciation doivent être envisagés. Il est souhaitable qu'ils aient les mêmes unités et le même repérage temporel que ETPH1, ce qui se traduit par l'identité de clé, à savoir m*, dans les relations correspondantes.

En pratique, on peut très bien envisager un scénario du type suivant: dans un premier temps ETPH1 est mis en place. Puis, pour les besoins des systèmes de reconnaissance par exemple, est créé ETPH2. Enfin, si une demande donnée justifie le besoin de telle ou telle précision supplémentaire, ETPH3 est créé et ainsi de suite. Bien entendu ces étiquetages supplémentaires peuvent être traités comme de simples ajouts de diacritiques phonétiques complémentaires: par exemple assimilations de labialité, palatalisation, etc.

Lorsque le corpus est muni a priori de transcriptions phonétiques, il est naturel qu'elles constituent les étiquettes des ETPHi. Inversement en rassemblant les étiquettes des ETPHi on obtient autant de transcriptions phonétiques. On observera toutefois que la transcription phonétique provenant d'un étiquetage résulte à la fois du jugement auditif et de celui de l'œil en ce sens que l'étiqueteur peut afficher le signal vocal et le sonagramme de l'UE.

En ce qui concerne les questions de localisation une controverse s'est instaurée à partir de la position exprimée dans [Abry,85] et des pratiques très souvent en vigueur, à savoir: une étiquette doit-elle renvoyer à un segment repéré par ses frontières ou à un pseudocentre ?

D'un point de vue purement informatique on peut mettre en avant que les accès sont grandement facilités si les segments sont déjà délimités dans la base. On notera d'ailleurs que les frontières peuvent être déduites à partir des centres consécutifs que de manière très grossière et surtout bien plus imprécise que par le marquage direct.

Quant aux concepteurs de systèmes de reconnaissance, ils souhaitent souvent extraire des ensembles de segments exactement cadrés si possible sur les unités sous-jacentes. Ici l'étiquetage segmental est préférable, même si des révisions de frontières sont parfois nécessaires; elles seront bien moins importantes que dans le cas de repérage par les centres.

Reste le point de vue phonétique. Tout le monde convient que des frontières peuvent être incertaines, voir structurellement inexistantes —transitions par liquides intervocaliques ou par semi-voyelles par exemples—. Le pseudocentre, à supposer qu'une définition convenable lui soit trouvée, n'est pas non plus quelque chose de facile à déterminer. Au plan pratique les étiqueteurs commencent d'ailleurs généralement par évaluer les frontières afin de mieux placer le pseudocentre. De plus, au plan phonétique les vibrantes, les semi-voyelles, etc, posent souvent des problèmes embarrassants.

Si l'on voulait satisfaire tout le monde, il faudrait sans doute marquer les centres et les frontières (voir par exemple [Takeda,87]), le tout assorti d'indices de fiabilité comme dans l'exemple donné en annexe. En effet la précision avec laquelle les frontières et les centres peuvent être localisés est très variable. Par ailleurs les deux types d'information sont parfois complémentaires.

Dans l'exemple traité en annexe où, sauf spécification contraire, le pseudocentre est le milieu du segment, l'étiqueteur n'a marqué explicitement qu'un pseudocentre; sans doute aurait-il pu le mettre pour deux autres. Mais de toute manière l'information sur les frontières reste largement prépondérante.

Un poste de travail doit permettre diverses variantes de visualisations où, par exemple, les étiquettes pourraient apparaître au centre des réalisations ou face aux pseudocentres.

CONCLUSION: STRUCTURATION DE BDAP

Une version de BDAP, incluant une interface d'étiquetage, est actuellement opérationnelle sur PDP 11/73 au laboratoire CERFIA. Une nouvelle version pour MASSCOMP 5600 est en cours de conception.

Compte tenu des principes retenus concernant, d'une part, la dualité entre noyau et vues particulières, d'autre part, l'étiquetage à profondeur variable selon les besoins et les contraintes de coût, les partages à effectuer dans notre BDAP se présentent comme suit:

3.1. Le noyau

Il contient les informations standards:

- le signal numérisé et sous forme de sonagramme avec possibilité d'affichage,
- les descripteurs de corpus, les descripteurs orthographiques, les transcriptions phonémiques et, dans certains cas, une transcription phonétique,
- les tables d'étiquetage en groupe (s'il y a lieu), en mots (s'il y a lieu), phonémique et, dans certains cas, phonétique.

Ce noyau doit permettre des échanges commodes avec BDSON: réception de corpus et fourniture de résultats d'étiquetage.

3.2. Une vue particulière pour la conception de systèmes de reconnaissance

Elle contiendra des informations dérivées de celles du noyau et d'autres totalement nouvelles, par exemple des corpus particuliers. Des procédures spécifiques y sont attachées. Dans les grandes lignes nous y trouverons:

- le signal dans les formes acoustico-phonétiques nécessaires au projet de système de reconnaissance en cours d'élaboration,
- les tables d'étiquetage appropriées aux unités de décision et aux statistiques que l'on souhaite faire,
- les algorithmes d'apprentissage et de décision tels que ceux de Viterbi et Baum-Welch qui opèrent tantôt sur des mots, tantôt sur des allophones,

• les tables des transformations phonologiques,

• les accès à BDLEX, à GEPH et aux compilateurs de modèles de mots.

3.3. Corpus de données physiologiques

En collaboration avec l'Institut de Phonétique d'Aix-en-Provence, le laboratoire étudie actuellement des corpus de données physiologiques [Auteserre,87] qui demande l'introduction d'indices et d'étiquettes particulières en vue d'étudier des hypothèses phonologiques ayant trait à l'assimilation. Ceci constituera une vue particulière de la base.

4. REFERENCES

[Abry,85] C. Abry, & coll. " Propositions pour la segmentation et l'étiquetage d'une Base de Données des Sons du Français," 14e JEP, GALF-CNRS, Paris, pp. 156-163.

[Autesserre,87] D. Autesserre, C. Barrera, R. Espesser, G. Pérennou, M. Rossi, B. Teston, N. Vigouroux, "Acoustic-Articulatory Information in Data Base,"European Conference on Speech Technology, Edinburgh, pp.125-7.

[Autesserre,88] D. Auteserre, G. Pérennou, M. Rossi, "Méthodologie de transcriptions et d'étiquetage des corpus de parole", rapport GRECO-SAM, juin 1988.

[BDSON,86] "BDSON. Base de Données de Sons de français," Rapport GRECO.

[Baker,83] J.M. Baker, D.S. Pallett, J.S.Bridle, "Speech recognition performance assessments and available databases", Proc. Icassp Boston, paper 12.2.

[Carlson,86] R. Carlson, B. Granstom, "A search for durational rules in a real-speech data base", Phonetica 43, pp. 140-154.

[Carré,84] R. Carré, R. Descout, M.Eskenazi, J. Mariani and M. Rossi, "The French Language database: defining, planning and recording a large database", Proc. Icassp 84, Paper 42.10.

[Dell,86a] F. Dell, "Les règles et les sons," Hermann.

[Dell,86b] F. Dell, "Deux nasalités en français," Séminaire GRECO/GALF, Toulouse, pp 187-192.

[Itahashi,86] S. Itahashi, "A japanese language speech database", Pro. Icassp 86, Tokyo, p. 321-324.

[Lambert,86] M. Lambert, "Codage phonétique du lexique dans une base de données lexicales. Thèse de Doctorat de 3ème Cycle Aix, 1986.

[Leonard,84] R.G. Leonard, "A database for speaker-independant digit recognition", Proc. Icassp 84, paper 42.12, San Diego.

[Malécot,72] A. Malécot, "Progressive Nasal Assimilation in French," Phonetica, Vol. 26, n°4, pp. 193, 209.

[Morin,71] Y.Ch. Morin, "Computer Experiments in Generative Phonology: Low-Level French Phonology," Natural Langage Study n°11, Univ. of Michigan.

[Pérennou,87] G. Pérennou & M. De Calmes,"BDLEX Lexical Data and Kn. Base of Spoken and Written French," European Conf. on Speech Tech., Edinburgh, pp.393-6.

Plénat,86] M.Plénat, " Labialisation et syllabation en français standard," Rapport interne GRECO-CP.

[Rossi,86] M. Rossi, M.L. Lambert, "Représentation et traitement des voyelles à timbre multiple", Actes du Séminaire GRECO-GALF, Toulouse 1986, pp. 141-162.

[SAM rpt,87] ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Schane,1968] S. A. Schane, "French Phonology and Morphology," MIT.

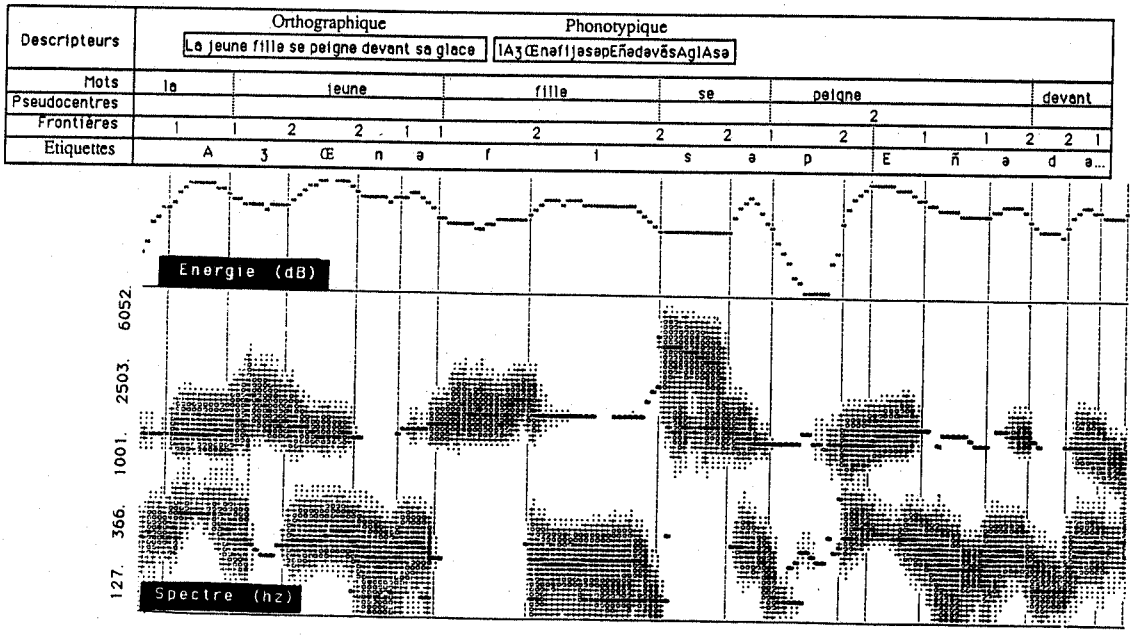
[Simon,70] P. Simon, "A propos de la désarticulation de la consonne palatale phonétique et linguistique romane," Mélanges offerts à G. Straka, Lyon-Strasbourg, pp.67-98.

[Takeda,87] K. Takeda, Y. Sagisaka, S. Katagiri, "Acoustic Phonetic labels in a japanese speech database","European Conference on Speech Technology, Edinburgh, Vol 2. pp.13-16.

[Vigouroux,87] N. Vigouroux, "Une base de données acoustico-phonétiques", Congrès RFIA, Antibes, pp. 367-380.

[Walter,76] H. Walter, "La dynamique des phonèmes dans le lexique français contemporain," France Expansion.

Annexe: Etiquetage de la séquence "La jeune fille se peigne devant sa glace" extraite du corpus PEQ04A.LC de BDSON



Visualisation des relations d'étiquetage lexical et phonétique large. Les frontières et les pseudocentres sont munis d'un indice de fiabilité: 0 si la frontière est très mal définie, 1 si elle est assez bien définie, 2 si elle est précise. Les pseudocentres sont indiqués lorsqu'ils s'écartent significativement du milieu de segment. Pour cette option de visualisation, les unités phonétiques apparaissent en leur centre.

Conception et réalisation d'un module de reconnaissance des voyelles

M.GRENIE & M. ROSSI

Institut de Phonétique L.A. 261 Parole et Langage
29 avenue R. Schuman, 13621 Aix-en-Provence cedex

ABSTRACT

This paper discusses the conception and production of a module for fine discrimination of vowels. Several approaches are presented and an original solution based on the interpretation of acoustic variability is proposed, which has the advantage of being closer to phonological analysis of speech than is the case with the techniques more commonly used.

INTRODUCTION

En raison de la variabilité acoustique de la parole, la mise au point de modules spécialisés de reconnaissance fine multilocuteur des voyelles s'avère délicate. Les difficultés généralement rencontrées s'articulent autour de trois questions essentielles :

1°) quelle est la nature des connaissances acoustiques à mettre en oeuvre;

2°) comment faut-il les extraire ;

3°) par quels formalismes peut-on les représenter.

La conception d'un module de décodage acoustico-phonétique implique nécessairement une réflexion sur l'organisation acoustique de la parole et sur les structures de représentation les plus appropriées pour la respecter. La première partie de cette communication traite de quelques points mis en évidence lors de l'analyse acoustique d'une base de donnée conséquente, la deuxième passe en revue les avantages et inconvénients de plusieurs systèmes développés.

1. CONSTITUTION D'UNE BASE DE DONNEES

Une base de données a été constituée en analysant à l'aide d'un vocodeur à 14 canaux (250-4200 Hz) un premier corpus de 330 mots de type CVCV prononcés par 4 locuteurs masculins (2650 voyelles) et un second corpus d'une quarantaine de nombres prononcés par 7 femmes et 7 hommes (1000 voyelles). Pour faciliter l'analyse automatique des résultats, une transcription large est associée à chaque mot. L'encodage des occurrences vocaliques s'est fait selon un système réduit : i (424), y (185), u (167), e (626), a (484), ɔ (391), ɔ̃ (334), ɑ̃ (449), ɛ̃ (316), ə (130), ø (154). C'est-à-dire que les oppositions e/ɛ, ə/ɑ, o/ɔ, ø/œ, œ̃/ɛ̃ ne sont pas dinstinguées. Un algorithme de cadrage automatique se fonde sur la transcription pour localiser le début et la fin des voyelles.

2. ANALYSE QUANTITATIVE

Afin d'évaluer l'efficacité de ce type d'approche dans le cadre d'une identification fine des voyelles, plusieurs techniques d'analyse quantitative furent mises en oeuvre. Aucune n'a cependant permis d'aller au-delà d'une reconnaissance de macro-classes vocaliques.

2.1. Agrégation aux spectres moyens

Après normalisation des spectres, de manière à obtenir une représentation indépendante du gain, les spectres moyens des 11 classes vocaliques ont été calculés selon cinq localisations temporelles : initiale de la voyelle, milieu de la voyelle, fin de la voyelle, spectre correspondant au maximum d'énergie de la voyelle, moyenne spectrale du début à la fin de la voyelle. Pour chaque occurrence de la base et chacune des 5 positions retenues, on calcule la distance euclidienne entre les spectres relevés et les spectres moyens. Ces distances sont alors ordonnées de façon croissante et ce sont les rangs obtenus entre l'occurrence à identifier et les 11 spectres moyens qui sont pris en compte. Le cumul des écarts sur les 5 positions constitue une sixième valeur.

Les résultats schématisés dans la figure 1 apportent plusieurs enseignements. Tout d'abord, ils confirment le phénomène bien connu suivant lequel c'est le coeur de la voyelle qui est le moins sensible à la coarticulation. Cependant, la prise en compte pour cette position des deux premiers candidats conduit à un score de reconnaissance inférieur à 80%. La technique d'agrégation aux spectres moyens ne convient donc pas pour une reconnaissance fine. D'autre part, le fait que le calcul d'un spectre moyen, pour toute la durée de la voyelle, n'améliore aucunement les performances par rapport aux autres positions, laisse penser que le calcul arithmétique d'une moyenne ne respecte pas la

structuration de la parole dans la mesure où il melange des spectres successifs qui subissent des contraintes différentes et doivent, à ce titre, être interprétés différemment. Enfin, le faible écart entre la médiane et la combinaison des cinq positions prouve que les candidats correctement classés en chaque position sont en partie décorrélés. Il ne faut donc pas négliger l'information apportée par les parties non stables des voyelles.

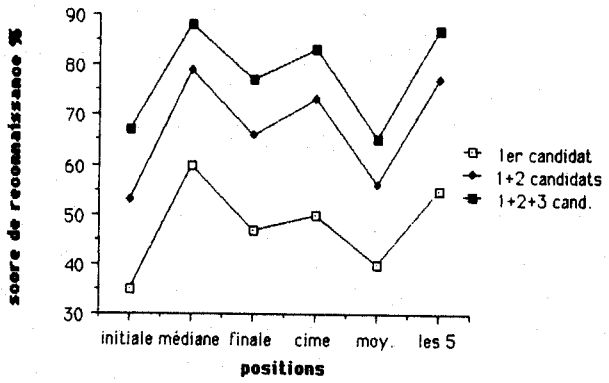


Figure1 : Scores d'agrégation aux spectres moyens pour les voyelles de la base selon la position et les candidats

2.2. Agrégation multi-référence

Généralement, les systèmes quantitatifs de reconnaissance multi-locuteur mettent en oeuvre des techniques d'agrégation multi-références. Des méthodes de classification automatique sont utilisées pour constituer le dictionnaire des spectres de référence représentatifs de chaque phonème. Afin de montrer les limitations d'une telle procédure, nous avons appliqué, aux spectres de la position médiane, un programme de classification hiérarchique ascendante qui constitue les spectres de référence. Cet algorithme regroupe successivement et deux à deux, voyelle par voyelle, les spectres occurrents les plus proches. Quand le nombre de classes initialement prévu est atteint, l'agglomération s'arrête et chaque classe est alors définie par les spectre de son centroïde. Lors de la reconnaissance, les distances entre chaque trame médiane de toutes les occurrences contenues dans la base et tous les centroïdes sont calculées. C'est la classe dont le centroïde est le plus proche qui est reconnue. Deux points ont retenu notre attention : l'importance de l'étendue des classes (1 à 30 classes ont été calculées), la sensibilité de cette méthode vis à vis des unités phonétiques considérées.

La figure 2 représente les résultats des agrégations pour les voyelles /i,e,a,ɔ/ en retenant les 3 premiers candidats. Il apparaît de manière générale que les scores n'évoluent que très peu quand on passe de 1 à 30 classes. D'autre part, des écarts notables s'observent

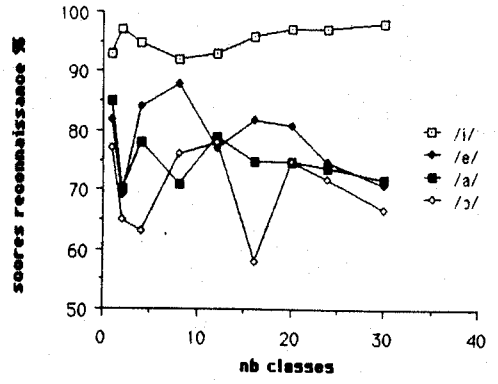


Figure 2 : Scores d'agrégation aux spectres moyens pour les voyelles /i, e, a , ɔ/ de la base en fonction du nombre de classes pour les 3 premiers candidats

selon les voyelles. Si une telle technique est éventuellement compatible avec une reconnaissance fine de /i/, elle donne pour /ɔ/ un score médiocre. Ces différences s'expliquent selon nous par la variabilité acoustique spécifique de ces voyelles : /ɔ/ connaît une plus grande variabilité acoustique que /i/. On constate également une interdépendance de certains scores : le sommet atteint pour /e/ avec huit classes et les trois premiers candidats correspond exactement aux plus mauvais résultats de /i/ et de /a/ qui sont, selon le contexte, ses plus proches voisins. De même, l'apogée de /a/ coïncide avec le plus bas résultat de /ɔ/. Ce phénomène est liée au fait qu'une technique d'agrégation n'interprète pas les distances spectrales de manière absolue mais en fonction des rangs. Or, dans le cas d'une reconnaissance fine des voyelles, en raison des nombreuses proximités acoustiques, il est à craindre que la seule prise en compte du rang soit insuffisante et que, de ce fait, on ne puisse dépasser le score de 85 % avec un ou deux candidats, même avec des apprentissages étendus.

Quelle que soit la technique de classification mise en oeuvre, le positionnement au sein de l'espace acoustique d'interfaces séparant les phonèmes est une opération délicate. Une classification hiérarchique aura tendance a donner trop d'importance à des réalisations très particulières, peu importantes statistiquement dans le corpus considéré et pas nécessairement représentatives de la population parente. A l'inverse, un algorithme de nuées dynamiques réduira le poids des réalisations particulières au profit du plus grand nombre. Malheureusement, à partir d'une base de données rassemblant quelques milliers d'occurrences, il est impossible de conclure que tout ce qui est rare est improbable. Dans la mesure où elle ne conduit pas à une appréciation linguistique qualitative des distributions acoustiques (locuteur typé, prononciation défectueuse,

contexte rarissime,...) l'utilisation d'une technique de classification ne garantit pas l'obtention de partitions optimales.

Les résultats présentés ci-dessus montrent que des techniques purement quantitatives sont insuffisantes pour parvenir à une reconnaissance fine des voyelles. En effet, la qualité de l'analyse acoustique effectuée par le vocodeur ne saurait être la seule explication de ces résultats. A l'évidence, des problèmes de fond s'opposent au bon fonctionnement de ces méthodes.

2.3. Critique de la notion de distance

La notion de distance est inadéquate pour rendre compte de l'organisation phonologique de l'espace acoustique car elle n'apprécie nullement la pertinence linguistique des écarts spectraux. Un exemple caricatural suffit pour illustrer ce point. Voici les spectres relevés au centre de deux /i/ prononcés par un même locuteur, après normalisation:

i1 *loti* GRM 20 10 03 03 02 03 07 07 02 02 05 12 12 12
i2 *gidon* GRM 12 08 06 06 09 07 05 06 08 07 07 09 12 08

La distance euclidienne entre ces deux vecteurs (i1, i2) vaut 246. Comparons ces deux occurrences avec un spectre fictif X :

X 12 07 06 12 06 06 08 05 02 02 04 10 10 10

La distance euclidienne entre i1 et X vaut 206. Cela signifie que la trame fictive X est plus proche de i1 que i2 ne l'est lui-même de i1. Mais la forte valeur du canal 4 de X indique clairement que cette voyelle n'est pas assez diffuse et ne peut en aucun cas être assimilée à un /i/. Il ne faut pas espérer inverser ces résultats par le choix de telle ou telle métrique. Plusieurs études se sont attachées à comparer les mérites relatifs de différents modes de calcul de distances spectrales. Dans tous les cas, il n'existe pas de distance miracle.

En reconnaissance des formes, le calcul de distance est associé à la notion de ressemblance. Plus deux éléments sont semblables et plus la distance qui les sépare est petite. Appliquée au signal de parole, cette notion peut se définir simultanément à deux niveaux qui ne sont pas en totale correspondance, celui de la substance physique et celui de la forme linguistique.

Un critère purement quantitatif tel que le calcul d'une distance, à partir de paramètres physiques, est inadéquat pour rendre compte des ressemblances qui existent à ce second niveau car il méconnaît la variabilité et l'organisation acoustique du signal. Tout dans un spectre n'a pas la même pertinence : c'est seulement au travers une interprétation des contraintes et des degrés de liberté qui s'exercent dans la chaîne parlée que les écarts observés sont analysables. Certaines variations acoustiques ne seront pas à prendre en considération par le DAP car elles seront libres ou aléatoires alors que des variations contraintes en apparence infimes correspondront parfois à des changement de type catastrophique.

Un autre défaut du calcul de distances est qu'il présuppose que l'ensemble des spectres de parole constitue un espace acoustique homogène et de même nature : entre deux spectres quelconques une distance est toujours calculable. Or, il n'est absolument pas certain qu'au sein du système phonétique d'une langue donnée, tous les éléments entretiennent des liens entre-eux qui puissent s'exprimer sous la forme d'une variable numérique continue et unique.

Enfin, dans la mesure où l'appréciation qualitative de ces valeurs quantitatives est pratiquement impossible, il est nécessaire de fixer des critères de classement et de décision tels que des seuils de rejet et d'émergence ainsi que le nombre de candidats retenus. De ce fait, les résultats sont extrêmement inter-dépendants et peu perfectibles : l'amélioration de la reconnaissance d'une unité perturbera la reconnaissance des autres. Ce phénomène, qui n'est pas gênant dans le cas d'une reconnaissance globale de mots en raison des dispersions acoustiques, paraît incompatible avec une reconnaissance fine de phonèmes.

Ces constatations nous conduisent à envisager une approche qualitative.

3. APPROCHE QUALITATIVE : TRAITS ET INDICES

3.1. Principes

La première approche qualitative que nous avons retenue en vue d'une reconnaissance fine des voyelles s'inspirait directement de la démarche indicielle proposées par Rossi et al. (1983), Mercier et al. (1985) et Caelen et al. (1981). Selon ces auteurs, il est possible de réduire la redondance du signal par l'utilisation de cinq ou six combinaisons linéaires de canaux sans perte notable d'information. Les indices spectraux non formantiques définis par Rossi et al. (1983) aboutissent à une identification satisfaisante des macro-classes vocaliques quels que soient les locuteurs et les contextes. Par une hiérarchisation de ces indices nous avons tenté de parvenir à une reconnaissance fine. La valeur des indices , au nombre de 47, se calcule simplement par comparaison de canaux ou de bandes spectrales, sur tout ou partie de la voyelle. Elle s'exprime sous forme binaire. La reconnaissance s'opère, sur des voyelles préalablement segmentées, selon trois étapes : 1°) segmentation, 2°) calcul des indices acoustiques, 3°) identification des unités phonétiques en fonction d'une base de connaissances définie

antérieurement qui encode les combinaisons d'indices associées aux candidats.

Pour passer des indices acoustiques aux unités phonétiques, on associe des candidats à des combinaisons d'indices sous la forme suivante :

Si - Ouve 01 - Aigu 05 + Aigu 06 + Aigu 10 + Aigu 11 alors /i/
Si - Ouve 01 - Aigu 05 + Aigu 06 - Aigu 10 - Aigu 11 alors /ɔ,u/

Au total, une centaine de règles sont ainsi définies, à chacune correspond un ou plusieurs candidats. Lors de la reconnaissance, on explore les règles dans l'ordre décroissant du nombre d'indices mis en jeu.

3.2. Résultats

Le score de reconnaissance phonétique obtenu en moyenne ne dépasse pas 85%. Plutôt que de discuter de ces résultats qui dépendent étroitement des corpus mis en oeuvre, il importe de faire plusieurs observations sur certains défauts de ce module. Car, sur plusieurs points, l'architecture proposée est en contradiction avec ce que l'on sait de la fonction distinctive et oppositive du phonème :

- dans la mesure où les indices sont calculés séparément les uns des autres, le phonème est assimilé à un conglomérat de propriétés acoustiques indépendantes ; c'est oublier que le tout est différent de la somme des parties ;
- la comparaison de formes mise en oeuvre lors de la reconnaissance réduit le phonème à un objet acoustique puisque pour qu'il soit reconnu il doit y avoir identité entre les indices calculés et les combinaisons constituées lors de l'apprentissage ;
- appliquer systématiquement toutes les règles pour toutes les occurrences revient à considérer que celles-ci sont issues d'une unique paradigme. Or, la substance acoustique est organisée ; elle ne se réduit pas à la concaténation d'unités qui forment un inventaire plus ou moins étendu. Des métaconnaissances sont nécessaires pour parvenir à interpréter spécifiquement la diversité des locuteurs et des contextes.

C'est seulement en intégrant explicitement la variabilité au coeur du module que l'on peut espérer respecter la nature oppositive du phonème et l'effet spécifique du contexte sur chaque partie de la voyelle.

4. VERS UNE ANALYSE DE LA VARIABILITE

4.1. Formalisation des connaissances

Les difficultés rencontrées dans la première approche laissent supposer que, pour espérer fonctionner correctement, tout système doit respecter impérativement les principes suivants :

- utiliser un nombre fini et restreint de connaissances car c'est un gage de perfectibilité ;
- dans la mesure où l'extraction des connaissances se fait par l'intermédiaire d'un expert, celles-ci doivent renvoyer à une réalité physiquement appréhensible. L'utilisation simultanée de plusieurs plans de description tels que traits, indices ou paramètres est à éviter, à ce niveau, car elle complique l'identification des erreurs au sein de la base de connaissances. Comment déterminer en effet si c'est la définition des indices, la détermination des combinaisons ou le choix des candidats qui sont à l'origine des confusions ?
- respect de l'organisation de la parole : ne retenir du signal que des propriétés indépendantes (position des formants, rapports d'énergie,...) qui ne disent rien de la globalité du phénomène, c'est procéder de façon aveugle et briser l'organisation. Car, chaque réalisation acoustique résulte d'une combinaison de contraintes et de degrés de liberté. Toute analyse qui tient compte de l'un sans tenir compte de l'autre ne permet pas d'interpréter convenablement la variabilité. Selon nous, l'invariant gage de la communication entre les hommes n'est ni absolu ni relatif mais organisationnel. C'est au travers d'une analyse de l'organisation du signal que la reconnaissance de la parole est possible. Le concept d'organisation appliqué au signal de parole correspond au constat qu'en raison des relations qu'entretiennent entre-eux les différents éléments mis en jeu (traits,phonèmes,...), leur agencement produit, au niveau acoustique, une unité qui possède des qualités en partie autres que celles de ses composantes. La seule réalité communément observable de la parole étant le signal physique, nous croyons fortement qu'il contient en lui-même tous les éléments nécessaires à son décodage. Les difficultés rencontrées jusqu'à ce jour dans la recherche d'invariants acoustiques proviennent, selon nous, du fait que l'on a cherché des invariants isolables physiquement. Or, la variabilité acoustique ne s'oppose pas nécessairement à l'invariance qu'implique la communication. Le terme d'invariance organisationnelle suppose que l'invariance n'existe pas en tant que tel dans le signal mais s'établit dans les relations que les différents termes entretiennent entre-eux. Ainsi le passage de la substance acoustique à la forme phonologique se ferait par l'intermédiaire de l'extraction d'une forme au niveau acoustique.

4.2. Des principes organisateurs ?

L'observation d'un grand nombre de documents montre que l'organisation du signal de parole résulte principalement de deux principes :

- en tout point de la chaîne parlée se définissent, selon les contraintes du moment, différents univers des possibles. Ces contraintes sont d'origine articulatoire, acoustique, phonétique ou phonologique. Elles sont donc prises en compte lors de l'encodage réalisé par le

locuteur. De même que la phonologie indique qu'en une position donnée ne commutent que certains phonèmes, de même l'analyse acoustique montre qu'en un endroit donné du signal seul un petit nombre de réalisations sont effectivement possibles. Ce constat implique que pour interpréter correctement une portion de signal, il faut la comparer uniquement au sous-ensemble constitué des réalisations compatibles avec les mêmes contraintes.

- d'autre part, du fait de la continuité des mouvements articulatoires, les contraintes ne demeurent identiques que pour des intervalles de temps très courts. A condition de respecter les contraintes locales en ne comparant entre-eux que des événements suceptibles d'appartenir au même paradigme, nous pensons qu'il est possible d'exploiter les informations acoustiques contenues dans les zones fortement coarticulées des voyelles. A l'inverse, le calcul de spectres moyens dégrade les scores car il génère des spectres chimériques issus du mélange de spectres instantanés soumis à des contraintes dissemblables. De même, l'utilisation d'une segmentation avant la reconnaissance, de façon à faire cadrer des portions de signal avec des unités préalablement constituées, revient à considérer qu'il s'agit de portions uniformes. C'est ignorer les informations syntagmatiques apportées par la modification des contraintes de part et d'autre du centre de la voyelle.

4.3. Réalisation

Nous avons entrepris de réaliser un système conforme aux points énumérés ci-dessus. La parole est analysée par un vocodeur et , dans la mesure où il s'agit d'identifier uniquement les voyelles, elle est représentée sous la forme des contributions de chaque canal à l'énergie totale. Par rapport à d'autres techniques de normalisation de l'énergie, le choix d'un espace fermé (le cumul des contributions vaut toujours 100) a le mérite de faire clairement ressortir et l'allure globale du spectre et les contraintes ou degrés de liberté qui affectent chaque canal. Pour respecter la nature locale des contraintes, la reconnaissance se déroule dans un premier temps trame par trame, sans utiliser de segmentation préalable. Ces résultats sont ensuite validés temporellement en fonction de la durée minimale de chaque voyelle. La prise en compte des contraintes s'effectue en fonction du numéro du canal dont la contribution est la plus forte. Bien évidemment ce critère est insuffisant pour apprécier dans toute son ampleur l'organisation du signal mais des relevés effectués sur notre base de données ont montré qu'il permettait cependant de dégager localement des typologies cohérentes de contextes et de locuteurs.

L'organisation des connaissances que nous avons adoptée est à l'image de l'organisation supposée de la parole (cf. figure 3):

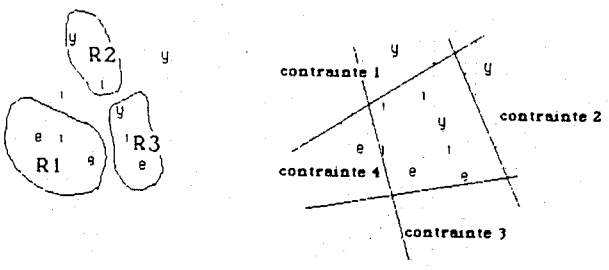
- les connaissances acoustiques ne forment plus des granules éparses mais elles sont structurées en blocs cohérents qui expriment, par classe paradigmatique, les contraintes et les degrés de liberté ;
- pour chaque phonème, on définit autant de blocs, c'est-à-dire autant de types de contraintes possibles, qu'il y a de canaux pouvant contenir la contribution maximale ;
- pour être conforme à la nature des oppositions phonologiques, les connaissances sont exclusivement définies négativement ;
- l'association de connaissances grossières au sein des blocs permet d'aboutir à une reconnaissance fine ;
- les connaissances spectrales mises en oeuvre sont en nombre restreint, elles se réfèrent directement aux valeurs des contributions du vocodeur et s'expriment sous forme d'inégalités.

```
Si canal 01 < 7 ou canal 01 > 11  pas /a/
Si canal 02 < 8 ou canal 02 > 11  pas /a/
Si canal 03 < 7 ou canal 03 > 12  pas /a/
Si canal 04 < 6 ou canal 04 > 11  pas /a/
Si canal 05 < 7 ou canal 05 > 11  pas /a/
Si canal 06 < 7 ou canal 06 > 12  pas /a/
Si canal 07 < 6 ou canal 07 > 09  pas /a/
Si canal 08 < 5 ou canal 08 > 09  pas /a/
Si canal 09 < 5 ou canal 09 > 10  pas /a/
Si canal 10 < 5 ou canal 10 > 09  pas /a/
Si canal 11 < 4 ou canal 11 > 08  pas /a/
Si canal 12 < 4 ou canal 12 > 09  pas /a/
Si canal 13 < 5 ou canal 13 > 09  pas /a/
Si canal 14 < 4 ou canal 14 > 09  pas /a/
Si canal 01 > canal 02            pas /a/
```

Figure 3 : Connaissances définies pour les /a/ ayant leur sommet en canal 6

Une telle approche a de nombreux avantages :

- la méthodologie définie pour constituer la base de connaissance est simple et explicite. Quiconque traite un corpus d'apprentissage donné aboutira aux mêmes connaissances. Il est possible d'imaginer des procédures automatiques pour extraire les connaissances.
- le fait d'exprimer les contraintes en blocs constitués d'inégalités grossières autorise un traitement rapide puisque seule un petit nombre de connaissances sont appliquées à chaque fois en moyenne.
- l'association de règles négatives robustes au sein des blocs donne une réelle possibilité de reconnaître des occurrences autres que celles contenues dans le corpus d'apprentissage. En effet, au travers des gabarits ou des inégalités entre canaux, ce sont les contraintes qu'on exprime et non des objets (cf. figure 4)



règles positives granulaires règles négatives structurées

Figure 4 : Résultats de la reconnaissance selon la nature des connaissances. La modélisation explicite de la variabilité à travers la prise en compte des contraintes exprimées sous forme de règles négatives permet de reconnaître potentiellement des réalisations qui ne sont pas contenues dans le corpus d'apprentissage mais qui y sont inscrites en filigrane comme des réalisation possibles.

4.4. Résultats

Des problèmes de logistique informatique nous ont empêché de tester ce système sur notre base de données. Une mise en oeuvre récente de cette architecture sur micro-ordinateur avec un vocodeur différent a permis d'obtenir des scores provisoires oscillants entre 90% et 97% de bonne reconnaissance moyenne des voyelles pour des locuteurs nouveaux sur un vocabulaire de commandes enchaînées avec un nombre moyen de candidats très légérement suprérieur à 2. Une évaluation plus rigoureuse des performances est en cours.

5. CONCLUSION

A la suite de diverses tentatives de réalisation de systèmes de reconnaissance fine des voyelles, il apparait qu'il n'existe pas dans le signal de parole d'invariants acoustiques suffisamment précis pour aller directement au-delà d'une identification de macro-classes vocaliques.

Une analyse de la variabilité destinée à mettre en évidence une invariance organisationnelle semble en revanche compatible avec une reconnaissance fine. Selon cette dernière perspective la variabilité acoustique de la parole ne constitue plus un obstacle au décodage acoustico-phonétique, c'est au contraire un élément porteur d'information. Pour être crédible, le concept d'invariance organisationnelle devra être à l'avenir affiné ; il devra notamment tenir compte de la structuration syntagmatique du signal.

6. BIBLIOGRAPHIE

Bonneau A., Rossi M., Mercier G. (1986) *Hierarchical representation of French vowels by expert system*, Proc. of Montreal Symposium on Speech Recognition, McGill University, 20-21.

Caelen G., Caelen J. (1981) *Indices et propriétés dans le projet Arial II*, sém. Processus encodage-décodage phonétique, GALF, Toulouse, 128-144.

Grenié M. (1987) *Nature et hiérarchie d'indices acoustiques indépendants du locuteur : application à la reconnaissance automatique des voyelles du français* Thèse 3ème cycle, Aix-Marseille I.

Perkell J. S. et Klatt D. H. (1985) *Variability and invariance in Speech Processes*, Eds. Hillsdale, Lawrence Erlbaum.

Liénard J. S. (1984) *Une approche globaliste de la variabilité acoustico-phonétique la parole*, 13ème JEP, Bruxelles, 57-68.

Mercier G., et al. (1985) *Acoustic phonetic decoding in the SERAC expert system*, actes Séminaire franco-suédois, Grenoble.

Petitot-Cocordia J. (1985) *Les catastrophes de la parole : de Roman Jakobson à René Thom*, Maloine, Paris.

Rossi M. et al. (1983) *Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance de la parole*, Speech Communication,2, 215-217.

DISTANCES INTERSPECTRALES ET MACROSENSIBILITE DES VOYELLES FOCALES

Haïyan YE[1] - Marie-José CARATY[3] - Denis TUFFELLI[1] - Louis-Jean BOE[2]

INSTITUT DE LA COMMUNICATION PARLEE - UA CNRS n°368
1. Laboratoire de la Communication Parlée
46, Avenue Félix-Viallet - 38031 Grenoble Cedex
UUCP YE%SAPHIR.DECNET@CIME.IMAG.FR
BITNET YE@FRCIME51

2. Institut de Phonétique de Grenoble
Université Stendhal
B.P. 25 - 38040 Grenoble Cedex
BITNET BOE@FRCICG71

LAFORIA - UA CNRS n° 1095
3. Laboratoire de Traitement de la Parole
Université Pierre et Marie Curie (PARIS 6)
4, place Jussieu - 75232 Paris Cedex 05

ABSTRACT

In a previous study, 16th JEP [YE & al., 87], we evaluated the behaviour of different dissimilarities (Plomp, LLR, MFCC, WSM,...) used in Speech Recognition. Our criteria were based on the capability of these processes to furnish vocalic distances that could be interpreted in terms of articulatory acoustic and perceptual phonetic knowledge. This paper deals presently with capability of such dissimilarities to obtain satisfactory discrimination results d_{inter}/d_{intra} for synthetic stimuli generated around focal points for [i, y, u, a].

1. INTRODUCTION : LA NOTION DE CONVERGENCE

On peut présenter le processus de la Communication Parlée comme une **divergence-convergence** qui assure le transfert du code linguistique du locuteur à l'auditeur (cf. figure 1). La divergence peut être associée, pour l'essentiel, aux contraintes de la coarticulation qui conduisent à des jeux de commandes combinant les possibilités des différents articulateurs (larynx, mâchoire, langue, lèvres,...). La convergence peut déjà être repérée au niveau de points "cruciaux" de la fonction d'aire, caractérisée par exemple pour les voyelles orales, par le lieu de constriction (abscisse X_c), sa dimension (aire A_c), et l'aperture aux lèvres (aire A_l). Ces points de passage obligés peuvent être mis en évidence soit par des perturbations (les "bite blocks") [GAY & al., 81] ou par des simulations extensives [ATAL & al., 78 ; RAJAA & al., 86 ; BOE & PERRIER, 88] avec des modèles de plus en plus réalistes tels que des fonctions d'aires paramétrisées [STEVENS & HOUSE, 55] ou des coupes sagittales générées à partir de cinq commandes articulatoires [MAEDA, 79]. La convergence se poursuit au niveau perceptif : le produit acoustique subit une intégration perceptive, par exemple de type F_2' [CARLSON & al., 75 ; MANTAKAS & al., 86] qui, en reproupant les masses d'énergie, "gomme" les différences intra-classe.

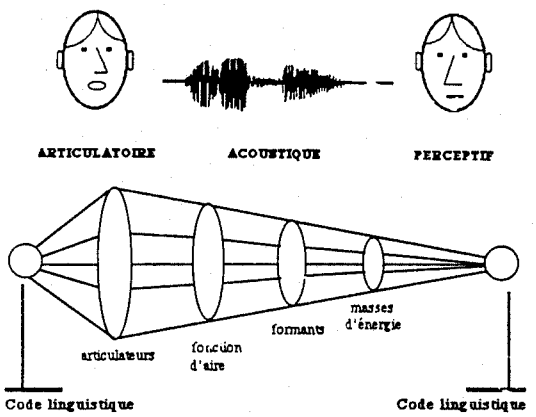


Figure 1.
Du code linguistique du locuteur à celui de l'auditeur : un schéma de la convergence de l'espace articulatoire à l'espace perceptif.

Notre étude utilise cette notion de convergence pour évaluer des distances (17 au total, cf. tableau 5 pour la liste et Annexe pour leur spécification) dans le cadre du traitement automatique de la parole (codage, reconnaissance,...). Nous avons choisi des stimuli vocaliques de type [i, y, u, a] qui présentent, pour une latitude articulatoire du lieu de constriction (X_c), des divergences bien particulières : croisement(s) formantique(s) dus à l'échange des affiliations cavité-résonance que l'on peut caractériser comme des phénomènes de catastrophe [PETITOT- COCORDA, 84].

A partir de 11 stimuli par voyelle, générés avec le modèle à 4 tuyaux [FANT, 60], ont été calculées 946 dissimilarités (symétrisées au besoin) qui servent d'entrée à une analyse multi-dimensionnelle (MDS), permettant d'obtenir une projection de l'espace reconstruit sur deux dimensions. Nous évaluerons ces distances en fonction de leur **aptitude à faire "converger" le système vers le code**, avec un critère qui reflète en partie leur capacité à **minimiser les distances intra-classe** tout en **maximisant les distances inter-classes.**

2. GENERATION DES STIMULI : MACROSENSIBILITE X_C DES VOYELLES FOCALES

Avec ce même modèle à 4 tuyaux, des études précédentes [BOE & ABRY, 86 ; BADIN & BOE, 87] nous ont permis de mettre en évidence et de localiser les voyelles focales. Nous les caractérisons comme des repères topologiques pour lesquels il y

a échange d'affiliation entre les formants et les cavités avant-arrière correspondantes. On obtient selon l'aperture des lèvres, ouvertes ou fermées (cf. tableau 2), des voyelles de type :

- ⇨ [a] : échange F1/F2,
  [i] : échange F2/F3 ou F3/F4.
- ⇨ [u] : échange F1/F2,
  [y] : échange F2/F3.

	Xc (cm)	Al (cm2)
[a]	3.2	4.0
[i]	12.9 13.8	4.0
[u]	7.3	0.16
[y]	12.5	0.16

Tableau 2.
Les paramètres des points focaux pour le modèle à 4 tuyaux [FANT, 60].
Aire des cavités avant et arrière : 8 cm^2, aire de la constriction : 0.65 cm^2.
Longueur totale : 16 cm.
Longueur de la constriction : 5 cm, longueur des lèvres : 1 cm.
X_c représente l'abscisse du milieu de la constriction.

Nous avons généré 11 stimuli, entre ou de part et d'autre de chaque point focal, en faisant varier le lieu de constriction X_c avec un pas de 1 mm pour [y, u, a] et de 2 mm pour [i] (cf. tableau 3).

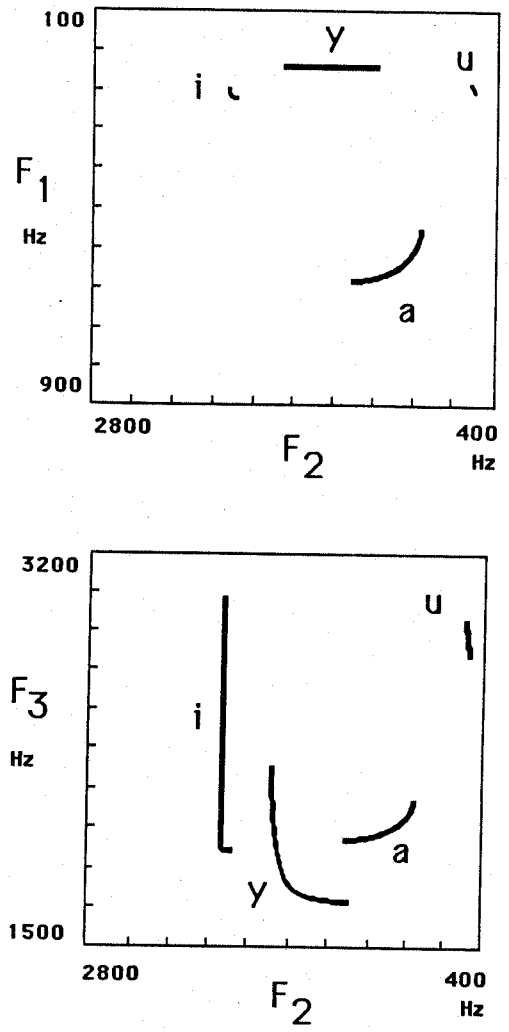
	Xc (cm) min	max
[a]	2.7 -	3.7
[i]	12.9 -	13.8
[u]	6.8 -	7.8
[y]	12.0 -	13.0

Tableau 3.
Les plages du lieu de constriction (X_c) utilisées pour les stimuli.

Une simulation acoustique (incluant les pertes par viscosité, chaleur, vibration des parois et rayonnement aux lèvres) nous a permis de calculer les formants de chaque configuration. Les signaux ont été synthétisés par un modèle tout pôles [FENG, 85] excité par une onde glottique réaliste (voix d'homme) et prenant en compte les pertes à l'ouverture de la glotte.

Pour [i] nous avons tenu compte des résultats obtenus par simulation "humaine" [LADEFOGED & BLADON, 82] et par un modèle réaliste [PERRIER & al., 87] en stabilisant le second formant après dépassement du point focal correspondant à l'échange F2/F3. Nous simulons ainsi les macrosensibilités* pour de larges variations du lieu de constriction (1 cm). Les figures 4 présentent ces macrosensibilités dans les plans F1/F2 et F2/F3.

Note*. En utilisant une méthode Lagrangienne il est possible de calculer directement [CHARPENTIER, 86], même dans le cas d'une simulation du conduit vocal avec pertes, la **sensibilité** de l'amplitude de **toute** la fonction de transfert (et en particulier des zones formantiques) pour de **petites** variations (transversales ou longitudinales) de la fonction d'aire. Dans le cas où les variations introduites à l'entrée sont importantes, on quitte le domaine de validité du calcul des sensibilités, aussi CHARPENTIER a proposé le terme de **macrosensibilité** pour les résultats qui sont alors obtenus par simulation extensive.



Figures 4.
Les macrosensibilités des voyelles focales dans les plans F1/F2 et F2/F3.

3. EVALUATION DE LA CONVERGENCE

3.1. Le critère

Pour chaque distance, nous avons calculé la matrice symétrisée des dissimilarités correspondant à l'ensemble des stimuli. Pour évaluer comparativement la **convergence** des distances nous avons choisi le critère R, rapport des distances inter-classes et intra-classe ($R = d_{inter}/d_{intra}$), calculé par voyelle et sa valeur moyenne Rm.

Soit une distance D_n ($n \in \{1,\dots 17\}$), R^n le vecteur de représentation associé et $C_{[V]}$ la classe vocalique considérée ($V \in \{i, y, u, a\}$) :

- la distance intra-classe est définie par :

$$d_{intra}^{n\,[V]} = \sum_{i \in C_{[V]}} \left[\sum_{j \in C_{[V]},\, j \neq i} D_n\,(R_i^n, R_j^n) \right]$$

- la distance inter-classes est définie par :

$$d_{inter}^{n\,[V]} = \sum_{V' \neq V} \left[\sum_{i \in C_{[V]}} \left[\sum_{j \in C_{[V']}} D_n\,(R_i^n, R_j^n) \right] \right]$$

Ce critère (R) peut être appliqué directement aux dissimilarités (R_d) ou aux projections dans le plan des deux premières dimensions (R_p) de l'espace reconstruit par l'analyse MDS-KRUSKAL [KRUSKAL, 64], procédure qui avait été adoptée lors d'une étude précédente [YE & al., 87]. Pour le calcul de R_p, la distance inter-points considérée est la distance euclidienne. Pour chaque distance, les résultats des rapports R_d et R_p sont présentés au tableau 5, par voyelle, par valeur moyenne et par rang (classement établi selon les valeurs décroissantes de R).

n°	Voyelles	[i]		[y]		[u]		[a]		moyenne		rang	
	Distances	Rd	Rp	Rd	Rp	Rd	Rp	Rd	Rp	Rd	Rp	Rd	Rp
1	PLOMP-FFT-KLA	4	5	6	6	5	4	6	9	5	6	16	17
2	PLOMP-FFT-ZWI	5	5	6	7	6	6	7	10	6	7	13	14
3	PLOMP-LPC-KLA	7	10	7	10	10	17	9	12	9	13	7	6
4	PLOMP-LPC-ZWI	8	14	9	15	12	21	10	19	10	17	6	2
5	LLR-LPC	7	7	10	12	16	14	11	11	11	11	2	8
6	CEP-LPC	6	10	6	6	8	10	5	4	6	7	13	14
7	CEP-LPC-LIN	7	11	5	6	9	12	4	4	7	8	10	12
8	CEP-LPC-QUA	7	12	5	6	9	13	4	4	7	9	10	10
9	MFCC-FFT-KLA	8	14	13	12	11	16	12	8	11	13	2	6
10	MFCC-FFT-ZWI	7	8	9	13	9	11	8	12	8	11	9	8
11	MFCC-LPC-KLA	15	13	11	11	22	25	10	12	14	15	1	4
12	MFCC-LPC-ZWI	9	11	10	14	15	22	11	18	11	16	2	3
13	WSM-FFT-KLA	4	4	5	8	6	8	6	9	5	7	16	14
14	WSM-FFT-ZWI	5	5	6	8	7	8	7	11	6	8	13	12
15	WSM-LPC-KLA	5	5	6	10	9	12	7	10	7	9	10	10
16	WSM-LPC-ZWI	7	11	8	21	11	17	8	13	9	15	7	4
17	APS-LPC	9	19	10	17	13	22	12	20	11	20	2	1
	R_{moy}	7.1	9.6	7.7	10.7	10.6	13.9	8	11				

Tableau 5.
Le critère d'évaluation R pour [i], [y], [u] et [a] avant (R_d) et après (R_p) l'analyse MDS-KRUSKAL pour les distances D_n.

Les distances intra-classe nous permettent de porter un jugement sur les stimuli eux-mêmes : c'est-à-dire sur la façon dont évoluent les voyelles lorsque leur lieu de constriction X_c varie. Si l'on se place à un niveau strictement acoustique il est difficile, à partir des trajectoires dans les plans F1/F2 et F2/F3 d'évaluer précisement les stabilités (ou instabilités) relatives de chaque voyelle. On peut dire tout au plus que [u] est la voyelle la plus stable au regard des trois autres. Par ordre décroissant des rapports moyens (R_{moy}) par voyelle (cf. tableau 5), comme par ordre croissant des distances intra-classe moyennes, on retrouve bien [u] comme la voyelle la plus stable, suivie par ordre de stabilité décroissante de [a], [y] et [i], et ceci que l'on utilise les critères R_d ou les R_p.

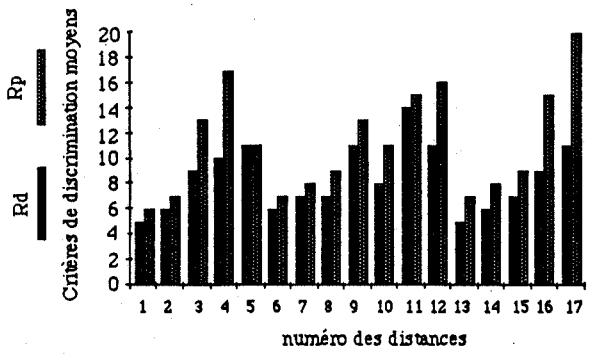


Figure 6.
Pouvoir de discrimination moyen (R_d et R_p) pour les distances D_n (numérotées de 1 à 17).

La figure 6 est une représentation par histogrammes des valeurs moyennes, par distance (numérotées de 1 à 17), des rapports R_d et R_p. Au sens de la **discrimination**, c'est-à-dire du rapport R le plus élevé, les distances les plus performantes sont :

- selon le critère $\mathbf{R_d}$,
  - [11] MFCC-LPC-KLA,
  - [5] LLR-LPC, [9] MFCC-FFT-KLA, [12] MFCC-LPC-ZWI, [17] APS-LPC ;
- selon le critère $\mathbf{R_p}$,
  - [17] APS-LPC,
  - [4] PLOMP-LPC-ZWI,
  - [12] MFCC-LPC-ZWI,
  - [11] MFCC-LPC-KLA, [16] WSM-LPC-ZWI.

3.2. Une interprétation des différences entre R_d et R_p

Notre étude précédente [YE & al., 87] avait montré, en accord avec des résultats obtenus par analyse MDS [LONCHAMP, 78] sur des données perceptives, que l'on pouvait associer le premier plan issu de l'analyse KRUSKAL avec le plan acoustique F1/F2. Sous cette hypothèse, il nous semble possible d'interpréter les différences entre R_d et R_p de la façon suivante :

⇨ Si $\mathbf{R_p \approx R_d}$,

vraisemblablement, la distance utilise bien non seulement la zone basse-fréquence F1-F3 (environ 250-2500 Hz) mais aussi la zone haute-fréquence F3-F5 (2500-5000Hz).

Parmi les distances les plus discriminantes, au sens des critères R_d et R_p, on note dans cette catégorie les distances :

- [11] MFCC-LPC-KLA,
- [9] MFCC-FFT-KLA,
- [5] LLR-LPC.

⇨ Si $\mathbf{R_p > R_d}$,

ce qui est le cas général, il est à remarquer, en effet, que toutes les distances voient leur pouvoir de discrimination augmenter dans le plan des deux premiers axes de projection de l'analyse MDS-KRUSKAL. On peut avancer l'hypothèse que les distances n'utilisent pas au mieux l'ensemble du spectre puisqu'une "limitation" à la zone F1-F2 les "rend plus performantes".

Parmi les distances les plus discriminantes selon R_p, les distances :

- [17] APS-LPC,
- [4] PLOMP-LPC-ZWI,
- [12] MFCC-LPC-ZWI,
- [16] WSM-LPC-ZWI,

semblent ainsi ne pas bien prendre en compte la zone de fréquence F3-F5 puisque leur pouvoir discriminant augmente très nettement après l'analyse MDS-KRUSKAL. Il faut néanmoins souligner que pour des stimuli générés à partir d'un modèle aussi simplifié, les valeurs de F_4 et F_5 ne sont pas très réalistes.

3.3. Représentation et discrimination

Dans l'étude précédente nous avions classé les distances en fonction de leur capacité à respecter la structure du système vocalique du français [i, e, ɛ, a, y, ϕ, œ, u, o, o]. Les **critères** de bonne **représentation** nous avaient permis de noter les distances de -4 à +4 : respect ou non des oppositions antérieur/postérieur, ouvert/fermé, labialisé/non labialisé, extrême/non extrême. Le tableau 7 présente les résultats pour les **critères** de **représentation** et de **discrimination**.

n°	Distances	Représentation [YE et al., 87]	Discrimination moyenne		Discrimination rang	
			Rd	Rp	Rd	Rp
1	PLOMP-FFT-KLA	-3	5	6	16	17
2	PLOMP-FFT-ZWI	-1	6	7	13	14
3	PLOMP-LPC-KLA	3	9	13	7	6
4	PLOMP-LPC-ZWI	3	10	17	6	2
5	LLR-LPC	1	11	11	2	8
6	CEP-LPC	1	6	7	13	14
7	CEP-LPC-LIN	3	7	8	10	12
8	CEP-LPC-QUA	-1	7	9	10	10
9	MFCC-FFT-KLA	1	11	13	2	6
10	MFCC-FFT-ZWI	2	8	11	9	8
11	MFCC-LPC-KLA	-1	14	15	1	4
12	MFCC-LPC-ZWI	2	11	16	2	3
13	WSM-FFT-KLA	-3	5	7	16	14
14	WSM-FFT-ZWI	-4	6	8	13	12
15	WSM-LPC-KLA	-2	7	9	10	10
16	WSM-LPC-ZWI	0	9	15	7	4
17	APS-LPC	4	11	20	2	1

Tableau 7.
Capacité de représentation pour le système vocalique français et pouvoir de discrimination R_d et R_p, par valeur moyenne et par rang, des distances D_n.

La figure 8 représente le comportement des dissimilarités (repérées en abscisse par leur numéro) à la fois en représentation et en discrimination (par rang). A priori, on peut s'attendre à des différences entre les performances de représentation et de discrimination, mais dans l'ensemble, les comportements des distances sont assez bien corrélés pour ces deux types d'évaluation. Quelques exceptions à cette tendance :

- [11] MFCC-LPC-KLA, efficace en discrimination et faible en représentation.
- [7] CEP-LPC-LIN et [6] CEP-LPC, respectivement performante et moyenne en représentation et peu efficaces en discrimination.

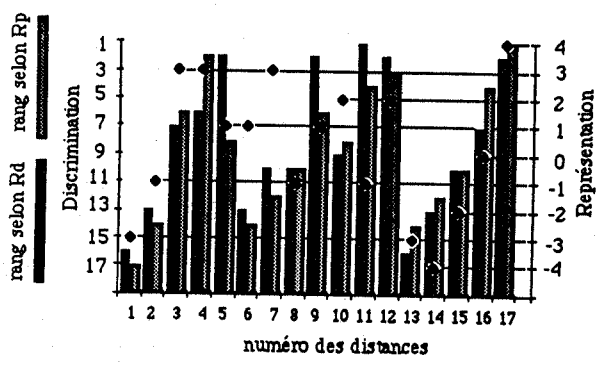


Figure 8.
Les distances présentées en fonction de leur capacité de représentation et de discrimination (par rang selon R_d et R_p).

4. CONCLUSION

Cette étude menée à partir des macrosensibilités des voyelles de type [i, y, u, a] par rapport au lieu d'articulation complète notre premier travail sur les distances [YE & al., 87]. Nous avions testé les distances sur leur capacité à bien respecter les relations structurelles d'un système vocalique. Cette fois-ci la mise à l'épreuve est différente : une mesure objective a été utilisée pour procéder à une évaluation par rapport à un critère de convergence locuteur-auditeur, défini dans le système de communication. Pour obtenir un bon résultat, les distances sont jugées sur leur efficacité à minimiser les distances intra-classe tout en respectant les distances inter-classes. En d'autres termes, le premier test était de **représentation** alors que celui-ci est de **discrimination**. Nous pouvons donc porter un jugement global sur deux comportements essentiels, l'un renvoyant à des exigences phonétiques et l'autre à des exigences de reconnaissance automatique.

Notre interprétation sur les différences entre R_d et R_p doit être confirmée. Il est en effet indispensable d'opérer à partir de stimuli générés à partir d'un modèle plus réaliste. Si les hypothèses d'interprétation après l'analyse MDS-KRUSKAL sont correctes et les tendances en discrimination maintenues, il serait envisageable d'améliorer les distances (en particulier celles présentant un écart important entre les critères R_d et R_p) par une meilleure intégration de l'information en haute fréquence (au-delà du troisième formant) dans la définition même des opérateurs.

On remarquera, enfin, le bon comportement en discrimination des distances dérivées de la LPC relativement à leur homologue dérivée de la FFT et l'efficacité des opérateurs APS et MFCC.

ANNEXE : SPECIFICATION DES DISTANCES

On notera, tout d'abord, l'abus de langage qui consiste à assimiler une mesure de dissimilarité au terme de distance défini en Mathématiques. On se limitera, dans cette spécification des mesures étudiées, à l'espace de représentation considéré et à l'opérateur associé. On indiquera, dans la nomenclature des mesures, l'analyse dont dérive l'espace de représentation (**FFT** ou **LPC**) et, s'il y a lieu, le type d'intégration spectrale (**KLATT** ou **ZWICKER**) [KLATT, 79 ; ZWICKER & FELDTKELLER, 81].

Pratiquement : la fréquence d'échantillonnage est 10 kHz, l'analyse FFT sur 512 points et l'analyse LPC d'ordre 8 (sauf pour la distance APS où l'ordre est 16).

A.1. Distance de PLOMP [PLOMP, 70]

4 variantes : [1] PLOMP-FFT-KLA
[2] PLOMP-FFT-ZWI
[3] PLOMP-LPC-KLA
[4] PLOMP-LPC-ZWI

Espace de représentation : énergie spectrale par bande de fréquences $\{L_i\}$, i=1,...Q.

Opérateur :

$$D_{plomp}(X,X') = \left[\sum_{i=1}^{Q} | L_i - L'_i |^p \right]^{\frac{1}{p}}$$

où, L_i représente le SPL (Sound Pressure Level) correspondant approximativement à l'énergie dans la $i^{ème}$ bande critique, Q est le nombre de bandes critiques et p=1.

A.2. Distance LLR (Log Likelyhood Ratio) [ITAKURA, 75]

[5] LLR-LPC

Espace de représentation : coefficients de prédiction $\{a_k\}$, k=1,...8.

Opérateur :

$$D_{llr}(X,X') = \log \left[\frac{a'^T R a'}{a^T R a} \right]$$

où, a et a' sont les vecteurs des coefficients de prédiction des séquences de test et de référence et R la matrice d'autocorrélation de la séquence-test.

A.3. Distances cepstrales [GRAY & MARKEL, 76]

3 variantes : [6] CEP-LPC
[7] CEP-LPC-LIN
[8] CEP-LPC-QUA

Espace de représentation : coefficients cepstraux $\{c_k\}$, k=1,...8.

Opérateurs :

- **CEP-LPC :**

$$D_{cep}(X,X') = \sum_{k=1}^{N} (c_k - c'_k)^2$$

- **CEP-LPC-LIN** : pondération linéaire ($\alpha = 1$),
- **CEP-LPC-QUA** : pondération quadratique ($\alpha = 2$) :

$$D_{cep^\alpha}(X,X') = \sum_{k=1}^{N} k^\alpha (c_k - c'_k)^2$$

A.4. Distance MFCC (Mel Frequency Cepstrum Coefficients) [DAVIS & MERMELSTEIN, 80]

4 variantes : [9] MFCC-FFT-KLA
[10] MFCC-FFT-ZWI
[11] MFCC-LPC-KLA
[12] MFCC-LPC-ZWI

Espace de représentation : coefficients cepstraux déterminés à partir d'une échelle de fréquences Mel $\{c_k\}$, k=1,...8.

Opérateur : distance euclidienne.

A.5. Distance WSM (Weighted Slope Metric) [KLATT, 82]

4 variantes : [13] WSM-FFT-KLA
[14] WSM-FFT-ZWI
[15] WSM-LPC-KLA
[16] WSM-LPC-ZWI

Espace de représentation : pentes spectrales des spectres lissés $\{S_i\}$, i=1,...Q.

Opérateur :

$$D_{wsm}(X,X') = k_e | E - E' | + \sum_{i=1}^{Q} k_i [S_i - S'_i]^2$$

où, S_i est la dérivée du spectre dans la $i^{ème}$ bande critique, K_e et K_i sont des coefficients pour lesquels nous avons pris (conformément aux auteurs) : $K_e = 0$ et $K_i = 1$. Q est le nombre de bandes critiques.

A.6. Distance APS (Ajustement de Pics Spectraux) [CARATY & RODET, 85 ; CARATY, 87]

[17] APS-LPC

Espace de représentation : ensemble des maxima d'un spectre LPC $\{P_k(f_k,l_k,a_k)\}$, k=1,...8. Les maxima spectraux sont caractérisés par leur fréquence centrale (f en Hz), leur largeur de bande (l en Hz) et leur amplitude (a en dB).

Opérateur : la mesure APS est calculée par Ajustement de Pics Spectraux à partir de distances inter-pics définies par :

$$D_{aps}(P,P') = 2\, p_f(P') \left| \frac{f' - f}{f' + f} \right| + 2\, p_l(P') \left| \frac{l' - l}{l' + l} \right| + p_a(P')\, |a' - a|$$

où, P(f,l,a) et P'(f',l',a') sont deux pics appartenant respectivement au spectre de test et de référence, et p_f, p_l et p_a sont des coefficients de pondération (fonction du pic-référence P') fondés sur les seuils différentiels perceptifs des formants (en fréquence, largeur de bande et amplitude).

REFERENCES BIBLIOGRAPHIQUES

ATAL B.S., CHANG J.J., MATHEWS M.V. & TUCKEY J. (1978)
"Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Sorting Technique." JASA, n° 63, pp. 1535-1555.

BADIN P. & BOE L.J. (1987)
"Vocal Tract Vocalic Nomograms : Acoustic Considerations." XIth ICPS, pp. 352-355.

BOE L.J. & ABRY C. (1986)
"Nomogrammes et systèmes vocaliques." 15èmes JEP, Aix-en-Provence, pp. 303-306.

BOE L.J. & PERRIER P. (1988)
"C.F. HELLWAG 200 ans après ou les éléments d'une fibre conductrice." 17èmes JEP, Nancy.

CARATY M.J. & RODET X. (1985)
"Distance interspectrale à critères perceptifs." 14èmes JEP, Paris, pp. 87-90.

CARATY M.J. (1987)
"Contribution au décodage acoustico-phonétique : études de distances inter-spectres et reconnaissance de cycles vocaliques." Thèse de Doctorat de l'Université Paris 6.

CARLSON R., FANT G. & GRANSTROM B. (1975)
"Two-Formant Models, Pitch and Vowel Perception."
in Auditory Analysis and Perception of Speech. Academic Press, eds. FANT & TATHAM, pp. 55-82.

CHARPENTIER F. (1986)
"Fonctions de sensibilité d'un modèle dissipatif." Bulletin de l'Institut de Phonétique de Grenoble, n° 15, pp. 1-33.

DAVIS S. B. & MERMELSTEIN P. (1980)
"Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences." IEEE ASSP-24, pp. 357-366.

FANT G. (1960)
Acoustic Theory of Speech Production. Mouton, The Hague.

FENG G. (1986)
"Modélisation Acoustique et Traitement du Signal de Parole." Thèse de Doctorat de l'INP Grenoble.

GAY T., LINDBLOM B. & LUBKER J. (1981)
"Production of Bite-Block Vowels : Acoustic Equivalence by Selective Compensation." JASA, n° 69, pp. 802-810.

ITAKURA F. (1975)
"Minimum Prediction Residual Principle Applied to Speech Recognition." IEEE ASSP-23, pp. 67-72.

KLATT D. H. (1979)
"A Digital Filter Bank for Spectral Matching." IEEE ICASSP, pp. 573-576.

KLATT D. H. (1982)
"Prediction of Perceived Phonetic Distance from Critical-Band Spectra : a First Step." IEEE ICASSP, pp. 1278-1281.

KRUSKAL J.B. (1964)
"Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." Psychometrika, n° 29, pp. 1-27.
"Multidimensional Scaling : A Numerical Method." Psychometrika, n° 29, pp. 115-129.

LADEFOGED P. & BLADON A. (1982)
"Attempts by Human Speakers to Reproduce Fant's Nomograms." Speech Comunication 1, pp. 185-198.

LONCHAMP F. (1978)
"Recherche sur les indices perceptifs des voyelles orales et nasales. Application à la structure du système vocalique français et de diverses autres langues." Thèse de Doctorat de 3ème cycle en Phonétique, Nancy II.

MAEDA S. (1979)
"Un modèle articulatoire de la langue avec des composantes linéaires." 10èmes JEP, Grenoble, pp. 152-162.

MANTAKAS M., SCHWARTZ J.L. & ESCUDIER P. (1986)
"Modèle de prédiction du deuxième formant effectif F_2' et application à l'étude de la labialité des voyelles avant du français." 15èmes JEP, Aix-en-Provence, pp. 157-161.

PERRIER P., BADIN P. & BOE L.J. (1987)
"Nomogrammes du conduit vocal par modèlisation articulatoire." 16èmes JEP, Hammamet, pp. 124-127.

PETITOT-COCORDA J. (1984)
Les catastrophes de la parole. Maloine, Paris.

PLOMP R. (1970)
"Timbre as a Multidimensional Attribute of Complex Tones."
in Frequency analysis and periodicity detection in hearing. Ed by PLOMP R. & SMOORENBURG G. F, Sijthoff, Leiden, pp 397-414.

RAJAA M., BOE L.J. & PERRIER P. (1986)
"Fonction de sensibilité, modèle articulatoire et voyelles du français." 15èmes JEP, Hammamet, pp.59-63.

STEVENS K.N. & HOUSE A.S. (1955)
"Development of a Quantitative Description of Vowel Articulation." JASA, n° 27, pp. 484-493.

YE H., TUFFELLI D. & BOE L.J. (1987)
"Etude du comportement phonétique des dissimilarités." 16èmes JEP, Hammamet, pp. 5-9.

ZWICKER E. & FELDTKELLER R. (1981)
Psycho-Acoustique. "L'oreille, récepteur d'information." Traduit par C. SORIN, Masson, Paris.

CONVERSION GRAPHEMIQUE-PHONETIQUE AVEC VARIANTES DU FRANCAIS PAR REGLES

A.Dujour, M.Eskénazi

LIMSI/CNRS-BP3091406 ORSAY CEDEX-FRANCE

ABSTRACT

It cannot be reasonably expected that, on a phonological level, users of speech recognition systems will adapt themselves to a "standard" pronunciation. Phonological variability must therefore be taken into account. We have used a rule-based text-to-phoneme translator to predict possible phonological variants in a speaker-independant, continuous speech mode.
This system, VARION, was tested on transcriptions of recorded speech (BDSONS) to verify if the variations found were correctly predicted by the system. Results for 30 speakers are given, showing a very high correlation between predicted and spoken strings of phonemes.

1. INTRODUCTION

Parmi les problèmes qui rendent le traitement automatique de la parole délicat, les difficultés liées à la variabilité, omniprésente en parole continue, ne sont toujours pas résolues à l'heure actuelle. Or, étant donné qu'il est irréaliste de penser pouvoir imposer à un utilisateur de technologie vocale une prononciation standard, un système de reconnaissance doit être doté de moyens nécessaires pour la prise en compte des variantes de prononciation. De même en synthèse, il est souhaitable de pouvoir reproduire, du moins partiellement, cette variabilité de la parole. Pour ce faire, différents niveaux d'analyse peuvent intervenir : niveau acoustico-phonétique, niveau syntaxique, niveau phonologique...
Dans les laboratoires travaillant sur la parole, ont pu être réalisés des systèmes de transformations graphémique-phonétique que nous appellerons déterministes : à une entrée graphémique donnée correspond une et une seule suite phonétique possible. Ainsi, il existe pour la langue française: GRAPHON au LIMSI [Prouts 80], TOPH à l'Institut Phonétique de Grenoble [Aubergé 85], le système de Divay et Guyomard au CNET [Divay 77], etc. Si, dans une optique orientée synthèse, ce déterminisme est nécessaire, il ne semble pas réaliste en reconnaissance de la parole.
Nous présentons dans cet article un module de règles de transformation graphémique-phonétique dont l'objectif est de rendre compte de variantes de prononciation en parole continue. Nous décrivons les outils utilisés pour la création de ce module, son fonctionnement et ses limites.

2. LE MODULE DE TRANSCRIPTION GRAPHEMIQUE-PHONETIQUE

Deux modules de transcription graphémique-phonétique ont été réalisés au LIMSI, l'un GRAPHON, pour la synthèse à partir du texte, l'autre, GRAPHER, dans le cadre des recherches sur les systèmes de compréhension de la langue parlée [Néel 82]. Conçu principalement pour pallier les erreurs de segmentation des systèmes de reconnaissance, telles que les élisions, les substitutions et les insertions, ce module ne présentait pas de façon exhaustive toutes les variantes phonétiques possibles.
Il était donc nécessaire, dans le cadre de la reconnaissance de la parole continue multilocuteur, qu'un troisième module soit créé, généralisant et affinant l'étude de ces variantes. Nous avons réalisé ce module de 1330 règles, VARION, fonctionnant sur le logiciel FONPARS (contrat Esprit 860) [Senders 86].

2.1. Outils préliminaires : constitution d'une base de donnée

Notre dessein étant de rendre compte des variantes de prononciation dans une communauté langagière donnée, il nous a paru tout à fait arbitraire de prendre comme références linguistiques uniquement notre prononciation et celle de notre entourage. Nous avons donc constitué un corpus de 150 phrases axé sur les problèmes que nous voulions traiter. Nous avons fait lire cet ensemble de phrases à 63 locuteurs natifs de la région parisienne et représentatifs d'un français "non marqué", 33 hommes et 30 femmes, de 10 à 70 ans. Sachant que tout est relatif et que le niveau social n'est pas toujours représentatif d'un certain niveau de langue, nous avons essayé dans la mesure du possible, de choisir différentes catégories sociaux-professionnelles (professions libérales, ouvriers, cadres etc) illustrant trois niveaux de langue : niveau familier, niveau normal et niveau soutenu. Nous avons indiqué à chaque sujet que ces enregistrements allaient être utilisés par un système de reconnaissance, nous avons donc exigé d'eux une lecture claire et soignée tout en restant naturelle et expressive.
A partir de l'examen de ces enregistrements (environ 20 heures de parole) nous avons établi des matrices fréquentielles sur les différentes variantes étudiées en fonction des contextes. C'est à partir de ces matrices que nous avons écrit les règles de VARION.

2.2. Le module de règles

Alors que GRAPHON et GRAPHER ont été écrits sur un support identique, à base de règles de production [Prouts 80], nous avons employé, pour écrire VARION, le programme FONPARS, logiciel réalisé à l'université de Nijmegen (Hollande) dans le cadre du contrat Esprit 860 (analyse linguistique des langues européennes).
D'une part, ce logiciel offrait la possibilité de tester notre module rapidement, d'autre part une partie du présent travail s'inscrivait dans le cadre de ce contrat.
La notion de prononciation normative nous paraissant relativement arbitraire, nous avons préféré choisir pour notre dictionnaire des entrées graphémiques plutôt que des segments phonétiques à partir desquels pourraient être engendrées plusieurs formes dérivées. Aussi, chaque élément du dictionnaire se compose d'une chaîne graphémique de longueur variable en entrée, d'une chaîne phonétique en sortie et, éventuellement, des contextes gauches et droits conditionnant la sortie phonétique. Comme pour GRAPHON et GRAPHER, l'exploration des règles se fait séquentiellement. Toute règle dont la chaîne graphémique est un sous ensemble de la chaîne de graphèmes d'une autre règle, ne doit pas être classée avant cette dernière. Mais, contrairement à GRAPHON et GRAPHER où la phase de traduction consiste à examiner de gauche à droite la chaîne à traduire, ici la transcription est effectuée dans l'ordre des règles rencontrées, quelle que soit la place de l'unité graphémique sur laquelle opère la règle dans l'ensemble du syntagme à traduire.

Ainsi, étant donné les règles:

1- le graphème *t* tombe quand il est suivi de *t*

2- le graphème *e* devient /ɛ/ quand il est suivi d'une double consonne ou de deux consonnes différentes

3- le graphème *e* devient e caduc quand il n'est pas en finale de mot

et le mot à phonétiser *cette*. Du fait de l'application de la règle n°1, qui transforme *cette en cete*, la règle n°3 sera déclenchée au lieu de la règle n°2. On aura donc une représentation phonétique erronée du déterminant, qui est /sət/ au lieu de /sɛt/.

Ceci pose bien sûr un important problème d'ordonnancement des règles et nous avons eu beaucoup de mal à éviter cet écueil. Nous avons été obligée de changer l'ordre des règles imposé par GRAPHON et GRAPHER. Le nombre de règles a été augmenté de façon non négligeable, c'est d'ailleurs pourquoi le chiffre de 1330 règles dont nous avons parlé plus haut ne doit pas être considéré comme un chiffre absolu, mais comme un chiffre relatif au fonctionnement de FONPARS. Enfin, nous avons dû parfois avoir recours à un contexte semi-graphémique, semi-phonétique dans l'écriture des règles.

Par exemple, soient les règles:

1- *ai* devient /ɛ/

2- *o* devient /ɔ/ quand il est suivi de *mme* et blanc

3- *c* devient /k/ quand il est suivi de o,a,u

4- *c* devient /s/ dans tous les autres cas

et les mots à phonétiser *caisse* et *comme*, pour qu'ils soient bien transcrits; il faudrait que la règle n°3 ait comme contexte droit : a, u, o, /ɛ/, /ɔ/.

2.3. Pourquoi un module et non un lexique

Contrairement à l'anglais par exemple, où la variabilité se manifeste principalement à l'intérieur des mots, dans la langue française elle se produit aussi bien en frontière de mots qu'à l'intérieur des lexèmes. Or, dès qu'on s'intéresse à la parole continue, la plupart des variations de prononciation concernent trop de mots à la fois pour qu'on puisse facilement les répertorier dans un lexique. Cette représentation serait lourde à mettre en oeuvre, redondante, coûteuse en mémoire et en temps de calcul et de ce fait sans doute trop restrictive. Il nous paraît plus adéquat de faire correspondre à une chaîne graphémique de longueur variée une ou plusieurs sorties phonétiques. Par exemple, la suite graphémique *dre* admet différentes prononciations selon le phonème qu'il précède. Les règles :

- *dre* en finale de mot, devient /drə/ ou /d/ quand il précède une sonore, /drə/ ou /t/ quand il précède une sourde, /dr/ ou /d/ quand il précède une voyelle,

permettent de traiter correctement aussi bien les suites *prendre du pain*, *descendre ta valise* que *prendre un avion* et *descendre à Paris*. Ainsi, ne considérer comme unité graphémique à traiter que la suite *dre* permet de limiter considérablement le nombre de cas à entrer dans le dictionnaire.

2.4. Les variantes étudiées

Afin de compléter GRAPHER, qui prenait en compte certains cas de variantes libres liées aux habitudes langagières des locuteurs comme la liaison facultative, la chute du e caduc ainsi que le traitement de certains homographes hétérophones, nous avons axé notre étude sur l'assimilation consonantique en frontière et en milieu de mot (*maintenant* -> /mɛ̃tənа/, /mɛ̃tna/, /mɛ̃dna/ ou /mɛ̃nna/) , la chute de certaines consonnes en finale absolue (*week end*-> /wik end/ ou /wikɛn/), les prononciations de l'archiphonème *gn*, (*oignon*->/onjõ/ ou /oɲõ/), le *l* final de certains mots comme *nombril*, *persil*, certains cas de diérèse ou de synérèse du yod (*il y a*-> /iliya/, /ilya/ ou /ya/), la liaison facultative devant un h aspiré (*des handicapés* -> /deza.../ ou /deã.../), enfin l'harmonisation vocalique des archiphonèmes *o* et *e* en position inaccentuée (*étude*-> /etyd/ ou /ɛtyd/.

3. LES TESTS DU MODULE

Considérant la nature de notre étude, il nous est apparu essentiel de tester VARION sur des réalisations individuelles et non pas simplement sur des transcriptions normatives comme c'est souvent le cas.

Pour ce faire, nous avons choisi dans un premier temps de tester VARION sur les prononciations de *La bise et le soleil*, corpus utilisé par le GRECO COMMUNICATION PARLEE dans le cadre de la base de données BDSONS. Nous avons retenu ce texte car, d'une part, il s'agit de parole continue, d'autre part, il a été prononcé par 32 locuteurs, enfin il est disponible dans tous les laboratoires francophones, ce qui permet d'éventuelles comparaisons entre différents systèmes. Sur les 32 prononciations de *La bise et le soleil* uniquement 30 étaient disponibles au LIMSI, lors de nos tests. Les notations ont été effectuées manuellement, notons toutefois leur limites. En effet, seulement quatre prononciations ont fait l'objet de comparaisons entre plusieurs transcriptions, les autres ayant été traitées par une seule spécialiste. Nous envisageons donc, d'une part, de soumettre ces enregistrements à différents experts afin de systématiser les comparaisons et, d'autre part, d'avoir recours à des analyses temporelles et fréquentielles des signaux enregistrés dans le but de compléter les informations fournies par l'oreille humaine.

Nous avons également testé nos règles sur cinq textes, articles de journaux et récits littéraires, de 131 à 348 mots dans le but de comparer la transcription de VARION à celle de GRAPHON et GRAPHER, c'est-à-dire afin de valider la transcription de VARION.

3.1. Résultats

Dans *La bise et le soleil* nous avons considéré qu'un graphème ou qu'une suite de graphèmes étaient mal transcrits si, parmi les suites possibles fournies par VARION, on ne trouvaient pas la suite de phonèmes prononcée par un locuteur.

Pour les textes écrits, nous avons comptabilisé les substitutions, les insertions et les élisions.

Les combinaisons phonétiques exactes ont été calculées de la manière suivante:

- taux de reconnaissance dans *La bise et le soleil* = nombre de phonèmes bien transcrits/nombre total de phonèmes pour chaque réalisation (fig. 1 et 2)
- taux de bonne transcription dans les textes écrits = nombre de mots bien transcrits/ nombre total de mots pour chaque texte (les erreurs répertoriées ne concernent jamais plus qu'un graphème du mot mal transcrit).

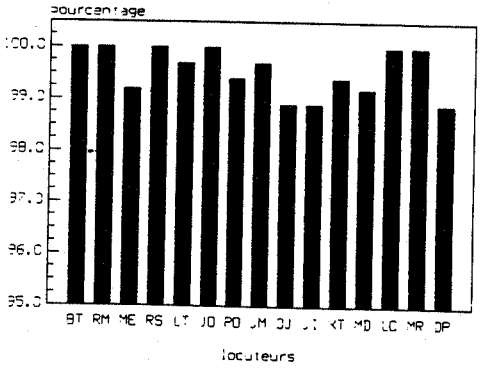


Figure 1 : test1 locuteurs BDSONS

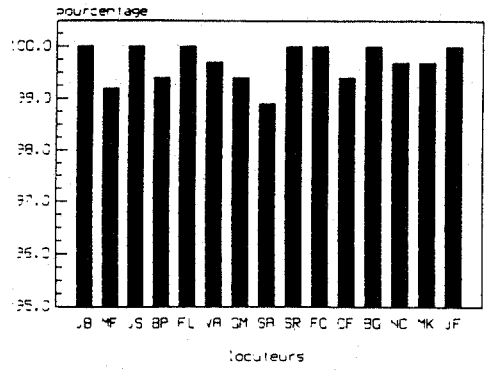


Figure 2 : test2 locuteurs BDSONS

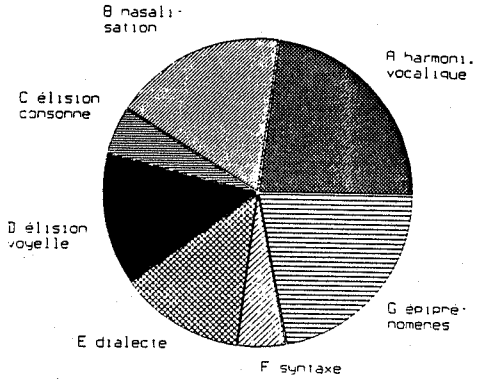


Figure 3 : classes d'erreurs répertoriées dans *La bise et le soleil*

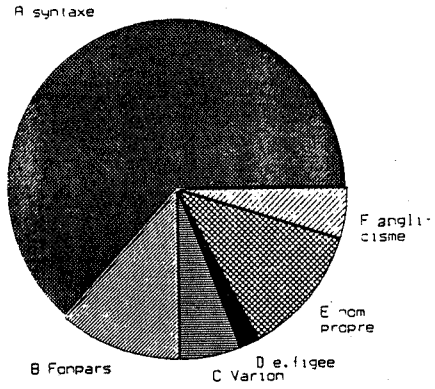


Figure 4 : classes d'erreurs répertoriées dans les textes écrits

3.2. Discussion des résultats

A la lumière des résultats, plusieurs remarques s'imposent. D'une façon générale, dès qu'on se situe en parole continue, le nombre de combinaisons phonétiques, c'est-à-dire le nombre de suites phonétiques possibles générées par VARION est très important. Or, il apparaît que l'occurrence d'un allophone varie en fonction de deux facteurs : d'une part des phénomènes indépendants des locuteurs, telles que les contraintes syntaxiques, prosodiques (nous parlerons dans ce cas de variabilité extralocuteur), d'autre part des phénomènes liés aux locuteurs et à leurs habitudes langagières.

Parmi les 30 réalisations de *La bise et le soleil* (fig.1 et 2), les règles fournissent des résultats peu satisfaisants pour quatre locuteurs, dont trois sont originaires du sud de la France (DP, DJ, JI). Leur prononciation et certaines variantes qu'ils emploient sont des caractéristiques régionales et ne sont pas traitées par notre module. Rappelons en effet que nous nous sommes limitée exclusivement lors de l'élaboration de notre dictionnaire à traiter des cas de variantes propres à "la communauté" parisienne. Ainsi, la suite *quand ils* est prononcée /kɛ̃til/ par ces trois locuteurs. Quant au dernier locuteur, SA, précisons qu'il est connu pour être un locuteur difficile, et posant de sérieux problèmes, dans d'autres travaux ayant utilisé cette base de données.

Concernant les classes d'erreurs répertoriées dans la fig.3, aucune d'entre elles n'est attribuable au déclenchement de mauvaises règles. Ces erreurs, ou plutôt ces omissions, sont liées aux limites que nous nous sommes imposée quant aux variantes à étudier. Parmi ces omissions, 23% relèvent de l'harmonisation vocalique et principalement nasale . Ainsi, la séquence *son manteau* est prononcée par trois locuteurs /sɔ̃mɔ̃to/. 18% concernent la nasalisation, par exemple le participe passé *commencé* est réalisé /kɑ̃mɑ̃se/ par deux locuteurs . 19% sont liées aux élisions de consonnes et de voyelles . Ainsi, la suite *à le lui* peut être prononcé /allɥi/ et le syntagme *briller et au* /brijeo/. Par ailleurs, une bonne partie des variantes omises sont liées à des mécanismes dialectaux précis (partie E). Ainsi, la locutrice ME, originaire des Etats-Unis a tendance à prononcer ses e caduc /e/ ou /ɛ/, ce qui est le cas dans *arriverait* et *enveloppé*. Les locuteurs du sud peuvent faire varier le lieu d'articulation de la nasale /ɑ̃/. (*quand ils* réalisé /Kɛ̃til/). Enfin, nous considérons les autres variantes non prévues (hormis la partie G) comme des épiphénomènes, c'est-à-dire non explicables phonologiquement et ne faisant pas intervenir de règles phonologiques précises ; par exemple, le dévoisement d'une sonore alors qu'elle n'est pas en contact avec un phonème sourd (*la bise a* -> /labisa/) ou au contraire la sonorisation d'une sourde dont le contexte est vocalique (*oté son* -> /otezɔ̃/).

Ici, une question se pose : ces variantes de prononciation sont-elles des accidents de prononciation non reproductibles systématiquement, ou sont elles caractéristiques de certains locuteurs?

La partie F de la fig.3 et la partie A de la fig.4 reflètent les limites d'un module uniquement phonologique dont sont absentes les considérations syntaxiques lexicales et prosodiques. En effet, FONPARS ne permet pas d'avoir recours à des procédures spéciales pour traiter certains lexèmes dont la prononciation varie en fonction de leur rôle syntaxique et/ou sémantique dans des occurrences données. Ainsi, dans le corpus du GRECO, l'adverbe *plus* est mal transcrit dans les deux occurrences du syntagme *le plus fort*: /ply/ */plys/. De même, les tests effectués sur les textes écrits révèlent 62% de fautes liées à l'absence de traitement syntaxique. Parmi ces erreurs 70% portent sur le lexème *est* qui, en fonction de sa nature (nom masculin ou auxiliaire) se prononce /ɛst/, /e/ ou /ɛ/.

En examinant plus précisément la fig.4 on voit que sur les 41 erreurs enregistrées dans l'ensemble des textes, certaines d'entre elles sont imputables au mécanisme de déclenchement des règles (partie B). Soient les règles:

1- eau devient /o/ quand il est suivi de x en finale
2- c devient /s/ quand il précède la voyelle *e*

et le mot à phonétiser *pinceaux*, on aura comme sortie phonétique /pɛ̃ko/. 13% des mots mal transcrits sont des noms propres, 2% sont des expressions figées et 5% sont des anglicismes. Enfin, 6% sont la conséquence du déclenchement d'une mauvaise règle (partie C): le graphème *T* a comme correspondant phonétique /t/ dans des mots tels que *sommations, éditions...*

4. CONCLUSION

Nous avons présenté un module de transformation graphémique-phonétique qui semble relativement satisfaisant quant aux résultats observés. Malgré tout, le nombre trop important de suites phonétiques possibles généré par VARION révèle qu'il est irréaliste, dans l'état actuel des connaissances, de penser pouvoir traiter une variabilité multilocuteur absolue, il serait donc nécessaire d'analyser avec précision quelles sont les sources de variation en parole, ce qui distingue les différents types de variantes entre elles et de quelle façon elles se manifestent. Ceci dans le but sinon d'éliminer les épiphénomènes, du moins de réduire le nombre de variantes envisagées. Une étude sur le débit et les dialectes semble être un bon point de départ pour réaliser ce dessein, notamment afin de contribuer à faire la part entre les variantes individuelles et les variantes dialectales. Pour cela, des descriptions exhaustives des variantes de prononciation, ainsi que des données statistiques nombreuses devraient faciliter le choix des compromis.

BIBLIOGRAPHIE

Aubergé, V., Contini, M., Maret, D., Schnabel, B., 1987 : "TOPH : un outil de phonétisation multilangue". Bulletin de l'Institut Phonétique de Grenoble, volume 16, pp. 155-176.

Cohen, M., Bernstein, J., Murveit, H., 1987 : "Pronunciation variation within and across speakers". (Speech research program, SRI international, Menlo Park, CA 94025) in THE JOURNAL OF THE ACOUSTICAL SOCIETY, volume 81, supplement 1, pp.S6-S7.

Divay, M., Guyomard., M., 1977 : "Contribution et réalisation d'un programme de transcription". Thèse de troisième cycle, Université de Rennes, 219 pages.

Dujour, A., 1987 : "Conversion graphème-phonème avec variantes du français par règles". Mémoire de DEA, Université de Paris VII, 63 pages.

Kerkhoff, J., Wester, J., 1987 : "Fonpars1 user manual". Internal publication for Esprit-project 860, Institute of Phonetics, Nijmegen University, 44 pages.

Néel, F., Eskénazi, M., Mariani, J., 1982 : "Etiquetage automatique du signal de parole continue à partir de sa transcription orthographique". dans FASE-DAGA 82. Troisième congrès de la fédération européenne des sociétés d'acoustique, pp. 919-922, Gottingen.

Néel, F., Eskénazi, M., Mariani, J., 1986 : "Module de traduction phonétique avec variantes". dans LEXIQUE ET TRAITEMENT AUTOMATIQUE DES LANGAGES, séminaire GALF GRECO, Toulouse, pp.129-137.

Laporte, E., 1987 : "Prise en compte des variations phonétiques en reconnaissance de la parole". 16 ème JEP du GALF, Hamammet, pp.153-156.

Pérennou, G., De Calmès, M., 1987 : "Bdlex (base de données lexicales du français écrit et parlé)", travaux du laboratoire Cerfia (UA CNRS 824), Action du Greco de la communication parlée. 25 pages.

Prouts, B., 1980 : "Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur". Thèse de docteur ingénieur, Université de Paris XI, 148 pages.

Senders, S., Bloemberg, V., 1987 : "An architecture for a phoneme to grapheme system based on hidden chains". Esprit Report UN-AR2, project 860. Institute of Phonetics University of Nijmegen. 51 pages.

Analyse de deux enquêtes sur l'évaluation des systèmes de reconnaissance à grand vocabulaire et du décodage phonétique de la parole continue.

C. Bourjot*, A. Boyer*, G. Pérennou**,
J.-P. Tubach***, N. Vigouroux**

* Laboratoire CRIN-INRIA, BP 239, Bd des Aiguillettes, 54 506 Vandœuvre - les -NANCY cédex,
** Laboratoire CERFIA, 118, Rte de Narbonne, 31062 TOULOUSE cédex,
*** Laboratoire ENST, 46, rue Barrault, 75013 PARIS cédex.

ABSTRACT

The ESPRIT SAM project n°1541 plans to define common tools and standard databases to assess speech input or output systems. Three french laboratories, CERFIA, CRIN, ENST, are involved in large vocabulary recognition and phonetic decoding assessment. Therefore, they design 2 questionnaires in order to overview the state of the art in these fields. These questionnaires aim also to precise the needs for a reliable and secure assessment and to precise it in terms of databases, criteria and methodologies. The following paper reports the main results obtained in the collected answers.

INTRODUCTION

Le GRECO de la Communication Parlée, en tant que partenaire du projet SAM —Multi-Lingual Speech Input-Output Assessment Methodolgy and Standardisation, ESPRIT Project 1541— participe à la spécification et à la mise en place d'un ensemble de moyens (outils normalisés et bases de données multilingues) pour l'évaluation des interfaces vocales [Dolmazon,1987a].

Les rapports suivants, établis lors de la phase de définition du projet, font le point sur les divers aspects du problème à résoudre : Synthèse à partir du texte [Pols,1987], Postes de travail [Dolmazon,1987b], Bases de données [Winski,1987], Alphabet phonétique pour l'évaluation [Wells,1987], Méthodes d'étiquetage [Tomlinson,1987], Méthodes d'évaluation des systèmes de reconnaissance [Campo,1987], Méthodologies d'évaluation proposées par le Danemark [Dalsgaard,1987a].

L'évaluation des cartes de reconnaissance existantes a déjà donné lieu de nombreux travaux. L'approche la plus directe consiste à définir une méthodologie de tests de reconnaissance globale et à tenir compte de divers critères économiques, ergonomiques et techniques pour vérifier l'adéquation au poste de travail envisagé.

L'article de [Lea,1983] *"Selecting the Best Speech Recognizer for the Job"*, est particulièrement représentatif de cette lignée de travaux (qui constitue aussi l'un des axes du projet SAM). Connaître le comportement d'un système de reconnaissance pour une tâche donnée ne suffit pas toujours. Il faut aussi pouvoir comparer les systèmes concurrents sur des bases objectives. Dans le *Journal of the National Bureau of Standard*, [Pallett,1985] indique qu'il n'y pas encore de consensus sur les vocabulaires de test permettant d'effectuer de telles comparaisons (la situation n'a pas changé à cet égard). L'une des tendances actuelles consiste à tester l'aptitude des systèmes de reconnaissance à discriminer les oppositions phonétiques fondamentales.

Une autre possibilité consiste à trouver des mesures objectives de la difficulté des vocabulaires. [Moore,1977] suggère de procéder en utilisant des mesures de rapport signal/bruit pour une intelligibilité équivalente. Il pourrait ainsi être possible de donner une mesure de difficulté spécifique de chaque opposition phonétique et, peut-être, de définir une approche multilingue du problème.

Des lacunes importantes ressortent de l'examen de ces travaux en ce qui concerne l'évaluation des systèmes dès qu'il ne s'agit pas de cartes "classiques" de reconnaissance de mots isolés ou enchaînés. Par ailleurs, les performances des étages de décision phonétique ne semblent pas faire l'objet d'évaluations normalisées. Certes, les tests de discrimination sur des vocabulaires où sont représentées les oppositions phonétiques fondamentales apportent des informations intéressantes à cet égard [Simpson ,1987a,b]. Mais ce ne sont que des informations indirectes ne renseignant pas de manière précise sur la nature des fautes commises —insertion, omission, substitution—.

Il a donc semblé qu'une enquête serait utile pour préciser ces différents points. C'est le but d'un premier questionnaire "Large Vocabulary Recognition Assessment" établi par le CRIN, qui développe une méthodologie d'évaluation des *Systèmes de Reconnaissance à Grand Vocabulaire* (SRGV) [Bourjot,1988]. L'analyse de ce premier questionnaire fait apparaître que les problèmes d'évaluation soulevés doivent faire l'objet d'une enquête plus détaillée.

A priori, il y a tout lieu de penser que les moyens d'évaluation phonétique nécessaires pourront s'adresser à divers types de systèmes, qu'ils soient ou non à grand vocabulaire. Il a donc paru nécessaire de compléter notre information au moyen d'un deuxième questionnaire "Phonetic Recognition of Continuous Speech and its Assessment", joint au premier, établi par le CERFIA et l'ENST. Les résultats du dépouillement de ce deuxième questionnaire sont discutés en deuxième partie de cette communication.

Les deux questionnaires ont été présentés au meeting SAM du 4 novembre à Turin; le principe de l'enquête y a été confirmé et les différents partenaires se sont engagés à diffuser les questionnaires auprès des laboratoires de recherche industriels et publics dans leur pays. Au 31 décembre 1987, le nombre de réponses reçues était de 16 pour les laboratoires publics et 5 pour les industriels.

La répartition des réponses par pays est la suivante: 9 pour la France, 3 pour le Royaume-Uni, 1 pour l'Italie, 2 pour la Belgique, 1 pour l'Espagne, 1 pour le Danemark, 2 pour l'Allemagne et 2 pour les Pays-Bas. On observe qu'elles sont proportionnellement plus importantes en France où la diffusion et le suivi ont sans doute été mieux assurés.

L'enquête devra donc certainement être complétée pour toucher plus de laboratoires. Par ailleurs, l'examen des réponses suggère de nouvelles questions plus précises sur la définition de corpus de test et les protocoles d'évaluation, ... Il fait aussi apparaître la nécessité de s'adresser à une classe plus grande de systèmes de reconnaissance. Mais tels qu'ils sont, les résultats actuels sont déjà intéressants par rapport aux questions que nous nous posions.

1ère PARTIE: EVALUATION DES SYSTEMES A GRAND VOCABULAIRE

1. Le questionnaire

Une évaluation de qualité des SRGV ne peut être réalisée de la même manière que pour les systèmes à petit vocabulaire compte-tenu des spécificités de ces systèmes : organisation en différents niveaux de connaissances (phonétique, lexical, syntaxique, sémantique, ...) et taille des bases de données nécessaires pour obtenir un taux de reconnaissance significatif.

Les trois aspects de l'évaluation : comparaison de systèmes, diagnostic des erreurs commises au cours du processus de reconnaissance, prédiction des performances dans des conditions autres que celles où le système de reconnaissance a été testé, imposent une connaissance précise et détaillée des mécanismes mis en oeuvre lors des processus d'apprentissage et de reconnaissance. Une étude bibliographique fournit une première base de connaissances des SRGV (voir la bibliographie "Systèmes de reconnaissance à grand vocabulaire").

Le questionnaire "Large vocabulary recognition assessment " visait à compléter ces informations en permettant de répondre à trois questions majeures :

- Quels sont actuellement les systèmes existant ou en cours de développement en Europe ?
- Qu'est-ce qu'une évaluation fiable et efficace des performances de ces systèmes ?
- Est-il intéressant d'évaluer séparément le niveau de décodage phonétique dans le cas où il fait partie du processus de reconnaissance ?

Ces trois problèmes importants étaient développés dans quatre sections du questionnaire qui sont :
- description du système existant ou en phase de développement,
- description du protocole d'évaluation utilisé actuellement,
- description d'une proposition d'évaluation,
- nécessité d'une évaluation du décodage phonétique.

2. Synthèse des résultats

2.1. Description des systèmes existants

Parmi les réponses obtenues, 13 laboratoires travaillent sur des vocabulaires de plus de 1000 mots. Voici les résultats qui les concernent.

Type du système

Type	Réponses
finalisé	8
langage non restreint	5

Taille du vocabulaire :
de 1000 à 5000 mots

Unité de base pour la Reconnaissance
(réponses multiples)

phrase	3
mot	7
syllabe	2
demi-syllabe	0
diphone	1
phoneme	9

Type de parole acceptée
(reponses multiples)

parole continue	9
mot connecté	3
mot isolé	4
syllabe isolée	1

Locuteur

mono-locuteur	5
multi-locuteur	5
independant du locuteur	3

Representation des unites de reconnaissance
(réponses multiples)

HMM	6
trait acoustique	3
forme	2
automate	1
autre	1

Methodes de reconnaissance
(réponses multiples)

système à base de connaissances	5
algorithme de Viterbi	5
DTW	4
classification statistique	2
autre	1

Paramétrisation

FFT	6
cepstre	4
banc de filtres	1
LPC	1
autre	1

Codage

quantification vectorielle	8
pas de quantification	4
quantification matricielle	1

Clustering

K-means	4
non	9

Sources de connaissances utilisées
(réponses multiples)

phonétique	10
syntaxique	10
lexicale	10
phonologique	7
sémantique	5
pragmatique	4
prosodique	2

Apprentissage
nombre de locuteurs de 0 à 20
durée de parole de 400 à 2000 s

Taux de reconnaissance
de 50% à 100 % (bien sûr)

commentaire

- Les systèmes finalisés sont les plus nombreux. Les applications recensées sont très variées (militaire, machine à dicter, interrogation de base de données, ...). Il s'avère donc délicat de spécifier des corpus de parole normalisés pour la comparaison des systèmes.
- Actuellement, la taille des vocabulaires est comprise entre 1000 et 5000 mots. Quelques équipes envisagent de travailler sur 20000 mots. Ainsi il sera nécessaire de prendre en compte la taille du vocabulaire, en distinguant au moment de l'évaluation, les grands vocabulaires (jusqu'à 5000 mots) des très grands vocabulaires (au delà de 5000 mots).
- Le phonème et le mot apparaissent comme les unités de base prépondérantes pour la reconnaissance.

• La multiplicité des types de parole implique la diversité des corpus de parole pour l'évaluation.

• L'enquête ne nous permet pas de savoir quelle est la population "ciblée" lorsque le système est multilocuteur.

• La plupart du temps, les organisations n'ont pas répondu à la question concernant l'adaptation au locuteur.

• Les 4 réponses multiples sur les méthodes de reconnaissance concernent toutes l'algorithme de Viterbi qui est associé à la programmation dynamique, à la classification statistique ou à un sytème à base de connaissances.

• On observe un découpage usuel en 3 sources de connaissances principales: phonétique, lexicale, syntaxique. Il est donc intéressant de procéder à une évaluation analytique qui évaluera entre autres le taux de reconnaissance pour chaque unité de base. En ce qui concerne la syntaxe, les chercheurs fournissent tantôt la perplexité qui varie de 60 à 500, tantôt le facteur de branchement moyen qui va de 25 à 1000, tantôt le nombre de règles qui est compris entre 200 et 50000.

• Les réponses traitant de l'apprentissage sont souvent imprécises voire incomplètes. Notamment, la durée de parole nécessaire à l'apprentissage n'est pas toujours donnée en secondes comme demandé dans le questionnaire. Le temps d'apprentissage exprimé en temps CPU varie de quelques heures à quelques jours.

• 7 équipes ne sont pas encore en mesure de donner des taux de reconnaissance. 3 indiquent un taux de reconnaissance sans aucune précision. 4 détaillent leur réponse en indiquant par exemple qu'il s'agit d'un taux de reconnaissance des phonèmes, ou des mots, avec syntaxe ou sans syntaxe. Une seule indique comment a été calculé ce taux.

• Le même phénomène s'observe concernant le temps de reconnaissance et le temps d'apprentissage. Tantôt on fournit le temps de reconnaissance par phrase, tantôt par mot, tantôt par seconde de parole, ...

2.2 Description de l'évaluation de ces systèmes

Les 13 laboratoires précédents ont répondu aux deux rubriques concernant les données et les méthodes de leur évaluation mais dans la plupart des cas, leurs réponses sont partielles, imprécises ou incomplètes.

DONNEES DE L'EVALUATION

Nombre de locuteurs
de 1 à 20, en moyenne 10,
plus d'hommes que de femmes

Corpus de parole du test
(réponses multiples)

Possibilités	Réponses
parole continue	8
mot isolé	5
syllabe isolée	1

Environnement

non bruité	6
bruité	5

Support du corpus

PCM vidéocassettes	4
bandes informatique	4
autres	3

METHODES D'EVALUATION

Critères utilisés

Calcul des taux :	
- types d'erreur	9
- taux d'erreur de chaine	7
- taux moyen	6
- évolution avec le nombre de modèle	6
- taux d'erreur sur l'unité de base	5
Taux de reconnaissance /locuteurs :	
- expérimentés	6
- de l'ensemble d'apprentissage	5
- hors ensemble d'apprentissage	3
- non experimentés	1
- non natifs	1
temps de calcul	5
reconnaissance en/hors ligne	4
besoins matériels	2
effort d'apprentissage pour un nouveau locuteur	2
taille du logiciel	0

L'analyse des résultats montre que :

• Lorsque le nombre des locuteurs est précisé, il n'est jamais accompagné des caractéristiques de ceux-ci : âge, accent, condition sociale, ...

• Ces corpus sont réalisés dans une langue donnée (anglais, français, ...), aucun pour l'instant n'est multilingue.

• En général ces corpus, soit sont finalisés, soit comprennent les 5000 mots les plus fréquents de la langue, soit sont constitués de textes de la langue courante.

• 8 laboratoires ne précisent pas la difficulté de leur vocabulaire d'évaluation. Les autres réponses font ressortir un manque de norme sur la définition de la difficulté.

• Les réponses ne précisent pas le type du bruit et ses caractéristiques.

Ces résultats montrent la multiplicité des données utilisées pour évaluer les systèmes. Les corpus décrits sont très différents : par le nombre de locuteurs, par la taille et la complexité du vocabulaire. Ces quelques remarques montrent le manque de bases de données standard pour l'évaluation. On peut noter que chaque équipe utilise des critères spécifiques. Les critères communs les plus utilisés sont le taux de reconnaissance et le type d'erreurs, mais même pour ces critères, les définitions varient suivant les centres de recherche. Le besoin d'une norme définissant la méthodologie pour l'évaluation des performances est par conséquent vivement souhaitable.

2.3. Description des propositions pour l'évaluation

11 laboratoires ont répondu à cette partie de l'enquête concernant des propositions pour l'évaluation.

CORPUS D'EVALUATION

Nombre de mots
de 1.000 à 20.000

Langue
toutes les langues européennes
corpus multilingues

Type de vocabulaire
mots les plus fréquents du langage
textes non restreints
textes spécifiques à la tâche

Nombre de locuteurs
de 10 à 100
moitié hommes moitié femmes

Type de parole

parole continue : 9
mot isolé : 4

CRITERES D'EVALUATION

type d' erreurs
résistance aux dégradations
effort d'apprentissage pour un nouveau locuteur
nombre de locuteurs
performances globales : temps de rec. ,taux de rec.
evolution du temps de rec. avec la taille du vocabulaire
ergonomie du dialogue
taux de recouvrement du vocabulaire
caractéristiques des locuteurs
bases de données "standards"
taille du corpus en octets
vitesse d'élocution

commentaire

• L'utilisation de corpus multilingue nécessite d'évaluer la difficulté relative d'une langue par rapport à une autre : difficulté phonétique, difficulté syntaxique.

• En ce qui concerne les corpus de parole continue, aucune spécification sur la syntaxe n'a été précisée.

• Parmi les critères les plus importants pour l'évaluation des algorithmes de reconnaissance figurent :
- le pourcentage de reconnaissance des grandes classes phonétiques,
- le taux de bonne détection de mots,
- la bonne compréhension d'un message,
- l'ergonomie du dialogue, le nombre moyen de phrases prononcées pour atteindre un but élémentaire,
- la facilité de l'apprentissage,
- la résistance aux dégradations (bruit, téléphone,...).

On peut ajouter les remarques suivantes :
- les industriels sont intéressés par une évaluation "globale", taux de reconnaissance, temps de cacul, ...
- les universitaires souhaiteraient également une évaluation plus précise, influence du lexique, de la syntaxe ...

4 laboratoires n'effectuent pas de reconnaissance phonétique. Parmi les 9 laboratoires restant, un seul n'est pas intéressé par l'évaluation phonétique. Le problème de l'évaluation du décodage acoustico-phonétique fait l'objet de la partie suivante.

2e PARTIE : EVALUATION DE LA RECONNAISSANCE PHONETIQUE

1. Les objectifs du questionnaire

Le questionnaire "Phonetic Recognition of Continuous Speech and its Assessment" visait à cerner les tendances relativement aux points suivants:

• Quelles méthodes de décodage phonétique sont utilisées actuellement en Europe, pour quels types de Systèmes de Reconnaissance de la Parole Continue (SRPC) —qu'ils soient au stade de la recherche ou en cours de développement—?

• Quels critères de mesure de performance, quelles évaluations des stratégies de décodage phonétique existent et/ou sont souhaitées?

• Quels corpus sont nécessaires pour constituer des tests d'évaluation standard?

Par rapport à ce dernier point, nous souhaitions dégager les implications éventuelles en matière de bases de données acoustiques-phonétiques, en particulier pour tout ce qui touche à la transcription et à l'étiquetage. Enfin, l'une des questions concernait les corpus multilingues.

2. Les résultats

Nous synthétisons les résultats par types de réponses. Celles-ci ont été parfois schématisées pour tenir compte de redondances (qui ont été utiles pour contrôler la cohérence).

2.1. SRPC et décodage phonétique: méthodes utilisées

—Les caractéristiques majeures qui se dégagent à l'examen des tableaux sont les suivantes:

Adaptation au locuteur

Possibilités	Réponses
indépendant	5
apprentissage court	5
apprentissage long "offline"	11

Paramétrisation et codage
(réponses multiples)

Quantification vectorielle	6
LPC	6
Banc de filtres	6
modèle d'oreille	6
FFT	9

Méthodes de reconnaissance
(réponses multiples)

modèles connectionnistes	2
DTW	4
modèle discret de Markov	6
autres	6
modèle continu de Markov	8
système à connaissance	9

Unités de reconnaissance
(réponses multiples)

phonème	14
allophone	5
diphone	4
demi-syllabe	2
syllabe	3
autres	8

commentaire

• Aucun des systèmes décrits dans les réponses qui nous sont parvenues n'est un système commercial. Beaucoup sont au stade de la recherche. On notera cependant que 9 d'entre eux sont en phase d'évaluation.

• Les systèmes sont généralement monolingues sauf 2.

• La plupart des systèmes prévoient une adaptation au locuteur. Dans 11 cas, il s'agit d'un apprentissage "offline" considéré comme long. Cinq laboratoires envisagent une reconnaissance indépendante du locuteur; il s'agit souvent de projets fondés sur des connaissances d'experts.

• En reconnaissance, les modèles du type HMM semblent très utilisés, bien plus que les méthodes du type "DTW" (cet algorithme était indiqué comme cas particulier de "*template matching*"). Les approches basées sur les connaissances d'experts sont également très fréquentes. On notera que deux laboratoires développent des modèles connectionnistes.

• Aucune méthode de traitement du signal ne semble prépondérante.

• Les phonèmes ou les allophones sont généralement les unités de décision de l'étage phonétique. Moins fréquemment, des unités plus grandes (diphones, demi-syllabes, syllabes) sont mentionnées, et quand elles le sont, c'est —excepté un cas— conjointement au phonème et/ou à l'allophone.
Sous la rubrique "*Autres*", les réponses mentionnent souvent "traits et événements phonétiques".

• Le questionnaire portant sur la parole continue, il n'est pas surprenant que le contrôle par un niveau supérieur soit généralement utilisé ou prévu.

2.2. Evaluation phonétique

2.2.1. Les questions — Une question portait sur le besoin de disposer de méthodes d'évaluation phonétique standard, une autre sur l'utilisation actuelle de méthode d'évaluation phonétique. Les autres questions visaient à faire préciser les critères d'évaluation des performances au niveau phonétique.

2.2.2. Analyse des résultats —

Besoin d'évaluation "standard"

oui	15
non	3

Utilisation d'évaluation phonétique

oui	13
non	4

Critères d'évaluation

Taux de reconnaissance phonétique	
- 1er candidat	10
- plusieurs candidats	13
Type d'erreurs	
- insertion	13
- omission	13
- autres	7

commentaire

• En grande majorité, 15 laboratoires, sont intéressés par une méthode d'évaluation "standard". 13 d'entre eux pratiquent déjà une évaluation phonétique sur leur propre système de reconnaissance phonétique.

• Les informations attendues d'un système d'évaluation phonétique sont : le taux de substitution, le taux d'insertion, le taux d'omission, le taux global d'erreur. Sous la rubrique *"autre"* les réponses mentionnent aussi les matrices de confusion, les classes d'erreurs, ...

2.3. Corpus de test et bases de données acoustique-phonétiques

Besoins en corpus de test et en transcription phonétique

Type de corpus

Phrases phonétiquement équilibrées	17
Phrases compactes	4
Autres	9

Transcription
(réponses multiples)

Transcription "Standard"	9
Transcription précise	15
Pas de transcription	1

Etiquetage

Tous les sons	13
Sons stables	2

commentaire

• Le nombre de locuteurs souhaité varie considérablement selon les laboratoires (souvent plus de 50 locuteurs, hommes, femmes et enfants).

• Pour constituer des corpus de test, les phrases phonétiquement équilibrées sont généralement mentionnées (17 laboratoires), les phrases "compactes" (4 réponses). Les phrases compactes sont des phrases phonétiquement déséquilibrées de façon à représenter toutes les combinaisons d'unités dans un corpus faible. Dans les 9 réponses "autres" sont exprimés des besoins en parole naturelle, en parole dictée, ...

• La plupart des laboratoires expriment un besoin en corpus transcrits: 9 souhaitent une transcription "standard" et 15, une transcription précise (ce que l'on peut sans doute traduire par transcription phonétique au sens du rapport SAM [Wells,1987]).

• L'étiquetage n'est pas demandé explicitement par tous mais 15 le souhaitent. En ce qui concerne la question de savoir si l'on doit étiqueter tous les segments phonétiques ou simplement ceux qui sont stables, une forte majorité est en faveur de la première des deux solutions. 4 laboratoires seulement ne sont pas intéressés par des bases de données multilingues.

3. CONCLUSION GENERALE

Au stade actuel, l'enquête permet de se faire une idée, partielle certes, des tendances actuelles de la recherche sur les SRGV et les SRPC en Europe.

D'une manière générale se dégage un besoin de méthodes d'évaluation standard, tant au plan lexical que phonétique. On notera que les SRGV demandent très souvent les deux niveaux d'évalution et que l'évaluation phonétique semble davantage répondre à des besoins de recherche et de conception qu'à des préoccupations industrielles.

Sur les critères nécessaires à la mesure des performances, il y a un assez large consensus, de telle manière qu'il ne semble pas déraisonnable d'envisager l'automatisation de méthodes d'évaluation standard aux niveaux lexical et phonétique sur le poste de travail prévu dans le cadre du projet SAM.

Leur mise en œuvre suppose des corpus de test standard. Les renseignements obtenus ne suffisent évidemment pas à définir exactement ces corpus standard ni leur priorité. L'enquête montre cependant le besoin d'enregistrements de corpus divers —phrases phonétiquement équilibrées, parole naturelle, mots fréquents, nombres,...— provenant d'une cinquantaine de locuteurs au moins, avec une dizaine de répétitions. Le projet SAM vise à développer des méthodologies multilingues. Les réponses obtenues témoignent de l'intérêt qu'y attachent généralement les équipes consultées.

Enfin, de manière très nette apparaît un besoin de transcription précise au niveau phonétique ainsi que, la plupart du temps, celui d'un étiquetage, si possible de toutes les unités phonétiques.

Jusqu'à présent il semble que les efforts en matière d'évaluation et de corpus de test aient porté essentiellement sur les cartes de reconnaissance classiques de mots isolés ou enchaînés. L'enquête, au stade actuel, montre qu'il sera nécessaire de développer des actions semblables pour la nouvelle génération de cartes vocales dont certaines arrivent maintenant à maturité. Les méthodologies en question devront évidemment être adaptées aux nouveaux besoins —l'enquête apporte déjà quelques éléments d'appréciation dans ce sens— et pour celà il nous semble nécessaire de développer des actions dans le cadre du GRECO de la Communication Parlée et/ou du projet Européen SAM-ESPRIT II.

4. BIBLIOGRAPHIE

Bibliographie évaluation

[Baker,1983] J. Baker, D.S.Pallet, J.S.Bridle, Speech recognition performance assessments and available databases, IEEE-ICASSP, Vol 2, pp. 527-530.

[Bourjot,1988] C. Bourjot, A. Boyer, J.F. Mari, "A methodologie about assessment of large vocabulary systems", Seventh FASE symposium, Edinburgh, August 1988, à paraître.

[Campo,1987] R. Campo, G. Castagneri, L. Vachetta, R. Moore, Recognition Evaluation Methods, ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Carlson,1986] Carlson, Elenius, Granstrom, Hunnicut, "Phonetic properties of the basic vocabulary of five european languages: implications for speech recognition", IEEE-ICASSP, Vol. 4, pp. 2763-2766.

[Chollet,1981] G. Chollet, A. Astier, M. Rossi, "Evaluating the performance of speech recognizers at the acoustic-phonetic level", IEEE-ICASSP,pp. 758-761.

[Chollet,1982] G. Chollet, C. Gagnoulet, "On the evaluation of speech recognizers and data bases using a reference system", IEEE-ICASSP, 2026-2029.

[Dalsgaard,1987a] P. Dalsgaard, S. Danielsen, "A first Danish approach to standard assessment methodologies", SAM-SC-06, Octobre 1987.

[Dalsgaard,1987b] P. Dalsgaard, S. Danielsen, J.M. Dolmazon , M. Taylor, R. Winski, "Visit Report of SAM USA Survey Group", SAM Report LO-01.

[Dolmazon,1987a] J.M. Dolmazon, J. Caelen, "Survey on Speech Workstation", ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Dolmazon,1987b] J.M. Dolmazon, C. Benoît, J.L. Gauvain, G. Pérennou, "Le projet européen "SAM": Evaluation mulilingue des dispositifs d'entrée-sortie vocale", 16ème JEP, Hammamet, pp. 314-317.

[Lea,1980] W.A. Lea, J.E. Shoup, "Specific contribution of the ARPA SUR project", Trends in Speech Recognition, Prentice Hall, Signal Processinf series 1980.

[Lea,1982] W.A. Lea, "Available speech databases for evaluating speech recognizers",Workshop on Standardization for Speech I-O Technology, Washington.

[Lea,1982] W.A. Lea, "What causes speech recognizers to make mistakes", IEEE-ICASSP, Vol. 3, pp. 2030-2033.

[Lea,1983] W.A. Lea, "Selecting the best speech recognizer for the job", Speech Technology, Vol.1, No 1, pp. 10-29.

[Montacié,1987] C. Montacié, G. Chollet, "Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole",16ème JEP 5-9 oct Hammamet, 1987.

[Moore,1977] R.K. Moore, Evaluating speech recognizers, IEEE transactions on ASSP, Vol . 25., No 2., pp. 178-183.

[Palett,1985] D.S. Palett, Automatic Speech recognition performance assessments, NBS Special publication, Washington.

[Pols,1987] L.C.W. Pols, "Quality of Text-to-Speech Synthesis Systems", ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Roach,1987] P.Roach, P. Rowlands, A.M. Dew, "Assessment of Accuracy in Automatic Phonetic Analysis, Speech Technology, Edinburgh, Vol. 2, pp. 158-160.

[Simpson,1987a] C. Simpson, J. Ruth, "The phonetic Discrimination Test for Speech Recognizers: Part I", Speech technology, Vol 3, N° 4, pp. 48-54.

[Simpson,1987b] C. Simpson, J. Ruth, "The phonetic Discrimination Test for Speech Recognizers: Part II", Speech Technology, Vol 3, N° 7, pp. 48-54.

[Thomas,1987] T.J. Thomas, R. Winski, "Speech Recogniser Assessment in the Laboratory, Not in the field", Speech technology, Vol 3, N° 4, pp. 88-93.

[Tomlinson,1987] M. Tomlison, "Labelling Methods Session", ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Tubach,1984] J.P. Tubach, "Problèmes et méthodes en évaluation de reconnaissance phonétique", 13ème JEP, Bruxelles.

[Wells,1987] J. Wells, "SAM-PA, Speech Assessment Methods Alphabet", ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88.

[Winski,1987] R. Winski, "Survey Report on Data Management", ESPRIT Project 1541, Final Report, Definition Phase: 2-2-87//31-1-88 .

Bibliographie grand vocabulaire

[Adda,1987a] G. Adda, "Reconnaissance de grands vocabulaires : une étude syntaxique et lexicale", Thèse de Doct. Ing. Info, Université Paris XI dec. 1987.

[Adda,1987b] G. Adda, M. Averbuch et al., "Experiments with the TANGORA 20.000 word speech recognizer", Proc. IEEE ICASSP 1987, pp. 17-3.

[Bahl,1983] L.R. Bahl, F. Jelinek, R.L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Trans. Pattern Analysis and Machine Intelligence 5(2), Maarch 1983, pp. &179-190.

[Baker,1987] J.K Baker, J.M. Baker, "Large vocabulary natural language speech recognition in software", European Conference on Speech Technology, sept 1987, vol. 2, p. 440.

[Billi,1986] R. Billi, G. Massia, F. Nesti, "Word preselection for large vocabulary speech recognition", Proc. IEEE ICASSP 1986, pp.2-7.

[Bonneau,1987] H. Bonneau, J.L. Gauvain, "Vector quantization for speaker adaptation", Proc. IEEE ICASSP 1987, pp. 1434.

[Carbonnel,1987] N.Carbonnel, J.P. Haton, J.M. Pierrel, "A knowledged based approach to the design of a Man-Machine Dialog system by voice", European Conference on Speech Technology, sept 1987.

[Cerf,1986] H. Cerf et al., "Speech recognition using 10.000 word vocabulary", Nato Advanced Institute on Pattern Recognition, Brussels 1986.

[Charpillet,1985] F. Charpillet, "Un système de reconnaissance de parole continue pour la saisie de textes lus", thèse de l'Université de Nancy 1, 1985.

[Chow,1987] Y.L. Chow et al., "BYBLOS : the BBN continuous speech recognition system", Proc. IEEE ICASSP 1987, pp. 3-7.

[Derouault,1987] A.M. Derouault, "Context-dependent phonetic Markov Models for large vocabulary speech recognition", Proc. IEEE ICASSP 1987, Dallas.

[Dumouchel,1988] P. Dumouchel et al, "Three probabilistic language models for a large vocabulary recognizer", IEEE ICASSP 1988, pp. 513-516.

[Fujisaki,1987] H. Fujisaki, "Overview of the Japanese national project on advanced man-machine interface through spoken language", European Conference on Speech Technology, sept 1987.

[Kaneko,1983] T. Kaneko, N.R Dixon, "A hierarchical decision approach to large-vocabulary discrete utterance recognition", IEEE Trans. ASSP, oct 1983, num. 5, pp. 1061-1066.

[Kohonen,1987] T. Kohonen et al., "Microprocessor implementation of a large vocabulary speech recognition and phonetic typewriter for Finnish and Japanese", European Conference on Speech Technology, sept 1987, vol. 2, pp. 377-380.

[Kubala,1988] F. Kubala et al, "Continuous speech recognition results on the DARPA 1000 words resource management database", IEEE ICASSP 1988, pp.291-294.

[Laface,1987] P.Laface, G. Micca, R Pieraccini, Experimental results on a large lexicon access task", Proc. IEEE ICASSP 1987, pp. 20-4.

[Lagger,1985] H. Lagger, A Waibel, "A coarse phonetic knowledge source for template independent large vocabulary word recognition", *Proc. IEEE ICASSP 1985*, pp. 23-6.

[Levinson,1988] S.E. Levinson, A. Ljolje, L.G. Miller, "Large vocabulary speech recognition using a HMM for acoustic/phonetic classification", IEEE ICASSP 1988, pp. 505-508.

[Lowerre,1980] B. Lowerre, R. reddy, "The HARPY speech understanding system, in Trends in speech recognition, W.A. LEA (ed.), Englewood Cliffs, N.J. : Prentice Hall, 1980, pp. 340-360.

[Mari,1984] J.F Mari, J.P Haton, "Some experiments in automatic recognition of a thousand word vocabulary", Proc. IEEE ICASSP 1984, pp.26-6.

[Mariani,1987] J. Mariani, "A prototype of a voice activated typewriter", European Conference on Speech Technology, sept 1987, vol. 2, pp. 222-225.

[Merialdo,1987] B. Merialdo, "Speech recognition using very large size dictionnary", Proc. IEEE ICASSP 1987, pp. 10-2.

[Pierrel,1987] J.M. Pierrel, "Dialogue oral homme-machine", Hermes, 1987, Chapitre 7 et 8.

[Quénot,1986] G.Quénot et al., A dynamic time warp VLSI processor for continuous speech recognition", Proc. IEEE ICASSP 86, Tokyo avr 1986.

[Rosenberg,1982] A.E. Rosenberg, L.R Rabiner, J.G Wilpon, "Speaker trained recognition of large vocabularies of isolated words", Proc. IEEE ICASSP 1982, pp. 2018-2021.

[Shipman,1982] D.W. Shipman, V.W Zue, "Properties of large lexicons : implications for advanced isolated word recognition systems", Proc. IEEE ICASSP 1982, pp. 546-549.

[Shirai,1984] K. Shirai, T. Kobayashi, "Phrase speech recognition of large vocabulary using features in articulatory domain", Proc. IEEE ICASSP 1984, pp. 26-9.

[Wolf,1980] J.J. Wolf, W.A. Woods, "The WHIM speech understanding system", in Trends in speech recognition, W.A. LEA (ed.), Englewood Cliffs, N.J. : Prentice Hall, 1980, pp. 316-339.

ETUDE COMPARATIVE DE PLUSIEURS MODELES D'ANALYSE ACOUSTIQUE EN PRESENCE DE BRUIT

Jean-Claude JUNQUA* et Hisashi WAKITA

Speech Technology Laboratory - Division of Panasonic Technologies, Inc.
3888 State Street - Santa Barbara, California 93105, USA
* aussi CRIN/INRIA - BP 239 - 54506 Vandoeuvre les Nancy, France

ABSTRACT

The performances of automatic speech recognition systems decrease dramatically in the presence of noise. We show in this paper that the recognition of noisy speech can be significantly improved by a proper selection of the analysis method and of the weights associated with the cepstral coefficients of the acoustic model. Best results are obtained with the PLP-RPS acoustic model and the GEL distance.

1. INTRODUCTION

En reconnaissance automatique de la parole en présence de bruit les performances se dégradent rapidement suivant le modèle d'analyse acoustique utilisé. Notre définition du modèle d'analyse acoustique intègre la représentation de la parole par un vecteur de paramètres mais aussi la modification de ce vecteur de paramètres par la distance utilisée par l'algorithme de comparaison (dans notre cas la programmation dynamique). Les propriétés de l'analyse et de la distance interagissent dans le modèle d'analyse acoustique. Ils doivent donc être étudiés ensemble.

Afin d'améliorer les performances des systèmes de reconnaissance automatique de la parole en présence de bruit, un grand nombre d'études ont été entreprises. Citons en particulier :

a) l'étude de nouvelles distances |Hanson 85|, |Itakura 87|, |Mansour 88|, |Matsumoto 86|, |Soong 87| présentées le plus souvent en conjonction avec l'analyse par prédiction linéaire (linear prediction ou LP),

b) le développement de nouvelles techniques d'analyse plus robustes en présence de bruit |Hunt 86|, |Ghitza 87|,

c) les essais de compensation ou de suppression de bruit |Cox 81|, |Ephraim 87|, |Ephraim 88|, |Kay 80|, |Neben 83|, |Petersen 81|.

Le premier objectif de notre étude était d'évaluer l'analyse par prédiction linéaire perceptivement fondée (perceptually based ninear predictive analysis ou PLP) en présence de bruit, par comparaison aux techniques d'analyse par banc de filtres critiques et prédiction linéaire. La distance euclidienne appliquée aux coefficients cepstraux avec différents poids (unité, root power sums (RPS)) |Hanson 85|, |Paliwal 82|, et general exponential lifter ou GEL |Hermansky 88| a été utilisée dans le domaine cepstral. Cette étude a été suivie par le développement d'une nouvelle méthode d'analyse utilisant un modèle auditif intégrant des concepts physiologiques : analyse par prédiction linéaire avec synchronisation temporelle (time synchronous linear predictive analysis ou SLP).

Dans cet article, nous montrons que les performances des systèmes de reconnaissance automatique de la parole en présence de bruit peuvent être améliorées par une sélection judicieuse de la méthode d'analyse et des poids appliqués aux coefficients cepstraux dans le modèle d'analyse acoustique. Parmi tous les modèles d'analyse acoustique utilisés PLP-RPS donne les meilleurs résultats en reconnaissance inter-locuteur (cross-speaker). Nous proposons l'utilisation de la distance GEL |Hermansky 88|. Celle-ci améliore les scores de reconnaissance, pour le cas inter-locuteur, par rapport à la distance RPS qui avait déjà donné de bons résultats en présence de bruit avec les techniques d'analyse LP et PLP |Hanson 85|. En combinaison avec la distance, general exponential lifter donne le meilleur résultat pour un faible rapport signal sur bruit (SNR).

2. CONDITIONS EXPERIMENTALES

Nous avons utilisé le vocabulaire alphanumérique (36 mots) qui est un vocabulaire difficile. 10 locuteurs américains (6 masculins et 4 féminins) ont prononcé deux fois le vocabulaire. Les mots ont été enregistrés isolément et les frontières de mots ont été identifiées manuellement. Dans tous les tests une référence par mot a été sélectionnée et aucune technique de classification n'a été utilisée. Dans le but de simuler différentes conditions de bruit, du bruit blanc et du bruit blanc filtré passe-bas (appelé dans la suite "bruit blanc filtré") ont été alternativement ajoutés au signal de parole. Le bruit blanc filtré a une pente spectrale moyenne similaire au signal de parole |Hanson 85|. De plus, ce type de bruit est une bonne approximation des conditions que l'on a enregistrées en conduisant sur l'autoroute avec la ventilation et les fenêtres ouvertes (cf. figure 1). La fonction de transfert du filtre passe-bas qui a été utilisée dans notre évaluation pour obtenir le bruit blanc filtré est $1/(1-0.92z^{-1})$. Les mots de référence ont toujours été considérés dépourvus de bruit en tenant compte de l'hypothèse qu'aucune indication sur les caractéristiques du bruit n'était disponible. Le rapport signal sur bruit a été défini comme le rapport entre la puissance du spectre de fréquence du signal correspondant au mot prononcé (moyennée sur tout le mot) sur la puissance du spectre de fréquence du bruit. Les tests ont été effectués pour trois valeurs du rapport signal sur bruit : 25, 15 et 5 dB. Lors de précédentes études, nous avons remarqué que, en reconnaissance multi-locuteur, le modèle d'analyse acoustique conjugue les propriétés du modèle observées lors de la reconnaissance mono-locuteur et inter-locuteur. Par conséquent nous avons effectué nos tests de reconnaissance en mono-locuteur et inter-locuteur.

3. VUE GENERALE DES MODELES D'ANALYSE ACOUSTIQUES ETUDIES

3.1. Les modèles d'analyse acoustique utilisant la technique LP

Au cours de nombreuses études des systèmes de reconnaissance en présence de bruit, la technique d'analyse LPC a été très utilisée. Une de ces études |Tierney 80| montra qu'un ordre du modèle d'analyse suffisamment élevé est nécessaire pour modéliser à la fois la parole et le bruit. Dans nos tests nous avons donc choisi d'utiliser un modèle d'analyse acoustique d'ordre 14.

Les distances associées à l'analyse LPC ont aussi fait l'objet de beaucoup d'études. Il a été observé que, en présence de bruit, des distances sensibles aux pics du spectre de fréquence |Matsumoto 86| amélioraient les scores de reconnaissance. De plus, plusieurs études ont

montré que des distances sensibles à la pente du spectre de fréquence et plus généralement que des distances pondérant les coefficients cepstraux amélioraient les systèmes de reconnaissance en présence de bruit |Hanson 85|, |Itakura 87|, |Juang 86|, |Paliwal 82|, |Tohkura 85|, |Yegnanarayana 79|. Lors des tests d'évaluation nous avons utilisé, pour comparer les coefficients cepstraux obtenus à partir de l'analyse LPC, deux distances : euclidienne et RPS. La distance euclidienne (S=0) et RPS (S=1) sont deux cas particuliers de la distance GEL définis par

$$E_n = n^S S > 0 \qquad (1)$$

3.2. Le modèle d'analyse acoustique utilisant un banc de filtres critiques

Nous avons utilisé un modèle d'analyse acoustique similaire à celui proposé par Klatt |Klatt 82| (critical-band slope metric ou CB-SM). Le banc de filtres critiques est formé de 17 filtres espacés de 1 bark sur une échelle graduée en barks allant de 0 à 5000 Hz. Une pré-accentuation par des courbes d'égale sonie est effectuée en même temps que le filtrage. Une tranformation logarithmique transforme d'intensité en sonie la sortie de chaque filtre. Enfin, la pente du spectre fréquentiel, à la fréquence centrale de chaque filtre, est calculée en prenant la différence

$$SL[i] = dB[i+1] - dB[i] \quad (2)$$

où dB[i] représente la sortie en décibels du i-ème filtre. 17 coefficents sont alors fournis à l'algorithme de programmation dynamique qui utilise la distance euclidienne.

3.3. Les modèles d'analyse acoustique utilisant la technique PLP

La méthode d'analyse PLP modélise le spectre auditif par celui d'un modèle tout pôle d'ordre réduit. Le spectre auditif est dérivé du signal de parole par filtrage en bandes critiques, pré-accentuation par courbes d'égale sonie et conversion d'intensité en sonie par extraction de la racine cubique (loi de puissance de stevens). Les différentes étapes de cette méthode d'analyse utilisent des concepts psycho-acoustiques bien établis. Des modèles d'analyse acoustique établis autour de cette méthode d'analyse ont déjà donné de bons résultats dans le cadre de systèmes de reconnaissance testés sur de la parole ayant un grand rapport signal sur bruit |Hermansky 87|, |Hermansky 88|, |Junqua 87|. Le modèle d'analyse acoustique PLP-RPS, dont un des avantages est de fournir un vecteur de paramètres de faible dimension, s'est révélé particulièrement intéressant. Au travers de cette étude, nous avons utilisé trois distances avec la méthode d'analyse PLP : euclidienne, RPS et GEL.

3.4. Les modèles d'analyse acoustique utilisant la technique SLP

Afin d'améliorer les performances des systèmes de reconnaissance en présence de bruit, nous avons développé une nouvelle méthode d'analyse appelée SLP. Cette méthode modélise les vibrations mécaniques intervenant dans la membrane basilaire (BM) et transforme ces vibrations dans une représentation simulant l'activité de neurones. SLP repose sur un modèle du système périphérique humain et en particulier sur un modèle du cochlear. La figure 2 présente un schéma fonctionnel de cette méthode d'analyse. La partie filtrage est formée d'un banc de 99 filtres espacés de 1 bark sur une échelle graduée en barks (de 0 à 5000 Hz). les filtres sont symétriques. Leur pente (10 dB) a été ajustée en optimisant les scores de reconnaissance lors de tests préliminaires. La principale fonction de cette section filtrage est de séparer un mélange complexe de sons en régions ayant un grand rapport signal sur bruit.

La seconde section du modèle est constituée de trois étapes :

a) détection d'enveloppe, utilisant une méthode de passage par zéro et construction d'histogrammes dans chaque canal |Allen 85|, |Ghitza 87|,

b) un filtre passe-bas dont le but est de réduire la synchronicité dans les hautes fréquences afin d'obtenir un spectre plus lisse,

c) un mécanisme permettant de corréler les canaux adjacents |Allen 85|, |Hunt 86| afin de mettre en valeur les pics du spectre fréquentiel (en particulier en présence de bruit). Cette dernière étape simule certaines propriétés attribuées au mécanisme d'inhibition latérale.

La troisième section définit une mesure de synchronicité (en sommant tous les canaux) dont le but est de combiner tous les canaux qui fournissent des informations à une fréquence donnée en même temps. La largeur de chaque région (correspondant à un intervalle d'histogramme) est utilisée comme une estimation de l'intensité spectrale dans cette région |Ghitza 87|.

La dernière section est un modèle tout pôle réduit (similaire à celui utilisé dans les méthiodes d'analyse LP et PLP) suivi par une paramétrisation en coefficients cepstraux. Un des avantages du modèle tout pôle réduit est de mettre en valeur les pics du spectre fréquentiel. Au travers des différents tests, nous avons utilisé trois versions de cette méthode d'analyse :

a) SLP1 qui n'utilise pas le mécanisme de corrélation et le filtre passe-bas,

b) SLP2 qui n'utilise pas le mécanisme de corrélation,

c) SLP3 qui est la méthode d'analyse telle qu'elle est décrite ci-dessus.

Comme pour la technique d'analyse LP, les deux distances : euclidienne et RPS ont été utilisées en combinaison avec cette nouvelle méthode d'analyse.

4. EVALUATION EN PRESENCE DE BRUIT DES MODELES D'ANALYSE ACOUSTIQUE UTILISANT LES TECHNIQUES LP, PLP ET BANC DE FILTRES CRITIQUES

4.1. Etude du modèle d'analyse acoustique PLP-RPS

Une série de tests ont été effectués afin d'évaluer le modèle d'analyse acoustique PLP-RPS pour différents ordres du modèle. Les modèles d'ordre 5 et 8 qui, lors de précédents tests, ont donné les meilleurs résultats |Hermansky 87|, |Junqua 87| ont été comparés au modèle d'ordre 14. Les résultats présentés aux figures 3 et 4 montrent que, comme pour la technique d'analyse LP, il est souhaitable d'utiliser un ordre du modèle élevé lorsque le rapport signal sur bruit est faible. Cependant nous pouvons remarquer que, pour la reconnaissance inter-locuteur, le modèle d'analyse acoustique d'ordre 8 est une bonne alternative au modèle d'ordre 14 (ceci est particulièrement visible pour le cas du bruit blanc filtré). Lorsque la parole est perturbée par du bruit blanc filtré (le gain se situe entre 5 et 10 dB).

Dans les prochains tests nous avons considéré uniquement l'ordre 11 pour les différents modèles d'analyse acoustique étudiés.

4.2. Comparaison des modèles d'analyse acoustique utilisant les techniques LP, PLP et banc de filtres critiques

Dans le but de comparer le modèle d'analyse acoustique CB-SM avec les modèles d'analyse acoustique utilisant les techniques LP et PLP nous avons effectué de nouveaux tests de reconnaissance. Les figures 5 et 6 présentent les résultats obtenus pour les modèles d'analyse acoustique CB-SM, LP-CEPS et PLP-CEPS dans le cadre de perturbations de la parole par du bruit blanc et du bruit blanc filtré. En reconnaissance mono-locuteur le modèle d'analyse CB-SM donne les meilleurs résultats. En reconnaissance inter-locuteur le modèle d'analyse CB-SM fournit les plus hauts scores de reconnaissance pour un rapport signal sur bruit faible. Les bons résultats du modèle d'analyse acoustique CB-SM peuvent vraisemblablement être attribués au fait que ce modèle est sensible aux variations de pente du spectre fréquentiel. Cette hypothèse s'est vue confirmée par d'autres tests d'évaluation qui ont permis de comparer les modèles d'analyse acoustique

CB-SM, LP-RPS et PLP-RPS. Les résultats de ces tests sont présentés aux figures 7 et 8. La distance RPS qui est une distance sensible aux variations de pente du spectre fréquentiel permet d'améliorer les scores de reconnaissance obtenus avec les modèles d'analyse acoustique utilisant les techniques d'analyse LP et PLP. Ceci avait déjà été observé lors de précédentes études |Hanson 86|. Le modèle d'analyse acoustique PLP-RPS donne des résultats similaires à ceux du modèle CB-SM. Enfin, dans le cas de la reconnaissance inter-locuteur, le modèle d'analyse acoustique PLP-RPS donne de meilleurs résultats que ceux obtenus par un modèle d'analyse acoustique utilisant la technique d'analyse LP.

Les résultats utilisés dans les précédentes discussions sont aussi présentés sous forme de tableaux (Table 1., Table 2., Table 3., Table 4.).

5. OPTIMISATION DU MODELE D'ANALYSE ACOUSTIQUE UTILISANT LA TECHNIQUE PLP

Nous avons montré précédemment que dans le cadre de la reconnaissance inter-locuteur le modèle d'analyse acoustique PLP-RPS fournit les meilleurs résultats (en adaptant l'ordre du modèle suivant les conditions de bruit). Dans cette section nous présentons une optimisation du modèle d'analyse acoustique PLP-RPS en présence de bruit.

Lors de précédentes études |Hermansky 88| utilisant de la parole ayant un grand rapport signal sur bruit, nous avons proposé une optimisation du modèle d'analyse acoustique PLP-RPS dans le cadre de la reconnaissance inter-locuteur. Nous avons montré que la distance RPS est trop sensible aux pics du spectre fréquentiel et pas assez sensible à la pente de celui-ci. Nous avons proposé la distance GEL qui optimise le modèle d'analyse acoustique en termes de sensibilité par rapport aux pics et à la pente du spectre fréquentiel. Les résultats de cette optimisation sont rappelés à la figure 9. Un optimum a été trouvé pour S=0.6. En reconnaissance mono-locuteur la distance RPS fournit les meilleurs résultats.

Nous avons appliqué cette nouvelle distance au cadre de notre étude. Une série de tests ont été effectués pour un rapport signal sur bruit de 15 dB et différentes valeurs de S (exposant de la distance GEL). Les résultats de ces tests sont présentés aux figures 10 et 11. Nous pouvons remarquer que, comme pour de la parole ayant un grand rapport signal sur bruit, un optimum est trouvé entre la distance euclidienne (S=0) et la distance RPS (S=1). La valeur optimale du paramètre S est semblable à celle trouvée lors de nos précédentes études (S=0.5-0.6). En reconnaissance mono-locuteur la distance RPS est très proche de la distance optimale. Ces résultats sont en accord avec ceux de nos précédentes études.

6. EVALUATION DU MODELE D'ANALYSE ACOUSTIQUE UTILISANT LA TECHNIQUE SLP

Au vu de nos précédents tests, nous pouvons remarquer que, en présence de bruit, les scores de reconnaissance obtenus par notre meilleur modèle d'analyse acoustique sont beaucoup plus bas que ceux obtenus dans des conditions sans bruit. Nous avons donc décidé de développer une nouvelle méthode d'analyse avec comme objectif l'obtention de meilleures performances en présence de bruit que les précédentes méthodes étudiées. Récemment, beaucoup d'efforts ont été consacrés au développement de modèles d'analyse acoustique articulés autour de modèles auditifs utilisant des concepts physiologiques |Ghitza 87|, |Lyon 82|, |Petersen 81|. De tels modèles ont déjà donné de bons résultats en présence de bruit |Ghitza 87|, |Hunt 86|. Le modèle auditif que nous avons utilisé a été décrit dans la section 3.4. Nous avons évalué ce modèle en reconnaissance mono-locuteur en présence de bruit blanc (comme pour toutes les autres évaluations uniquement pour les mots-tests). La figure 12 présente les résultats des tests effectués pour le modèle d'analyse acoustique SLP1. Le modèle d'analyse acoustique SLP1 donne les meilleurs résultats, par comparaison avec les techniques LP et PLP, lorsque la distance euclidienne appliquée sur les coefficients cepstraux est utilisée. Par contre, lorsque la distance RPS est utilisée le modèle PLP-RPS fournit les meilleurs scores de reconnaissance. Ces résultats illustrent le fait que la technique d'analyse et la distance utilisée lors de la comparaison ne doivent pas être étudiées séparément. Notons enfin que la distance RPS ne semble pas bien adaptée au modèle SLP1.

Dans le but de rendre cette méthode d'analyse plus robuste en présence de bruit, nous avons ajouté dans la méthode d'analyse deux étapes destinées à donner plus d'importance aux pics du spectre fréquentiel (notamment en présence de bruit) et à diminuer la synchronicité dans les hautes fréquences afin d'obtenir un spectre plus lisse. Ces deux étapes sont décrites dans la section 3.4. Des tests effectués en présence de bruit blanc (SNR = 5 dB) permettent de constater (Table 5.) que chacune des étapes améliore les scores de reconnaissance. Le modèle d'analyse SLP3 donne le meilleur résultat pour un rapport signal sur bruit de 5 dB.

7. CONCLUSIONS

Au cours de cette analyse nous avons évalué des modèles d'analyse acoustique utilisant la technique d'analyse PLP (par comparaison avec d'autres modèles d'analyse acoustique). De plus, nous avons présenté une nouvelle méthode d'analyse utilisant des concepts physiologiques. Nous avons montré que :

a) pour le modèle d'analyse acoustique PLP-RPS un ordre faible du modèle (8) donne de bons résultats en reconnaissance inter-locuteur alors qu'il est souhaitable d'utiliser un ordre du modèle plus grand en reconnaissance mono-locuteur,

b) le modèle d'analyse acoustique PLP-RPS donne les meilleurs résultats, en reconnaissance inter-locuteur, comparé aux modèles étudiés utilisant la technique d'analyse LP ou un banc de filtres critiques,

c) dans le cadre du modèle d'analyse acoustique utilisant la technique d'analyse PLP, la distance GEL améliore les scores de reconnaissance (par rapport à la distance RPS) en reconnaissance inter-locuteur,

d) une nouvelle méthode d'analyse, SLP, utilisant des concepts physiologiques donne le meilleur résultat en reconnaissance inter-locuteur pour un rapport signal sur bruit de 5 dB.

Nous avons présenté le début d'un travail concernant le développement d'une nouvelle méthode d'analyse. Les résultats obtenus sont déjà encourageants. Toutefois, cette technique d'analyse nécessite certains ajustements et d'autres tests d'évaluation. Une partie des efforts doit aussi porter sur le développement ou l'adaptation d'une distance appropriée à la technique d'analyse. Ceci constitue la direction de nos travaux actuels.

BIBLIOGRAPHIE

|Allen 85| Allen, J.B., "Cochlear Modelling", IEEE ASSP Magazine, pp. 3-29, 1985.

|Cox 81| Cox, R.V. and Malah, D., "A Technique for perceptually reducing periodically structured noise in speech", Proc. ICASSP-81, pp. 1089-1092, 1981.

|Ephraim 87| Ephraim, Y., Wilpon, J.G. and Rabiner, L.R., "A linear prediction front-end processor for speech recognition in noisy environments", Proc. ICASSP-87, pp. 1321-1327, 1987.

|Ephraim 88| Ephraim, Y., Malah, D. and Juang, B.H., "On the application of Hidden Markov Models for enhancing noisy speech", Proc. ICASSP-88, pp. 533536, 1988.

|Ghitza 87| Ghitza, O., "Robustness against noise : the role of timing-synchrony measurement", Proc. ICASSP-87, pp. 2372-2375, 1987.

|Hanson 85| Hanson, B.A., Hermansky, H. and Wakita, H., "Root-power sums and spectral slope distortion measures for all-pole models of speech", JASA 78, S1, p. S49, 1985.

also see Hanson, B.A. and Wakita, H., "Spectral slope based distortion measures for all-pole models of speech", Proc. ICASSP-86, pp. 757-760, 1986.

|Hermansky 85| Hermansky, H., Hanson, B.A. and Wakita, H., "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain", Speech Communication 4, pp. 181-187, 1985.

|Hermansky 87| Hermansky, H., "An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception", Proc. ICASSP-87, pp. 1159-1162, 1987.

|Hermansky 88| Hermansky, H. and Junqua, J.C., "Optimization of perceptually-based front-end", Proc. ICASSP-88, pp. 219-222, 1988.

|Hunt 86| Hunt, M.J. and Lefebvre, C., "Speech recognition using a cochlear model", Proc. ICASSP-86, pp. 1979-1982, 1986.

|Itakura 87| Itakura, F. and Umezaki, T., "Distance measure for speech recognition based on the smoothed group delay spectrum", Proc. ICASSP-87, pp. 1257-1260, 1987.

|Junqua 87| Junqua, J.C., "Evaluation of ASR front-ends in speaker-dependent and speaker-independent recognition", JASA, S1, p. S93, 1987.

|Juang 86| Juang, B.H., Rabiner, L.R. and Wilpon, J.G., "On the use of bandpass liftering in speech recognition", Proc. ICASSP-86, pp. 765-768, 1986.

|Kay 80| Kay, S.M., "Noise compensation for autoregressive spectral estimates", IEEE ASSP-28, pp. 292-303, 1980.

|Klatt 82| Klatt, D.H., "Prediction of perceived phonetic distance from critical-band spectra : a first step", Proc. ICASSP-82, pp. 1278-1281, 1982.

|Lyon 82| Lyon, R.F., "A computational model of filtering, detection and compression in the cochlear", Proc. ICASSP-82, pp. 1282-1285, 1982.

|Mansour 88| Mansour, D. and Juang, B.H., "A family of distortion measures based upon projection operation for robust speech recognition", Proc. ICASSP-88, pp. 36-39, 1988.

|Matsumoto 86| Matsumoto, H. and Imai, H., "Comparative study of various spectrum matching measures on noise robustness", Proc. ICASSP-86, pp. 769-772, 1986.

|Neben 83| Neben, G., Mc Aulay, R.J. and Weinstein, C.J., "Experiments in isolated word recognition", Proc. ICASSP-83, pp. 1156-1159, 1983.

|Paliwal 82| Paliwal, K.K., "On the performance of the frequency-weighted cepstral coefficients in vowel recognition", Speech Communication 1, pp. 151-154, 1982.

|Petersen 81| Petersen, T.L. and Boll, S.F., "Acoustic noise suppression in the context of a perceptual model", Proc. ICASSP-81, pp. 1086-1088, 1981.

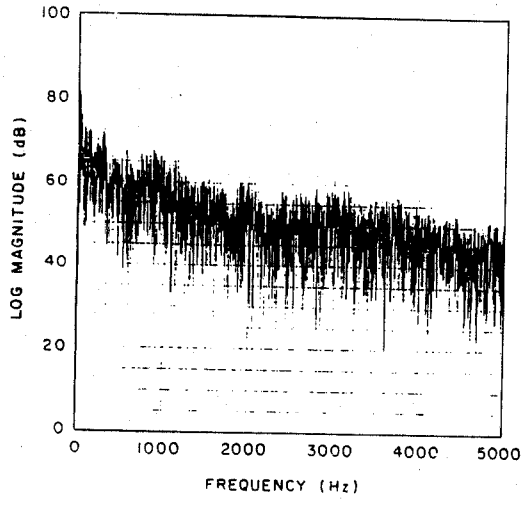
|Seneff 87| Seneff, S., "A model for the trandcution stage of auditory speech processing", JASA 82, S1, p. S83, 1987.

|Soong 87| Soong, F.K. and Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition", Proc. ICASSP-87, pp. 625-628, 1987.

|Tohkura 85| Tohkura, Y., "Speaker-independent ASR of isolated digits using a weighted cepstral distance", JASA 77, S1, p. S11, 1985.

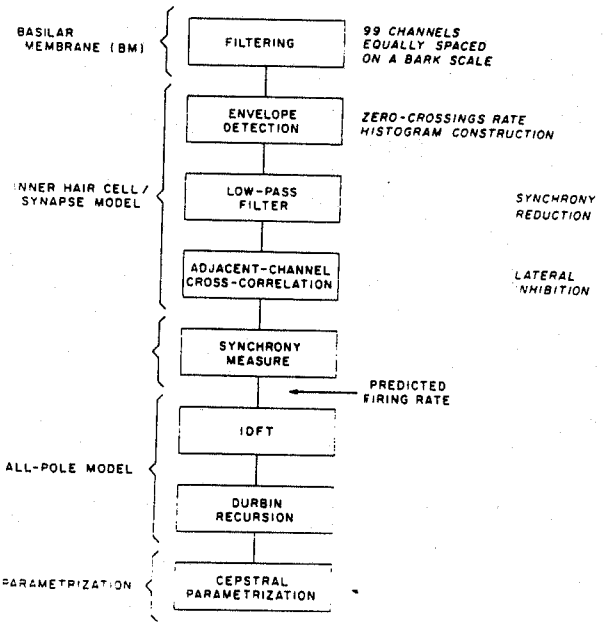
|Tierney 80| Tierney, J., "A study of LPC analysis of speech in additive noise", IEEE ASSP-28, pp. 389-397, 1980.

|Yegnanarayana 79| Yegnanarayana, B. and Reddy, R., "A distance measure based on the derivative of linear prediction phase spectrum", Proc. ICASSP-79, pp. 744-747, 1979.



LOG MAGNITUDE OF NOISE SIGNAL RECORDED IN A CAR

FIG 1.



BLOCK DIAGRAM OF THE SLP ANALYSIS

FIG 2.

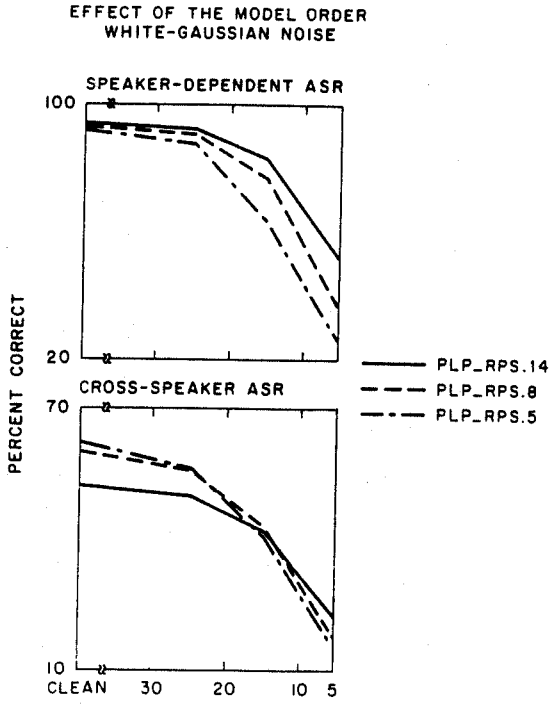


FIG 3.

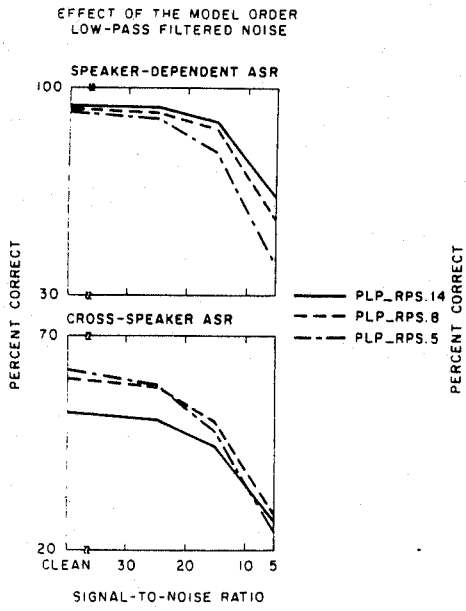


FIG 4.

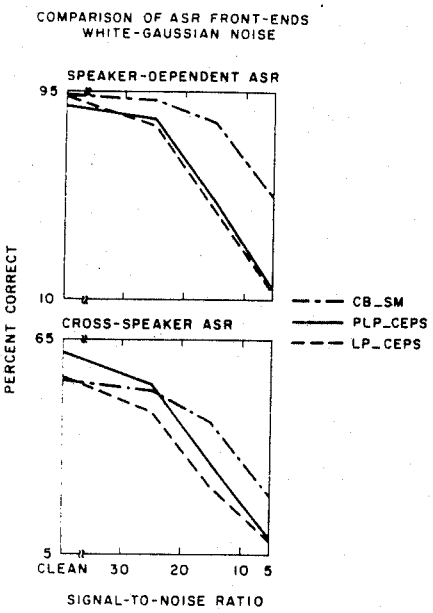


FIG 5.

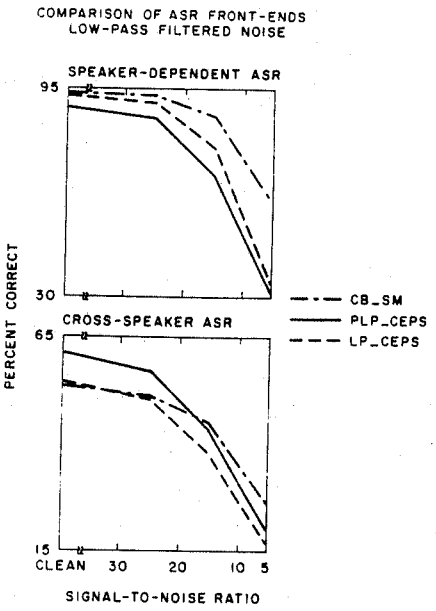


FIG 6.

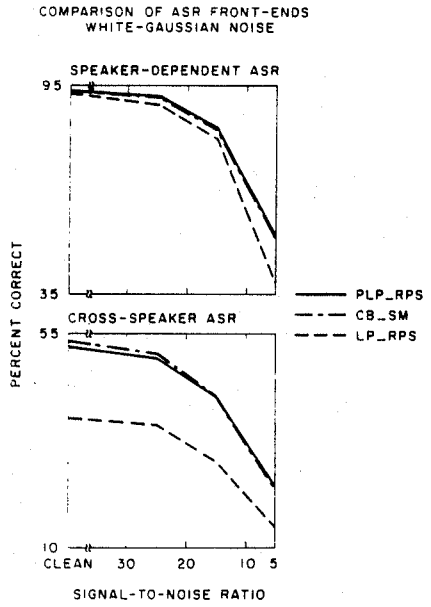


FIG 7.

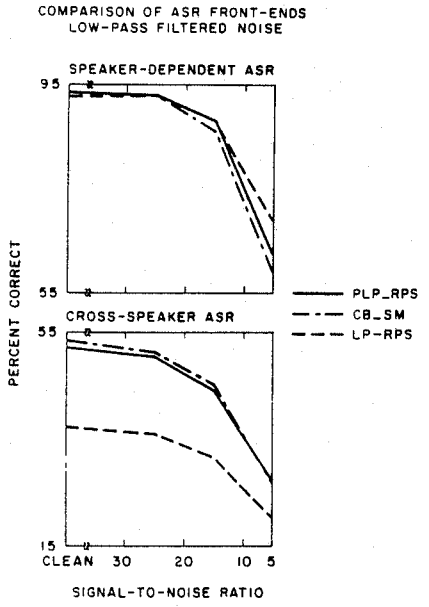


FIG 8.

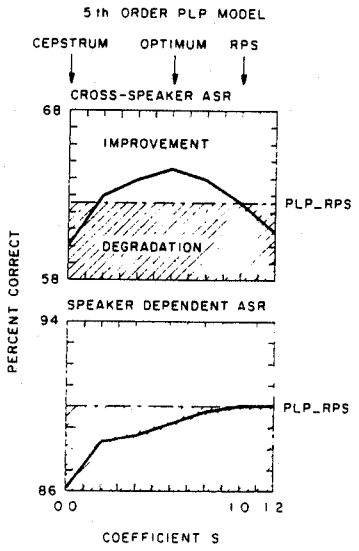


FIG 9.

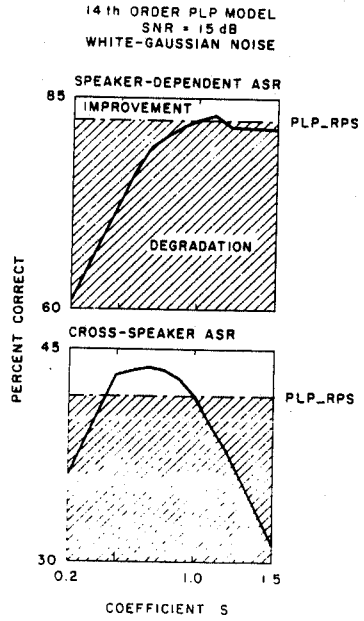


FIG 10.

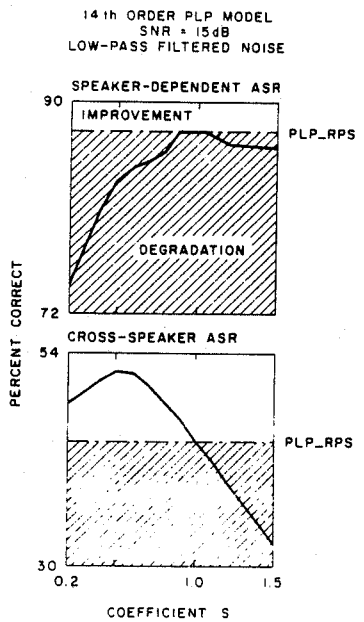


FIG 11.

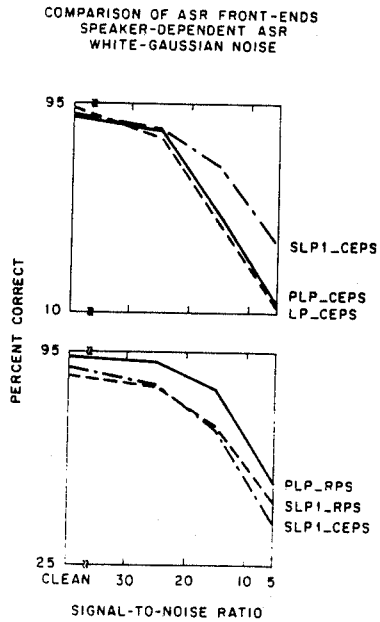


FIG 12.

Speaker-dependent recognition

low-pass filtered noise

Front-ends studied	order of the model	SNR			
		clean	25	15	5
PLP_RPS	5	91.11	88.89	77.50	39.72
PLP_RPS	8	92.22	90.83	85.28	54.72
PLP_RPS	14	93.33	92.78	87.78	62.22
PLP_CEPS	14	88.89	85.28	67.50	30.28
LP_CEPS	14	92.50	90.0	76.11	33.06
LP_RPS	14	92.50	92.50	87.78	68.61
CB_SM	17	93.33	92.50	85.83	58.89

Table 1.

Speaker-dependent recognition

white-Gaussian noise

Front-ends studied	order of the model	SNR			
		clean	25	15	5
PLP_RPS	5	91.11	86.94	62.78	25.56
PLP_RPS	8	92.22	89.72	76.11	35.83
PLP_RPS	14	93.33	91.67	82.50	51.39
PLP_CEPS	14	88.89	83.33	50.0	13.61
LP_CEPS	14	92.50	80.28	46.11	11.91
LP_RPS	14	92.50	88.89	79.17	37.78
CB_SM	17	93.33	91.11	81.91	50.83

Table 2.

Cross-speaker recognition

low-pass filtered noise

Front-ends studied	order of the model	SNR			
		clean	25	15	5
PLP_RPS	5	62.10	58.52	47.65	24.35
PLP_RPS	8	59.97	57.87	49.91	28.33
PLP_RPS	14	52.19	50.34	41.07	26.80
PLP_CEPS	14	61.36	57.10	43.55	20.06
LP_CEPS	14	54.51	50.31	37.96	16.79
LP_RPS	14	37.16	35.80	31.39	19.94
CB_SM	17	53.49	51.20	45.12	26.30

Table 3.

Cross-speaker recognition

white-Gaussian noise

Front-ends studied	order of the model	SNR			
		clean	25	15	5
PLP_RPS	5	62.10	56.05	40.37	14.88
PLP_RPS	8	59.97	55.52	42.72	17.31
PLP_RPS	14	52.19	49.78	41.70	22.01
PLP_CEPS	14	61.36	52.35	30.43	9.51
LP_CEPS	14	54.51	44.48	23.89	8.49
LP_RPS	14	37.16	35.59	27.96	13.64
CB_SM	17	53.49	50.71	41.91	21.39

Table 4.

Speaker-dependent recognition
14 th models order

Front-ends studied	white-Gaussian noise SNR=5dB
PLP_RPS	51.39
LP_RPS	37.78
SLP1_RPS	45.00
SLP2_RPS	51.67
SLP3_RPS	58.33

Table 5.

Bases de données et de connaissances
et
outils

CONSULTATION GENERALISEE DE LA BDSON A L'AIDE DE LA THEORIE DES GRAPHES.

H. BELBACHIR* J.F. SERIGNAT* O. CERVANTES* **

* LCP-INPG GRENOBLE ** LGI-IMAG GRENOBLE.

ABSTRACT

In this paper we describe the implementation of an interface for a general consultation of French GRECO Speech BDSON. A graph theory approach is used for the interface designing. The exploitation of BDSON is made easy by means of this interface : the relational structure of the base does not appear to the user. A greater flexibility is provided by this interface : the user formulates any selection query by indicating only the desired informations and conditions by means of two menus. We show that this approach is still valid through further extensions of the base.

I.INTRODUCTION

La BDSON est une base de données des sons du français supportée par le GRECO-CNRS "Communication Parlée" [CARRE et al, 84]. Elle est gérée par un système de gestion de base de données (SGBD) relationnel. Le langage de manipulation des bases de données relationnelles oblige en général l'utilisateur à faire référence à la structure relationnelle de la base lorsqu'il formule une requête. Pour la BDSON, les utilisateurs éventuels se heurteraient ainsi à la difficulté d'interroger la base en faisant intervenir sa structure interne. Il est intéressant de leur fournir une interface qui leur permette d'exploiter le contenu de la BDSON sans connaître obligatoirement la structure interne de la base.

Une première interface a été réalisée [CERVANTES et al, 86], [DESCOUT et al, 86], et livrée, avec une licence d'une version "run time" du gestionnaire relationnel Knowledgeman (KMAN), à une douzaine de laboratoires français. Elle permet à l'utilisateur de formuler certaines requêtes de sélection à partir de "menus" prédéfinis.

L'interface que nous proposons est plus générale dans la mesure où l'utilisateur peut poser n'importe quelle requête de sélection ou de statistique et d'obtenir seulement les informations désirées. De plus la requête est formulée trés simplement à partir de deux "menus" en mentionnant seulement les attributs et en spécifiant les conditions. Cette interface convient aussi bien pour les requêtes les plus courantes que pour des requêtes particulières formulées par un utilisateur occasionnel.

De nombreux travaux [BOTTINI et al, 81], [DE ARTENOLIS et al, 79], [PRAMANCK et al, 85] ont montré que l'on pouvait exploiter avantageusement la théorie des graphes pour la définition du schéma conceptuel des bases de données relationnelles et pour l'optimisation des techniques d'accès aux données.

Nous utilisons la théorie des graphes pour élaborer une méthode qui traduit la requête utilisateur en une requête exécutable par le SGBD existant. Cette méthode consiste principalement à :
- Représenter le schéma conceptuel de la base sous la forme d'un graphe où chaque sommet représente une relation et les arêtes correspondent aux attributs communs entre deux relations.
- Appliquer un algorithme sur ce graphe permettant la recherche optimale des relations nécessaires à l'exécution de la requête.

II.FORMULATION DE LA REQUETE UTILISATEUR :

Tout d'abord, nous donnons le schéma conceptuel de la BDSON ; pour plus de précision on peut se référer à [CERVANTES et al, 87], [DESCOUT et al, 86].

CORPUS : (*codcor, titre, type-corpus, nbre-élément, type-élément, fréq-échantillonnage, numcor)*.

LOCUTEUR : (*codloc, nom, sexe, âge, caract-linguist, adresse, téléphone)*.

REALISATION : (*codcor, codloc, répétition, date, lieu, longueur, no-bande, no-cass, position, taille-signal, durée, cdrom)*.

CONTENU-CORPUS : (*codcor, no-élém, descr-ortho, descr-phoné)*.

ELEMENT : (*codcor, codloc, répétition, no-élém, position-ds-réalis, longueur, durée-sec, étiq-large, étiq-fin-temp, étiq-fin-spect)*.

Les attributs soulignés constituent les clés des relations.

L'interface que nous proposons permet à l'utilisateur de formuler sa requête en deux étapes :

a) Dans la première étape, un menu présente

tous les attributs disponibles dans la base (voir fig.1). L'utilisateur sélectionne ceux qui l'intéressent, et précise la forme sous laquelle il désire obtenir les résultats :

- affichage ou impression de l'ensemble des données sélectionnées.
- ou seulement des informations statistiques sur ces données.

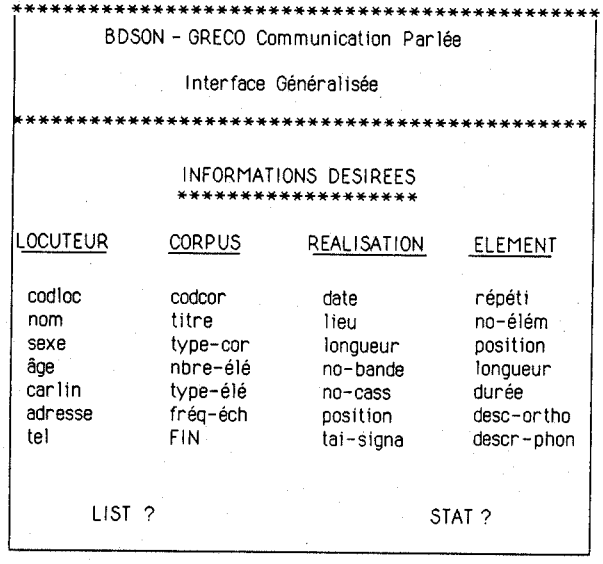


figure 1 : Menu permettant de sélectionner les informations désirées dans une requête.

b) Dans la deuxième étape tous les attributs de la base sur lesquels peuvent porter des conditions apparaissent à l'écran (voir fig.2).

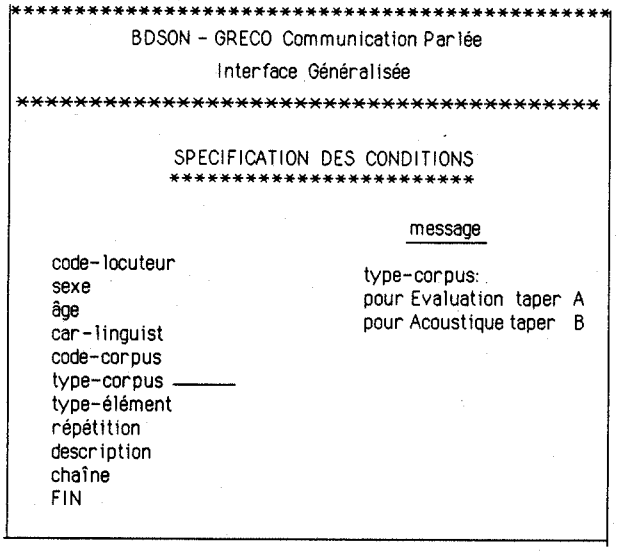


figure 2 : Menu permettant de préciser les conditions de la requête.

Pour chaque attribut sélectionné, une fenêtre est ouverte pour indiquer à l'utilisateur la façon d'introduire les conditions sur cet attribut (par exemple, dans la fig2, l'attribut *type-corpus* est choisi, le message correspondant s'affiche sur l'écran).

Une fois que l'utilisateur a formulé sa question (mentionné les attributs et leurs conditions), le système doit la traduire automatiquement dans le langage de requêtes du SGBD c'est à dire :
- définir les relations de la base qui contiennent les attributs référencés,
- définir les jointures entre ces relations.

Les relations contenant les attributs référencés et les jointures forment l'ensemble des relations nécessaires à l'exécution de la requête. Il peut exister plusieurs solutions. Pour optimiser le temps d'exécution, il est intéressant de considérer l'ensemble donnant le minimum de relations à parcourir lors de la recherche des items demandés.

Nous proposons ci-aprés une méthode qui permet de trouver effectivement l'ensemble minimal de relations nécessaires à l'exécution de la requête.

III. METHODE PROPOSEE :

Comme nous l'avons dit précédemment, l'utilisateur exprime sa requête en mentionnant seulement les attributs et leurs conditions par l'intermédiaire des écrans successifs qui se présentent à l'utilisateur.

Le langage d'interrogation du système de gestion de base de données gérant la BDSON étant de type SQL [ADIBA et al, 86], [DELOBEL et al, 82], la requête exécutable doit être de la forme :

SELECT (attributs)
FROM (relations)
WHERE (conditions)

Le problème est donc de traduire la requête de l'utilisateur en une requête exécutable de type SQL. Il s'agit plus précisément de trouver l'ensemble minimal de relations nécessaires pour son exécution.

III.1. Graphe associé à la BDSON :

D'une façon générale, on considère une base de données relationnelle formée par les relations (R1, R2, ..., RP). Chaque relation comporte plusieurs attributs, un attribut peut être commun à plusieurs relations (jointure). On associe à cette base, un graphe (G = X, U) où l'ensemble des sommets X est formé par l'ensemble des relations de la base :

X = { R1, R2, ..., RP }
et l'ensemble des arêtes U est tel que :

U = { (RI, RJ) où RI et RJ appartiennent à X, et sont telles qu'il existe une jointure entre les relations RI et RJ. }

Nous représentons sur la fig.3, le graphe associé à la BDSON où :

X = { CORPUS, LOCUTEUR, CONTENU-CORPUS, REALISATION, ELEMENT}

Une chaîne dans le graphe est un ensemble d'arêtes consécutives. La longueur de la chaîne est le nombre d'arêtes la constituant. Elle est dite élémentaire si elle ne passe pas plus d'une fois par le même sommet. Soit C un sous ensemble de X tel que : C = (R1,R2,...,Rj), on dit qu'une chaîne couvre le sous ensemble C si elle passe par tous les sommets de C. Elle est dite minimale si c'est la plus petite chaîne (du point de vue longueur) couvrant C.

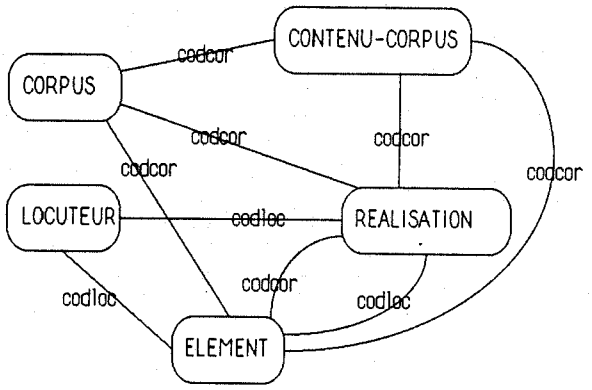


Figure 3 : Graphe associé à la BDSON

III.2. Algorithme traduisant la requête utilisateur en une requête exécutable.

Cet algorithme s'applique sur le graphe associé à la base. Il consiste à affecter à chaque attribut de la requête une couleur différente, et ensuite à marquer, dans le graphe, chaque relation contenant un attribut de la requête par la couleur correspondante.

Les sommets marqués (ou colorés) correspondent à toutes les relations contenant les attributs de la requête. Il s'agit ensuite de choisir un ensemble minimal de relations et de jointures contenant tous les attributs mentionnés. Ce qui revient à choisir dans le graphe une chaîne minimale passant par des sommets couvrant toutes les couleurs.

1) Associer à chaque attribut i de la requête une couleur ci.

2) Colorer dans le graphe le sommet xj par la couleur ci si i est un attribut de la relation Rj correspondante à xj. On remarque qu'un sommet peut avoir zéro, une, ou plusieurs couleurs.

3) Rechercher la chaîne minimale couvrant toutes les couleurs. Soit (R1, R2, ...,RK) cette chaîne. Si plusieurs chaînes existent, le choix optimal (du point de vue du temps d'exécution) dépend de la structure de la base.

4) Formation de la requête exécutable :

SELECT (attributs mentionnés)
FROM R1, R2, ..., RK
WHERE (conditions sur les attributs mentionnés de RI) $\forall I \in [1, K]$
AND (conditions de jointures entre RI et R(I-1)) $\forall I \in [2, K]$

On remarque que pour être valide une requête doit être telle qu'il existe au moins une chaîne regroupant les relations qui contiennent les attributs mentionnés. Dans le cas de la BDSON, le graphe étant connexe, il existe toujours une chaîne couvrant les sommets mentionnés, donc au moins une chaîne minimale [BERGE, 70]. Bien évidemment, il peut arriver que la requête formulée par l'utilisateur n'ait pas de sens pour l'interrogation de la BDSON, de la même façon qu'une requête formulée directement en langage SQL peut ne pas avoir de sens. Cependant, lors de la phase de saisie des conditions de la requête, certains contrôles de cohérence de la requête sont effectués et peuvent se traduire par des messages d'avertissement à l'utilisateur.

IV. QUELQUES EXEMPLES.

Nous présentons ci-aprés deux exemples de requêtes formulées en langage naturel. Nous précisons, pour chacune la façon de l'exprimer à l'aide de l'interface, et nous donnons leur traduction en requêtes exécutables (résultat de l'algorithme).

1) "DONNER LA LONGUEUR ET LA POSITION DES REALISATIONS PRODUITES PAR DES LOCUTEURS PARISIENS, FEMININS, AYANT PRONONCE LES CORPUS DE TYPE EVALUATION".

Les informations désirées portent sur les attributs *Longueur* et *position* que l'utilisateur sélectionne sur le premier menu. Les conditions portent sur les attributs *Carlin="paris", sexe="f", type-corpus="A"* que l'utilisateur précise à l'aide du deuxiéme menu. Cette requête concerne donc les attributs *Longueur, Position, carlin, sexe, type-corpus*. L'algorithme traduit cette requête en la requête exécutable de type SQL suivante :

SELECT *REALIS.Longueur, REALIS.Position*
FROM *REALIS*
FROM *CORPUS* WHERE *CORPUS.type-corpus = "A"*
AND *CORPUS.codcor = REALIS.codcor*
FROM *LOCUTEUR* WHERE *LOCUTEUR.sexe = "f"*
AND *LOCUTEUR.carlin = "paris"*
AND *LOCUTEUR.codloc = REALIS.codloc*

2) "DONNER L'EMPLACEMENT, LA DUREE ET LA LONGUEUR DES ELEMENTS PARMI LES CORPUS DE TYPE MOTS, CONTENANT LA CHAINE DE CARACTERES (DESCRIPTION ORTHOGRAPHIQUE) "EST", PRONONCES PAR DES LOCUTEURS MASCULINS, AVEC UN ACCENT GRENOBLOIS".

Cette requête concerne les attributs *position-ds-réalis, longueur, durée-sec, carlin, sexe, type-élément, descr-ortho*. Les conditions portent sur les attributs : *descr-ortho* = "*est*", *sexe* = "*m*", *type-élément* ="*mots*", *carlin* ="*Grenoble*". Les informations souhaitées portent sur *position-ds-réalis, longueur* et *durée* des éléments correspondants. L'algorithme traduit cette requête utilisateur en la requête exécutable de type SQL :

```
SELECT ELEMENT.position-ds-réalis,
       ELEMENT.longueur, ELEMENT.durée
FROM ELEMENT
FROM CORPUS WHERE CORPUS.type-élément = "mots
            AND CORPUS.codcor = ELEMENT.codcor
FROM LOCUTEUR WHERE LOCUTEUR.sexe = "m"
            AND LOCUTEUR.carlin = "grenoble"
            AND LOCUTEUR.codloc = ELEMENT.codloc
FROM CONTENU-CORPUS WHERE
       CONTENU-CORPUS.descr-ortho IN ["*est*"]
       AND CONTENU-CORPUS.codcor =CORPUS.codcor
```

V.CONCLUSION

Nous avons proposé une méthode pour concevoir une interface d'interrogation généralisée de la BDSON à l'aide de la théorie des graphes. Cette interface a été developpé à l'aide du SGBD KMAN, elle permet d'élargir les possibilités de consultation de la BDSON et n'impose plus à l'utilisateur le contexte restreint (sélection des corpus, sélection des locuteurs, sélection des réalisations ou des éléments) inhérent à l'interface simplifiée de la version initiale [CERVANTES et al, 87]. Ainsi l'utilisateur bénéficie donc :
- d'une souplesse accrue dans la formulation de sa requête et une liberté de choix plus grande des attributs de celle-ci.
- d'une transparence de la structure de la base : il n'est plus confronté au problème des jointures.

En plus, cette interface se prête plus facilement à une extension de la base (ajout d'autres relations et/ou d'autres attributs).

BIBLIOGRAPHIE

[ADIBA et al, 86] M. ADIBA, R. DEMOLOMBE, G. GARDARIN, M. SCHOLL, C. ROLLAND, J. ROHMER, (1986)
"Nouvelles perspectives des bases de données", Eyrolles.

[BERGE, 70] C. BERGE, (1970)
"Graphes et Hypergraphes", Dunod, Paris.

[BOTTINI et al, 81] C. BOTTINI, A. D'ATRI, (1981)
"Schema Hypergraphs : A formalism to Investigate Logical Data Base Design", Graph Theoritic Concepts in Computer Science, LNCS, Vol.100.

[CARRE et al, 84] R. CARRE, R. DESCOUT, M. ESKENAZ, J. MARIANI, M. ROSSI, (1984)
"The French Language Database : Defining, Planning and Recording a Large Database" I.E.E.E - ICASSP, 42.11 San Diego.

[CERVANTES et al, 86] O. CERVANTES, J.F. SERIGNAT, R. DESCOUT, R. CARRE, (1986)
"Définition et réalisation d'une base de données des sons du français", 15èmes JEP GALF, 213-216, Aix-en-Provence.

[CERVANTES et al, 87] O. CERVANTES, J.F. SERIGNAT, J. CAELEN, (1987)
"D"une Base de Données-Sons vers une Base de Données Parole", 3ème Journées B.D.Avancées, AFCET, pp 429-446, Port Camargue.

[DE ARTENOLIS et al, 79] V. DE ARTENOLIS, F. DE CINDIO, G. DEGLI ANTHONI, G. MAURI, (1979)
"Use of Bipartite Graph as a Notation for Data Base", Information Systems, Vol.4.

[DELOBEL et al, 82] C. DELOBEL, M. ADIBA, (1982)
"Bases de données et systèmes relationnels" Dunod.

[DESCOUT et al, 86] R. DESCOUT, J.F SERIGNAT, O. CERVANTES, R. CARRE, (1986)
"BDSON : une base de données des sons du français", 12th International Congress on Acoustics, A4-7, Toronto.

[PRAMANK et al, 85] S. PRAMANCK, D. ITNER, (1985)
"Use of Graph Theoritic Models for Optimal Relational Data Base Accesses to Perform Join", ACM Transactions on Data Base Systems, Vol.10, No.1.

LE CODAGE DE L'ALPHABET PHONETIQUE INTERNATIONAL

Gilbert PUECH

Centre de Recherches Linguistiques et Sémiologiques
Université LUMIERE-LYON 2 - F 69500 Bron

ABSTRACT

The International Phonetic Alphabet is used by linguists to transcribe languages or dialects into phonetic symbols which are independent of orthographic conventions. The ASCII code assigns computer representations to letters of the Roman alphabet but cannot accommodate the large collection of sounds which are attested in natural languages. This paper presents a computer oriented coding system of IPA symbols, in association with their graphic representation. It allows an alphaphonetic classification of words in lexical data bases and the analysis of sounds into their constituents for expert systems.

0. Introduction

Le développement des bases de données sur des langues sans tradition écrite pour lesquelles on ne dispose que de transcriptions phonétiques pose le problème d'une norme de codage dont le rôle serait comparable à celui que joue le code A.S.C.I.I. pour l'alphabet latin. L'alphabet phonétique International (A.P.I.) est le seul système de notation auquel peuvent se référer tous les linguistes indépendamment de la famille de langues sur laquelle ils travaillent. Mais il n'existe pas pour cet alphabet d'ordre conventionnel des symboles. On ne peut donc pas classer les formes transcrites comme on procède à un classement alphabétique - selon un principe qui s'impose à tous - ni créer des index qui s'appuient sur la structure interne des sons pour retrouver une forme dans une base. L'objet de cette communication est de s'appuyer sur l'organisation de l'A.P.I. pour proposer un codage numérique permettant un classement alphaphonétique.

1. La représentation des symboles phonétiques

Chaque symbole phonétique est associé à un code unique qui constitue sa représentation interne pour la base de données. Le symbole a par ailleurs une représentation graphique. Cette représentation, destinée à l'écran et à l'imprimante, inclut les diacritiques d'usage pour les tons, l'accent et certaines caractéristiques surimposées. Dans la mesure du possible, il est préférable de s'en tenir aux symboles sanctionnés par l'A.P.I.

* Je tiens à remercier Pierre Bancel et Pierre Dupont, Université Lumière-Lyon 2, qui m'ont suggéré d'importantes améliorations et mettent à l'épreuve l'efficacité du codage en l'utilisant pour un système expert de comparaison des langues bantu.

Un filtre permet d'associer à une entrée au clavier un code donné et la représentation graphique qui lui correspond. Ces filtres sont programmables de manière à obtenir les correspondances les plus simples possibles, compte tenu de l'inventaire des symboles à générer pour la langue considérée. Ainsi en français on aura :

Touche clavier :	N
Symbole graphique :	ɲ
Code interne :	B908

Dans une langue qui comprendrait aussi une nasale vélaire et une nasale uvulaire, la convention la plus attendue serait au contraire que le filtre associe les séquences ny et ng à [ɲ] et [ŋ] respectivement et la majuscule N à la nasale uvulaire [ɴ]. Pour certaines applications, on peut également avoir intérêt à afficher tous les symboles utilisés sur une partie de l'écran et sélectionner celui qui est choisi à l'aide d'une souris.

La base de données linguistique peut être couplée avec une banque de sons. Dans ce cas, le code, associé à d'autres informations sur la langue, le locuteur, le protocole etc., établit un lien entre un symbole transcrit et l'archivage du signal qui constitue sa réalisation dans un contexte déterminé. A l'intérieur de la base de données, un mot ou un morphème sont représentés par la suite de codes qui correspondent aux symboles constituant leur transcription. Nous joignons en annexe un exemple de mots classés alphaphonétiquement pour une langue du sud du Soudan, le viri.

2. Principes généraux du codage

Les segments sont répartis en consonnes et voyelles ; ils peuvent être marges de syllabe (cas non marqué pour les consonnes) ou centres de syllabe (cas non marqué pour les voyelles). Les consonnes sont organisées dans une table dont les lignes correspondent aux modes et les colonnes aux lieux d'articulation. Pour la table des voyelles, les lignes correspondent au degré d'aperture et les colonnes à leur qualité palatale, centrale ou vélaire. Les voyelles sont en outre caractérisées par la valeur qu'elles prennent pour les traits d'arrondissement, de nasalité et de pharyngalisation. Les segments ainsi définis et illustrés dans les tables 1 et 2 sont dits primaires. Un octet suffit à coder les consonnes et les voyelles primaires ; un octet complémentaire prend en charge les articulations secondaires des consonnes et les informations suprasegmentales portées par les voyelles. Nous

avons donc retenu l'entier court (16 bits) comme format du codage. Trois principes nous ont guidé dans les choix qui ont été faits :

1) l'interprétation phonétique du code doit être immédiate et transparente ;
2) l'organisation du codage doit rester indépendante de toute langue particulière même si certaines options doivent être offertes pour éviter une surcharge inutile ;
3) le codage doit se suffire à lui-même pour permettre un classement alphaphonétique des formes contenues dans la base.

Chaque code constitue une clé d'accès à un segment unique et on peut créer des masques permettant de sélectionner un ensemble de segments partageant les mêmes traits.

3. Les voyelles

Pour signaler le renvoi à la table des voyelles, le bit 12 été mis à zéro ; en numérotation hexadécimale, le code des voyelles se termine donc toujours par une valeur comprise entre 0 et 7. La table 1.a correspond aux voyelles non arrondies et la table 1.b aux voyelles arrondies. Chacune comprend 15 symboles répartis dans un espace à deux dimensions, d'interprétation articulatoire et acoustique : les lignes correspondent aux 5 degrés d'aperture pris en compte (corrélat acoustique F1) ; les colonnes correspondent aux trois positions de la langue dans la cavité orale sur un axe Avant/Arrière (corrélat acoustique F2). Les bits de poids fort 0 à 3 codent ces 15 positions selon un ordre croissant par colonnes :

APERTURE	AVANT	CENTRAL	ARRIERE
	1	6	B
	2	7	C
	3	8	D
	4	9	E
	5	A	F

Le bit 7 permet de choisir entre la table 1.a (valeur 0 pour les voyelles non-arrondies) et la table 1.b (valeur 1 pour les voyelles arrondies). Voici quelques exemples pour les voyelles non arrondies :

i	10	(1er degré, Avant)
æ	50	(5ème degré, Avant)
a	A0	(5ème degré, Central)
ɑ	F0	(5ème degré, Arrière)

et pour les voyelles arrondies :

y	11	(1er degré, Avant)
u	B1	(1er degré, Arrière)
ɒ	F1	(5ème degré, Arrière)

L'ensemble de ces symboles est reproduit en annexe : on trouvera d'une part les matrices correspondant au codage (tables 1.a et 1.b), d'autre part les trapèzes de l'A.P.I..

Le bit 6 est réservé à la nasalité, trait qui peut se combiner avec chacune des possiblités précédentes et est représenté par le diacritique ~ (placé au dessous ou au dessus du symbole primaire suivant que l'on a des marques accentuelles ou non) :

ɛ̃	42	(4ème degré, Avant, Non-arrondi, Nasalisé)
œ̃	43	(4ème degré, Avant, Arrondi, Nasalisé)
ɑ̃	F2	(5ème degré, Arrière, Non-arrondi, Nasalisé)
ɔ̃	F3	(5ème degré, Arrière, Arrondi, Nasalisé)

Le bit 5 servira à coder la pharyngalisation, qu'il s'agisse d'opposer, comme dans de nombreuses langues, notamment africaines, deux ensembles de voyelles entretenant des relations d'harmonie vocalique qui relèvent du trait RLA (Racine de la Langue Avancée) ou les voyelles emphatiques aux voyelles non emphatiques sur le domaine arabe et berbère. Il est en outre fréquent que la pharyngalisation soit elle-même liée à une différenciation du timbre primaire :

i	10	(1er degré, Avant, Non-pharyngalisé)
ɪ	24	(2ème degré, Avant, Pharyngalisé)
u	B1	(1er degré, Arrière, Arrondi, Non-pharyngalisé)
ụ	C5	(2ème degré, Arrière, Arrondi, Pharyngalisé)

Le bit 4 est réservé pour la marque d'accent :

i	10	(Voyelle non accentuée)
'i	18	(Voyelle accentuée)

L'approche retenue permet de coder 15 positions vocaliques de base, pouvant se combiner librement avec 3 caractéristiques de coarticulation et celle d'accent, soit 240 combinaisons offertes. Il existe toutefois d'autres propriétés qu'on peut vouloir coder parce qu'importantes pour telle ou telle langue. Nous proposons de laisser l'interprétation du bit 8 (premier bit du deuxième octet) libre. A titre d'exemple, il peut servir à noter la qualité rétroflexe des voyelles dans une famille de langues et leur dévoisement final ou leur réduction dans une autre.

4. Les mores vocaliques

Une more segmentale est représentée sur un octet ; il s'ensuit que le 2ème octet d'un entier peut servir à représenter une autre more ou porter de l'information suprasegmentale. Dans les langues non tonales, il est sans doute préférable de condenser les transcriptions en codant les voyelles longues et les diphtongues sur le même entier. Dans ce cas le bit 12, homologue du bit 4 réservé à l'accent dans le deuxième octet, restera toujours avec la valeur 0, puisque ce bit sert à distinguer voyelles et consonnes.

Autrement dit, dans cette approche, c'est le segment qui est ou n'est pas accentué et non l'une ou l'autre more :

ai	A010	(diphtongue non accentuée)
'ai	A810	(diphtongue accentuée)
uu	B1B1	(voyelle longue non accentuée)
'uu	B9B1	(voyelle longue accentuée)

Ce codage compacté est exclu pour les langues qui devraient faire appel au bit 8 pour noter telle ou telle propriété du système des voyelles.

5. Les tons

Dans les langues pour lesquelles on souhaite que les tons soient portés par les segments, le premier octet code une more segmentale et le deuxième octet une information tonale. Une diphtongue ou une voyelle longue sont alors transcrites sur deux entiers consécutifs.

Il existe plusieurs systèmes de notation des tons. Pour les langues asiatiques, le système généralement utilisé fait référence à une échelle de 5 paliers (croissant de 1 à 5) pour définir les contours phonétiques des tons phonologiques. Les 4 tons du chinois mandarin sont ainsi décrits comme des contours dont on note les paliers initial et final :

ton 1 :	55	(Haut)
ton 2 :	35	(Montant)
ton 3 :	214	(Descendant-montant)
ton 4 :	51	(Descendant

Un ton, dans cette approche, est noté par deux valeurs au moins. Dans les langues africaines l'unité de notation est le niveau, transcrit par une seule valeur. Toutefois une more peut porter un ton modulé qui requiert, au même titre qu'un contour, deux valeurs. Il est nécessaire de prévoir 4 niveaux au moins : Bas, Moyen, Haut et Suprahaut. De nombreuses langues opposent par ailleurs un ton Bas descendant à un ton Bas maintenu. La solution la plus simple est dès lors de décrire le Bas descendant comme un ton modulé de Bas à Infrabas, d'où 5 niveaux à interpréter ainsi :

1. Infrabas
2. Bas
3. Moyen
4. Haut
5. Suprahaut

Pour coder 5 valeurs de référence, il faut nécessairement 3 bits. On peut donc associer à une more segmentale deux valeurs tonales au plus. Un ton doublement modulé (Descendant-montant par exemple) devra être associé à deux mores. L'organisation proposée pour le 2ème octet est la suivante :

bits 9 à 11 :	valeur tonale 1
bit 12 :	choix tables V/C
bits 13 à 15 :	valeur tonale 2

Cette répartition permet une interprétation transparente du codage des tons en numérotation hexadécimale, comme le montrent les exemples suivants :

ĭ	1024	(Bas-Haut)
î	1021	(Bas-Infrabas)

Si la voyelle est accentuée, le bit 4 prend la valeur 1, d'où :

'ĭ	1840	(Haut, more accentuée)
'ĭ	1820	(Bas, more accentuée)
'î	1842	(Haut-Bas, more accentuée)
'ĭ	1824	(Bas-Haut, more accentuée)

L'information tonale est dissociable de l'information segmentale pour permettre notamment de distinguer suite segmentale et schème tonal. Une more à laquelle n'est pas encore affecté un ton prend pour le deuxième octet la valeur 00 :

i	1000
'i.	1800

Inversement un ton non projeté est transcrit avec la valeur 00 sur le premier octet et la valeur affectée aux bits 9 à 11 (ton 1) sur le deuxième octet :

H	0040	(Ton Haut non projeté)

Il est impératif de pouvoir coder une faille tonale (downstep), dont le statut est phonologique, et il est souhaitable de pouvoir faire appel au même codage pour la dérive tonale (downdrift) pour les transcriptions phonétiques étroites. La valeur 6 (E si le bit 8 est à 1) sur le troisième quartet sera affectée à la notation d'un abaissement (traité dans l'espace réservé au ton 1) du niveau tonal de référence suivi de la notation du ton abaissé (dans l'espace réservé au ton 2). Soit par exemple le mot [k í t↓á], dans lequel il y a une faille tonale entre les deux tons Haut consécutifs :

k	1B08
í	1040
t	1408
↓á	A064

Pour noter un profil tonal tel que [¯ _ —], dans lequel le deuxième ton phonologique Haut est séparé du premier par un ton Bas et présente une réalisation abaissée par rapport à ce dernier (dérive tonale), on aura pour un support vocalique tel que [i] la séquence :

1040 ..... 1020 ..... 1064.

Le cas inverse d'élévation tonale (upstep) sur des séquences de Haut ou de Suprahaut, bien que rare, a été documenté pour quelques langues. La valeur 7 (F si le bit 8 est à 1) est disponible sur le troisième quartet par un codage symétrique :

↑á	A074 (ton Haut avec élévation tonale)
↑á	A075 (ton Suprahaut avec élévation tonale)

6. Les consonnes

L'organisation générale proposée pour les consonnes est la suivante :

- Les deux premiers quartets donnent accès à un symbole primaire contenu dans la table 2 ;
- le troisième quartet code pour les géminées la duplication et pour les consonnes complexes le changement du mode d'articulation avec maintien du lieu ;
- le quatrième quartet code une articulation secondaire.

Les quinze lignes de la table des consonnes correspondent à

des modes et les quinze colonnes à des lieux d'articulations. La table 2 est sélectionnée lorsque le bit 12 a la valeur 1, d'où un code se terminant par 8 en l'absence d'articulation secondaire et une valeur comprise entre 9 et F sinon :

p	1108
b	4108
k	1B08
g	4B08
s	9408
z	A408

Pour les consonnes géminées, le troisième quartet a une valeur identique au premier :

bb	4148
ss	9498

Les affriquées peuvent aussi être géminées, d'où l'intérêt de les inclure dans la table des consonnes primaires bien qu'elles présentent un changement de mode d'articulation comme les consonnes complexes. Cette approche permet aussi d'inclure dans le codage les affriquées non homorganiques :

ts	7408
tts	7478
kf	7C08
kkf	7C78

Pour les autres segments complexes, réputés homorganiques, la deuxième valeur du mode d'articulation est différente de la première valeur. C'est le cas des prénasalisées :

ns	B498
nz	B4A8

C'est également le cas des consonnes avec relâchement nasal, latéral ou trillé homorganique :

tn	14B8
tl	14D8
dr	44F8

Les articulations secondaires non homorganiques sont codées sur le quatrième quartet, avec une valeur correspondant à celle des symboles de la table des consonnes :

9	palatalisation	– tʲ	1409
A	labiopalatalisation	– kᶣ	1B0A
B	vélarisation	– tˠ	140B
C	labiovélarisation	– kʷ	1B0C
D	uvularisation	– tˁ	140D
E	pharyngalisation	– tˤ	140E

Une articulation secondaire est compatible avec la gémination ou un mode d'articulation complexe :

ttʲ	1419
ntsʲ	B479

La glottalisation (ou la laryngalisation) qui correspondrait à la valeur F ne peut être mise sur le même plan que la vélarisation ou la pharyngalisation puisque ce qui est en jeu alors est un mode de phonation et non une configuration articulatoire. Nous utiliserons cette valeur laissée libre sur le quatrième quartet pour conférer au troisième quartet une autre signification que celle qui est la sienne précédemment. Il s'agit en effet de compacter, puisque l'option a été prise de travailler sur des mots de 16 bits, des informations parfois nécessaires que l'on doit pouvoir coder mais pour lesquelles aucun espace spécifique n'a pu être réservé. Ces informations concernent :

a) la syllabicité des consonnes ;
b) les modes de phonation ;
c) certaines co-caractéristiques articulatoires des symboles primaires.

Lorsque le quatrième octet a la valeur F, nous proposons que le troisième quartet s'interprète avec les conventions suivantes :

a)	0	consonne syllabique non tonale
	1 à 5	consonne syllabique portant un ton
b)	6 à 9	dévoisement, voisement, laryngalisation, murmure
c)	A à E	labialisation, latéralisation, chuintement, sifflement, apicalité
d)	F	rétention du relâchement

Les nasales et les liquides peuvent fonctionner comme centres de syllabe et porter un ton. Le plus simple est alors de reprendre pour le troisième quartet les conventions appliquées pour les voyelles, d'où :

n̩	B10F	(nasale syllabique non tonale)
m̀	B12F	(nasale syllabique avec ton Bas)
m̀	B14F	(nasale syllabique avec ton Haut)

La valeur 6 servira à transcrire le non voisement de consonnes réputées voisées :

n̥	B16F	(nasale non voisée)
l̥	D46F	(latérale non voisée)

et la valeur 7 le voisement de consonnes réputées non voisées :

ʘ̬	617F	(click bilabial voisé)

Les valeurs 8 et 9 ont été affectées à la laryngalisation (creaky voice) et à la voix murmurée (breathy voice) :

b̰	418F
b̤	419F

Le cas des fricatives latéralisées est particulièrement important comme le confirme le fait qu'il existe des symboles A.P.I. pour ces sons :

ɮ	A4BF	(Fricative alvéolaire latéralisée)

De même le non relâchement d'une occlusive sera noté

tˀ	14FF

7. Coloration consonantique

Certaines voyelles peuvent être influencées par un élément consonantique qui n'est pas pleinement réalisé comme segment. Il en va ainsi par exemple des voyelles rhotacisées de l'anglais ou des voyelles suivies d'un appendice nasal dans certaines variétés du français méridional. On notera cette coloration consonantique en donnant au premier quartet la valeur 0 (absence de mode d'articulation primaire) et en notant respectivement sur le deuxième et le troisième quartet le lieu et le mode (articulation secondaire) de la coloration consonantique :

ɘ^{ɹ}	8000	07C8
a^{n}	A000	04B8

8. Schèmes tonaux et prosodie

Dans nombre de langues une même entrée lexicale peut se réaliser sous différentes formes suivant le contexte. Dans certains parlers bantu par exemple, les tons portés par une forme nominale ou verbale dépendent des morphèmes constituants mais aussi de la position dans l'énoncé. Il est alors souhaitable de dissocier dans la base l'information sur le schème tonal du thème et l'information segmentale. Nous avons proposé de coder les tons associés en donnant la valeur 00 au premier octet. Des règles de projection permettent ensuite de réaliser une association :

B100 0040 > B140
(association à la voyelle [u] d'un ton Haut)

On peut aussi projeter un même ton sur plusieurs mores porteuses :

B100 1108 1000
> B130 1108 1030
0030
(association d'un ton Moyen aux deux mores vocaliques)

Des possibilités de codage analogues s'appliquent aux langues dont la phonologie requiert la définition d'un domaine prosodique. Ce domaine pourra alors être délimité par des marqueurs ; ainsi le domaine de l'emphase dans les parlers arabes ou berbères peut être formellement marqué en intercalant entre les segments les codes : 0400 ... 0400 dans lesquels le bit 12 a la valeur 0 (référence à la table des voyelles) et le bit 5 a la valeur 1 (pharyngalisation d'une voyelle). Dans les langues pour lesquelles il est nécessaire de distinguer un accent primaire d'un accent secondaire, on fera précéder la voyelle concernée du code 0800 si l'accent est primaire :

' t i	1408 / 0800 / 1800	(accent primaire)
, t i	1408 / 1800	(accent secondaire)

9. Les frontières

Les frontières de morphème donnent une information qui peut s'avérer indispensable dans l'exploitation d'une base. Il est le plus souvent nécessaire d'opérer avec deux types de frontières, faible et forte, correspondant généralement à la morphologie flexionnelle et à la morphologie dérivationnelle. Nous proposons de leur faire correspondre les codes 000F et 00FF respectivement. Il est évident que d'autres frontières peuvent être codées selon le même principe, notamment la frontière de syllabe.

10. Les masques et les tests de sélection

Il est nécessaire d'apparier la représentation interne (le code) et la représentation graphique des symboles API telle qu'elle figure dans les tables et est contenue dans le jeu de caractères disponible sur l'écran et l'imprimante. Le premier choix porte sur la sélection de la table des voyelles ou celle des consonnes. On utilise à cette fin un masque dans lequel seul le bit 12 est mis à 1, le ET logique et un algorithme interprétant le résultat d'une comparaison arithmétique. En prenant pour exemple le code correspondant au segment t (code 1408), on aura :

Masque de sélection des tables V / C	:	0000	0000	0000	1000
Représentation interne du segment	:	0001	0100	0000	1000
ET logique	:	0000	0000	0000	1000

Comparaison : SI résultat > 0 ALORS segment = consonne
SINON segment = voyelle

Décision : Le segment de code 1408 est choisi dans la table des consonnes

Par des opérations du même ordre, on sélectionne ensuite le symbole requis dans la table choisie. De même des tests opérant sur un ou plusieurs bits permettent de raisonner sur la structure phonétique des formes à des fins de comparaison interdialectales ou de reconstruction de proto-formes par des systèmes experts.

Conclusion

De même que l'Alphabet Phonétique International permet aux linguistes de partager l'interprétation des données quelle que soit la langue, de même le développement des bases de données linguistiques rend nécessaire l'émergence d'un standard dans le codage informatique des transcriptions, de manière à faciliter la communication et permettre un principe de classement alphaphonétique commun. Cette étude se veut une contribution en ce sens. Le codage proposé peut paraître trop lourd pour des langues comme le français qui n'ont ni tons, ni consonne à articulation secondaire. Pour de telles langues, le codage sur le premier octet suffit aussi bien pour les consonnes (puisqu'elles sont toutes contenues dans la table des consonnes primaires) que pour les voyelles. Cette organisation en deux niveaux permet de concilier les exigences d'économie et d'universalité.

REFERENCES

CATFORD J.C. [1977] *Fundamental problems in Phonetics*, Bloomington : Indiana University Press.

INTERNATIONAL PHONETIC ASSOCIATION [1949] *The principles of the International Phonetic Association*, London : Department of Phonetics, University College.

KOUSSANI R. [1983] *Esquisse de description du viri (Soudan)*, mémoire de DEA, Université Lyon 2.

LADEFOGED P. [1982] *A course in Phonetics*, 2ème édition, New-York : Harcourt Brace Jovanovich.

PULLUM G.K. et W.A. LADUSAW [1986] *Phonetic Symbol Guide*, Chicago, London : The University of Chicago Press.

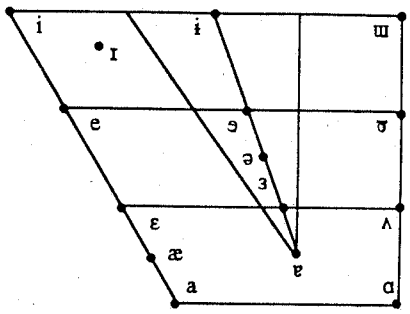
THOMAS J.M.C., BOUQUIAUX L. et F. CLOAREC-HEISS [1976] *Initiation à la Phonétique*, Paris : P.U.F.

ANNEXE

	Transcription glose	Code	
1	tū terre	1408	B140
2	tū sésame	1408	B142
3	dǎ nuit	4408	A024
4	dɔ̌ oignon	4408	E120
5	dɔ̄ mensonge	4408	E130
6	ɓī main	5108	1020
7	ɓá maison	5108	A040
8	sɛ̀ poison	9408	4020
9	sɛ́ escargot	9408	4040
10	lī rêve	D408	1030
11	lí fruit	D408	1040
12	lǐ sommeil	D408	2020
13	lī soleil	D408	2030
14	lí miel	D408	2040
15	lɛ̀ nom	D408	4020
16	lɛ́ coeur	D408	4040

i	ɨ	ɯ
ɪ	ɤ	
e	ə	ʊ
ɛ	ɐ	ʌ
æ	a	ɑ

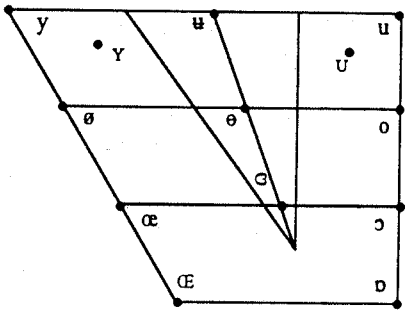
Table 1.a : *voyelles non arrondies*



Placement des voyelles non arrondies selon l'A.P.I.

y	ʉ	u
ʏ	ʊ	ʊ
ø	ɵ	o
œ	ɞ	ɔ
Œ	ɐ	ɒ

Table 1.b : *voyelles arrondies*



Placement des voyelles arrondies selon l'A.P.I.

Table 3 : *consonnes*

	bilabiale 1	labiodentale 2	dentale 3	alvéolaire 4	labioalvéolaire 5	rétroflexe 6	postalvéolaire 7	prépalatale 8	palatale 9	labiopalatale A	vélaire B	labiovélaire C	uvulaire D	pharyngale E	glottale F
Occlusives non voisées 1	p			t	p t	ʈ			c		k	kp	q		ʔ
Occlusives aspirées 2	pʰ			tʰ							kʰ				
Occlusives éjectives 3	p'			t'							k'				
Occlusives voisées 4	b			d	bd	ɖ			ɟ		g	gb	ɢ		
Occlusives implosives 5	ɓ			ɗ							ɠ				
Clicks (non voisés) 6	ʘ		ǀ	ƅ		ǃ	ƈ								
Affriquées non voisées 7		pf		ts			tʃ					kf			
Affriquées voisées 8		bv		dz			dʒ					gv			
Fricatives non voisées 9	ɸ	f	θ	s		ʂ	ʃ	ɕ	ç		x		χ	ħ	h
Fricatives voisées A	β	v	ð	z		ʐ	ʒ	ʑ	ʝ		ɣ		ʁ	ʕ	ɦ
Nasales B	m	ɱ		n		ɳ			ɲ		ŋ		N		
Approximantes centrales C	ʋ						ɹ		j	ɥ		w			
Approximantes latérales D				l		ɭ			ʎ		ɬ				
Approximantes battues E				ɾ		ɽ									
Trilles F	ʙ			r									ʀ		

Lexiques et groupes consonantiques.*

AUBERGE V. [1,2,3], BOE L.J. [1], LEFEVRE J.P. [2].

1 : Institut de la Communication Parlée - Grenoble.
2 : Société OROS - Meylan.
3 : CRISS - Université des Sciences Sociales - Grenoble.

ABSTRACT

Even if of practical interest for a lot applications, exhaustive lists of the consonant clusters allowed in French are not currently available.

In the first part of this paper, we derive such lists from existing material. On one side, we have extracted consonant clusters with their frequencies of occurence (lexical frequencies) from a large and representative dictionary. On the other side, the exploitation of a broad corpus elaborated from real recordings gave us more information on the statistical repartition of the clusters in natural speech.

Then, from these results, we suggest a classification of the consonants clusters for the French language. Lastly, in the framework of a text-to-speech synthesis application, we propose an automatic syllabification approach, based on the use of both lexicons of clusters and rules of segmentation.

1. INTRODUCTION.

Les descriptions phonotactiques du français offrent peu de résultats concernant les groupes consonantiques. C'est aussi le cas d'ailleurs au niveau de l'acoustique et de la production. Les travaux de ROCHETTE(1973) constituent une exception et ce n'est que très récemment que l'étude de l'organisation temporelle du signal de parole a permis de mettre en évidence des schémas moteur différents selon la nature des cycles considérés (VC ou VCC par exemple), et donc des consonnes composantes (SOCK & al, 1987).

Il faut voir dans cette lacune non pas un manque d'intérêt pour ce type de données, mais plutôt la conséquence de la difficulté à réunir un volume de données suffisamment important pour satisfaire aux exigences de la précision statistique**. La comparaison et/ou l'utilisation des groupes consonantiques trouvent de vastes champs d'application, à la fois dans les domaines de la description (WIOLAND, 1985), du classement de consonnes, de la syllabation (DELATTRE 1966), des liaisons, de l'élision du E latent (VAN EIBERGEN-THIEULLE, 1986), de l'établissement des règles de prononciation, de la construction de classes (CARLSON & al., 1985), ainsi que pour l'accès rapide au lexique en reconnaissance par exemple (HATON, 1984), pour la définition de dictionnaires acoustiques de segments élémentaires (di- tri- quadri- ... phones), et pour l'évaluation de la qualité de la parole synthétique pour ce type de segments (VAN BEZOOIJEN, 1988).

Nous appellerons groupe consonantique **GC** les séquences phonétiques de 2 consonnes et plus, comprises dans les frontières du mot, laissant ainsi de côté les groupes résultant de jointures entre mots.

Dans le but premier de l'évaluation de l'intelligibilité d'un système de synthèse du français (VAN SON, 1987), nous avons cherché vainement une liste exhaustive des groupes consonantiques phonétiques du français. En l'absence d'une telle liste, nous avons cherché à l'établir à partir de lexiques disponibles pour le français. Ainsi, notre matériau est tiré du dictionnaire **DELA** du Laboratoire d'Automatique Documentaire et Linguistique à Paris VII, ainsi que du lexique phonétique **BDPHO** (TUBACH & BOE, 1985a).

A partir du DELA, nous avons constitué un lexique phonétique en utilisant une phonétisation automatique mise en oeuvre auparavant pour des études linguistiques (AUBERGE & al., 1987), lexicales (PERENNOU & DE CALMES, 1987), ainsi que pour la synthèse. Il s'agit donc d'un matériau phonétique quelque peu "artificiel", mais justement représentatif de ce que peut être l'entrée d'un synthétiseur. Les 50.000 entrées phonétisées du DELA sont suffisantes pour prétendre à un ensemble représentatif de GC du français. Particularité sur laquelle nous reviendrons par la suite, tous les E latents dans les frontières du mot ont été maintenus lors de la phonétisation, ce qui a limité la formation de certains GC (VAN EIBERGEN). Ce lexique est donc quantitativement exploitable au

* Ce travail a été partiellement réalisé dans le cadre du projet SPIN, sous financement de la Communauté Economique Européenne, par l'intermédiaire du programme ESPRIT.

** Lorsque l'on cherche à évaluer une probabilité p par des comptages correspondant à un nombre d'occurrences n pour un corpus de longueur N, soit une fréquence f = n/N, pour un intervalle de confiance à 95% l'incertitude absolue sur p est donnée par :

$$|p-f| = 2\,[\,p(1-p)/N\,]^{1/2}$$

La précision augmente donc avec la racine carrée de la taille du corpus.

sens des fréquences d'apparition des GC dans le lexique, que nous appellerons fréquences lexicales.

BDPHO, quant à lui, a été élaboré à partir d'un corpus d'environ 300.000 sons enregistrés, correspondant à 28h 30 de parole transcrite finement, pour 70 locuteurs au total 102.000 mots issus de 7.000 formes différentes. Peu représentatif au niveau des fréquences lexicales des GC, ce lexique est en revanche riche en indications sur les fréquences d'apparition des GC dans le discours, que nous appellerons fréquences d'usage.

Le choix de ces deux types de matériau nous semble ainsi justifié :

- Un lexique comme DELA permet de tendre vers un repérage exhaustif de tous les GC du Français. La fréquence lexicale (c'est-à-dire la capacité d'un GC à construire les mots du lexique) est un indice du système consonantique. De telles données sont par exemple utiles à des études structurelles des GC, à l'établissement de listes exhaustives pour l'évaluation de systèmes de synthèse sur les difficultés propres aux GC , à la mise au point de procédures de syllabation automatique (nécessaires à la synthèse vocale depuis l'écrit ou bien à la césure des mots dans les traitements automatiques de l'écrit).

- Un lexique comme BDPHO fournit des indications sur l'usage des GC dans le discours. La fréquence d'usage est la projection consonantique de l'usage lexical des mots du discours. Ces données serviront en particulier à l'interprétation des résultats de l'évaluation de la synthèse.

A partir des résultats statistiques obtenus nous proposons une classification des consonnes et une approche concernant la syllabation automatique dans le cadre de la synthèse à partir du texte.

2. LES RESULTATS STATISTIQUES.

Il nous faut noter tout d'abord qu'en français les consonnes se taillent une part d'importance par rapport aux voyelles, environ 55%, moins qu'en allemand ou en anglais mais plus qu'en espagnol par exemple. Jusqu'en 1968, on ne disposait que d'informations éparses contenues dans des travaux consacrés essentiellement aux fréquences d'occurrences des sons (voir par exemple CHAVASSE, 1948). Depuis lors, ont paru des recherches spécifiques sur les GC (ROSSI, 1968 ; TUBACH, 1969 ; MALECOT, 1974 ; VAN EIBERGEN- THIEULLE, 1986 ; et surtout WIOLAND, 1985a,b), et d'autres plus générales mais incluant cependant des études utiles pour les GC (HATON & LAMOTTE, 1971 ; TUBACH & BOE, 1985b). Au niveau orthographique, il est important de noter le travail de LECLERE (1984).

Nous présentons ici une partie des résultats qui nous semblent importants d'un point de vue descriptif et qui éclairent, à notre avis, la structure et le fonctionnement des GC en tant que révélateurs systémiques de la syllabe. Les listes L1 (a,b et c), issues du DELA (respectivement L2 (a,b et c), issues de BDPHO) livrent les GC les plus occurrents au sens des fréquences lexicales (respectivement d'usage), quelque soit le nombre >= 2 de consonnes constituantes. Les listes L1&L2 (a) contiennent les GC initiaux, L1&L2 (b) les GC intervocaliques et L1&L2 (c) les GC finaux. Nous nous sommes limités ici à la présentation de GC de fréquence lexicale ou d'usage supérieure ou égale à 1 %. Notons cependant que le nombre total de GC extraits de DELA s'élève à :

- en initiale : 122 (dont 21 >= 1 %)
- en médiane : 417 (dont 33 >= 1 %)
- en finale : 104 (dont 13 >= 1 %)

Dans DELA, les 122 GC initiaux ont été extraits de 7017 mots (soit 14 % du lexique total DELA), les 417 GC intervocaliques de 26044 mots (soit 52 % du DELA), les 104 GC finaux de 4325 mots (soit 8.7 % du DELA).

Pour BDPHO, le nombre total de GC s'élève à :

- en initiale : 97 (dont 19 >= 1 %)
- en médiane : 218 (dont 28 >= 1 %)
- en finale : 61 (dont 23 >= 1 %)

Si on compare ces résultats avec ceux du DELA, ils confirment la représentativité de ce lexique.

Certains GC des listes L2 (a, b, c) ne se retrouvent pas dans L1 (a, b, c) ou bien apparaissent avec une fréquence très inférieure. Comme on peut le constater dans les listes L2 (a, b, c), la raison en est souvent l'emploi répété de certaines formes. A titre d'exemple, dans L2 (c), le GC lk provient uniquement de l'occurrence de la forme quelque (ou des ses dérivés de même prononciation).

Afin de simplifier la présentation, ces formes sont donnnées sous une seule représentation orthographique possible de leur prononciation proche ou immédiate. Entre "()" après un mot est donné son nombre d'occurrences. Ainsi, par exemple, le mot "exemple" (100), dans L2 (b) a été trouvé comme [Egzãpl](64), [Egzãplə](34), [egzãplə](2). Lorsque qu'une forme est suivie de "..." c'est que le GC corrrespondant provient de la même famille (par exemple : voir... = voir, vois, voyez, voyons.).

Ce type de renseignements doit trouver une application pour la constitution de corpus d'analyse acoustique (par exemple ils auraient sûrement été utiles pour la construction du corpus BDSON), ou bien dans le choix de segments élémentaires pour les bibliothèques de synthèse, ou encore dans l'établissement de listes de GC et de critères d'interprétation des résultats dans le cadre de l'évaluation de la parole synthétique. C'est de leur dépouillement que nous obtenons en particulier des indices sytémiques pur la classification des consonnes, et pour la syllabation.

Groupes consonantiques les plus "fréquents" du français.

Liste L1-a

GC	Fréquence lexicale
pR	13.17 %
tR	12.23 %
kR	6.34 %
gR	5.99 %
bR	5.14 %
pl	4.95 %
fR	4.70 %
st	4.52 %
kl	3.99 %
fl	3.46 %
dj	3.08 %
sp	2.71 %
gl	2.37 %
dR	2.14 %
bl	1.77 %
sk	1.61 %
stR	1.48 %
pj	1.17 %
ps	1.15 %
bj	1.11 %
pw	1.05 %

Liste L1-b

GC	Fréquence lexicale
sj	9.94 %
tR	5.11 %
st	5.00 %
kt	3.20 %
gR	3.03 %
Rt	2.78 %
Rm	2.47 %
kR	2.20 %
Rj	2.17 %
ks	2.13 %
Rs	1.96 %
Rd	1.90 %
nj	1.75 %
sk	1.69 %
pR	1.66 %
bl	1.62 %
lj	1.57 %
pl	1.52 %
Rn	1.52 %
bR	1.49 %
dj	1.43 %
dR	1.39 %
Rb	1.36 %
pt	1.23 %
kl	1.21 %
lt	1.19 %
Rk	1.17 %
stR	1.13 %
ksj	1.12 %
zj	1.11 %
tj	1.11 %
sp	1.09 %
fj	1.04 %

Liste L1-c :

GC	Fréquence lexicale
bl	20.94 %
sm	19.19 %
st	17.81 %
tR	7.78 %
dR	3.29 %
ks	2.88 %
Rm	2.53 %
sk	1.91 %
Rn	1.66 %
Rd	1.40 %
stR	1.24 %
kl	1.08 %
Rs	1.04 %

Liste L2-a

GC	Fréquence d'usage	
pR	16.32 %	
tR	14.62 %	
pl	8.74 %	
bj	7.82 %	bien (600) = 7.43 %
mw	7.04 %	moi (389) = 4.82 %
fR	5.80 %	
vw	4.58 %	voir ... (204), voilà (58) = 3.25%
gR	3.87 %	
kRw	3.32 %	crois (238) = 2.95%
vR	2.32 %	
sw	2.26 %	
tRw	1.85 %	
vj	1.33 %	
kR	1.30 %	
kl	1.20 %	
Rj	1.14 %	rien (90) = 1.12%
fw	1.08 %	
st	1.00 %	
pw	1.00 %	

Liste L2-b

GC	Fréquence d'usage	
sj	9.43 %	
tR	7.89 %	
Rt	6.18 %	
Rs	5.29 %	
st	3.94 %	
pR	3.60 %	entreprise (38), après (131), impression (32) = 1.57 %
bl	3.18 %	problème (104) = 0.81 %
kt	2.50 %	
Rm	2.15 %	
zj	1.98 %	télévision (16), plusieurs (19), deuxième (20), troisième (25), parisien...(73) = 1.2 %
Rj	1.88 %	
gz	1.87 %	exemple (100), exactement (24) = 0.97 %
pl	1.80 %	
vw	1.75 %	avoir (122), pouvoir (25), savoir (32) = 1.38 %
Rd	1.58 %	
lk	1.50 %	quelque (135) = 1.05 %
dR	1.46 %	
sp	1.38 %	
gR	1.35 %	
kR	1.33 %	
gl	1.27 %	
stR	1.20 %	
sk	1.20 %	puisque (44), discuter (12), presque (16) = 0.36 %
Rl	1.15 %	parle ...(116) = 0.91 %
lj	1.10 %	
dj	1.05 %	
ks	1.01 %	
Rn	1.01 %	

Liste L2-c

GC	Fréquence d'usage	
tR	21.67 %	être (98), entre (20), autre (90), mettre (22) = 12.52 %
bl	10.13 %	
st	7.84 %	
sk	5.39 %	
dR	5.23 %	
pl	4.25 %	exemple (64) = 1.9 %
Rt	4.19 %	
Rs	3.32 %	bourse (12), commerce (14) = 1.42 %
vR	3.16 %	livre (35), vivre (12) = 2.56 %
sm	2.94 %	
bR	2.72 %	libre (10), nombre (9) = 1.03 %
Rsk	2.56 %	parce que (36), lorsque (11) = 2.56 %
Rm	2.18 %	
kl	2.01 %	
RS	2.01 %	recherche (25) = 1.36 %
kt	1.91 %	architecte (9), contact (9) = 0.98 %
kst	1.74 %	texte (27) = 1.47 %
Rd	1.63 %	
lk	1.47 %	quelque (27) = 1.47 %
ks	1.36 %	
Rn	1.25 %	
Rl	1.20 %	parle ...(21) = 1.14 %
Rk	1.09 %	marque (12) = 0.65 %

3. UNE CLASSIFICATION DES CONSONNES:

Nous avons établi, pour chaque corpus, et pour les trois positions initiale intervocalique et finale, des matrices de distribution des consonnes pour chaque combinaison de deux consonnes dans les GC.

Nous donnons ici (figure 1) une matrice simplifiée ne contenant que des indications sur l'existence des GC de 2 consonnes en position initiale dans le lexique tiré de DELA. Nous avons choisi comme illustration les GC en position initiale, puisque ceux-ci ont la particularité d'appartenir obligatoirement à la même syllabe (la syllabe initiale). Nous nous sommes de plus restreints aux groupes de exactement deux consonnes afin de mettre en évidence dans une seule matrice la combinaison entre la première consonne du GC (la plus éloignée du noyau vocalique) et la dernière consonne du GC (la plus proche).

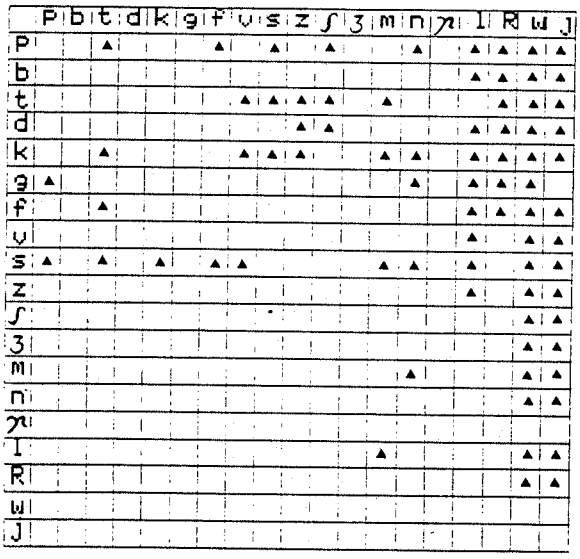


Figure 1 : matrice de distribution des GC initiaux de deux consonnes

A partir de cette matrice, il est possible d'établir une classification des consonnes (figure 2). Si GC = $C_1.C_2$, alors l'ensemble des consonnes C_2 (appelées classifieurs droits) est ordonné en comptabilisant pour chaque C_2 l'ensemble des consonnes C_1 qui peuvent distribuer C_2 à droite. Ainsi un niveau i du graphe sera supérieur à un niveau j quand C_2 pour i est plus distribuée que C_2 pour j. Le critère de dichotomie pour un niveau i du graphe est le suivant : les consonnes de type C_1 pour i qui admettent à droite la consonne C_2 pour i reçoivent le trait " + ", sinon elles reçoivent le trait " - ". Afin de séparer toutes les consonnes, le graphe présenté en figure 2 a été complété par le premier niveau du graphe des classifieurs gauches (même raisonnement que les classifieurs droits en considérant cette fois les consonnes de type C_1).

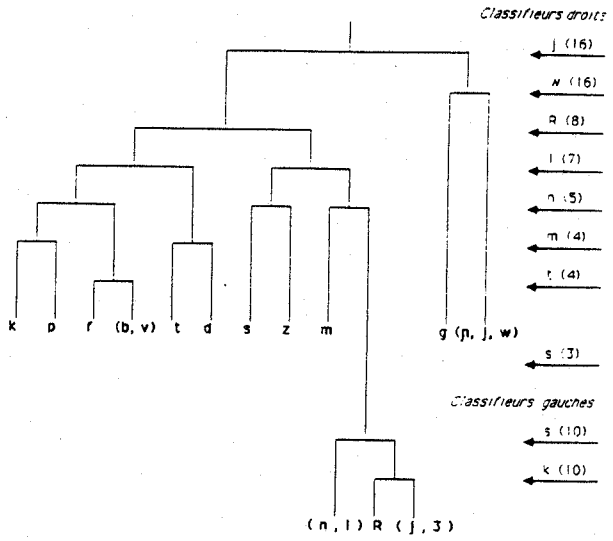


Figure 2 : Classification des consonnes.

Le graphe obtenu est tout à fait cohérent avec le système consonantique du français (BOE & TUBACH), exception faite pour le [g] qui n'accepte pas comme contexte droit le plus fort classificateur [j]. En chaînant les classifiés (figure 3), on peut s'apercevoir que le trait **[+/- voisé]** est transparent ([p] très proche de [b], [f] de [v],... ce qui met en évidence le fait que les consonnes sont séparées seulement à un niveau très bas dans l'arbre de classification).

Bien que peu classifiant, nous avons retenu [t], essentiellement comme intermédiaire jusqu'au classifieur [s] qui nous permet ensuite de distribuer l'ensemble [R, l, n, ʃ,]. Les classifieurs droits sont, dans l'ordre :

[j] , [w] , [R] , [l] , [n] , [m.]

Toutes ces consonnes possèdent le trait **[+ vocalique]** (BOE&TUBACH, p138). Il manque à l'appel [ɲ] qui ne figure jamais dans un GC (phénomène sans doute explicable par l'altération réciproque entre [nj] et [ɲ]), ainsi que [ɥ] qui n'a pas été produit par les règles de phonétisation. En tête de liste [j] et [w] ont effectivement ici un statut de consonnes puisque ne pouvant normalement pas constituer de noyau vocalique , mais ne peuvent pas apparaître en initiale d'un GC (sauf s'ils en sont le seul élément). Il n'est donc pas surprenant que ces deux consonnes soient associables à la plupart des autres consonnes, établissant ainsi une transition entre la consonne précédente et la voyelle suivante (phénomène de synérèse, comme dans " lion : [ljɔ̃]").

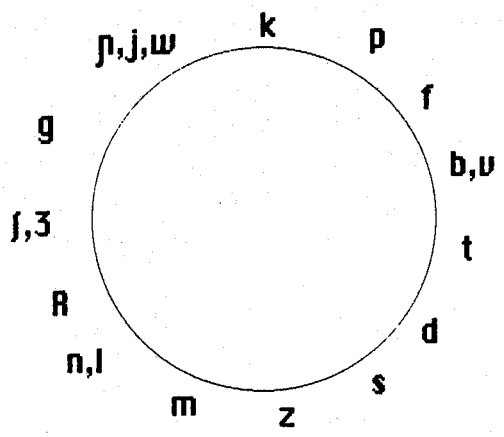


Figure 3 : Chaînage des consonnes classifiées.

4. PROPOSITION SUR LA SYLLABATION.

Adoptons la décomposition très classique de la syllabe :

attaque . noyau vocalique . coda

dans laquelle l'attaque et la coda peuvent être réduites à la chaîne vide.

Le GC initial d'un mot constitue donc l'attaque de la syllabe initiale, tandis que le GC final est la coda de la dernière syllabe du mot. Tout le problème de la syllabation réside ainsi dans la "segmentation" des GC intervocaliques. Nous ne reviendrons pas ici sur la difficulté à définir la syllabe et donc sur l'hésitation à reconnaître l'existence d'une coupe syllabique. Notre proposition s'inscrit spécifiquement dans le cadre de la synthèse, où la syllabe joue le rôle fondamental d'horloge rythmique, même si là encore l'analyse du discours oral peut porter à discussion quant à la propriété du français d'être "syllable-timed" (WENK & WIOLAND, 1982).

Le problème se limite donc aux GC intervocaliques. Puisque nous disposons d'une liste assez complète des GC, le type de solution vers lequel nous nous orientons consiste à trouver un compromis entre la syllabation par référence à une coupe syllabique inscrite dans le GC correspondant du lexique, et la syllabation par règles. La première solution a une exhaustivité équivalente à celle du lexique des GC intervocaliques. De plus, dans le cas où plusieurs solutions co-existent, il faut établir un choix a priori. Le problème est résolu lorsque ces solutions sont licites en même temps : il suffit d'en conserver une (par exemple : kstR -> k/stR ou ks/tR). En revanche, il est nécessaire d'ajouter des sous-lexiques d'exceptions dans les autres cas (exemple : mn -> m/n dans [am/nezik], et -> mn dans [a/mne]). Il nous faut donc définir une syllabation par règles, capable de déterminer une frontière possible pour tout GC intervocalique. Comme les règles phonotactiques habituellement énoncées (par exemple voir " Le dictionnaire de la prononciation française " par WARNANT, p. XCII-XCIII) ne permettent pas de résoudre tous les problèmes de syllabation pour notre lexique de GC intervocaliques (en particulier pour les groupes de trois consonnes et plus) nous avons appliqué une analyse distributionnaliste classique de la syllabe par rapport au mot :

$$GC_{intervocalique} = GC_1 \quad \text{ou}$$
$$GC_{intervocalique} = GC_1 . GC_2$$

tels que GC_1 et GC_2 appartiennent à l'ensemble des GC initiaux.
Par exemple :

- tR ne subit pas de coupe puisque existant en position initiale de mot.

- dst -> d/st car d et st existent en initiale alors que dst, ds et t n'y figurent pas

Comme nous avons d'ailleurs pu le constater précédemment, le problème reste ouvert lorsque plusieurs segmentations sont possibles selon ce critère. Nous pouvons en donner les exemples suivants :

- mn -> m/n ou -> mn où m,n,mn $\in$ [$GC_{initiaux}$]

Si l'on pondère ce critère par la fréquence lexicale, que nous interprétons ici comme le **rendement** de la construction GC, on s'aperçoit que mn a une fréquence lexicale de 0.03 %, donc très faible, parmi les GC initiaux. Nous choisirons donc ici de couper mn -> m/n.

- dR -> d/R ou dR où d,R,dR $\in$ [$GC_{initiaux}$]

Par contre, dans ce cas, $dR_{initial}$ a une fréquence lexicale de 2.44 %, c'est-à-dire un rendement déjà élevé. Nous choisirons donc de ne pas le couper.

- ktR -> k/tR ou -> kt/R
  où k,kt,tR,R $\in$ [$GC_{initiaux}$]

$tR_{initial}$ a un rendement de 12.23 %, $kt_{initial}$ a un rendement de 0.03 %. La coupe retenue est donc ktR -> k/tR.

Dans l'application de cette règle de distribution, [w] et [j] sont transparents, puisque ne pouvant pas apparaître à l'initiale d'un GC initial. Ainsi, par rapport aux exemples que nous avons déjà donnés, la syllabation de dRw et ktRw sera identique à dR et ktR, c'est à dire :

- dRw -> dRw

- ktRw -> k/tRw

6. CONCLUSION.

La motivation première qui nous a amenés à

constituer de tels lexiques de GC trouve son origine dans un projet d'évaluation de l'intelligibilité de systèmes de synthèse à partir du texte. Une telle procédure nécessite à la fois une liste pseudo-exhaustive des GC issue d'un dictionnaire orthographique représentatif le DELA, mais aussi des critères sur l'usage de ces GC extraits de données phonétiques : BDPHO, qui permettent de pondérer les résultats des tests d'évaluation.

Les données ainsi réunies nous ont permis de mener une analyse afin d'extraire des indices sur la structure des GC, pour lesquels, comme nous l'avons souligné, peu d'études avaient été réalisées.

Nous avons proposé une syllabation automatique : cette procédure ne prétend pas offrir une solution générale, notre approche, précisons-le bien, se limitant à la synthèse à partir du texte.

Remerciements.

Nous remercions tout spécialement M. GROSS, qui a mis à notre disposition le dictionnaire DELA, C. ABRY, qui nous a proposé la méthode de classification des consonnes, et M. CONTINI, pour son appui dans les analyses phonétiques.

REFERENCES

AUBERGE V. CONTINI M. MARET D. & SCHNABEL B. (1987)
TOPH : un outil de phonétisation multilangue.
Bulletin de l'Institut de Phonétique de Grenoble 16, 155-176.

BOE L.J. & TUBACH J.P. (1986)
Des matrices phonétiques aux matrices phonologiques et vice versa.
Bulletin de l'Institut de Phonétique de Grenoble, 15,135-155.

CARLSON R. ELENIUS K. GRANSTROM G.& HUNNICUTT S. (1985)
Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages.
French Swedish seminar on Speech 2, 511-547.

CHAVASSE P. (1948)
Essai sur la phonétisation statistique de la langue française et son application à l'étude de l'intelligibilité d'une conversation.
Cahiers d'acoustique, 1, 5-23.

DELATTRE P. (1966)
Studies in French and Comparative Phonetics.
Ed. Mouton & Co, London-The Hague-Paris, 150-167.

HATON J.P. (1984)
Accès lexical et reconnnaisssance de grands vocabulaires.
13èmes JEP du GALF, 89-96.

HATON J.P. & LAMOTTE M. (1971)
Etude statistique des phonèmes et diphonèmes dans le français parlé.
Revue d'Acoustique 16, 258-262.

JUBAN P. (1974)
Etudes des types et fréquences syllabiques dans le dictionnaire du français fondamental.
Note Techn. CNET-Lannion, CEI/CSE 58, 5 p.

LECLERE C. (1984)
Dictionnaire orthographique des groupes de consonnes du français.
Rapport Technique LADL. 311 p.

MALECOT A. (1974)
Frequency of Occurrence of French Phonemes and Consonant Clusters.
Phonetica, 29, 158-170.

PERENNOU G. & DE CALMES M. (1987)
BDLEX Base de données lexicales du français écrit et parlé. Volume 1, Lexique général.
Travaux du Laboratoire CERFIA.

ROCHETTE E. (1973)
Les groupes de consonnes en français. Etude de l'enchaînement articulatoire à l'aide de la radio-cinématographie et de l'oscillographie.
Klincksieck Paris, Presses de l'Université, 706 p.

ROSSI M. (1968)
Au sujet des groupes consonantiques du français.
Revue d'acoustique, 3-4, 306-311.

SOCK R. OLLILA L. DELATTRE C. & ZILLIOX C. (1987)
Intersegmental (VC & CC) and Intrasegmental (VOT & VTT) phasings in French.
11th Int. Cong. Phon. Sci., Tallin, Se 43.1.1-4.

TUBACH J.P. (1969)
Etude des contraintes statistiques des groupements phonématiques
Colloque de l'Informatique au Service de l'Homme, 31-48.

TUBACH J.P. & BOE L.J. (1985a)
De A à Zut. Petit dictionnaire phonétique du français. Vol. 1 - Classement alphabétique. Vol. 2 - Classement par fréquence.
Travaux et Recherches de l'Institut de Phonétique de Grenoble.

TUBACH J.P. & BOE L.J. (1985b)
Un corpus de transcriptions phonétiques : constitution et exploitation statistique.
Rapport ENST 85D0001., 56 p.

VAN BEZOOIJEN R. (1988)
Evaluation of two Synthesis Systems for Dutch : Intelligibility and Overall Quality of Initial and Final Clusters.
Publication de l'Institut de Phonétique d'Amsterdam, 88 p.

VAN EIBERGEN-THIEULLE J. (1986)
Réalisation et rôle du E bifide. Applications pédagogiques en français langue étrangère.
Thèse de 3ème Cycle, Phonétique. Grenoble III. 337 p.

VAN SON N. & POLS L.C.W. (1987)
Evaluation of a French Diphone-Based Synthesis System.
Rapport Techn. TNO, Amsterdam, IZF C-33.

UN MODELE DE LANGAGE UNIFIE DANS UN SYSTEME DE DIALOGUE ORAL PILOTE / AVION

S. Bornerand, F. Néel, G. Sabah

LIMSI/CNRS - BP30 91406 ORSAY CEDEX - FRANCE

ABSTRACT

This paper describes a language model formalism in a dialogue system, which is oriented towards the treatment of command languages. The language model is used by both the analysis stage and the dialogue unit. The interpretation of network transitions guides the analysis, and the exploitation of specific nodes directs the dialogue. The recognition level based on a word spotting algorithm provides a lexical lattice. An evaluation on a pilot/aircraft application allows to compare our system to connected word recognition systems.

I - INTRODUCTION

Ce document présente une mise en oeuvre de techniques d'intelligence artificielle dans un système de dialogue oral. Une collaboration avec la société CROUZET* a permis de définir une application pilote/avion utilisant un langage de commande muni de paramètres numériques. Le manque de fiabilité des systèmes de reconnaissance de formes acoustiques amène à introduire dans la phase d'analyse des activités de compréhension en concordance avec les activités de reconnaissance qui délivrent un treillis de mots affectés d'une note (fig.1) au lieu d'une séquence de mots. Comme dans les réalisations de [HAYES 86] et [GIACHIN 87], l'étage de compréhension utilisent la syntaxe et la sémantique comme contraintes et comme modèle de construction du sens de l'énoncé réduisant ainsi le champ des investigations qui est très vaste à l'oral à cause de l'ambiguïté de la langue et de l'incertitude des données.

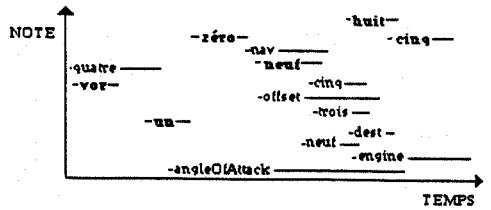


Figure 1: Treillis lexical de la commande *"vor 1 0 9 8 5"*

Un système de commande vocale opérationnel ne peut se limiter à une simple compréhension. Comme le montre [MORIN 87], il s'intègre dans un environnement de communication homme/machine et doit posséder, à ce titre, la capacité de gérer et structurer un dialogue. Dans le cadre d'applications avioniques [MATROUF 87], la tâche requiert un dialogue minimal de reprise sur erreurs pour pallier le manque de fiabilité du processus de reconnaissance.

* Cette étude a été effectuée grâce à un contrat DRET n° 86/307 (CROUZET-VECSYS).

II - EXIGENCES EN COMMANDE VOCALE

Le système de reconnaissance recherche à partir d'un treillis lexical la meilleure séquence de mots recouvrant tout l'énoncé et compatible avec le modèle de langage et le domaine. Pour ce faire, il doit posséder les caractéristiques suivantes:

. un critère de sélection dépendant des notes;

. un rattrapage des limitations inhérentes à l'étage de reconnaissance qui tient compte des recouvrements et des écarts permis entre deux hypothèses de mots consécutifs;

. une base de connaissances unifiée qui permet d'exploiter simultanément les contraintes linguistiques et pragmatiques;

. une stratégie de contrôle distribuée pour étudier les contraintes dans leur contexte d'utilisation uniquement;

. un traitement en parallèle de toutes les solutions. Plusieurs séquences de mots-hypothèses peuvent donner une interprétation complète et correcte sans être l'énoncé prononcé. Il faut attendre la fin de la phrase pour décider de la meilleure séquence;

. une analyse gauche-droite nécessaire dans un système temps réel;

. une répartition des tâches compatible avec une architecture parallèle comportant un étage qui fournit un treillis lexical, un étage qui combine les contraintes syntaxiques et sémantiques et qui construit les séquences de mots, et un étage qui les évalue et qui renvoie un message.

La tâche de dialogue de correction d'erreurs essaye de nuancer le comportement du système pour éviter le tout-ou-rien [BEROULE 84]. Le système peut percevoir un message suivant trois catégories: inintelligible, incomplet ou entier. Il faut aussi que le système de dialogue tienne compte de son rôle d'interface. La convivialité se traduit par la minimisation de l'effort d'expression du locuteur. Le système doit posséder plusieurs degrés de sensiblité afin d'être réceptif soit à une commande entière, soit à plusieurs éléments d'une commande ou bien à un seul élément d'une commande. Mais, cette dernière aptitude doit être accompagnée d'une souplesse d'utilisation qui laisse à l'utilisateur le choix du degré de précision qu'il veut mettre dans ses commandes.

III - ARCHITECTURE DE ADAS

Le synoptique de la fig.2 décrit l'interaction des bases de connaissances et des modules de traitement dans le système ADAS.

Le module de reconnaissance construit un treillis lexical à partir du signal acoustique. L'analyse traite chaque mot du treillis comme un ensemble de contraintes définies dans le modèle de langage que le séquenceur a la charge de vérifier. Le séquenceur construit dynamiquement le sens des phrases-hypothèses en mémorisant temporairement les structures des séquences de mots induites dans l'Environnement qui constitue l'espace de travail du système. A l'aide d'un ensemble de scénarios, le module de dialogue choisit un mode de réponse sous forme écrite en fonction de l'historique du dialogue et de la meilleure séquence de mots reconstituée dans l'Environnement.

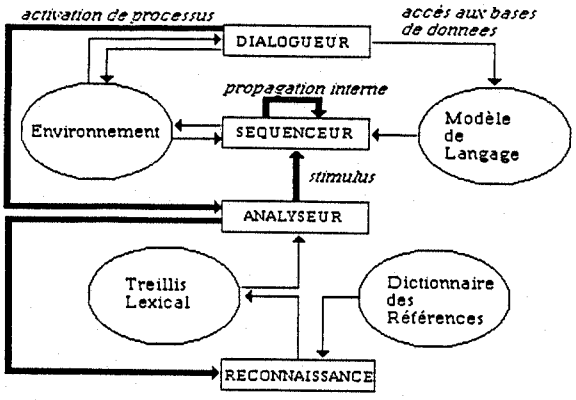


Figure 2: Architecture du système ADAS

IV - FORMALISATION DU MODELE DE LANGAGE

IV.1 Modèle de langage unifié

Le modèle de langage est capable d'enregistrer des connaissances acoustiques, linguistiques et pragmatiques lors de la phase de définition du langage de l'application, de les retrouver et de les exploiter lors de la phase d'analyse. Cette aptitude à accepter toutes sortes d'informations relève de l'utilisation d'un formalisme unique: un réseau d'états reliés par des transitions. Les états représentent les constituants syntaxico-sémantiques (unité lexicale, catégorie grammaticale, concept) et les éléments de contrôle (prédiction, vérification, attribut, question, interprétation,...). Les transitions entre deux états sont munies de conditions et d'actions qui traduisent les contraintes sous forme de règles de combinaison des données. Dans un système expert, une règle produit de nouveaux faits qui sont ajoutés à la base de faits ou complète des faits structurés dans la base de faits. De façon similaire, une transition produit des instances d'état qui sont créées dans l'Environnement, ou complète des instances d'états existant déjà dans l'Environnement. Une instance d'état dénote un emplacement mémoire structuré qui est destiné au stockage des données relatives à l'état. L'Environnement composé d'instances conserve en parallèle toutes les solutions de connexions possibles entre les instances reliées les unes aux autres par des pointeurs porteurs de sens.

Un dictionnaire d'états permet d'associer à chaque état un sous-réseau, et un seul, qui indique l'ensemble des contraintes associés à cet état.

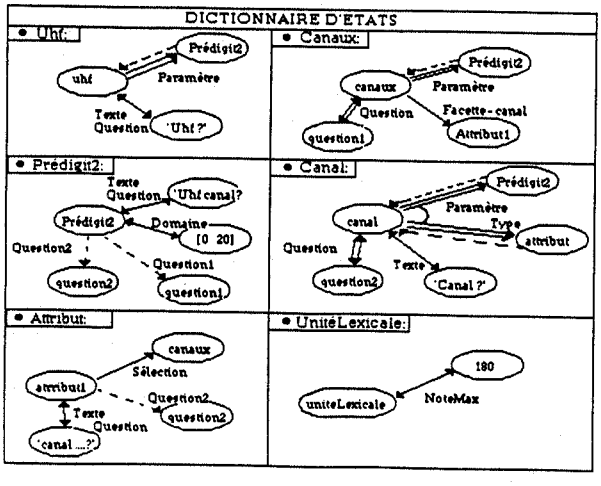


Figure 3: Quelques éléments illustrant une partie du réseau correspondant à une même commande du langage pilote/avion formulée indifféremment par: *"Uhf vingt", "Uhf deux zero", "Uhf canal vingt"*...

La déclaration d'un état en tant que sous-état définit une hiérarchie qui permet le partage de contraintes communes à plusieurs états grâce à un mécanisme d'héritage des transitions.

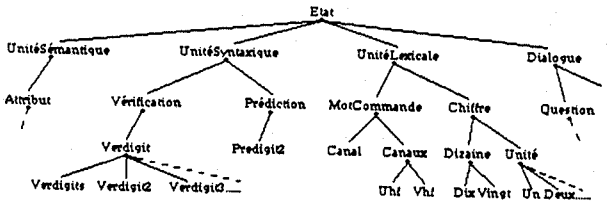


Figure 4: Un exemple de hiérarchie d'états

IV.2 Structure des données

Un état possède la liste des transitions qui l'admettent comme origine. Les conditions et actions d'une transition se définissent par des expressions procédurales, appelées respectivement "Condition" et "Action". Une liste de procédures est donc stockée dans chaque état.

Le réseau comporte des transitions de propagation qui sont celles citées tout au long de l'article, des transitions d'attachement de propriétés, qui permettent d'affecter une propriété à un état, et des transitions de déclenchement de question, qui pointent sur des entités de gestion du dialogue. Les transitions possèdent des champs qui contiennent l'adresse des procédures correspondant aux conditions et aux actions, et une liste de transitions à inhiber qui interdit l'utilisation de certaines contraintes et définit le contexte d'utilisation des contraintes exprimées par cette transition.

Une instance d'état possède un répertoire des connexions qui est accessible en écriture et lecture par le séquenceur. Des registres internes sont utilisés par le séquenceur pour décrire l'état correspondant à l'instance, le degré d'activité de l'instance, les transitions vérifiées, les transitions inhibées et d'autres informations.

Une instance est dite passive tant que sa structure de données est incomplète. Le succés de l'évaluation d'une fonction booléenne "Réaction", appliquée sur ses registres, rend l'instance active et correspond à une structure de données complète. Une transition entre deux états propage les données si et seulement si l'évaluation de sa fonction booléenne "Condition", à partir du contenu des registres de l'instance active, donne lieu à un succès. Sa fonction "Action" est aussitôt appliquée sur les registres de l'autre instance. Les registres d'une instance seront ainsi affectés par des actions de transitions, notées "propagation interne" sur la fig.2, ou par des opérations externes, notées "stimulus".

IV.3 Structure de contrôle

Dans les réseaux de transitions augmentées (ATN) présentés dans l'ouvrage de T. Winograd [WINOGRAD 83], l'auteur distingue la description du langage sous forme de graphe et l'exploitation de ces notations. L'exploitation se fait par un interpréteur qui scrute le réseau pour guider l'analyse d'une phrase selon une stratégie globale. Dans le cas présent, le séquenceur interprète localement le modèle de langage en fonction de l'état courant et du type de transitions explorées appartenant à un sous-réseau. Chaque mot du vocabulaire est associé à un état et l'analyse d'un mot consiste à passer au séquenceur le sous-réseau du dictionnaire correspondant à son état. La production et la propagation des données dans l'Environnement s'effectuent par l'exploitation des transitions du modèle de langage qui possèdent des indicateurs de contrôle. La stratégie dépend donc des transitions visitées.

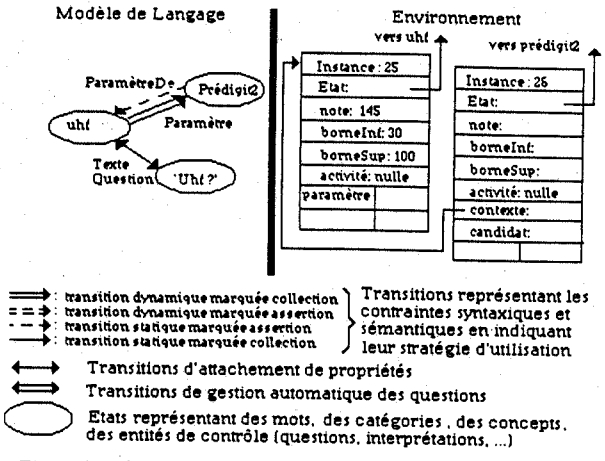


Figure 5: Création d'instances dans l'Environnement à partir du modèle de langage au cours de l'analyse du mot *"uhf"*

Le séquenceur prend en charge l'étude d'un état lors d'une création ou de la sélection d'une instance dans l'Environnement afin de vérifier les contraintes qui se rapportent à l'occurrence d'un mot ou à une réalisation d'un constituant du modèle de langage suivant l'algorithme:

```
si instance passive alors
[étude des transitions marquées collection;
 si l'évaluation Réaction est un succès alors
  instance devient active;
]
si instance active alors
 étude des transitions marquées assertion;
```

Lors de l'étude d'une transition, la marque "assertion" correspond à une propagation des données en mode chaînage avant et la marque "collection" correspond à un collationnement des données en mode chaînage arrière. L'algorithme d'exploitation des transitions est le suivant:

```
si transition marquée collection alors
[si instance destination passive alors
  étude de l'instance destination;
 si instance destination active alors
  si Condition vérifiée sur les registres de l'instance destination alors
  [transition vérifiée;
   Action complète les registres de l'instance origine;
  ]
]
si transition marquée assertion alors
 si Condition vérifiée sur les registres de l'instance origine alors
 [transition vérifiée;
  Action complète les registres de l'instance destination;
  étude de l'instance destination;
 ]
```

De plus, le marqueur "statique" d'une transition restreint l'étude d'une transition à deux instances existant dans l'Environnement alors que le marqueur "dynamique" autorise la création d'une nouvelle instance.

V - RECONNAISSANCE MULTI-MODE

Le module de reconnaissance utilise de façon concomitante deux algorithmes qui construisent un treillis lexical (fig.1) défini par un ensemble de mots qui possèdent quatre composantes: une étiquette, une note, un début et une fin de segmentation. A partir d'un dictionnaire de références de mots de commande, un premier algorithme [GAUVAIN 82] détecte toutes les occurrences de ces références dans le signal, quelle que soit leur position dans l'énoncé. La technique de détection de mots permet de filtrer les écarts de langage, les hésitations et les auto-corrections qui sont fréquents en situation de dialogue opérationnelle. Pour améliorer le taux de reconnaissance et s'adapter au type du langage de commande qui est constitué de plusieurs mots suivis de paramètres, les nombres (suite de chiffres) peuvent être détectés globalement à l'aide d'un algorithme de reconnaissance de mots enchaînés classique; ce qui réduit considérablement la combinatoire. A partir d'un dictionnaire de références de chiffres, ce deuxième algorithme donne la suite de références ressemblant le plus à l'énoncé.

De plus, les deux algorithmes acceptent la définition d'un seuil de détection relatif à chaque entité lexicale et l'utilisation de plusieurs références par mots.

VI - ANALYSE DU TREILLIS AVEC FILTRAGE

En début d'analyse, l'Environnement est considéré comme vide. L'analyseur extrait du treillis les mots un par un dans un ordre chronologique respectant leur position dans l'observation acoustique. La phase de filtrage permet d'éliminer des mots n'appartenant pas au vocabulaire courant et ne respectant pas les contraintes sur les paramètres caractérisant le stimulus (seuils sur la note et sur la longueur de l'occurrence d'un mot). A chaque mot, l'analyseur sélectionne l'état et le sous-réseau associé dans le modèle de langage, crée l'instance de l'état dans l'Environnement et déclenche l'étude de l'état via le séquenceur.

En fin d'analyse, l'Environnement contient l'ensemble des énoncés partiels et complets présents dans le treillis.

- vor 1 0 9 8 5
- vor 1 0 9 5 5
- vor 1 0 9 3 5
- nav
- Offset
- engine

Figure 6: Ensemble de séquences de mots obtenu à partir du treillis présenté en fig.1

Les énoncés complets sont singularisés par une affiliation à un état-interprétation et les énoncés incomplets le sont par une affiliation à un état-question.

VII - GESTION DES ECHANGES HOMME/MACHINE

A chaque nouvel échange, le dialogueur appelle l'analyseur qui traite un message à la fois. Grâce aux points d'entrée définis par les états-interprétations et les états-questions, il a directement accès aux énoncés (interprétation, question) contenus dans l'Environnement. Un critère permet de classer les énoncés dans un ordre préférenciel en fonction de la note A qui correspond à la distances cumulées en programmation dynamique entre la séquence des références acoustiques et la forme d'entrée. La note A est donc calculée à partir de la note et de la longueur des mots appartenant à l'énoncé, et des espaces intermots pondérés par la valeur du seuil de détection.

$$A = \sum_i (N_i * L_i) + C * \sum_i E_i$$

avec N_i : note du mot i
L_i : nombre de trames du mot i
E_i : nombre de trames séparant la fin du mot i du début du mot i+1
C : valeur du seuil de détection

En fonction de la note, de l'émergence et du type du meilleur énoncé, le dialogueur choisit d'accepter le message, de le refuser, de demander une confirmation ou de demander une précision selon un scénario procédural déterminé.

```
H> uhf canal seize                    H> la fréquence uhf sur le canal quinze
M> uhf canal 13. confirmer ?          M> uhf canal 15. confirmer ?
H> seize                              H> confirmation
M> uhf canal 16                       M> uhf canal 15
⋮
H> uhf vingt
M> uhf ?
H> uhf deux zéro
M> uhf 20
⋮
```

Figure 7: Exemples de dialogue Homme/Machine

Avant chaque analyse, le dialogueur initialise l'Environnement et les paramètres du modèle de langage en fonction de l'historique du dialogue pour introduire les restrictions de vocabulaire et les relâchements de contraintes topographiques. Un historique de dialogue est utile au niveau des enchaînements question/réponse mais non justifié au niveau de la succession des commandes qui sont indépendantes entre elles.

Dans le cadre d'une demande supplémentaire, le dialogueur garde en mémoire dans l'Environnement le constituant incomplet qui a engendré cette question. Grâce à quoi, l'utilisateur peut répéter toute la commande, seulement donner l'élément absent ou erroné, ou utiliser des mots-clés d'annulation ou de confirmation.

VIII - EVALUATION

Le système ADAS fonctionne sur un IBM AT. Les algorithmes de reconnaissance ont été développés en langage C et le système de dialogue a été écrit en SMALLTALK. Une série de tests a permis d'évaluer la différence de comportement entre un système de reconnaissance par mots enchaînés traditionnel [GAUVAIN 83] et le système ADAS face à deux langages d'application de complexité graduée. Un langage dit "strict" constituait le corpus 1 qui comportait 70 chaînes de 1 à 6 mots pris dans un vocabulaire d'une quarantaine de mots intégrant les chiffres de 0 à 9. Un langage dit "souple" qui accepte les silences, les hésitations et les formulations variables, tant pour les commandes entières que pour l'expression des paramètres, constituait le corpus 2 qui contenait 100 chaînes de 1 à 6 mots pris dans un vocabulaire d'une soixantaine de mots acceptant les nombres de 0 à 9999.

Dans le cas d'un langage strict, les systèmes classiques tirent avantage de leur algorithme de prédiction pour contraindre fortement la reconnaissance et affichent ainsi des résultats meilleurs de 10% par rapport au système ADAS. Par contre, les langages souples paraissent plus favorables au système ADAS qui maintient son taux de reconnaissance, alors que les performances du système classique chutent sensiblement; ce qui donne un léger avantage au système ADAS.

Cependant, la capacité de dialogue a totalement été ignorée et aucune reprise sur erreurs n'a été exploitée au cours de ces essais pour améliorer le taux de reconnaissance du système ADAS.

IX - CONCLUSION

L'analyse de certaines limites qui sont essentiellement dues à des simplifications d'implantation peut être riche en information:

. l'étage de compréhension est indépendant du système de reconnaissance. Dans les cas où le treillis ne contient pas les bonnes hypothèses, le système n'offre pas la possibilité de les récupérer directement à partir du signal acoustique;

. le recours à un Environnement unique pour mémoriser l'état des recherches et pour partager les données interdit l'utilisation de prédictions récursives qui nécessitent une structure de stockage arborescente;

. la stratégie de contrôle distribuée, qui utilise le type des transitions pour guider l'analyse, impose un traitement unique des contraintes quel que soit le contexte de réalisation;

. attendre qu'une instance soit complète pour qu'elle puisse déclencher les transitions de type "assertions" (chainage-avant) interdit de faire la distinction entre une hypothèse incohérente et pas d'hypothèse du tout;

Cependant, le formalisme de représentation et de manipulation des connaissances, commun aux étages de compréhension et de dialogue, confère au système:

- une puissance d'expression qui ramène au même niveau de description toutes les connaissances. Toutes sortes de relations peuvent être représentées et traitées sans être uniquement séquentielles ou hiérarchiques;
- une uniformité des types de connaissances car il accepte aussi bien les descriptions procédurales que les représentations déclaratives;
- un pouvoir d'extension et de modification du langage qui lui permet d'accepter une gamme d'applications;
- une compatibilité totale avec une architecture parallèle sur support matériel;
- un aspect évolutif car le modèle de langage est modifiable dynamiquement en cours de dialogue.

Enfin, l'évaluation a bien fait ressortir que la réaction d'un système dépend essentiellement de la complexité du langage d'application qui lui est soumis. Les systèmes de reconnaissance de mots enchaînés classiques rendus optimaux par l'utilisation de technologies de pointe [QUENOT 86] évoluent vers une plus grande tolérance aux hésitations, aux reprises et autres bruits. De ce fait, ces systèmes deviennent performants pour les applications fermées utilisant des langages de commande de quelques centaines de mots. En revanche, lorsque l'application nécessite des dialogues complexes, un vocabulaire d'un millier de mots et une évolution du langage au cours des échanges, notre approche semble être en mesure d'apporter une nouvelle solution aux problèmes de développement des futurs systèmes capables d'intégrer et de gérer toutes ces composantes.

BIBLIOGRAPHIE

[BEROULE 84] **D. Béroule, F.Néel,**
"Une approche des problèmes liés à la communication parlée homme-machine", RFIA Afcet-Inria, 1984, pages 53-63.

[COZANNET 87] **A.Cozannet,**
"ALOEMDA, un analyseur linguistique pour l'oral et l'écrit", RFIA Afcet-Inria, 1987, pages 381-389.

[GAUVAIN 82] **J.L. Gauvain,**
"Reconnaissance de mots enchaînés et détection de mots dans la parole continue", Thèse de 3 cycle, Paris XI, 1982.

[GAUVAIN 83] **J.L. Gauvain, J.J Gangolf,**
"Terminal Integrates Speech Recognition and Text-to-Speech Synthesis", Speech Technology, Sept/Oct, 1983, pages 25-38.

[GIACHIN 87] **E. Giachin, C. Rullent,**
"A Control Strategy for a Knowledge-Based Approach to Signal Untersdanting", ESPRIT'87, Tome 1, pages 836-849.

[HAYES 86] **P.J. Hayes, A.G. Hauptmann, J.G. Carbonell and M. Tomita,**
"Parsing Spoken Language: a Semantic Caseframe Approach", COLING86, pages 587-592.

[MATROUF 87] **A.K. Matrouf, F. Néel et J.Mariani,**
"Système de dialogue orienté par la tâche: une application en avionique", JEP87, 1987, pages 187-190.

[MORIN 87] **P. Morin, J.M. Pierrel,**
"Partner: un système de dialogue oral homme-machine", Cognitiva87, 1987, pages 355-361.

[QUENOT 86] **G. Quenot, J.L. Gauvain, J.J. Gangolf and J. Mariani,**
"A dynamic time warp VLSI processor for continuous speech recognition", ICASSP'86, Tokio, 1986, pages 1549-1552.

[WINOGRAD 83] **T. Winograd,**
"Language as a Cognitive Process", Addison-Wesley, 1983.

ESSAI D'ANALYSE DU SYSTEME ACCENTUEL DU FRANCAIS : DISTRIBUTION DE L'ACCENT SECONDAIRE

Valérie Pasdeloup

Institut de Phonétique, UA CNRS 261 Parole et Langage,
Université de Provence, 13621 Aix-en-Provence Cedex

ABSTRACT

We can infer from our study of secondary accent distribution rules in French that the stress system of that language results from the combination of two stress mechanisms :

(1) a right to left mechanism (counting syllables from the end of a word or a group of words to the beginning)

(2) a left to right mechanism (counting syllables from the beginning of a word or a group of words to the end)

This paper presents stress rules, a three level prosodic structure and phonotactic principles; various acceptable prosodic structures can be determined from these elements for a given sentence (lexical basis).

INTRODUCTION

Il ne semble pas possible de définir le système accentuel du français en ne prenant en compte que la limite et la hiérarchie des constituants syntaxiques, et en négligeant les informations qui concernent le mot : ses limites, son nombre de syllabes, sa place dans l'énoncé et son poids sémantique.

Les modèles accentuels ou intonatifs, dont l'organisation accentuelle ou intonative découle directement ou indirectement de la structure syntaxique, et optionnellement de la structure énonciative, sont en général obligés d'introduire, à un niveau ou à un autre, des principes tels que les principes phonotactiques (Rossi 1985) ou les principes d'eurythmie (Dell 1984; Martin 1987), afin d'expliquer des phénomènes dont les organisations syntaxique et énonciative ne peuvent pas rendre compte.

Par ailleurs, certains travaux mettent en évidence l'existence dans la parole de phénomènes dont la fonction principale ne semble pas être linguistique : principes d'alternance rythmique fort/faible ou faible/fort (Liberman et Prince 1977; Verluyten 1983; Dell 1984), principes d'alternance de durée syllabique (Duez et Nishinuma 1985). Ces principes et plus généralement tout phénomène dans la parole plus ou moins périodique ou récurrent semblent avoir à l'origine une fonction psychophysiologique et cognitive (Fraisse 1956), avant d'être réinvestis par le code linguistique - intégrés dans une structure d'ordre linguistique, ou à l'inverse maintenus perceptivement à une valeur seuil -.

L'accent secondaire apparaît en français être un élément susceptible de remplir une fonction phonotactique et/ou linguistique.

A partir de l'analyse de la distribution de l'accent secondaire (étude acoustique et perceptive d'un corpus lu (Pasdeloup 1988)), nous proposons un modèle d'accentuation des phrases en français qui organise l'énoncé selon des principes linguistiques et phonotactiques.

1. ORGANISATION LINGUISTIQUE ET PHONOTACTIQUE DE L'ENONCE

Les productions langagières tendent à être à la fois économiques sur le plan de la production motrice (Fraisse 1956) et rentables sur le plan linguistique. L'organisation phonotactique d'un énoncé a pour fonction d'en réguler sa production, par le biais entre autres de la structuration accentuelle et intonative de ses unités linguistiques, de telle sorte que l'énoncé soit produit selon les normes imposées par les possibilités et les contraintes psychophysiologiques et cognitives.

1.1 Organisation linguistique

Elle correspond, d'une part, à la structuration syntactico-sémantique et énonciative (organisation thème/rhème, focalisation etc...) et au découpage de l'énoncé en mots lexicaux et grammaticaux et, d'autre part, à la structuration prosodique (accentuation et intonation).

La prosodie de l'énoncé s'organise hiérarchiquement en séquences rythmiques, en mots rythmiques et en pieds accentuels :

- séquence rythmique: unité intonative majeure qui organise un groupe syntactico-sémantique majeur et qui est délimitée par un contour mélodique marqué et un fort allongement (ce qui correspond approximativement à la continuation majeure de Delattre, à l'intonème continuatif majeur ou conclusif de Rossi, au tronçon intonatif de Dell et à l'unité intonative de Hirst et Di Cristo).

- mot rythmique : le mot rythmique coïncide rarement avec le mot lexical ou grammatical, mais il en comprend au moins un. Le mot rythmique est une unité intonative mineure qui organise un groupe syntactico-sémantique mineur et qui se caractérise par la présence à la finale d'un accent primaire (le mot rythmique correspond approximativement au mot phonologique de Milner et Regnault). Le regroupement des mots en mots rythmiques est soumis à des contraintes phonotactiques : le mot rythmique tend à être ni trop long (dans ce cas, constitution d'un autre mot rythmique), ni trop court (dans ce cas, intégration des éléments environnants dans le même mot) (Vaissière 1971); dans le cas d'un syntagme nominal Adj. + Nom, on aura ainsi

tendance à constituer un seul mot rythmique s'il s'agit du groupe (ce petit chien), mais deux s'il s'agit du groupe (ce gigantesque)(chimpanzé).

Exemples :
(Ce musicien) (Ce grand musicien)
(Ce musicien)(talentueux)

Un mot rythmique peut se comporter du point de vue accentuel et intonatif comme un mot.

- pied accentuel ou rythmique : suite de syllabes inaccentuées suivies d'un accent; le pied accentuel ne correspond souvent pas à des groupes de sens mais peut parfois coïncider avec un mot.

1.2 Organisation non linguistique

Elle concerne les principes phonotactiques. Ces principes jouent à différents niveaux d'organisation de l'énoncé et se combinent différemment suivant la stratégie discursive adoptée. Ils interviennent, entre autres, dans la distribution de l'accent secondaire et dans l'organisation générale de la structure rythmique.

L'accent secondaire se réalise dans un mot selon les critères phonotactiques suivants :

- Position du mot dans l'énoncé : l'accent secondaire apparaît plus facilement en début d'énoncé, surtout s'il s'agit du premier mot de la phrase, qu'en fin d'énoncé.

- Nombre de syllabes du mot : l'accent secondaire a tendance à se réaliser lorsque le mot (mono ou polymorphémique) comporte une suite de syllabes inaccentuées supérieure ou égale à 3.

- Contexte rythmique immédiat (inter-mot) : dans le mot rythmique, l'accent secondaire a tendance à se réaliser de telle sorte qu'une suite de syllabes inaccentuées supérieure ou égale à 4 soit évitée (sauf stratégies contraires : effet de parenthésage, débit très rapide etc...).

- Structure rythmique de l'ensemble de l'énoncé : suivant la stratégie discursive adoptée il est possible, entre autres, de sur-accentuer, de sous-accentuer (lors d'un débit très rapide par exemple), de n'utiliser qu'un seul type d'accent secondaire, ou d'avoir tendance à répéter les mêmes structures rythmiques (principe de récurrence).

De façon plus générale, des marques prosodiques ont tendance à être utilisées, à tous les niveaux de l'organisation linguistique, pour sur-segmenter les longs groupes syntactico-sémantiques et pour sous-segmenter les groupes syntactico-sémantiques courts.

Les diverses phénomènes d'équilibrage, qui se manifestent sous des modes variés (Wioland 1984) sont également des principes phonotactiques qui opèrent à différents niveaux de l'organisation linguistique :

- principe d'équilibrage par le nombre de syllabes (De Cornulier 1979) : des groupes prosodiques de différents tailles sont composées d'un nombre de syllabes proche ou identique.

- principes d'équilibrage temporel (Wenk et Wioland 1982; Martin 1987) : isochronie des groupes prosodiques (chronométrage accentuel ou "sress-timing") ou isochronie syllabique (chronométrage syllabique ou "syllable-timing").

2. DISTRIBUTION DE L'ACCENT SECONDAIRE

De nombreux travaux tendent à prouver l'existence en français d'un accent secondaire facultatif (appelé parfois accent rythmique) situé en dehors de la finale et distinct des accents rhétoriques et énonciatifs (focalisation). Selon les auteurs, l'accent est situé à l'initiale (Hirst et Di Cristo 1984; Rossi 1985; Milner et Regnault 1987), sur l'antépénultième (Mazaleyrat 1974; Verluyten 1983), ou même de façon probabilitaire (Fonagy 1980).

A partir de l'analyse acoustique et perceptive d'un corpus lu (Pasdeloup 1988), nous présentons une description exhaustive de la distribution de l'accent secondaire.

Nous appelons accents primaires les accents réalisés à la finale d'un mot ou d'un groupe de mots, et accents secondaires les accents réalisés à l'initiale d'un mot ou d'un groupe de mots, sur l'antépénultième d'un mot lexical et à la finale d'un morphème interne dans un mot polymorphémique. L'accent primaire se distingue de l'accent secondaire non seulement par sa position, il est le seul à être situé à la frontière droite d'un mot, mais aussi par son caractère intonogène (Rossi 1985) : il semble être le seul accent porteur d'une marque intonative. L'accent secondaire ne peut porter une marque intonative et se caractérise sur le plan acoustique par un contour mélodique toujours montant et, en général, moyennement marqué (les corrélats acoustiques de ces accents sont présentés dans Pasdeloup 1988).

Observation 1 : l'accent secondaire peut être situé sur l'antépénultième d'un mot portant un accent primaire à la finale; il annonce et renforce la perception de l'accent primaire.

Exemple : les formalités administratives
- - ' - ' - - ' - '

(- désigne une syllabe inaccentuée, ' une syllabe accentuée)

Observation 2 : l'accent secondaire peut être situé sur la première syllabe (comprenant un support consonnantique) d'un mot ne portant pas obligatoirement d'accent primaire à la finale. Le support consonnantique peut faire partie de la composition syllabique intrinsèque du mot ou être ajouté : coup de glotte ou /z/ de la liaison avec le mot précédent intégré à la syllabe accentuée.

Exemples : les formalités administratives
- ' - - ' - ' - '
les formalités_administratives
- ' - - - ' - - - '
un mauvais traitement
- ' - - '

Observation 3 : en début de phrase on accentue habituellement sur la première syllabe (qui comprend un support consonnantique) du premier mot accentuable de la phrase, même si celui-ci ne porte pas d'accent primaire à la finale.

Exemple : La librairie du boulevard
- ' - ' - - '

Observation 4 : dans un mot polymorphémique portant un accent primaire à la finale, l'accent secondaire peut être situé à la frontière des morphèmes (à la finale d'un morphème non terminal).

Exemple : la mélodramatisation
- - ' - - - - '

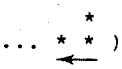
3. DESCRIPTION DU SYSTEME ACCENTUEL

Nous présentons deux procédés d'accentuation complémentaires qui coexistent dans le système accentuel du français, des règles d'accentuation et des contraintes imposées à la constitution des pieds. Cette description rend compte de la distribution de l'accent secondaire et de l'accent primaire. La présence de l'accent secondaire est conditionnée par les critères phonotactiques énoncés précédemment.

3.1 Deux procédés accentuels

- un procédé d'accentuation droite-gauche (procédé D) gouverne les règles d'accentuation selon un décompte syllabique de droite à gauche. Un accent primaire est placé sur la première syllabe à droite accentuable de chaque mot rythmique, et un accent secondaire est placé optionnellement sur l'antépénultième de chaque mot lexical de plus de trois syllabes; ces deux accents délimitent ainsi la fin, la droite, d'un mot rythmique.

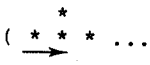
L'accentuation d'un mot rythmique selon le procédé D peut se représenter ainsi :



( ) désignent les limites du mot rythmique. Le nombre de * dans une colonne représente le degré d'accent, relativement à l'accentuation de la ligne où ce signe se trouve (Halle et Vergnaud 1987).

- un procédé d'accentuation gauche-droite (procédé G) gouverne les règles d'accentuation selon un décompte syllabique de gauche à droite. Un accent secondaire est placé optionnellement sur la première syllabe à gauche accentuable de chaque mot rythmique et délimite ainsi le début, la gauche, d'un mot rythmique.

L'accentuation d'un mot rythmique selon le procédé D peut se représenter ainsi :



La première syllabe à gauche désigne une syllabe inaccentuable (clitique par exemple).

Le procédé D met en oeuvre un principe d'alternance accentuelle droite-gauche (accentué-inaccentué) (Verluyten 1983), principe qui ne semble pas s'appliquer au delà de quatre syllabes (à compter de la droite).

Le procédé G est moins complexe que le procédé D puisqu'il ne semble pas mettre en oeuvre un principe d'alternance accentuelle gauche-droite. On remarquera cependant la symétrie de ces deux procédés qui démarquent, par l'accent, le début et la fin d'une unité linguistique.

3.2 Règles d'accentuation

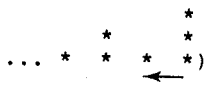
Toutes les règles, excepté la première (R.1), sont facultatives. L'organisation linguistique et phonotactique d'un énoncé, de même que la stratégie discursive employée déterminent leur application.

R.1 : Accentuer la fin d'un mot rythmique (procédé accentuel D).

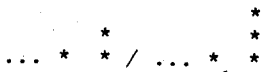
R.2 : Accentuer le début d'un mot rythmique sur la première syllabe accentuable, généralement la première syllabe du premier mot lexical (procédé accentuel G).

La première syllabe à gauche désigne une syllabe inaccentuable

R.3 : Accentuer l'antépénultième d'un mot lexical d'au moins 4 syllabes, à condition qu'il soit déjà accentué à la finale, c'est-à-dire que la dernière syllabe du mot lexical soit la dernière syllabe du mot rythmique (procédé D).



R.4 : Accentuer la syllabe finale d'un morphème non terminal dans un mot polymorphémique, à condition qu'il soit accentué à la finale (procédé D).



/ désigne une limite morphémique interne dans un mot polymorphémique.

Les règles d'accentuation déterminent le domaine des propriétés de l'accent. Dans les règles précédentes, l'accent délimite soit la fin, la droite, d'un mot rythmique, d'un mot lexical ou d'un morphème interne dans un mot polymorphémique (procédé D), soit le début, la gauche, d'un mot rythmique (procédé G). Ces règles accentuels sont indépendantes des contraintes imposées à la constitution des pieds accentuels (Halle et Vergnaud 1987).

3.3 Constitution des pieds accentuels

Le pied accentuel en français à la tête à droite : il est constitué d'une suite de syllabes inaccentuées suivies d'un accent; suivant les cas, le pied accentuel est binaire ou illimité.

Contrainte 1 : toute syllabe doit être dans un pied.

Contrainte 2 : chaque pied en français à la tête à droite.

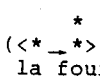
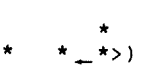
R.5 : Constituer des pieds binaires ou illimités avec tête à droite.

3.4 Exemples d'accentuation dans des mots lexicaux

L'accentuation d'un mot lexical selon ces règles d'accentuation et les contraintes imposées à la constitution des pieds peut se représenter ainsi :

- Dans un mot de 3 syllabes :

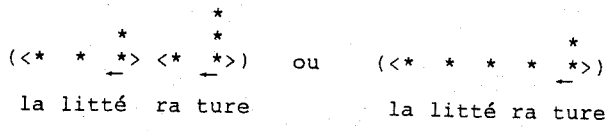
Procédés accentuels :

D + G	ou	D seul
		
la fourgonnette	ou	la fourgonnette

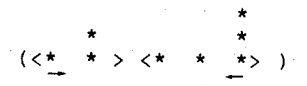
< > désignent les limites des pieds accentuels et ( ) les limites des mots rythmiques

- Dans un mot de 4 syllabes :

Procédé accentuel D :



Procédés accentuels D + G :



4. MODELE D'ORGANISATION RYTHMIQUE D'UN ENONCE

A partir des règles accentuels, de la structuration intonative des unités syntactico-sémantiques (séquences rythmiques, mots rythmiques) et des principes phonotactiques présentés ici, nous proposons un exemple des opérations à effectuer afin d'obtenir les différentes structures accentuelles et intonatives acceptables pour une même phrase (support verbal).

- Découpage de la phrase en séquences rythmiques : unités intonatives majeures.

Exemples :
1 / Le boulanger du coin /
/regardait ce gros artificier. /

2 / Le boulanger, / du coin, /
/ regardait ce gros artificier. /

/ désignent les limites des séquences rythmiques

- Découpage en mots rythmiques : unités intonatives mineures.

Exemples :
1a / (Le boulanger du coin) /
/ (regardait) (ce gros artificier.) /

1b / (Le boulanger) (du coin) /
/ (regardait) (ce gros artificier.) /

2a / (Le boulanger,) / (du coin,) /
/ (regardait) (ce gros artificier.) /

( ) désignent les limites des mots rythmiques

- Accentuation selon les règles accentuels :

- dans un premier temps, tous les accents primaires et secondaires possibles sont placés : en fonction des limites des mots rythmiques pour les accents à l'initiale et à la finale, en fonction des limites des mots lexicaux pour les accents sur l'antépénultième et en fonction des limites des morphèmes pour les accents situés à la finale des morphèmes non terminaux dans les mots polymorphémiques (cf. exemple 1).

- dans un second temps, les structures accentuelles possibles sont constituées en fonction (cf. exemple 2) :

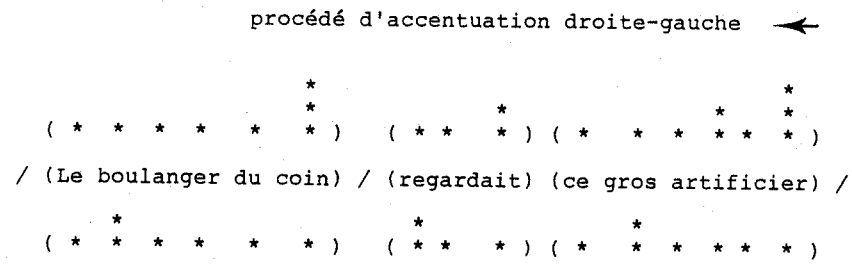
- des principes phonotactiques énoncés précédemment.

- de la règle du premier accent : en début d'énoncé on se dépêche d'accentuer; le premier accent possible, même s'il s'agit d'un accent facultatif, devient généralement obligatoire.

- de la règle de collision d'accent : deux syllabes accentuées ne peuvent pas être contiguës dans une même séquence rythmique; elles ne peuvent être contiguës que dans le cas où elles se situent de part et d'autre de la frontière d'une séquence rythmique.

EXEMPLE 1 : ACCENTUATION DE L'ENONCE 1A SELON LES REGLES ACCENTUELS

Tous les accents primaires et secondaires possibles sont plaçés;
/ désignent les limites des séquences rythmiques, ( ) celles des mots rythmiques

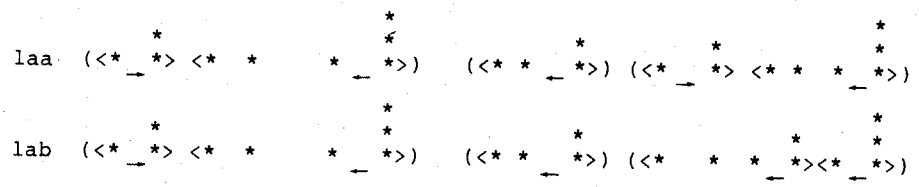


EXEMPLE 2 : STRUCTURES PROSODIQUES POSSIBLES

< > désignent les limites des pieds accentuels

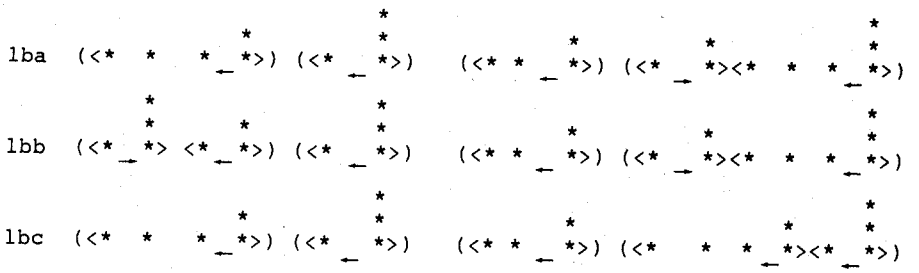
Enoncé 1a, 3 mots rythmiques :

/ (Le bou langer du coin) / (regardait) (ce gros arti ficier) /



Enoncé 1b, 4 mots rythmiques :

/ (Le bou langer) (du coin) / (regardait) (ce gros arti ficier) /



Certaines structures répondent mieux à certaines stratégies discursives que d'autres. Les structures prosodiques 1aa et 1bb favorisent la présence d'accents secondaires à l'initiale des mots rythmiques, tandis que 1bc évite les accents secondaires à l'initiale; la structure 1bc satisfait au principe de récurrence (répétition de la structure accentuelle (- - - ' - ')); la structure 1bb comporte le plus grand nombre d'accents, ce qui peut être acceptable à un débit assez lent, de lecture par exemple.

CONCLUSION

Les règles accentuelles présentées ici ont l'avantage de rendre compte de tous les cas d'accents secondaires recensés dans les travaux antérieurs. Bien que ces règles accentuelles soient facultatives (excepté celle qui détermine la présence de l'accent primaire à la finale du mot rythmique), leur application n'est pas probabilitaire mais est conditionnée par l'organisation linguistique de l'énoncé et par les contraintes phonotactiques, de même que par la stratégie discursive utilisée (débit et style employés, état émotif du locuteur etc ...).

La coexistence dans le système accentuel du français de deux procédés accentuels différents droite-gauche et gauche-droite peut s'interpréter diachroniquement si l'on considère que le français est à une période charnière de son évolution (Fonagy 1980). On peut alors interpréter l'existence d'un accent secondaire sur la première syllabe comme le résultat de la neutralisation, de la "banalisation" de la fonction remplie par l'accent d'insistance (accent rhétorique et accent énonciatif); comme dit Jakobson (1929) " (...) la forme caractéristique de la projection de la synchronie dans la diachronie, c'est la généralisation d'un style."

REMERCIEMENTS

Je tiens à remercier D. Hirst, chercheur au CNRS, pour son "aide phonologique" dans la réalisation de ce travail, ainsi que le Professeur M.Rossi pour ces précieux conseils.

BIBLIOGRAPHIE

DE CORNULIER, B. (1979), Problèmes de métrique française, Thèse de Doctorat, Aix-Marseille I.
DELATTRE, P. (1966), Les dix intonations de base en français, French Review, 40-1, 1-14.
DELL, F. (1984), L'accentuation dans les phrases en français, Forme sonore du langage , Hermann, Paris.
DUEZ, D.; NISHINUMA, Y. (1985), Le rythme en français : alternance des durées syllabiques, Travaux de l'Institut de Phonétique d'Aix, 10.
FONAGY, I. (1980), L'accent français : accent probabilitaire, L'accent en français contemporain, Studia Phonetica, 15, Didier.
FRAISSE, P. (1956), Les structures rythmiques, Studia Psychologica, Publications universitaires de Louvain.
HALLE, M.; VERGNAUD, J.R. (1987), An essay on stress, MIT Press.
HIRST, D. et DI CRISTO, A. (1984), French Intonation : A Parametric Approach, Die Neueren Sprachen, 83:5.
JAKOBSON, R. (1971), Selected writings I (2ème éd.), Mouton, The Hague.
LIBERMAN, M.; PRINCE, A. (1977), On Stress and Linguistic Rhythm, Linguistic Inquiry, 8 (2).
MARTIN, Ph. (1987), Structure rythmique de la phrase française, statut théorique et données expérimentales, XVI JEP, SFA, , Hammamet.
MAZALEYRAT, J. (1974), Eléments de métrique française, Colin, Paris.
MILNER, J.C.; REGNAULT, F. (1987), Dire le vers, Seuil, Paris.

PASDELOUP, V. (1988), Essai de modélisation du rythme du français, Revue d'acoustique, SFA (à paraître).
ROSSI, M. (1985), L'intonation et l'organisation de l'énoncé, Phonetica, 42.
VAISSIERE, J. (1971), Contribution à la synthèse par règles du français, Thèse de Doctorat, Université de Grenoble.
VERLUYTEN, S.P. (1983), Phonetic Reality of Linguistic Structures : the Case of (Secondary) Stress in French, Proceedings of the Tenth International Congress of Phonetic Sciences, Utrecht, M.P.R. Van den Broecke et A.Cohen eds.
WENK, B.J.; WIOLAND, F. (1982), Is French really syllable-timed ?, Journal of Phonetics, 10.
WIOLAND, F. (1984), Organisation temporelle des structures rythmiques du français parlé, Bul. des rencontres régionales de linguistique, Lausanne.

Snorri
Un système d'étude interactif de la parole

Y. Laprie

CRIN INRIA Lorraine
BP 239 54506 Vandœuvre les Nancy.
laprie@crin.crin.fr tel 83 91 20 00 poste 2880

Abstract

Snorri has been developed in CRIN; this is a speech analysis tool which helps the researcher to acquire acoustic-phonetic knowledge. It allows him to record sentences , to play back speech signal, to compute and to display spectrograms and to label speech utterances. Snorri differs from the other speech analysis packages by providing tools to investigate speech corpus and to display formant tracking in F1-F2 plane. Using multi-windowing and mouse Snorri is an easy-to-use system. All our researches at CRIN in acoustic-phonetic decoding uses Snorri facilities.

1 Introduction

Les progrès en décodage automatique de la parole dépendent étroitement des moyens d'analyse et d'observation mis à la dispostion des chercheurs. Le CRIN, après avoir acquis une nouvelle machine destinée au traitement de la parole (Masscomp 5600), a développé un logiciel, Snorri, facilitant l'acquisition des connaissances acoustico-phonétiques.

Les motivations qui ont amené la conception et le développement de Snorri sont donc semblables à celles qui ont conduit à la création de Spire [8], de ILS [7], ou de Audlab [5].

Le but de cet article est de présenter sommairement ce logiciel (pour connaître toutes les fonctions disponibles se reporter à [4]), qui est à la disposition gratuite des universités.

Snorri est un outil d'observation destiné aux spécialistes comme aux non spécialistes de la parole. Il possède les outils classiques pour acquérir et restituer le signal de parole, le visualiser et calculer le spectrogramme, ainsi qu'un certain nombre de fonctions permettant d'analyser plus finement la parole et de manipuler des corpus.

2 Configuration matérielle et logicielle

Snorri fonctionne sur un Masscomp 5600. Il utilise une carte d'acquisition 12 ou 16 bits. Tous les calculs de traitement du signal sont effectués sur un processeur vectoriel VA-1 qui multiplie par un facteur supérieur à 10 la vitesse d'exécution. Les fichiers de parole sont stockés sur un disque de 400 Mo.

L'affichage a lieu sur une console bitmap couleur (4 ou 12 plans, 1152x910 pixels). Les copies d'écran sont faites sur une imprimante Postscript avec une résolution de 400 points par pouce.

Snorri est programmé en C et fait appel au multifenêtrage et aux primitives graphiques de Masscomp. Toutes les primitives de traitement du signal sont programmées avec les commandes du processeur vectoriel mais une version sans appel au processeur vectoriel (sauf pour les cepstres) est aussi disponible.

3 Organisation de Snorri

Snorri est un système dont toutes les fonctions sont accessibles avec la souris. Elles sont regroupées par grandes classes. Chaque classe est liée soit à une représentation du signal de parole (signal temporel, spectrogramme), soit à une des utilisations possibles de Snorri (étiquetage de la parole, copies de fenêtres et multifenêtrage ...).

Il faut discerner deux types d'utilisation de Snorri:

1. études de phrases isolées,
2. étude d'un corpus déja enregistré et commun à plusieurs chercheurs; l'utilisateur dispose alors d'outils d'exploration du corpus.

Nous n'avons pas limité la taille des fichiers de parole contrairement à certains environnements de traitement de parole comme dans le projet SPAR [2]. Snorri étant développé, depuis le début, à l'intérieur de notre équipe n'a pas eu à intégrer des modules d'origines diverses imposant des contraintes de longueur sur les fichiers manipulés. Néanmoins nous avons défini toute une classe de fonctions qui permettent au programmeur de s'affranchir de la gestion des déplacements dans le signal de parole.

Snorri exploitant largement les ressources du multifenêtrage associe une fenêtre au signal temporel, au spectrogramme, mais aussi à certaines fonctions (zoom, suivi de formants ...). L'utilisateur peut facilement modifier l'environnement dans lequel il évolue en changeant la disposition des fenêtres, en en créant de nouvelles, ou en sélectionnant une autre palette de couleurs.

Snorri dispose aussi d'une fonction de copie de fenêtres qui permet de conserver une trace écrite de toute fenêtre existante. L'utilisateur peut règler de nombreux paramètres (copie noir et blanc, en niveaux de gris, dimensions de l'image, titre, modification du contraste, affectation d'un niveau de gris à une couleur quelconque). Les images produites sont au format Postscript. Toutes les figures de cet article ont été obtenues avec cette fonction.

4 Signal temporel

La figure 1 montre le signal temporel tel qu'il est acquis par Snorri [1]. La fréquence normale d'échantillonnage est 16 kHz, pour l'acquisition comme pour la restitution, mais il est possible de l'augmenter jusqu'à 330 kHz. Chaque acquisition est automatiquement sauvegardée. L'utilisateur peut redéfinir les début et fin du signal à étudier et le mettre en mémoire après lui avoir donné un nom.

L'utilisateur peut examiner avec précision une partie du signal grâce à un zoom dont il précise l'étendue avec la souris. De même le début et la fin de phrase sont marqués par Snorri mais peuvent être déplacés pour n'étudier que la partie la plus intéressante du signal.

Les fichiers enregistrés au CRIN ne contiennent que le signal de parole mais Snorri peut aussi lire les fichiers au standard GRECO ce qui lui permet d'avoir accès au corpus BDSONS. Quand Snorri est utilisé pour étudier un corpus, la liste des fichiers disponibles apparait à l'écran et l'utilisateur sélectionne le fichier à étudier avec la souris.

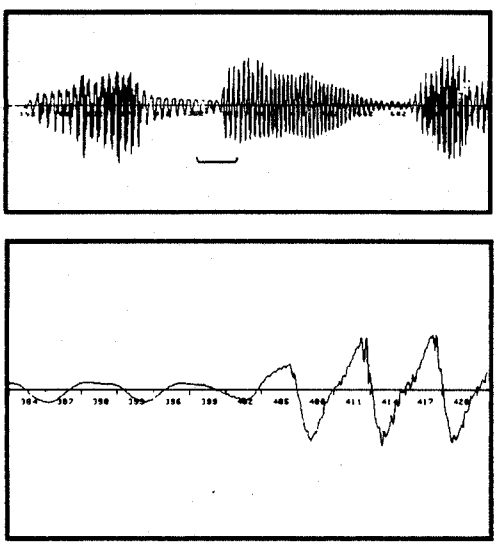


Figure 1: signal temporel et zoom sur une partie du signal

5 Spectrogramme

Snorri calcule et affiche un spectrogramme [2] en 13 niveaux de couleurs (l'utilisateur a le choix entre plusieurs palettes). Pour 2 secondes de parole, il faut 0.5 seconde de calcul et 8 secondes d'affichage. Il est aussi possible de choisir entre un spectrogramme calculé par FFT (Fig. 2) et un spectrogramme calculé par FFT sur les coefficients de la LPC (Fig. 3) sur lequel les formants apparaissent plus nettement.

L'utilisateur peut demander à Snorri de recalculer localement un spectrogramme de plus haute définition (Fig. 4) sur lequel les détails apparaissent plus clairement.

6 Etiquetage de la parole

Snorri avec les visualisations de la parole décrites précédemment est un bon outil de segmentation et d'étiquetage. Le phonéticien pose ses marques temporelles avec la souris, soit sur le spectrogramme initial (Fig. 2), soit sur le zoom d'une partie du spectrogramme (Fig. 4). Pour l'aider le phonéticien peut écouter un morceau de parole autant de fois qu'il le veut. D'autres options permettent, en cas d'erreur, de modifier la limite du segment ou son étiquette, et même de supprimer le segment créé.

L'utilisateur peut se déplacer dans le signal avec la souris et se positionnner sur une étiquette existant déja.

Les marques temporelles et les étiquettes phonétiques sont sauvegardées dans un fichier qu'il est ensuite possible de rappeler pour le consulter ou le modifier.

7 Utilisation de corpus de parole

Snorri est avant tout conçu comme outil d'acquisition de connaissances acoustico-phonétiques et c'est donc dans cet esprit que nous avons organisé les accès aux corpus étiquetés.

Afin de bien étudier les variations inhérentes au locuteur et celles liées au contexte (ex: plusieurs réalisations du phonème /b/) il faut pouvoir avoir accès rapidement à toutes les occurences d'un phonème à travers tout le corpus, et pouvoir repérer facilement le contexte dans lequel ce phonème a été extrait. Il faut disposer de corpus communs à l'ensemble des chercheurs, mais aussi de corpus de test directement liés à l'étude menée.

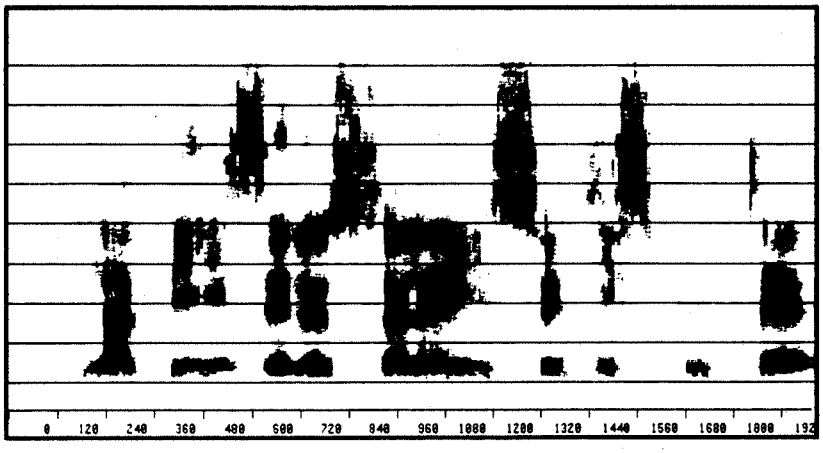


Figure 2 : spectrogramme normal ("La bise et le soleil se disputaient")

[1] L'unité de temps dans toutes les figures est la milliseconde.
[2] Les traits horizontaux indiquent les fréquences en kHz.

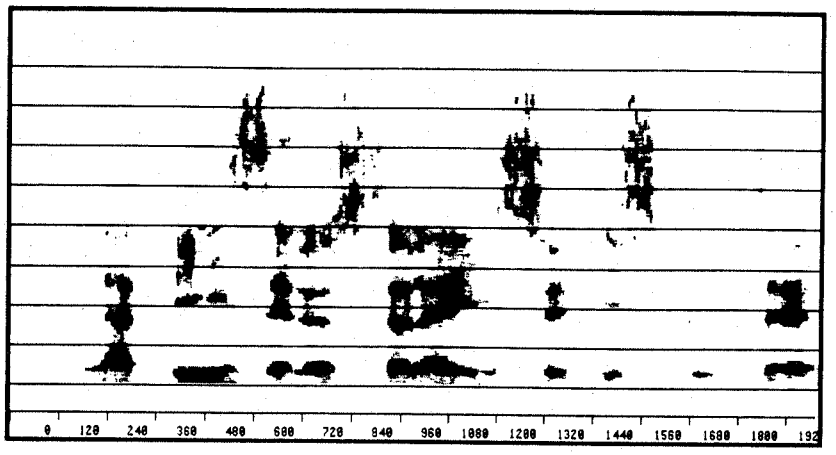


Figure 3 : spectrogramme calculé par FFT à partir des coefficients de LPC ("La bise et le soleil se disputaient")

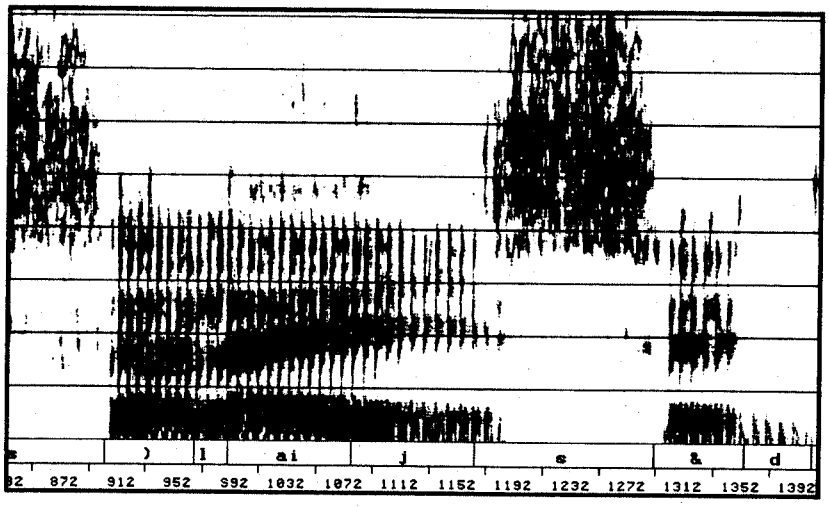


Figure 4 : spectrogramme de haute définition et exemple d'étiquetage

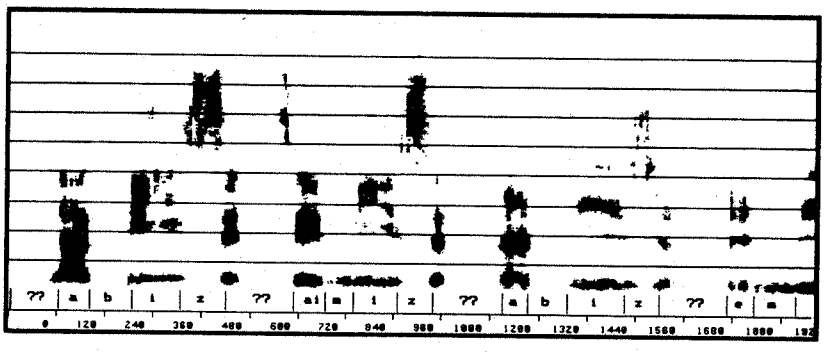


Figure 5 : extraction du contexte (labiale + /i/)

Ces exigences nous ont donc amené à choisir l'organisation suivante:

1. L'utilisateur peut choisir soit le corpus "public" soit son propre corpus. Il lui suffit ensuite de cliquer pour choisir le fichier de parole sur lequel il souhaite travailler.
2. Le phonéticien peut extraire, soit un phonème particulier (ex: les /i/), soit une séquence de phonèmes (ex: les /i/ précédés d'une labiale), soit une séquence de classes phonétiques (ex: les voyelles antérieures) dans tout le corpus étiqueté. Snorri construit alors un fichier de parole avec les séquences correspondantes présentes dans le corpus ainsi que l'étiquetage qui va avec (ex: la figure 5 montre les /i/ précédés d'une

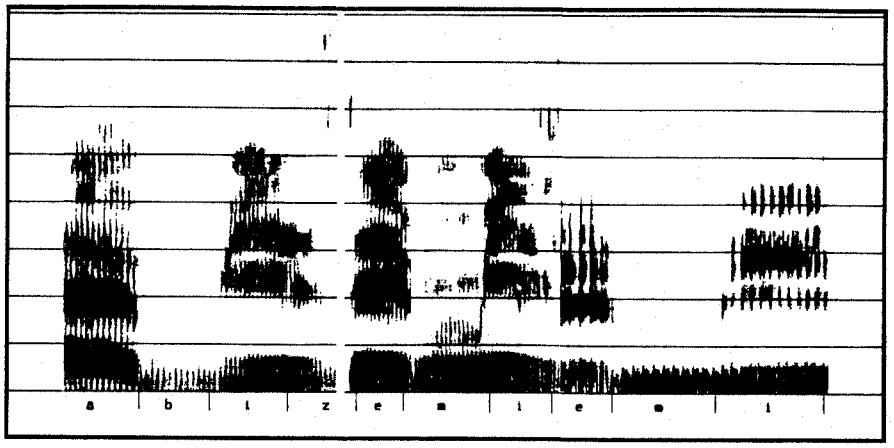


Figure 6 : spectrogrammes fins du contexte (labiale + /i/)

labiale). Ces deux fichiers (signal et étiquetage) sont placés dans le corpus du phonéticien. Il est alors plus facile de découvrir les corrélats acoustiques avec ces fichiers grâce à la visualisation successive de toutes les séquences (l'écoute est également possible).

La figure 6 montre les spectrogrammes fins de 3 occurrences du contexte (labiale + /i/). Une observation rapide permet de tirer les conclusions suivantes:

- élévation rapide (dans le sens du temps) des formants F2 et F3 au début de la voyelle.
- apparition plus rapide de F2 que de F3 au début de /i/.

8 Outils de phonétique

1. Outre les outils généraux de manipulation et d'étiquetage de parole, Snorri offre d'autres outils destinés à l'étude détaillée d'un spectre. Le spectrogramme ne faisant pas apparaître suffisamment clairement les niveaux d'énergie, Snorri calcule et affiche (Fig. 7) pour une fenètre choisie sur le signal temporel, la FFT, la LPC et le cepstre (échelle linéaire).
2. Comme ce sont en général les maxima de la LPC ou du cepstre qui sont intéressants le phonéticien peut afficher les pics de LPC ou de cepstre dans un domaine

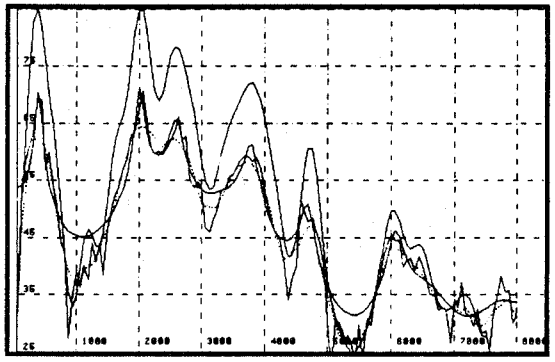


Figure 7: FFT, FFT lissée, LPC et cepstre. La courbe en dents de scie représente la FFT, la courbe en pointillés la FFT lissée, la courbe dont les premiers maxima sont renforcés, le cepstre (avec une échelle verticale différente), et la dernière la LPC.

du signal temporel délimité par la souris. Snorri permet aussi de suivre les pics de cepstre et de les organiser en lignes de pics (Fig. 8). La recherche des lignes de pics conduit à un algorithme de suivi de formants: cet algorithme doit éliminer certains pics (dans le cas d'un lissage cepstral) n'apportant aucune information formantique. D'autre part il tente de propager les solutions locales en émettant des hypothèses de jonction

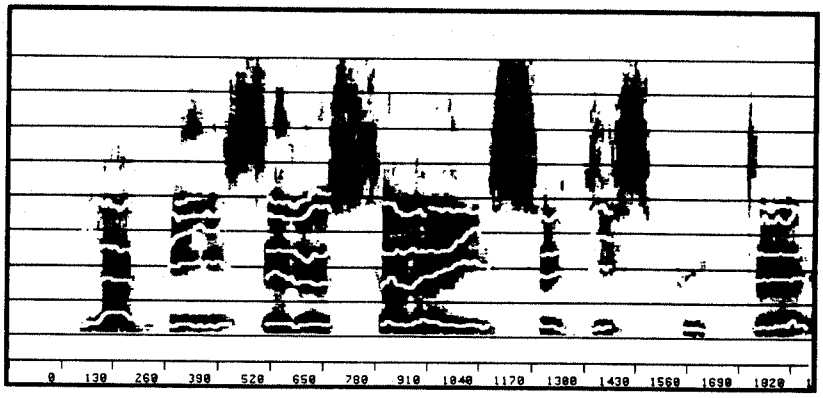


Figure 8 : suivi de pics de cepstres sur les segments vocaliques

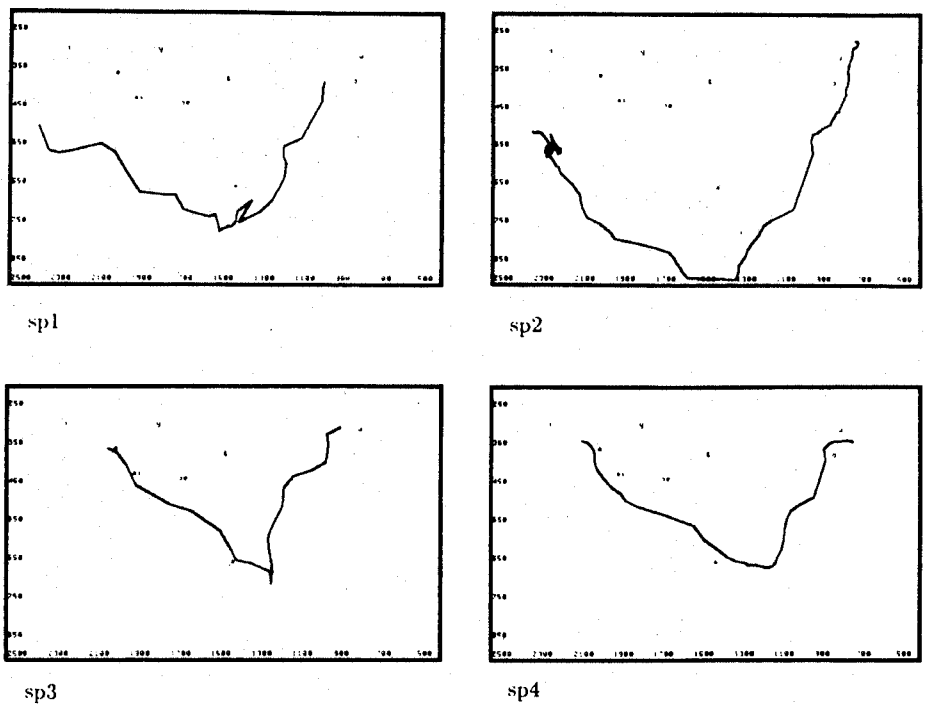


Figure 9 : suivis de formants dans le plan F1F2

ou d'élimination de certaines lignes de pics. Nous disposons d'un autre algorithme travaillant avec la LPC: les racines de LPC conduisent en effet assez directement aux formants. Le calcul de LPC ne modèlisant pas correctement le couplage entre le conduit vocal et les fosses nasales conduit à des erreurs notamment sur les voyelles nasales.

3. Le phonéticien a souvent besoin de décrire les phénomènes de coarticulation en étudiant le chemin que suivent les formants et en le situant par rapport aux cibles que sont les phonèmes qui auraient du apparaître. Snorri permet donc de suivre dans le plan F1-F2 les deux premiers formants (estimés grâce aux pics de LPC) sur un segment de parole (Fig. 9). Les cibles de référence des phonèmes du français pour un locuteur masculin sont indiqués et permettent de se repérer.

   La figure 9 montre le chemin que suivent F1 et F2 lors de la prononciation en série des voyelles /e//ɛ//a//o//u/ pour deux locutrices (sp1 sp2) et deux locuteurs (sp3 sp4). On découvre clairement sur la figure 9 que la position du triangle dépend très sensiblement du locuteur. Ainsi pour la voyelle /e/ les formants des locutrices sont nettement supérieurs à ceux des locuteurs (400 Hz pour F2 et 100 Hz pour F1). Le triangle vocalique apparaît donc comme un outil efficace de normalisation des références vocaliques.

9 Comparaison de Snorri avec les autres systèmes d'analyse de la parole

L'auteur n'a hélas pas pu voir fonctionner d'autre système d'analyse de la parole et les articles disponibles sont trop imprécis pour évaluer ces logiciels.

L'ambition de Snorri est d'accroître considérablement le volume de parole que le chercheur peut étudier. Snorri permet donc, à la différence de Audlab et d'ILS, d'explorer facilement tout le corpus étiqueté, sans faire de statistiques comme le permet Spirex [6]. Snorri se distingue aussi des autres systèmes par l'affichage d'un spectrogramme calculé sur les coefficients de LPC et par la visualisation dans le plan F1-F2 du suivi de formants d'un segment de parole. Par contre l'utilisateur ne peut pas modifier Snorri aussi facilement que dans Spire puisqu'il doit pour cela modifier directement le code.

10 Utilisation de Snorri au CRIN

Snorri sous la forme qui vient d'être présentée est utilisé comme outil d'étiquetage (corpus GRECO "La bise et le soleil" et un corpus de phrases météorologiques) et comme outil d'analyse acoustico-phonétique de corpus. Par ailleurs il sert de base à plusieurs projets:

1. Dominique Fohr a réimplanté le système expert en décodage acoustico-phonétique Aphodex [1] en utilisant les modules de Snorri.
2. Les résultats du décodage acoustico-phonétique sont exploités par le système de reconnaissance lexicale de Bernard Mangeol [3].
3. Une version de Snorri (intégrant le calcul de F0) est destinée à l'étude de la structure prosodique de la phrase et permet de localiser les fins de mots et les fins d'unités syntaxiques.

11 Conclusion

Snorri a demandé environ 1 an de travail: de nombreuses améliorations sont en cours de développement (notamment un éditeur de signal). Actuellement ce programme nous sert comme outil d'expérimentation pour la mise au point

de notre système de décodage. Par vocation Snorri est donc un outil d'acquisition de connaissances phonétiques et non pas un outil de traitement du signal comme l'est ILS [7]. Snorri était initialement destiné au CRIN. De nombreux visiteurs ayant montré un net intérêt pour ce logiciel, il a été donné à plusieurs équipes en France (CNET à Lannion, ICP à Grenoble, prochainement le CERFIA à Toulouse). Ces équipes contribueront par leur travail à faire évoluer Snorri (l'adjonction d'un module statistique est déja prévu).

Remerciements

Je tiens à remercier Dominique Fohr qui est à l'origine de Snorri, François Lonchamp et Jacqueline Vaissière qui par leurs conseils ont contribué à l'élaboration de Snorri.

Références

[1] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, and J. M. Pierrel. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986.

[2] M.A. Huckvale et al. The Spar speech filing system. In *European Conference on Speech Technology*, 1987.

[3] Jean-Paul Haton, Bernard Mangeol, and Jean-Marie Pierrel. Organisation et fonctionnement d'une composante lexicale dans un syste'me de dialogue homme machine. In *Actes du se'minaire "Lexique et traitement automatique des langages"*, pages 259–267, Toulouse, 1986.

[4] Y. Laprie. *Notice d'utilisation de Snorri.* Technical Report, CRIN, 1988.

[5] University of Edinburgh. *The AUDLAB Interactive Speech Analysis System.* Technical Report, 1986.

[6] D.W. Shipman. Spirex: statistical analysis in the spire acoustic-phonetic workstation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, 1983.

[7] Signal Technology. *Introduction to ILS Interactive Laboratory System.* Technical Report, Signal Technology, Inc., 1986.

[8] Victor W. Zue and D. Scott Cyphers. *The MIT Spire System.* Technical Report, 1985.

Contour de la communication parlée, hier aujourd'hui et demain

LA COMMUNICATION PARLEE EST-ELLE UNE SCIENCE ?
En doutiez-vous ?

ELEMENTS DE DISCUSSION ET DE REFLEXION SUIVIS DE REPERES CHRONOLOGIQUES.

Louis-Jean BOE
Institut de Phonétique de Grenoble
Institut de la Communication Parlée
UA CNRS n° 368

&

Jean-Sylvain LIENARD
LIMSI
Orsay
LP CNRS n° 3251

RESUME

Nous nous proposons de montrer comment la Communication Parlée s'est constituée en véritable domaine scientifique par une réorganisation de savoirs partagés essentiellement depuis le siècle dernier entre la physiologie, la phonétique et l'acoustique. Après un rappel historique concernant ces différents champs, sont analysées quatre ruptures épistémologiques provoquées par les développement de la linguistique structurale puis de la phonologie, les projets des "Preliminaries to Speech Analysis" et les difficultés rencontrées par la reconnaissance automatique de la parole. La Communication Parlée apparait comme un système spécifique, opérant avec un objet spécifique dans la biologie des espèces, avec comme pôles principaux la cognition, la production, l'acoustique et la perception gouvernés par le code.

0 - INTRODUCTION

Plus de vingt ans après la première réunion nationale sur la parole*, il est temps de s'interroger sur la nature du domaine de la Communication Parlée et cela d'autant plus que certaines disciplines connexes auraient quelques tendances expansionnistes. (c.f. la note finale). Sans que les critères soient explicites, les acteurs de ce champ de recherche savent bien s'identifier, se différencier des spécialistes des autres disciplines et même préciser le chemin parcouru par rapport à leur formation d'origine.

Historiquement, la parole est un objet d'étude pluridisciplinaire** à l'intersection de trois grands secteurs*** : les Sciences Physiques (et de l'Ingénieur), la Médecine, et les Sciences Phonétiques avec des apports spécialisés de l'acoustique, de l'électronique, de l'informatique, du traitement du signal, de l'analyse des données, de la reconnaissance des formes, de l'intelligence artificielle, des sciences cognitives, du neuromimétisme, de l'anatomie, de la physiologie, de l'électrophysiologie, de l'acquisition et de la pathologie de la parole et du langage, de la psychomotricité, de la psychoacoustique, de la psycholinguistique ...

Mais en focalisant ainsi tant de faisceaux de recherche, la parole n'est-elle pas devenue plus qu'une application pluridisciplinaire ? Peut-on dire, comme nous l'avons déjà avancé (L.J. BOE, 1985, 86; J.S. LIENARD, 1985) que s'est constitué un véritable corps de connaissances visant à l'établissement de lois et de modèles qui régissent un sytème bien spécifique, celui de la Communication Parlée.

Sans laisser de côté les aspects internationaux (ils sont en permanence sous-jacents dans notre exposé et explicites dans les repères chronologiques), nous avons centré notre discussion au niveau français, estimant que les développements et la structuration des recherches au sein de la communauté hexagonale ont été et restent suffisamment importants pour être révélateurs de l'évolution que nous avons voulu cerner.

I - LA PAROLE ECLATEE

Dès le début du XIXe siècle la parole relève essentiellement de trois grands secteurs : la physiologie, la physique et la phonétique.

I.1. La physiologie

La parole y est étudiée comme caractéristique spécifique de l'homme. Elle fait partie des fonctions de relation avec le mouvement et les cinq sens.

---

* Préfigurant les premières Journées d'Etude sur la Parole, le Colloque sur les "Structures Acoustiques de la Parole" a eu lieu du 10 au 12 avril 1967 à Grenoble sour l'égide du Groupement des Acousticiens de Langue Française (le GALF). Cette manifestation a regroupé 126 participants parmi lesquels on peut noter la présence de nombreux acousticiens, audiophonologues, électroniciens, informaticiens, ingénieurs des télécommunications, linguistes, neurologues, neurophysiologues, orthophonistes, phonéticiens, physiologues, psychologues, psycho-phonéticiens (organisation : R. GSELL, P. MOUNIER-KUHN, B. VAUQUOIS, J.C. LAFON).

** On peut considérer avec P. DELATTRE (pas le phonéticien, mais le spécialiste des systèmes) que la pluridisciplinarité est une association de disciplines qui concourent à une réalisation commune, mais sans que chacune d'entre elles n'ait à modifier sensiblement ni sa propre vision des choses ni ses propres méthodes.

*** Il faudrait compléter ces trois pôles en nous penchant du côté de la psychologie et de l'acoustique/informatique musicale.

En France la psychologie s'est, pour l'essentiel, intéressée à la langue (système) en tant qu'objet cognitif et pratiquement pas à la parole. Qui plus est, si la parole relève du champ de la psychoacoustique, elle fait appel, vraisemblablement, à des modes de perception bien spécifiques. Nous noterons les travaux sur la perception du nourisson et l'acquisition du langage (J. MEHLER & J. BERTONCINI, B. DE BOYSSON-BARDIES, 1980) et les liens qui se sont noués tout récemment entre la parole et la psychologie dans le domaine de la psychomotricité ; ces travaux s'inscrivent dans le cadre des études consacrées aux coordinations temporelles des différentes phases des segments corporels. Il est peut-être trop tôt pour préjuger de l'avenir, mais une voie s'ouvre pour préciser le poids des invariances et des paramétrisations dans le contrôle moteur de la parole (C. ABRY, J.P. ORLIAGUET & R. SOCK, 1988).

La parole et l'acoustique musicale ont, c'est certain, entretenu des relations croisées, souvent au sein des mêmes laboratoires : aux USA (M. MATHEWS, J.C. RISSET, J.P. OLIVE) à la BELL, à Stockholm (J. SUNDBERG) au KTH, en France avec A. MOLES (problèmes généraux de la communication), au Laboratoire d'Acoustique avec le Groupe d'Acoustique Musicale à Paris (E. LEIPP, M. CASTELLENGO, J.S. LIENARD), à l'IRCAM à Paris (J.C. RISSET, X. RODET, J.P. JULLIEN), à l'ACROE en relation avec le LCP (C. CADOZ) à Grenoble.

Les manuels d'anatomie et surtout ceux de physiologie (voir par exemple celui de J. BECLARD, 1869) présentent, non seulement des descriptions détaillées du larynx et des articulateurs, mais de véritables explications sur le fonctionnement de la source (comparaison avec les instruments à vent et/ou à cordes). L'influence des physiologistes hollandais, autrichiens et surtout allemands, est considérable : J. MULLER, K. LUDWIG, E.W. BRÜCKE, F.C. DONDERS, J.N. CZERMARK, C.L. MERKEL. C'est l'aspect phonation qui y est certainement le plus détaillé : grâce au chanteur M. GARCIA l'observation laryngoscopique s'est développée dès 1835, et c'est à cette époque que le montage de larynx de cadavres et le développement de modèles mécaniques vont permettre de faire progresser les connaissances sur la source vocale.

La classification des voyelles et des consonnes est approfondie (voyelles extrêmes i, a, u, opposition oral/nasal). C'est aussi à cette époque que la physiologie met en place un ensemble de procédés d'enregistrement graphique (C. LUDWIG en Allemagne, J.E. MAREY en France).

D'ailleurs, en parallèle avec la physiologie, la médecine s'est dotée d'une discipline, la "physique médicale" qui fait une place, dans le domaine de l'acoustique, à la parole (V. DESPLATS & C.M. GARIEL, 1870; W. WUNDT, 1884).

Aux frontières de la pathologie et de la prévention il existe aussi, à cette époque, une tradition "hygiène de la voix parlée et chantée" avec un enseignement et des manuels (voir par exemple A. CASTEX, 1894).

Dans le champ relevant plus ou moins directement de la pathologie, nous situerons, par simplification, tous les travaux effectués dans le cadre de la rééducation des sourds-muets. Les "silencieux", après avoir été quelque peu maltraités par la société, ont trouvé des défenseurs, au siècle précédent, notamment avec l'Abbé DE L'EPEE (1712-1789), mais ce n'est qu'avec BONAPARTE que va leur être octroyée l'entière capacité civile.

En 1880, le Congrès de Milan va proclamer la supériorité de la méthode orale pure, c'est-à-dire de "l'ensemble des procédés ayant pour but de faire de la parole lue sur les lèvres et articulée par le sourd, l'instrument de communication de la pensée [en] supprimant l'intervention des signes" (G.M. DEMEYER, 1885). C'est dire toute l'importance des travaux qui vont être centrés autour de l'articulation et de la lecture labiale (voire oro-faciale). A partir du siècle suivant les écoles d'orthophonie vont, par leur enseignement, diffuser (essentiellement dans le cadre des facultés de médecine) les connaissances sur la parole normale et pathologique (cf. les travaux de S. BOREL-MAISONNY).

I. 2. La physique

Dans les ouvrages de physique, la parole apparaît dans l'acoustique, au chapitre de la production et de la propagation des sons, après les notions sur les vibrations des corps, sur les échelles musicales et les tuyaux sonores. Grâce aux travaux de H.L.F. HELMOLTZ et aux recherches du constructeur R. KOENIG, sont présentées l'analyse des sons par résonateurs couplés à des flammes manométriques et la synthèse par sirènes acoustiques et même par diapasons électriques. C'est bien sûr l'aspect physique de la parole qui est évoqué mais, dans certains ouvrages, la classification des voyelles et des consonnes est présentée, bien que de manière moins approfondie qu'en physiologie. Très souvent le mécanisme de l'audition fait suite au chapitre sur l'acoustique mais, en fait, n'est pas établie de véritable relation entre la production du son et sa perception.

Mais le phénomène qui va marquer la deuxième moitié du XIXe siècle, c'est l'aboutissement d'un rêve qui a agité les générations précédentes : matérialiser le signal de parole, l'enregistrer et le reproduire. C'est un typographe, E.L. SCOTT, qui va réussir à fixer "les sons de l'air" en imaginant le phonautographe. Faute de moyens, il ne pourra réaliser cet appareil qu'en prenant un contrat avec le génial constructeur R. KOENIG.

Si E.L. SCOTT peut réaliser les premiers tracés du signal de parole. Il constate très vite que la deuxième tâche qu'il s'est fixée : "arriver ensuite par le secours de moyens mathématiques à déchiffrer cette sténographie naturelle" est hors de sa portée*. Ni le succès, ni le grand public ne vont être au rendez-vous de E.L. SCOTT. Si celui-ci s'est donné comme but de fixer la trace de la parole, il a clairement indiqué qu'il ne souhaitait pas la reproduire.

C'est T. EDISON qui en 1878 va y parvenir le premier, avec le phonographe que C. CROS imagine mais ne réalise pas. Il faut le souligner, le phonographe à feuille d'étain, malgré une énorme publicité, est une invention mort-née. Peut être à cause de la mollesse des compagnies d'exploitation. Il reste à T. EDISON à réinventer, 10 ans plus tard, le phonographe à rouleau de cire ; ce sera l'un des grands succès de l'Exposition Universelle de Paris en 1889.

Une autre invention, qui ne va pas directement faire progresser les recherches sur la parole, mais attirer bon nombre d'ingénieurs dans son champ, c'est le téléphone, présenté en 1876 à l'Expositon de Philadelphie, par G. BELL , son inventeur. En 1878, le premier standard téléphonique est installé à New Haven. Il y aura à Paris, à la veille de la première guerre mondiale, près de 100.000 abonnés et cinq fois plus à New York. La TSF va aussi permettre le transport de la parole. En 1898 est établie la première communication entre la France et l'Angleterre, en 1903 entre l'Angleterre et les USA.

La parole va petit à petit déserter les chapitres de l'acoustique, qui d'ailleurs a du mal à s'affirmer dans le cadre de l'enseignement supérieur. C'est au chapitre du téléphone et de la radioélectricité qu'elle va s'installer. Tout se met en place pour que, progressivement, la Communication Parlée attire les électriciens, les électroniciens (et plus tard les informaticiens) qui veulent en savoir plus sur l'objet qu'ils transmettent, traitent et analysent.

I. 3. Les sciences phonétiques

Lorsque les premiers Congrès de Phonétique se tiennent en 1886 à Paris et à Stockholm, il existe depuis 10 ans des manuels de cette discipline, qui date du milieu du siècle : ceux de E. SIEVERS et de H. SWEET, par exemple. La phonétique commence à être enseignée aux USA à Harvard en 1882 dans le cadre de la philologie.

C'est aussi en 1886 que se crée l'Association de Phonétique Internationale (l'API) avec sa revue, le "Maître Phonétique" dirigée par P. PASSY.

Les liens de la phonétique avec la linguistique, qui à l'époque est historique, restent faibles. En effet la grammaire comparée s'est affinée au cours de la première moitié du XIX° siècle. Ses fondateurs s'appuient surtout sur les théories grecques et opèrent à partir du niveau orthographique. SCHLEICHER sera le premier à proposer une reconstruction de l'indo-européen à partir de considérations articulatoires.

---

* Une lucidité qui va faire défaut à certains reconnaisseurs des années 1950, certains d'entre eux fixant à l'horizon 70 la fin de leurs peines. Il est vrai que d'autres, plus prudents, prophétisaient un débarquement sur la lune avant l'aboutissement de la reconnaissance automatique de la parole (cf. J. DREYFUS-GRAF, 1967).

Essentiellement instrumentale et axée sur la physiologie, la phonétique va atteindre avec O. JESPERSEN un "sommet et un achèvement" dans sa forme classique.

En 1896 est créée la première chaire de phonétique à l'Institut Catholique de Paris. En 1897 naît le laboratoire de Phonétique Expérimentale du Collège de France (P. ROUSSELOT), et c'est la même année que l'Alliance Française fonde un centre de correction d'accents étrangers (l'époque est à la norme !).

Le premier Institut de Phonétique universitaire voit le jour à Grenoble, en 1904. Le début du XXème va être incontestablement marqué par l'Abbé P. ROUSSELOT. Ses "Principes de phonétique expérimentale" constituent un véritable monument. Il couvre tous les domaines de l'articulatoire et l'analyse manuelle des tracés du signal de parole, en série de Fourier, ne le décourage nullement ! En 1911, l'Institut de Phonétique de Paris est, dès sa naissance, regroupé avec le Musée de la Parole qui compte alors plus de mille phonogrammes.

La phonétique va se diversifier : à l'étude historique et à la physiologie, ses axes principaux, elle va ajouter les aspects anthropologiques, acoustiques, biologiques, neurologiques, pathologiques, psychologiques, sociologiques, tout en intégrant certaines pratiques expérimentales. Si bien que les organisateurs chargés du deuxième Congrès International de Phonétique Expérimentale vont prendre une décision d'importance. A leurs yeux la phonétique ne constitue plus une seule science et ils décident de la tenue du premier Congrès International des Sciences Phonétiques.

Il aura lieu à La Haye, en 1932 et sera couplé à une réunion dont l'importance historique est fondamentale : la première Réunion Phonologique Internationale. Les deux congrès suivants, en 1935 et 1938 seront pour les dernières fois associés aux réunions de phonologie. Il faudra attendre 1961 pour que le IVe Congrès des Sciences Phonétiques puisse se tenir à nouveau, et la seconde guerre mondiale n'est pas l'unique raison de ce long silence. Comme nous l'analyserons au chapitre III, une rupture épistémologique d'importance va se produire en 1932.

II - LA PAROLE FOCALISANTE

Nous allons passer par un raccourci historique des années 30 à la fin des années 60. Nous avons voulu montrer comment, jusque là, par un énorme effort d'enseignement et de recherche, la physiologie avait fourni à la phonétique des bases solides de description articulatoire et comment, portés par les développements de l'enregistrement sonore, du téléphone, de la radio, des scientifiques (essentiellement électroniciens) vont tout naturellement se tourner vers la parole.

Dès le début des années 40 et en moins de 25 ans, vont se développer dans les laboratoires d'électronique (et plus tard d'informatique), souvent avec la participation de phonéticiens, une impressionnante quantité de travaux dans le domaine de l'analyse, de la modélisation du conduit vocal, de la synthèse et de la reconnaissance.

En consultant les repères chronologiques on pourra se faire une idée du chemin parcouru. En 1936, H. DUDLEY réalise, avec le vocodeur, le premier codeur moderne : la parole est bien redondante et la théorie de l'information de C. SHANNON permettra, 12 ans plus tard, de le préciser quantitativement. La production des voyelles est déjà bien modélisée dès 1941 par T. CHIBA & M. KAJIYAMA (un phonéticien et un physicien).

En 1946, le spectrographe acoustique (bien connu sous le nom commercial de "sonagraph") va être conçu et réalisé pour délivrer de "la parole visible" et tenter de permettre aux sourds de lire ce qu'ils n'entendent pas. Si cette tentative tournée vers les malentendants sera un échec, la représentation sonagraphique du signal de la parole se révèlera comme la plus riche*. La notion de formant devient explicatoire et permet de passer de l'articulatoire traditionnelle (lieu et aperture) à l'analyse acoustique avec vérification par la synthèse (P. DELATTRE, 1948). En 1950 un analogue électrique du conduit vocal est réalisé par H.K. DUNN et en 1960 G. FANT publie la célèbre "Théorie Acoustique de la Production" (l'analogie acoustique/électrique permet d'utiliser tous les concepts et méthodes de la théorie des lignes et du calcul matriciel des des quadripôles).

La synthèse à formants permet la réduction de la redondance (W. LAWRENCE, 1953) et devient un moyen de validation de règles issues de l'analyse (A.M. LIBERMAN & al., 1959).

La reconnaissance automatique de la parole débute aussi à cette époque, de façon presque inaperçue, en URSS (MJASNIKOV, 1943) et en Suisse (J. DREYFUS GRAF, 1950). D'ailleurs les rétrospectives historiques commencent souvent avec les travaux menés aux USA (C.P. SMITH, 1951; K.H. DAVIS & al., 1952, etc.). Avec l'informatique, la numérisation du signal de parole va ouvrir la voie au traitement du signal (FFT, LPC, cepstre) à l'analyse des données (analyse factorielle, des correspondances, multidimensionnelle...).

Il n'est donc pas étonnant qu'en 1967 on puisse réunir en France 126 participants, venus de tous les horizons (cf. introduction) sur le thème des structures acoustiques de la parole. Pas étonnant non plus que cette réunion constitue la générale des Journées d'Etude sur la Parole dont la première édition aura lieu à Grenoble en 1970 et qui sera suivie de 16 autres (à ce jour). Le Groupe Communication Parlée du GALF (premier président R. CARRE) voit le jour la même année et il se place tout naturellement sous l'égide du Groupement des Acousticiens de Langue Française, dont le fonctionnement est grandement facilité par la logistique du CNET de Lannion. Ingénieurs et phonéticiens vont constituer l'essentiel des troupes mais on trouve aussi des psychoacousticiens, des phoniatres, des orthophonistes, des psychologues, etc.

Depuis vingt ans le Groupe Communication Parlée du GALF joue le rôle d'un creuset dans lequel s'est construite, comme nous tenterons de le montrer, une nouvelle science. C'est grâce à cette structure, en dépit du cloisonnement universitaire et des distances liées aux formations différentes que vont, dès le début, s'établir des échanges pluridisciplinaires, se nouer des collaborations, se préparer des regroupements structurels.

- Au départ les finalités et les méthodes sont différentes. Les uns situent leurs recherches dans une problématique explicatoire : comment les locuteurs-auditeurs encodent et décodent au niveau cognitif les informations contenues dans le signal de parole via les étapes articulatoires, acoustiques et perceptives. Les autres, sans sous-estimer les aspects fondamentaux, sont souvent soumis à des contraintes de réalisation dans les domaines de la transmission, l'analyse, la synthèse et la reconnaissance.

Au milieu des années 70, comme nous le verrons, un premier bilan est tiré. Il est illusoire de vouloir traiter le signal de parole comme un "simple" signal physique, et il est impossible d'avancer des explications systémiques sans notions fondamentales ni méthodes de traitement adaptées aux systèmes.

---

* Comme mode de représentation on n'a pas trouvé mieux depuis ! Il faudra attendre presque quarante ans pour que l'on dispose, avec un système informatique, de documents ayant la même qualité que ceux qui étaient réalisés avec un appareillage électronique (J. MONNE, 1983).

Actuellement, le Groupe Communication Parlée de la Société Française d'Acoustique (nouvelle appellation du GALF) regroupe plus de 300 membres. Il ne fait aucun doute que la parole a eu un effet focalisant, mais a-t-elle changé les méthodes et les démarches scientifiques des participants ?

III - DES RUPTURES EPISTEMOLOGIQUES

Pour répondre à cette question , il est nécessaire de remonter dans le temps et d'essayer de cerner comment, par ruptures* épistémologiques successives ou réorganisations, tout s'est mis en place progressivement pour que se constitue une nouvelle science.

Ces ruptures vont se produire, côté sciences humaines, entre la linguistique et la phonétique et plus notablement entre la phonologie et la phonétique; côté sciences physiques, autour des problèmes fondamentaux rencontrés par la reconnaissance automatique de la parole et l'apport de l'intelligence artificielle.

Nous avons simplifié, voire grossi le trait, pour mieux souligner les grandes failles entre ces plaques tectoniques qui vont se déplacer en moins de deux demi-siècles.

1906 -1911 : Le cours de F. DE SAUSSURE

Dans cet intervalle, F. DE SAUSSURE va donner tous les deux ans un cours de Linguistique Générale qui va marquer fondamentalement les linguistes de l'époque et des générations suivantes : c'est la naissance du structuralisme linguistique. Il n'entre pas dans notre propos d'en discuter les principaux aspects théoriques, nous voudrions seulement en retenir ce qui va entraîner une cassure fondamentale entre la linguistique dont l'objet est l'étude du système (langue) et la phonétique qui est l'étude de la matérialisation de ce système (la parole).

Pour F. DE SAUSSURE l'indépendance entre les deux parties est totale puisqu'il s'agit d'une relation purement arbitraire. "Les organes vocaux sont aussi extérieurs à la langue que les appareils électriques qui servent à transmettre l'alphabet Morse sont étrangers à cet alphabet"**. Il suffit donc d'étudier le système, la façon dont il se réalise n'ayant que peu d'intérêt. L'exemple du jeu d' échec donné par F. DE SAUSSURE est hautement significatif : la description des pièces n'apporte rien à la théorie échiquéenne. La linguistique doit être "la science des systèmes des signes [sémiologie] au sein de la vie sociale".

Le conduit vocal et le système auditif n'ont qu'une importance tout à fait "secondaire" dans le processus de communication et la phonation, c'est à dire l'exécution des images acoustiques, n'affecte en rien le système lui-même. Une telle approche va reléguer l'étude de la parole aux confins de la linguistique, elle devient une science marginale et dévalorisée par essence, puisque ne pouvant éclairer la compréhension du système.

Il va falloir près d'un demi-siècle pour remettre en cause le principe d'indépendance entre le système et sa réalisation, d'autant plus que les phonologues vont enfoncer le clou avec un brio incontestable.

1932 : Sciences Phonétiques et Phonologie

Cette année est celle d'une réorganisation scientifique et d'une rupture épistémologique.

Le 1° Congrès International des Sciences Phonétiques

La phonétique devient plurielle. Les organisateurs chargés de mettre en place le 2° Congrès de Phonétique Expérimentale vont tirer les conséquences de l'éclatement de la phonétique ; descriptive jusque là, elle s'est ouverte à l'anthropologie, la psychologie, la neurologie, la sociologie, la biologie et même à la phonologie , ... au total plus de 12 ramifications. Avec réalisme, les phonéticiens vont prendre conscience du fait que la parole est une activité humaine fondamentale et qu'il va falloir maîtriser les approches différentes de cet objet complexe. Le mot "pluridisciplinarité" n'est pas prononcé, mais le concept est bien là ; il s'agit, en fait, d'une réorganisation sans véritable rupture conceptuelle.

La 2° Réunion Phonologique Internationale

Mais le fait le plus important, c'est la rupture qui va avoir lieu pendant ce même congrès. En effet les organisateurs ont invité les représentants du Cercle Linguistique de Prague (CLP) animé par N.S. TROUBETZKOY et R. JAKOBSON à tenir la 2° Réunion Phonologique Internationale. La phonologie (au sens moderne du terme) vient tout juste de naître, ses théories ayant été présentées au cours du 1° Congrès International des Linguistes, en 1928. Les phonologues du Cercle de Prague vont magistralement définir tout un cadre théorique et un ensemble de procédures permettant la mise en évidence, la classification et l'analyse systématique des unités abstraites élémentaires : les phonèmes, qui sont à la base de tout le fonctionnement du système linguistique.

Les aspects diachroniques et synchroniques y sont traités, tous les phénomènes étant, jusqu'à un certain point, déterminés par des lois de structures générales. Pour le CLP, l'étude du système relève de la phonologie, et donc de la linguistique, l'étude des sons relève des "sciences naturelles". La phonologie ne doit en aucun cas se sentir liée aux études articulatoires et acoustiques : "Les tâches de la phonologie ne sont en somme point affectées par ces méthodes, puisque la langue est en dehors de la mesure et du nombre" (N.S. TROUBETZKOY, 1939).

Les phonéticiens sont donc rejetés aux marges de la phonologie : "l'absence de distinction nette entre phonologie et phonétique [...] a eu une influence retardatrice sur le développement de la phonétique aussi bien que sur celui de la phonologie : il n'y a désormais aucun motif d'y persister" (N.S. TROUBETZKOY, 1939).

Bien entendu le concept de système est essentiel et les recherches effectuées par le Cercle de Linguistique de Prague, par les plus brillants linguistes de cette époque, vont transformer fondamentalement l'étude de la langue.

---

* C'est G. BACHELARD qui a le plus insisté sur cette notion épistémologique. Dans son esprit l'idée de rupture introduisait une discontinuité nette dans l'histoire conceptuelle. Il n'est pas toujours facile de faire la différence entre rupture et réorganisation (voir par exemple T. KUHN, Structure des révolutions scientifiques, 1956). Sans entrer dans les détails nous avons été très sensibles aux soubresauts, voire aux ruptures dans les domaines connexes (probabilité/ analyse des données, électronique/informatique, informatique/cybernétique, informatique/intelligence artificielle, intelligence artificielle/ sciences cognitives, automatisme/traitement du signal, traitement du signal/traitement des images ...).

** A y regarder de plus près il serait difficile de prouver qu'il n'y a pas de relations entre la structure du code Morse, d'une part, et les contraintes de l'appareillage de transmission électrique (utilisé à l'époque) et les contraintes psychomotrices et psychoacoustiques du manipulateur-récepteur, d'autre part.

Mais peut-on se désintéresser totalement* de la façon dont celle-ci se produit articulatoirement, se véhicule acoustiquement et se perçoit auditivement ?

Quarante ans plus tard B. LINDBLOM et J. LILJENCRANTS (1972) feront joliment la preuve que si une langue ne possède que trois voyelles ce sont toujours [i, a, u], et non point [e, a, o] ou [ɛ, a, ɔ] qui pourtant entretiennent les mêmes relations systémiques. Il existe donc bien quelque part une relation entre le système vocalique et la façon dont il se réalise et se perçoit**.

1952 : Les "Preliminaries to Speech Analysis"

Par un de ces retournements dont l'histoire a le secret, c'est un des fondateurs du Cercle Linguistique de Prague, R. JAKOBSON, alors au M.I.T. (Massachussetts Institute of Technology), qui va provoquer une nouvelle rupture, à première vue surprenante***. Comment, après avoir proclamé l'indépendance de la forme et de la substance, après avoir relégué la phonétique dans un exil peu gratifiant, la linguistique va-t-elle retourner aux enfers de la substance ?

En bref R. JAKOBSON, G. FANT et M. HALLE, un linguiste, un scientifique et un phonologue proposent, en 1952, un système universel pour décrire les langues du monde. Ils avancent que le phonème est constitué d'un faisceau de traits, résultant d'un choix binaire entre douze traits universels.

Jusque là, rien de surprenant : il est logique pour un théoricien de chercher des universaux et R. JAKOBSON réalise là le rêve que N.S. TROUBETZKOY a certainement nourri à l'instar de son compatriote D.I. MENDELEIEV : donner un tableau universel des éléments phoniques. A en croire G. MOUNIN (1972), R. JAKOBSON veut "rivaliser avec EINSTEIN en donnant au monde une théorie de la 'relativité linguistique' et non pas une linguistique générale mais une linguistique généralisée".

Ce qui est en totale rupture avec les principes de l'indépendance entre le système et sa réalisation, entre la forme et la substance, c'est la bi-univocité qui est faite par les auteurs des "Preliminaries" entre les traits et leur réalisation acoustique. Par exemple, le trait [± voisé] est décrit tout simplement comme présence/absence de vibrations laryngées. Après qu'ait été affirmée la totale arbitrarité du signifiant et du signifié, R. JAKOBSON, G. FANT & M. HALLE vont associer, de la façon la plus étroite qui soit, forme et substance.

Il y a là, il faut le souligner, une évolution très importante puisque les éléments du système phonologique sont réalisés par des invariants au niveau de la substance*.

La bi-univocité affirmée ne tardera pas à être remise en cause : le trait pouvant se réaliser par plusieurs indices (P. DELATTRE, 1958), une dizaine dans le cas du voisement, par exemple.

Mais par leur ambition, les "Preliminaries" vont rassembler les linguistes désireux de confronter à la réalité leurs modèles formels et les phonéticiens qui, dans leur très grande majorité, n'avaient jamais mis en doute la nécessité de devoir disposer d'un système formel sous-jacent. La phonétique va alors apparaître comme la pierre de touche des systèmes de formalisation et de leur limites ; et encore faut-il voir qu'il s'agit là d'un niveau bien abstrait au regard des contraintes physiques voire biologiques.

1976 : Les difficultés de la reconnaissance automatique de la parole

La rupture qui va se produire au milieu des années 70 est liée aux difficultés rencontrées par la reconnaissance automatique de la parole qui se trouve dans une impasse. Lorsqu'au début des années 50 les électroniciens, puis les informaticiens vont s'attaquer au problème de la reconnaissance, la plupart d'entre eux ne se doutent vraisemblablement pas qu'ils engagent un pari dont on ne sait toujours pas actuellement comment il pourra être tenu. A la fin des années 60, à partir des résultats obtenus par une cinquantaine de systèmes de reconnaissance, un premier bilan peut être dressé : il est possible d'obtenir un taux de reconnaissance de 95% pour des mots isolés, d'un vocabulaire limité (par exemple les chiffres), enregistré dans de bonnes conditions, avec un petit nombre de locuteurs. Avec un seul locuteur les résultats sont meilleurs, mais ils chutent notablement si le nombre de locuteurs est augmenté et terriblement avec un élargissement de la taille du vocabulaire. Une question fondamentale reste posée : comment dépasser cette première étape ? Comment passer de la reconnaissance de mots isolés à la parole continue ? Ce qui implique à la fois la résolution de problèmes acoustiques complexes, l'adoption d'un "modèle de langage" et le passage du stade de la reconnaissance à celui de la "compréhension".

En 1969, un chercheur de la BELL va faire une sérieuse mise en garde (elle est d'ailleurs intéressante à rapprocher de l'alarme sonnée 10 ans plus tôt par Y. BAR-HILLEL pour la traduction automatique). Dans "the most popular letter to the editor that was ever published in the Journal of the Acoustical Society of America" (W.A. LEA, 1979) J. PIERCE remet en cause, fondamentalement, la Reconnaissance Automatique de la Parole. Son argumentation repose sur le fait que la communication fonctionne entre locuteur et auditeur parce que tous deux ont en commun , non seulement la connaissance du langage, mais aussi l'intelligence de la situation. Compte tenu de l'état des connaissances théoriques, on ne peut être conduit qu'à croire que " a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English".

---

* On peut être étonné de voir avancées, 30 ans plus tard, des analyses sans nuances, ni justifications : "L'erreur fondamentale du phonéticien est de considérer les sons du langage comme des réalités physiques quelconques, que l'on pourrait observer et classer à la façon dont un botaniste classe les plantes. Allons même plus loin : le langage en tant que tel est indépendant de l'usage du son articulé et la possibilité de traduire terme à terme le langage parlé en langage écrit le prouve bien" (E. BOLTANSKI, présentation de "Linguistique" de E. SAPIR, ed. de Minuit, 1968).

** Pour ce qui concerne l'aspect perceptif on pourra se reporter aux travaux de L.A. CHISTOVITCH (1979) et son équipe, et de J.L. SCHWARTZ & P. ESCUDIER (1986).

*** En fait R. JAKOBSON a commencé dès 1939 à remettre en cause et à dépasser les thèses phonologiques du Cercle Linguistique de Prague. En 1949, il a déjà présenté un système de traits acoustiques binaires pour le français (R. JAKOBSON et J. LOTZ, 1949).

* Cette démarche qui consiste à vouloir organiser, décrire très précisément la matière, alors que son importance est minimisée, voire évacuée, quand il est difficile d'imaginer pouvoir la maîtriser, va être aussi, 16 ans plus tard, celle de N. CHOMSKY et M. HALLE. La phonologie générative, système abstrait s'il en est, va décrire la réalisation des traits en termes articulatoires spécifiques très précis dans "Sound Pattern of English" (1968).

Suivent des considérations quelque peu provocantes et fort peu amènes pour les reconnaisseurs ("mad inventors or untrustworthy engineers") et ceux qui les financent.

L'intervention de J. PIERCE, qui a le mérite de poser un problème fondamental, se situe à une étape charnière. Parmi les réponses que va susciter cette "lettre", celle de W.A. LEA (1970) nous semble importante. Après avoir noté qu'il est surprenant que "so little careful research has been devoted to evaluating the past, present, and future of speech recognition or speech communication with computers" W.A. LEA reconnaît qu'il y a très peu de chances d'aboutir à la reconnaissance d' "arbitrary sentences spoken by arbitrary speakers in arbitrary environments"; en conséquence, il propose de construire des systèmes de reconnaissance avec de la parole "limitée"*. Il faut noter ici, qu'en France, J.P. TUBACH avait déjà présenté, la même année, un système de reconnaissance, dans le cadre d'une tâche particulière, avec des règles morphologiques, syntaxiques et sémantiques.

En 1971 la preuve que la parole continue peut être reconnue n'a pas été faite, (les résultats de P.J. VICENS sont trop limités pour être extrapolables) ; aucune équipe ne s'est lancée dans un projet de reconnaissance d'un vocabulaire de 1000 mots et aucun système n'a été conçu pour intégrer, avec des structures de contrôle adaptées, des connaissances acoustiques, phonétiques, lexicales, prosodiques, syntaxiques, sémantiques ou pragmatiques. Aussi le programme ARPA SUR (Advanced Research Project Agency / Speech Understanding Research) lancé par le Département de la Défense des USA est ambitieux. Son but : accepter de la parole continue (avec un grand nombre de locuteurs coopératifs, dans de bonnes conditions d'enregistrement, avec un micro de bonne qualité), avec un vocabulaire de 1.000 mots, en utilisant une syntaxe artificielle, pour une tâche déterminée, et seulement "en plusieurs fois" le temps réel. De 1971 à 1976,15 millions de $ vont être utilisés pour le financement de ARPA SUR. Cinq laboratoires sont contractants, trois systèmes vont présenter leurs résultats. Un seul d'entre eux (HARPY de Carnegie-Mellon University) va atteindre des spécifications révisées : 184 phrases, 5 locuteurs, 1011 mots, 5% d'erreurs sémantiques (42 % de reconnaissance phonétique).

Si le projet ARPA est considéré, par certains, comme un demi-échec et, pour les plus optimistes, comme la démonstration que la reconnaissance automatique de la parole continue n'est pas un "unattained hope", il est surtout à l'origine de l'intégration massive des connaissances phonétiques et linguistiques, désormais considérées comme difficilement contournables, pour la mise en oeuvre de tout système de reconnaissance. Le projet ARPA confirme la prise de position formulée dès 1968 par S.R. HYDE et qui reste toujours d'actualité : "Significant advances in speech recognition are likely to come not from researches into signal analysis, adaptative pattern matching, or computer implementation (although these fields have valuable techniques to offer the speech researcher), but from studies of speech perception and generation, phonetics, linguistics and psychology."

En France, la Reconnaissance Automatique est très active. C'est l'époque de la réalisation de systèmes: MYRTILLE I (J.P. HATON et J.M. PIERREL, 1976) au CRIN-Nancy, KEAL (G. MERCIER et al., 1977) au CNET-Lannion, ESOPE (J. MARIANI et J.S. LIÉNARD, 1978) au LIMSI-Orsay, le système du LCP (B. GROC et D. TUFFELLI, 1980) à Grenoble, ARIAL I et II (G. PERENNOU et J. CAÉLEN, 1979) au CERFIA-Toulouse, MYRTILLE II (J.P. HATON et J.M. PIERREL, 1982) au CRIN-Nancy.

A partir de 1976, à la suite du rapport ARPA SUR, un certain nombre de certitudes ont commencé à s'installer dans la communauté scientifique internationale : le signal de parole n'est pas un objet physique comme les autres et les techniques classiques de traitement du signal, d'analyses de données, de reconnaissance des formes ne permettront pas de résoudre, seules, les problèmes.

Il faut obligatoirement prendre en compte les connaissances phonétiques, phonologiques, lexicales, syntaxiques, pragmatiques, voire sémantiques. Grâce à celles-ci, il est posible d'améliorer le taux de reconnaissance à la sortie du décodage acoustico-phonétique, mais la reconnaissance automatique bute à ce premier niveau de traitement. Une enquête effectuée aux USA auprès de 34 experts fait bien apparaître que les spécialistes de la reconnaisance placent le décodage acoustico- phonétique au premier rang de leurs préoccupations (W.A. LEA, 1979). Comme l'a fait fort judicieusement remarquer Mac KAY : "we cannot successfully mechanize what we do not fully understand". Nos connaissances phonétiques ne sont pas suffisantes : il existe pour le moment, un fossé , difficilement franchissable, entre le signal et sa représentation phonétique. L'accumulation des données n'est pas obligatoirement génératrice de connaissances, il faut connaître le système sous-jacent. Les apports de l'intelligence artificielle seront sûrement fructueux, mais seulement à partir des représentations phonétiques, voire orthographiques.

C'est à cette époque qu'en France le GRECO Communication Parlée met en place, entre autres actions, : le "Décodage Acoustico-Phonétique" et la "Base de données lexicale du français" (BDLEX) qui vont permettre de transfuser les connaissances phonétiques, phonologiques, morphologiques et lexicales dans le domaine de la reconnaissance.

En Grande Bretagne le Projet ALVEY débute en 1984, son financement va permettre un énorme investissement dans la recherche. C'est un phonéticien, J. LAVER, directeur du Centre "for Speech Technology" qui est nommé coordinateur d'un projet de reconnaissance.

En se fixant des contraintes : vocabulaire déterminé, nombre limité de locuteurs, apprentissage, structures syntaxiques formalisées, etc..., la reconnaissance automatique n'est pas tombée dans la catégorie des "problèmes mal posés". Elle a évité, certains diront de justesse, ce qu'une autre entreprise, la traduction automatique (qui a bien amorcé une reconversion en Traduction Assistée par Ordinateur) , n'a pas su faire : "History provides no better example of the improper use of computers than machine translation". Mais surtout elle a eu le mérite de mettre en évidence qu'il est difficile de vouloir décoder le produit d'un système si l'on ne possède pas la connaissance de son fonctionnement.

IV - LE TERRITOIRE "DE L'HOMME DE PAROLE"

Au terme de cette analyse il est possible de proposer une première interprétation : la Communication Parlée s'est constituée par une réorganisation de savoirs induite par une succession de ruptures et de crises affectant les domaines concernés. Notons, tout d'abord, qu'il n'apparaît pas de rupture due à l'émergence d'une nouvelle théorie qui partagerait ces champs en une histoire "périmée" et une histoire "sanctionnée" (G. BACHELARD). On pourrait considérer qu'il s'agit d'une succession de regroupement, segmentation , redéfinition et enfin réarticulation. Ainsi en 1932, à la veille du 1° Congrès des Sciences Phonétiques, on peut dire que celles-ci avaient réussi à focaliser, puis regrouper l'essentiel des recherches effectuées dans le domaine de la parole. La phonologie, ou plus exactement les phonologies (du Cercle Linguistique de Prague, du Cercle Linguistique de Copenhague, de l'école distributionnaliste, de l'école mentaliste) vont faire éclater ce bel ensemble, en fait fragile. A partir de là, la phonologie, les sciences phonétiques, l'analyse-synthèse (puis reconnaissance) vont se (re-)construire et se développer de façon plus ou moins indépendante.

En 1952, le projet des "Preliminaries", va provoquer un regroupement qui se parachève en 1976 avec l'intégration massive des connaissances phonétiques et linguistiques dans la reconnaissance automatique de la parole: la Communication Parlée s'est mise en place avec son théoricien/praticien "l'homme de parole".

Il n'est pas dans notre démarche de vouloir fixer les frontières de son territoire, mais plutôt d'en préciser la zone centrale. C'est un pentapode (figure), avec en son centre le code de la langue ; la cognition, la production, l'acoustique, la perception en constituent les repères cardinaux. La circulation entre ces aires est bien connue : par exemple les parcours code->acoustique, acoustique->code renvoyant à la synthèse et à la reconnaissance automatique de la parole.

"L'homme de parole" possède déjà une formation et des connaissances bien spécifiques qui comportent des aspects qu'il ne faut pas systématiquement renvoyer à une classification traditionnelle, sous peine de malentendu.

Les méthodes, les techniques, les outils qu'il utilise, les structures* dans lesquelles il travaille, les collaborations connexes**, s'organisent en fonction d'une finalité -l'étude de la parole- objet suffisamment spécifique dans la biologie des espèces.

Par rapport à tous les autres moyens de communications, la parole est le seul système pour lequel on possède une telle connaissance du code : le code linguistique au sens général du terme (oral ou écrit). C'est aussi le seul qui fonctionne, en temps réel, en interaction encodage/décodage. Peut être faut-il considérer l'étude de la Communication Parlée au sens très général de **l'analyse d'un système** telle que L. BERTALANFFY*** l'a définie.

La Communication Parlée n'est pas un champ théorique vide, il suffit de citer la "Théorie Quantique" de K.N. STEVENS " (1962) (pour une discussion générale, on pourra consulter le n° spécial du Journal of Phonetics, à paraître) ; la "Théorie Motrice de la Perception" (A.M. LIBERMAN , F.S. COOPER, K.S. HARRIS, P.F. Mc NEILAGE, 1962) ; la "Théorie de la Dispersion" de B. LINDBLOM et J. LILJENCRANTS (1972) : pour être distincts les sons doivent rester distants ; la "Théorie de la représentation interne à partir d'une intégration large bande" de J.L. SCHWARTZ et P. ESCUDIER (1986) , qui prolonge les travaux de P. DELATTRE (1952) sur les voyelles à deux formants, la première formulation du F'2 (G. FANT, 1959), et les travaux de L.A. CHISTOVITCH (1979) sur la notion de centre de gravité. Les théories développées dans le cadre de la Communication Parlée sont bien falsifiables, au sens de K. POPPER.

Que la Communication Parlée constitue une science, ceci ne veut pas dire qu'elle ne puise pas à d'autres sciences, bien au contraire . Comme mise à l'épreuve de théories venues d'ailleurs, citons, par exemple, la "Perception Catégorielle" (cf. B.H. REPP, 1982), la "Théorie de l'Action" (C. FOWLER, P. RUBIN, R.E. REMEZ, M. TURVEY, 1979) et récemment la "Théorie des Catastrophes" (de R. THOM à J. PETITOT-COCORDA, 1985). Mais comme les spécialistes de la vision ont des connaissances en neurophysiologie, en optique, en modélisation (des rétines, etc...), mais ne sont pas pour autant des opticiens, des mathématiciens car ils sont davantage des neurophysiologues ou psychologues, les "hommes de parole" doivent être davantage, et tout naturellement, des spécialistes du code phonétique et linguistique.

NOTE

Il est difficile de ne pas remarquer que la Communication Parlée est l'objet de nombreuses visées annexionnistes.

L'Intelligence Artificielle ne fait aucun mystère de ses ambitions globales : "Tout problème pour lequel aucune solution algorithmique n'est connue relève a priori de l'Inteligence Artificielle" (J.L. LAURIERE, 1987). Pour la parole, la tentative d'appropriation "descendante" (top-down") est manifeste. Dans la collection de "La Nouvelle Encyclopédie des Sciences et des Techniques", volume "Intelligence des mécanismes et mécanismes de l'intelligence", au chapitre "Intelligence Artificielle : Panorama des Techniques et des Domaines d'Application", la Communication Parlée figure, comme une simple application, à côté de la vision par ordinateur et de la commande des robots. Au glossaire de l'ouvrage on peut lire l'article suivant :

Reconnaissance de la Parole : discipline issue de l'I.A., dans les années 70, qui s'est vite scindée en plusieurs domaines; outre la synthèse de la parole (bien antérieure : cf. l'écouteur téléphonique), on distingue la reconnaissance du locuteur et la compréhension des énoncés. La reconnaissance du locuteur est un cas particulier important de la reconnaissance des formes, les programmes cherchant des correspondances entre les caractéristiques physiques des sons véhiculant la parole (fréquence, amplitude, timbre, etc.). En revanche, la compréhension des énoncés est un des domaines les plus difficiles des sciences de la cognition, nécessitant la compréhension de la compréhension ; qu'est-ce que comprendre l'énoncé d'un problème, ou le silence d'un partenaire ?

Définition d'autant plus étonnante que, dans le même ouvrage, J.L. LEMOIGNE reconnait que si, en France, la première thèse d'Etat en I.A. a été soutenue en 1966 (J. PITRAT), il a fallu attendre 10 ans pour voir présentée la seconde (J.L. LAURIERE).

Le Traitement du Signal est fortement tenté par une démarche "ascendante" (bottom-up) qui risque de se heurter, dès les premières tentatives de "segmentation" aux mêmes difficultés que le reconnaissance automatique de la parole dans les années 1970.

La Linguistique, nous l'avons montré, a toujours entretenu avec la parole des rapports complexes, évacuant son étude du champ des sciences humaines ou tentant purement et simplement de l'annexer. Il faut éviter toute appropriation abusive par les linguistes qui tentent de récupérer la parole tout en expliquant que sa description, aux niveaux articulatoire, acoustique et perceptif ne fait pas partie du système de communication. Par exemple C. HAGEGE persiste à considérer que "l'établissement, sur des bases articulatoires et acoustiques de catégories de sons tels que peut les produire l'appareil phonateur (des lèvres au larynx) et que l'oreille perçoit" relève des sciences de la nature, alors qu'étudier les classes de sons constitués dans la langue relève des sciences humaines ("L'homme de Paroles", Fayard, Paris, 1985; p. 101, pp. 130-131).

Les toutes récentes neuro-sciences, et notamment celles qui se reconnaissent dans l'approche connexioniste, s'appuient largement sur leurs capacités - potentielles ou supposées - à résoudre les problèmes de la parole (cf. NETTALK). Mais pour le moment, on ne peut pas dire que la preuve en ait été faite.

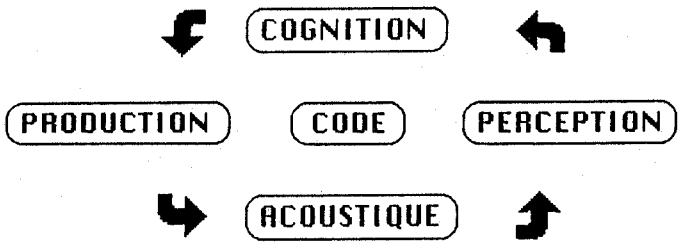
---

* La communauté parole s'est restructurée : sont apparus des chercheurs, souvent issus du secteur SPI, ayant une double formation universitaire ; les JEP, puis le GRECO n° 39, ont favorisé, provoqué des échanges ; des relations synergiques et/ou des regroupements se sont mis en place entre les secteurs SPI et SHS : GIA-IPA, CNET-IPG, LCP-IPG, CRIN-Lab. de Phonétique de Nancy, LIMSI-Lab. de Phonétique de Paris V et VI ; le recrutement essentiellement endogame s'est diversifié au CNET, au LIMSI, au LCP ; enfin le CNRS a créé une Unité Associée, l'ICP, relevant de deux secteurs (SPI et SHS).

** Les expériences accumulées par les laboratoires de recherche ont permis l'apparition et le développement d'entreprises qui travaillent en relation avec le monde de la recherche (FERMA, INFLUX, OROS, VECSYS, XCOM ...).

*** L'idée de construire une théorie générale, qui fournirait un formalisme de base pour l'étude des systèmes très divers, remonte aux premiers travaux de Von BERTALANFFY, vers 1925. En mettant l'accent sur l'aspect "organismique" des êtres vivants, c'est à dire sur la caractéristique systémique, ce chercheur attirait précisément l'attention sur la difficulté majeure que présente la reconstruction d'un tout à partir d'éléments constitutifs étudiés séparément.

Actuellement, en France, les chercheurs en Communication Parlée sont majoritairement rattachés à l'Informatique (dans le cadre du CNRS) , ce n'est d'ailleurs pas toujours le cas à l'étranger. L'essor de l'informatique a favorisé l'apparition de nouveaux champs de recherche, dont la communication Homme-Machine qui a drainé d'importants financements qui ont pu être consacrés à l'étude de la parole. Mais il faut se garder de penser, ou de laisser penser, que la Communication Parlée est un sous-ensemble de l'informatique. Comme nous avons essayé de le montrer, la Communication Parlée est un champ bien spécifique qui a de fortes résonances sociales, culturelles et politiques (au sens très général du terme). Il suffit pour s'en persuader de relire le Manifeste du Colloque International sur les Industries de la Langue (Tours, 28 février - 1 mars 1987).



Les cinq pôles de la Communication Parlée

REFERENCES

ARSAC J. (1987)
Les machines à penser. Seuil, Paris.
BECLARD J. (1870)
Traité élémentaire de physiologie humaine comprenant les principales notions de la physiologie comparée. Asselin, Paris, 6° ed.
BENZECRI J.P. (1982)
Histoire et préhistoire de l'analyse des données. Dunod, Paris.
BRETON Ph. (1987)
Histoire de l'informatique. La Découverte, Paris.
BOE L.J. (1985)
Ingénieurs et phonéticiens : la symbiose nécessaire. Machines parlantes 85, 1° Exposition Internationale Synthèse et Reconnaissance de la Parole. Le Carrefour International de la Communication, 28-29.
BOE L.J. (1986)
Rapport d'activité du Groupe Communication Parlée. Revue d'Acoustique 79, 4, 55-57.
BOUSSINOT R. (1980)
L'encyclopédie du cinéma. Bordas, Paris.
CASTEX A. (1894)
Hygiène de la voix parlée et chantée. Masson, Gauthier-Villars, Paris.
CHARBON P. (1981)
La machine parlante. Ed. J.P. Gyss.
DAHAN-DALMEDICO A. & PEIFFER J. (1986)
Une théorie des mathématiques. Seuil, Paris.
DELATTRE P. (1971)
Système, structure, fonction, évolution. Maloine, Paris.
De MEYER G.H. (1885)
Les organes de la parole et leur emploi pour la formation des sons du langage.
DESPLATS V. & GARIEL C.M. (1870)
Nouveaux éléments de physique médicale. F. Savy, Paris.
DREYFUS-GRAF J.A. (1967)
Spectres phonétographiques. Revue d'Acoustique 3-4.
DRION Ch & FERNET E. (1884)
Traité de physique élémentaire. 11e éd., Masson & Cie, Paris.
Du MONCEL Th. (1882)
Le microphone, le radiophone et le phonographe. Hachette, Paris.
FERRIEU G. (1986)
Paroles et Machines. Radome Revue d'information du CNET-Lannion 6, 5-9.
GANOT A. (1884)
Traité de physique élémentaire. 19e éd., Hachette, Paris.
GARRISON F.H. (1967)
An Introduction to the History of Medicine with Medical Chronology, Suggestions for Study and Bibliographic Data. Saunders, Phil. 4th ed.
GENTILLI A. (1882)
Der Glossograph. Leipzig.
LAURIERE J.L. (1986)
Intelligence Artificielle. Résolutions de problèmes par l'homme et la machine. Eyrolles, Paris.
LEA W.A. (1970)
Evaluating Speech-Recognition Work. J. Acoust. Soc. Am. 47,1612-1614 (L).
LEA W.A. (1979)
Review of the ARPA SUR Project and Survey of Current Technology in Speech Understanding. Final Report. Office Naval Research. Contract n° N00014-77-C-0570.
LEA W.A. & J.E. SHOUP (1980)
Trends in Speech Recognition. Prentice Hall, London.
LIENARD J.S. (1986)
Où en sont les machines parlantes ? Machines Parlantes 85, 1° Exposition Internationale Synthèse et Reconnaissance de la Parole. Le Carrefour International de la Communication 28-29.
MALMBERG B. (1968)
Les nouvelles tendances de la linguistique. PUF, Paris.
MALMBERG B. (1971)
Les domaines de la phonétique. PUF, Paris.
MOUNIN G. (1967)
Histoire de la linguistique. Des origines au XXe siècle. PUF, Paris.
MOUNIN G. (1972)
La linguistique du XXe siècle. PUF, Paris.
PIERCE J.R. (1969)
Wither Speech Recognition ? J. Acoust. Soc. Am. 46, 1049-1051 (L).
POP S. (1956)
Instituts de Phonétique et Archives Phonographiques. Publications de la Commission d'Enquête Linguistique, Louvain.
ROUSSELOT P.J. (1897-1901)
Principes de phonétique expérimentale. Tome I. Welter, Paris.
TSEMEL G.I. (1971)
Reconnaissance des signaux de parole. Ed. Science, Moscou [en Russe].
Van GINNEKEN J. (1932)
Allocution d'ouverture. [1rst] Int. Congr. Phonetic Sci. - In : Archives Néerlandaises de Phonétique Expérimentale 8-9, 8-17 (1933).
Von BERTALANFFY L. (1949)
Les problèmes de la vie. Essai sur la pensée biologique moderne.Trad. M. DEUTSH, Gallimard, Paris.
WUNDT W. (1884)
Traité élémentaire de physique médicale. Trad. par T. MONOYER, 2e ed. Baillère, Paris.

Remerciements

Ce travail n'a pu être mené que grâce à de multiples informations, discussions, critiques et encouragements. Nous remercions tout particulièrement J. LAMBERT, Philosophe et Historien des Sciences à l'Université Grenoble II, et C. ABRY sans qui nous n'aurions pu le mener à terme.

Un grand merci à C. BENOIT, J. CAELEN, R. CARRE, M. CARTIER, M. CONTINI, R. DESCOUT, G. FENG, Y. GRENIER, C. GUEGUEN, J.P. HATON, J.M. HOMBERT, G. KONOPCZYNSKI, M. MANTAKAS, L. MICLET, D. PASCAL, G. PERENNOU, P. PERRIER, H. RAKOTOFIRINGA, X. RODET, B. SCHNABEL, J.L. SCHWARTZ, P. SPECKER, G.S. SLUSTKER, B. TESTON, J.P. TUBACH, bien connus dans le domaines de la Communication Parlée et à S. BOE (lycée E. Mounier, Grenoble), H. HATWELL & J.P. ORLIAGUET (Lab. de Psychologie expérimentale, Grenoble II) E. WEIMANN (Direction de la Communication, Relations avec les Téléspectateurs, La CINQ).

REPERES CHRONOLOGIQUES

Cette première tentative dans le domaine de la Communication Parlée comporte des omissions et très certainement des inexactitudes. En l'établissant nous avons souhaité qu'elle suscite discussions, approfondissements et critiques dont nous tiendrons le plus grand compte. Un grand merci d'avance.

LA COMMUNICATION PARLEE

1781
*Le triangle vocalique articulatoire: C.F. HELLWAG (Allemagne).

1786
*Sir William Jones présente à la Société Asiatique de Calcutta les bases de la grammaire comparée qui vont permettre de mettre en évidence l'indo-européen.

1791
*Von KEMPELEN réalise une machine parlante.

1809
*Le triangle vocalique inspiré de l'acoustique : E. CHLADNI (Allemagne).

1835
*Synthétiseur mécanique du conduit vocal de J. FABER (Autriche).

1837
*1ère mesure de la pression sub-glottique pendant la phonation par CAGNIARD-LATOUR.

1853
*Modèle mécanique du larynx élaboré : M. HARLESS (Allemagne).

1855
*Invention du miroir laryngoscopique par M. GARCIA (Angleterre).

1856
*Manuel sur les fondements de la physiologie du langage: E.W. BRUCKE (Allemagne).

1857
*E.L. SCOTT invente le Phonautographe "écriveur de son" (il sera construit par R. KOENIG) et constate qu'il n'est pas évident de décoder le tracé du signal de parole !

1860
*Le Congrès de Milan consacre la "suprématie" de la parole lue sur les lèvres et articulée par les sourds-muets, opposée au langage des signes.

1863
*Acoustique et production des voyelles : par le médecin physiologue H.L.F. HELMHOLTZ (Allemagne).

1864
*Présentation par R. KOENIG au Conservatoire des Arts et Métiers d'un album d'enregistrements obtenus avec le phonautographe de E.L. SCOTT.

1874
*Enregistrements par F.C. DONDERS de H. SWEET qui servira de modèle à Bernard SHAW pour sa pièce Pygmalion (au cinéma : My Fair Lady de G. CUKOR, 1964).

1876
*Publication par SIEVERS des fondements de la physio-phonétique (Allemagne).

LES DOMAINES D'INFLUENCE

1793
*Rapport de l'abbé GREGOIRE à la Convention pour "L'universalisation du français et l'anéantisement des patois" : efficace mais terriblement réducteur !
*R. PRONY : calcul d'une fonction interpolant des mesures comme une somme d'exponentielles. Un précurseur du modèle ARMA déterministe.

1807
*Méthode des moindres carrés de C.F. GAUSS qui va être utilisée dans de multiples développements, dont le codage prédictif.

1811
* C.F. GAUSS définit les nombres imaginaires dans une lettre à F. BESSEL.

1812
*Transformation de P.S. LAPLACE : la base théorique du calcul symbolique.

1826
*N. NIEPCE invente la photographie

1827
*G.S. OHM publie sa célèbre loi.

1832
*Décomposition des signaux complexes : J.B.J. FOURIER.

1847
*Développement du kymographe par K. LUDWIG (Allemagne).
*La naissance de la logique symbolique mathématique, hors du champ de la philosophie, avec "Mathematical Analysis of Logic" de G. BOOLE.

1859
*Publication par Ch. DARWIN de "L'Origine des espèces". Cet ouvrage va avoir une forte influence sur les néo-grammairiens (la langue serait-elle un organisme vivant qui naît, se développe et décline ?), sur les biologistes et les statisticiens.
*A la suite des travaux de H.L.F. HELMHOLTZ le la$_3$ est fixé à 435 périodes par seconde.

1861
*P. BROCA localise dans le cerveau la zone correspondant à l'activité parole.
*G.R. KIRCHOFF invente l'analyse spectrale.

1868
*Le concept de "feed-back" analysé par J.C. MAXWELL.

1871
*Découverte de la loi périodique et construction d'un tableau des éléments : D.I. MENDELEIEV. Cette classification influencera N.S. TROUBETZKOY et R. JAKOBSON à la recherche d'éléments phoniques universels.
*Le dentiste J.O. COLES invente la palatographie directe.

1876
*Invention du téléphone par G. EELL.
*"Theory of Sound" : Lord RAYLEIGH.

1879
*Première utilisation de la palatographie en phonétique: N.W. KINGSLEY (USA).

1882
*Introduction de la phonétique à l'Université de Harvard dans le cadre de la philologie germanique (USA).

1883
*Reconnaissance automatique de la parole par enregistrement in situ du mouvement des articulateurs : A. GENTILLI (Allemagne).

1886
*1ers congrès de Phonétique à Paris et à Stockholm.
*Création de l'Association Internationale de Phonétique (API) et de sa revue : Le Maître Phonétique.

1889
*Première chaire de Phonétique à l'Institut Catholique de Paris. Elle va être occupée par l'Abbé P. ROUSSELOT.

1896
*Création d'un Laboratoire de Phonétique Expérimentale dans le cadre d'un Institut sur les troubles du langage à Copenhague.

1897
*Création à Paris par l'Abbé ROUSSELOT, à l'Alliance Française, d'un Centre de Correction d'Accents Etrangers.
*Création d'un laboratoire de Phonétique Expérimentale au Collège de France à Paris.

1897-1901
*Publication du tome I des Principes de Phonétique Expérimentale par l'Abbé Rousselot Le tome II paraîtra en 1908. Un monument !

1904
*Création du premier Institut de Phonétique Universitaire, à Grenoble (Directeur : Th. ROSSET).
*Création du Laboratoire de Phonétique de l'Université de Montpellier (Directeur : M. GRAMMONT jusqu'en 1936).

1906 - 1911
*Enseignement du cours de Linguistique Générale par F. DE SAUSSURE (Genève) : naissance du structuralisme en linguistique. (La première publication posthume du cours sera faite en 1916).

1911
*Création du Musée de la Parole à Paris.
*Création de l'Institut de Phonétique de Paris (Directeur: F. BRUNOT).

1914
*Publication de "Language" : L. BLOOMFIELD

1921
*Publication de "Langage" : E.SAPIR.

1922
*Analogue électrique schématique du conduit vocal : J.Q. STEWART (USA).

1923
*1ères statistiques importantes sur les sons, les syllabes et les mots de l'anglais : DEWEY et GODFREY.

1877
*Dépot du brevet du phonographe à cylindre à feuille d'étain par Ch. CROS et Th. EDISON que ce dernier réalisera l'année suivante.

1893
*1er oscillographe électrodynamique : A. BLONDEL.

1894
*Invention de l'oscillographe cathodique par K. BRAUN, il sera construit en 1897.

1898
*1ers enregistrements sur fil magnétique: V. POULSEN.

1900
*Théorie des quanta : M. PLANCK. Le terme sera repris pour l'étude des systèmes phonétiques par K.N. STEVENS en 1962.

1904
*Invention de la diode par J.A. FLEMING.

1906
*1ère émission radiophonique par TSF par R.A. FESSNDEN.

1907
*Brevet de la triode par Lee de FOREST : à la base de tout amplificateur.

1910-1911
*"Principia Mathematica" : 3 volumes de B. RUSSELL et A.N. WHITEHEAD. Les bases de la logique symbolique.

1912
*Naissance de la GESTALT théorie avec les travaux convergents de W. KöHLER et de M. WERTHEIMER, K. KOFFKA, K.LEVIN, L. METZGER : la réorganisation du champ perceptif dans le sens des formes les meilleures (Univ. Berlin).

1913
*Paris compte 92.000 appareils téléphoniques et New York 500.000.
*"Psychology as the Behaviorist Views it" de J.B. WATSON: la rupture behavioriste en psychologie de l'étude du comportement. Voir en linguistique l'école behavioriste, avec L. BLOOMFIELD.
*En examinant les 20.000 premières lettres du 1er chapitre et les 16 premiers sonnets du second chapitre d'Eugène Onéguine de Pouchkine, A.A. MARKOV développe la "théorie des chaînes", modèle stochastique du langage.

1919
*Equations de W.H. WEBSTER qui, reprises trente ans plus tard dans le domaine de la parole, permettront d'asseoir la théorie de la production.
*Fonctionnement de la première station de radiodiffusion aux USA.

1921
*Création de Radio France.

1925
*"Sound Patterns in Language" de E. SAPIR : des idées essentielles sur la phonologie avant les travaux du Cercle Linguistique de Prague.
1926
*Création du Cercle Linguistique de Prague CLP au sein duquel va se développer la phonologie, avec l'arrivée, en 1928, de N.S. TROUBETZKOY et de R. JAKOBSON.

1928
*1er Congrès International des Linguistes, La Haye. Présentation par R. JAKOBSON des thèses de la Phonologie.

1930
*1ère Réunion Phonologique Internationale animée par les représentants du Cercle Linguistique de Prague.
*1er film radiographique pour l'étude de la parole : H. GUNTZMAN et V. GOTTHEINER.

1932
*1er Congrès International des Sciences Phonétiques (La Haye). La phonétique ne se reconnaît plus comme une et indivisible mais plurielle. Congrès couplé avec la 2ème Réunion Phonologique Internationale.
*E. ESCLANGON : réalisation de l'horloge parlante.

1936
*Le vocodeur : H. DUDLEY. Premier codeur de la parole exploitant ses propriétés de redondance.
1937
*1er détecteur électronique du fondamental : GRÜTZMACHER et W. LOTTERMOSER.

1939
*Synthèse de la parole avec le VODER (Voice DEmonstratoR) : à partir d'une douzaine de touches l'opérateur (formé à raison de 3h par jour pendant 1 an) pouvait modeler un spectre instantané et la hauteur de la voix (Bell).
*Publication posthume des Principes de Phonologie de N.S. TROUBETZKOY.
1939-1944
*Publication par L. KAISER d'un travail fondamental sur les caractéristiques individuelles de la voix.
1940
*Film à grande vitesse des cordes vocales : R.W. FARNSWORTH (USA).
1941
*Etude et modélisation de la production. Les prémices de la théorie acoustique de la production : T. CHIBA et M. KAJIYAMA (Japon).
*"Langage enfantin et aphasie" : R. JAKOBSON.

1943
*URSS : premiers travaux sur la reconnaissance automatique de la parole : énergie dans 14 filtres, 75-80% de reconnaissance des voyelles (L.L. MJASNIKOV).

1946
*Invention du Sonagraphe : W. KOENIG, H.K. DUNN, L.Y. LACEY. (Bell)
1947
*Création de l'Institut de Phonétique de Strasbourg (Dir. G. STRAKA).
*Visible Speech : l'encyclopédie icono-sonagraphique : R.K. POTTER G.A. KOPP et H.C. GREEN.
1948
*Formants, voyelles et triangle vocalique du français : P. DELATTRE.

1925
*1er électrophone : gramophone avec amplificateur.

1927
*"The Jazz Singer" : 1er film parlant (A. CROSLAND).

1929
*Travaux de G.K. ZIPF, un précurseur en analyse statistique : la loi rang-fréquence.
1930
*Les principes de l'analyse multifactorielle : L. THURSTONE.
1931
*La limitation des méthodes formelles : "Toutes les formulations axiomatiques consistantes de la théorie des nombres incluent des propositions indécidables" K. GÖDEL.
*Création du Cercle Linguistique de Copenhague et développement de la glossématique (L. HJELMSLEV).

1933
*1er enregistrement sur bande magnétique par les laboratoires AEG en Allemagne.
*Courbes du seuil de l'audition : H. FLETCHER et W.A. MUNSON.
*Les quadripoles, application du calcul matriciel à l'étude des circuits : R. FELDTKELLER.
*Mise au point de l'iconoscope, première caméra de TV électronique, par V. ZWORYKIN

1936
*Conception par A.M. TURING de "machines" abstraites qui constituent la base de la théorie des automates et d'une manière générale de la théorie de la "calculabilité".
1937
*1er système téléphonique français à 12 voies avec des fréquences porteuses espacées de 4 kHz.
1938
*Naissance officielle du structuralisme linguistique : L. HJELMSLEV et V. BRÖNDAL.

1945
*ENIAC : Dernier grand calculateur électronique prédécesseur de l'ordinateur.
*Naissance de l'informatique aux USA et en Angleterre.
1946
*Développement du filtre de N. WIENER : N. LEVINSON.
1947
*Mise au point du transistor : J. BAARDEN, W. BRATTIN, et W. SHOCKLEY (Bell Telephone Co).
1948
*Fonctionnement du premier ordinateur, le MARK1, en Grande-Bretagne.
*Théorie de l'information par C. SHANNON : elle va permettre, entre autres, d'évaluer la quantité d'informations véhiculée par la parole et sa redondance.
*N. WIENER "Cybernetics or Control and Communication in the Animal and Machine" : les grandes idées de la cybernétique.
*P.M. MORSE : "Sound and Vibration".
1949
*Publication par D.O. HEBB de "The organization of Behaviour" contenant la première proposition d'apprentissage synaptique.
*1er usage spécifique du terme prédiction linéaire : N. WIENER.

1950
*Travaux sur la reconnaissance automatique de la parole par le suisse J. DREYFUS-GRAF (qui travaillera par la suite au CNET, Lannion).
*Analogue du conduit vocal : H.K. DUNN.
*Utisation de l'EMG dans la parole : R.H. STETSON.
1951
*G.A. MILLER "Language and Communication" : les bases de la psychophonétique et de la psycholinguistique, en rupture avec le behaviorisme.
1952
*"Preliminaries to Speech Analysis" : R. JAKOBSON, G. FANT et M. HALLE.
1953
*Synthétiseur à formants, le PAT d'Edimbourg de W. LAWRENCE.
*Théorie myoélastique de la vibration des cordes vocales : J.W. Van Den BERG.
*En montrant qu'il n'est pas possible de faire de la synthèse de la parole à partir d'unités acoustiques correspondant aux phonèmes, C.M. HARRIS apporte la preuve de l'impossibilité de découper la parole en segments, en "perles enfilées sur l'axe du temps" ou plus prosaïquement en" tranches de saucisson".
*"Speech and Hearing in Communication" : la bible par un ancien directeur de la Bell, H. FLETCHER.
*Analyse perceptive des confusions entre consonnes : G.A MILLER et E. NICELY (USA).
*1ère utilisation de contraintes phonétiques en reconnaissance automatique de la parole (probabilités cumulées des occurrences successives de deux sons) : D.B. FRY & P. DENES.

1956
*1er système de reconnaissance utilisant les traits de R. JAKOBSON, G. FANT et M. HALLE : J. WIREN et H.L. STUBBS.
1957
*Synthèse à partir d'une relecture de sonagramme et établissement de règles : J.M. BORST, F.S. COOPER, A.M. LIBERMAN et P. DELATTRE.
*Invention du glottographe : P. FABRE.
1958
*La définition du diphone et son utilisation en synthèse: G.E. PETERSON, W.S.Y. WANG et E. SIVERSTON.
*1er analogue dynamique du conduit vocal : G. ROSEN.
1959
*Synthèse par règles : A.M. LIBERMAN, F. INGEMAN, L. LISKER, P. DELATTRE, F.S. COOPER.
*1er système de reconnaissance à utiliser l'informatique: J.W. FORGIE et C.D. FORGIE.
*1ère formulation du F'2 : G. FANT
1960
*Théorie acoustique de la production de la parole par G. FANT.
1961
*Méthodologie de la radiocinématographie de la parole : P. SIMON (Institut de Phonétique de Strasbourg).
*G. FAURE lance, en France (Aix), les travaux sur l'intonation évacuée jusque-là par la linguistique dominante.
*Très large utilisation des traits distinctifs en reconnaissance automatique de la parole : G.W. HUGHES.
1962
*Théorie Quantique de K.N. STEVENS : une hypothèse pour l'organisation des systèmes phonétiques.
*Le Voice Onset Time, première structure coordinative : A. ABRAMSON et L. LISKER.
*Modèle de simulation temporelle du conduit vocal : J.L. KELLY et C. LOCHBAUM.
*CNET : débuts des travaux sur la compression de la parole (G. FERRIEU et G. BOUCHEZ).
1963
*1ère utilisation du codage vectoriel pour la transmission de données : C.P. SMITH.
1964
* Utilisation du cepstre dans le traitement de la parole: A.M. NOLL.

1950
*Publication par A. TURING de "Computing Machinery and Intelligence", l'article auquel vont pratiquement se référer tous les travaux en Intelligence Artificielle.
*Conception du tube TV couleurs.
1951
*Théorie générale et logique des automates : J. Von NEUMANN.

1956
*IPL : premier langage de traitement de listes, prédécesseur du LISP (A. NEWELL, J.C. SHAW, H. SIMON)
*J. Mc CARTHY : LISP un langage formel (fondé sur le λ calcul) utilisant systématiquement la récursivité.
1957
*Programmation dynamique : R. BELLMAN.
*A.N. CHOMSKY "Syntactic Structure" : l'ouvrage de base de la grammaire générative.
*Mise au point du FORTRAN : J. BACKUS (IBM).
1958
*Le "perceptron" de F. ROSENBLATT : le premier réseau neuromimétique pour l'apprentissage de formes visuelles.
*Les bases de l'analyse multidimensionnelle : W. TORGERSON.
*J.S. KILBY réalise le prototype du 1er circuit intégré (Texas Instrument).
1959
*Filtrage AutoRégressif sur des bases statistiques : J. DURBIN.
*Rapport de Y. BAR-HILLEL sur la Traduction Automatique et "A Demonstration of the Nonfeasability of Fully Automatic High Quality Translation" : la traduction automatique est limitée, faute d'une connaissance encyclopédique de l'univers non-linguistique.
*Travaux de Von BEKESY sur l'oreille.

1961
*1er Congrès "Calcul et Traitement de l'information" (Grenoble).
1962
*Version informatisée de l'analyse multidimensionnelle : R.N. SHEPPARD.
*1er article en français présentant l'Intelligence Artificielle (J. PITRAT).
*Création du mot "informatique" par J.A. DREYFUS.

1963
*L'Analyse des Correspondances de J.P. BENZECRI : une méthode puissante d'analyse de données permettant de croiser individus et caractéristiques.
*1er mini-ordinateur le PDP 8 de Digital : une révolution pour les utilisateurs.
1964
*Version MDSCAL par J.B. KRUSKAL : traitement multidimensionnel des dissimilarités.
*"Le geste et la parole" de LEROI-GOURHAN : une explication des relations geste technique et parole : prolongement des acquis de la biomécanique grâce à une vaste connaissance comparative préhistorique et ethnologique.

1965
*Création du département Etudes et Techniques d'Acoustique au CNET dans lequel vont se développer les travaux sur l'analyse et la synthèse de la parole.
*"Speech Analysis Synthesis and Perception" de J.L. FLANAGAN : une encyclopédie, après celle de G. FANT(1960).
*Publication de l'Album Phonétique de G. STRAKA : l'articulatoire et la tradition de la phonétique historique.
*Traduction anglaise de l'ouvrage de V.A. KOZHEVNIKOV, L.A. CHISTOVICH et al., "Articulation and Perception" (URSS).

1966
*ASPIC , vocodeur du CNET : G. FERRIEU.
*1ère étude en reconnaissance du CNET-Lannion : vocodeur+calculateur RAMSES+analyse discriminante linéaire (M. J. VINCENT CARREFOUR, M. PONCIN, M. ROUX).
*Codage prédictif : S. SAITO et F. ITAKURA.

1967
*Colloque du GALF organisé à Grenoble : "Les structures acoustiques de la parole" : la générale des premières JEP.
*1ère conférence ICASSP.
*Radiographie des consonnes du français : P. SIMON.

1968
*Modèle articulatoire : C.M. COKER (USA).
*Programmation dynamique pour la reconnaissance de la parole : G.S. SLUTSKER et T.K. VINTSJUK (URSS).
*Synthèse de textes à partir de diphones : E. LEIPP, M. CASTELLENGO, J.S. LIENARD.
*"The Sound Pattern of English" : les principes de la phonologie générative de N. CHOMSKY et M. HALLE.
*Création de l'Institut de Phonétique d'Aix en Provence (Dir. : G. FAURE).

1969
*Statistiques phonétiques sur le français (sur plus de 100.000 sons) : J.P. TUBACH.
*"Lettre à l'éditeur" de J. PIERCE, mettant en cause la reconnaissance automatique.
*95% de reconnaissance, plusieurs locuteurs, les chiffres : les performances d'un système de reconnaissance normalement constitué.
*1ers essais de reconnaissance de la parole continue (16 mots) : P.J. VICENS.

1970
*Constitution du Groupe Communication Parlée du GALF et 1ères Journées d'Etude (R. CARRE).
*Radiographie des voyelles du français : C. BRICHLER-LABAEYE.
*1ère thèse sur la reconnaissance automatique de la parole : J.P. TUBACH, (IMAG Grenoble).
*Salutaire protestation des chercheurs contre l'utilisation abusive des "empreintes vocales" par la justice aux USA : R.H. BOLT, F.S. COOPER, E.E. DAVID, P.B. DENES, J.M. PICKET et K.N. STEVENS.

1971
*Test de rime pour le français : M. CARTIER, J. PECKELS et M. ROSSI.
*1ère thèse de phonétique consacrée à la synthèse à partir du texte : J. VAISSIERE (Université de Grenoble III et IBM La Gaude).

1972
*Utilisation de l'analyse multidimensionnelle des similarités/confusions : R.N. SHEPPARD.
*1ère informatisation d'un Institut de Phonétique en France : Aix (avec un T1600)

1973
*Modèle de larynx à deux masses : K. ISHISAKA et J.L. FLANAGAN.
*1ères fonctions d'aire pour les voyelles du français : L.J. BOE.
*Premiers systèmes de reconnaissance automatique de mots commercialisés aux USA.
*Utilisation de la programmation dynamique en France : J.Y. GRESSER, G. MERCIER et J.P. HATON.
*1ère thèse de reconnaissance de parole utilisant en prétraitement une modélisation cochléaire : P. ALINAT.
*1ère thèse sur le codage prédictif de la parole : G. CARAYANNIS (ENST).

1965
*Algorithme FFT transformée de Fourier rapide: J.W. COOLEY et J.W. TUCKEY.
*Dans un épisode de la série STAR TREK, à la TV, les humains utilisent le paradoxe d'Epaminondas pour se débarasser de robots qui, décidément, ne connaissent pas le théorème de Gödel (J. RODDENBERRY).

1966
*1ère thèse en Intelligence Articielle, en France (J. PITRAT). Il faudra attendre 10 ans pour que soit soutenue la seconde (J.L. LAURIERE).

1967
*Thèse de J.C. RISSET en acoustique musicale: numérisation du signal, traitement et synthèse. L'informatique musicale et la parole vont se développer en entretenant des rapports pluridisciplinaires.
*Algorithme de A.J. VITERBI, utilisé pour le décodage.

1968
*"Introduction à la grammaire générative" de N. RUWET : la diffusion en France des travaux de N. CHOMSKY.
*HAL, l'ordinateur de bord de "2001 Odyssée de l'espace", qui comprend la parole et lit sur les lèvres, donne de la logique à retordre à ses interlocuteurs (S. KUBRICK).
*"French Phonology and Morphology" : S.A. SCHANE.
*Normalisation du MIC (PCM : Pulse Code Modulation) à 64 kbits/s.

1969
*1er Congrès International d'Intelligence Artificielle.
*MUSIC V : M.V. MATHEWS, F.R. MOORE, P. PIERCE et J.C. RISSET (Bell).

1970
*INDSCAL par J.D. CARROLL et J.J. CHANG.
*Naissance du PASCAL (dans la lignée de EULER, PL/60, ALGOL W), à Zurich, dans une équipe dirigée par N. WIRTH.
*1ère mémoire ROM intégrée (256 bits) : Sté FAIRCHILD et 1er RAM (1k) : Sté INTEL.

1971
*1er microprocesseur (Intel 4004).
*PROLOG, langage de programmation : l'unité n'est pas l'instruction mais un théorème de logique du 1er ordre (COLMERAUER, Luminy).

1972
*Pénétration dans le domaine public de la théorie des catastrophes élaborée par R. THOM pour décrire la naissance et l'évolution des formes, avec : " Stabilité structurelle et morphogénèse".
*1ère thèse (en France) en traitement de signal sur le codage prédictif : J. MENEZ (Univ. de Nice).
*Les débuts de l'IRCAM.
*Développement du micro-ordinateur et du floppy disque.
*Modèle de MARKOV caché : L.E. BAUM.
*1er système expert : DENDRAL (Université de Stanford).

1974
*Les fonctions de sensibilité formantiques du conduit vocal : G. FANT et S. PAULI.
*Classification et réduction des données vocales par analyse des correspondances : P. GRAILLOT (CNET).
*Segmentation automatique de la parole : J.S. LIENARD et M. MLOUKA (LIMSI).
*Modèle mathématique de la cochlée : J. CAELEN.

1975
*Le polyphonomètre : la mesure simultanée de pression et de débit lors de la production de la parole, B.TESTON.
*1ère thèse sur l'utilisation du codage prédictif pour l'analyse articulatoire : I. EL-MALAWANY (CNET).
*1er système de compréhension en ligne de la parole continue : MYRTILLE I (CRIN, Nancy).

1976
*Fin du projet ARPA (il avait débuté en 1972).
*Projet HEARSAY II de D.R. REDDY : système de reconnaissance le plus élaboré au niveau syntaxique et prenant en compte des techniques d'Intelligence Artificielle.
*Utilisation de la modélisation markovienne : F. JELINEK (IBM).
*J.D. MARKEL & A.H. GRAY :"Linear Prediction of Speech" ouvrage de référence du codage prédictif.

1977
*Démonstration de synthèse à partir du texte par liaison téléphonique Grenoble-Lannion : F. EMERARD et J. GENIN.
*Analyse et structuration du voisement du trait aux corrélats en passant par les indices et les propriétés : C. ABRY et L.J. BOE.
*Symposium International "Modèles Articulatoires et Phonétique" (R. CARRE, R. DESCOUTet M. WASJKOP).
*Implantation du modèle de J.L. KELLY et C. LOCHBAUM : R. ESPESSER (Institut de Phonétique d'Aix).
*1ère utilisation de INDSCAL pour les voyelles du français : R. BECKMANS (Bruxelles).

1979
*Création au CNET-Lannion de la Division Traitement du Signal de Parole et Services TSS avec 3 départements.
*1ère thèse française sur la reconnaissance automatique utilisant explicitement les techniques d'Intelligence Artificielle pour le traitement du parallélisme : P. SPECKER, Grenoble.
*Modélistion articulatoire du conduit vocal intégrant connaissances physiologiques et phonétiques : S. MAEDA.

1980
*Création du GRECO Communication Parlée associant secteur SPI et SHS (Dir. : J. P. HATON).
*Modèle d'oreille intégrant les bases physiologiques : J.M. DOLMAZON.
*Séminaire GRECO : "Processus d'encodage et de décodage phonétiques" (C. ABRY, J. CAELEN, J.S. LIENARD, G. PERENNOU, M. ROSSI). Structuration et intégration des propriétés, indices et traits dans la reconnaissance automatique de la parole.
*Reconnaissance de mots en chaîne avec une seule carte LIMSI. Première carte de reconnaissance réalisée en France (VECSYS-LIMSI). Le CNET et XCOM réaliseront SERAPHINE en 1983.
*1er système interpréteur pour la phonétisation du français mis au point avec les utilisateurs et pour les utilisateurs : M. LETY. Redéfini (TOPH) il servira à phonétiser BDLEX.

1981
*1ères Rencontres Franco-Soviétiques.

1982
*Création de la revue "Speech Communication" par M. WAJSKOP ancien président du GCP du GALF.

1983
*Création de l'Institut de la Communication Parlée : laboratoire interuniversitaire et UA CNRS regroupant deux laboratoires du secteur SHS et SPI (Dir. R. CARRE).

1985
*Application de la théorie des catastrophes et l'étude des systèmes vocalique et consonantique: J. PETITOT-COCORDA.
*1er Symposium Franco-Suédois.
*Machines parlantes 85 - 1ère Exposition Internationale synthèse et reconnaissance de la parole en France.
* Fin de l'enregistrement de BDSON (R. DESCOUT, CNET).

1986
*BDLEX la Base de Donnée Lexicale du GRECO opérationnelle (G. PERENNOU).

1974
*Développement du BASIC

1975
*Commercialisation du 1er micro-ordinateur (MARK8) : vente par correspondance par J. TITUS.

1976
*Développement des systèmes experts aux USA : MYCIN, le mieux testé, n'a pas été utilisé en clinique et a surtout servi de modèle.
*1er Congrès ICASSP.

1977
*Les deux Steve JOBS et WOZNIAK livrent leur 1er Apple II

1978
*Création du mot "télématique" par A. MINC et S. NORA.

1979
*1ères publications des "Annals of History of Computing".
*Invention du disque optique numérique, le CD-audio, par la Société Philips, il sera commercialisé en 1982-1983.

1981
*IBM se lance dans l'informatique personnelle avec le PC.
*Traduction française de l'ouvrage de E. ZWICKER et R. FELDTKELLER :"Psychoacoustique : l'oreille récepteur d'information" par C. SORIN.

1984
*F. DELL : "Les règles et les sons. Introduction à la phonologie générative".
*A l'occasion des 3° journées de l'ATALA une rencontre phonétique - phonologie (org. G. BOULAKIA).
*MICDA : 1er normalisation internationale du codage adaptatif (32 kbits).
*N. CATACH : "Les listes orthographiques de base".
*"Vocabulaire du roman français (1962-1968) Dictionnaire des fréquences" de G. ENGWALL : 500.000 mots.
*Le Macintosh d'apple.

Perception

CONTRIBUTIONS RELATIVES DES INDICES ACOUSTIQUES ET DES FACTEURS CONTEXTUELS A LA PERCEPTION DU TRAIT DE VOISEMENT DES OCCLUSIVES DU FRANCAIS DANS LA PAROLE SPONTANEE

SAERENS M., SERNICLAES W., BEECKMANS R.

Université Libre de Bruxelles. Institut de Phonétique.
50 av. F. Roosevelt, CP 110, 1050 Bruxelles

ABSTRACT

Our purpose here is to specify the relative contributions of bottom-up and top-down processing to the perception of the voicing feature of French stop consonants. Earlier studies have shown that voice onset time and other voice timing cues play a major role in bottom-up processing - at least as far as isolated syllables are concerned. The two first experiments in our study confirm the decisive importance of voice timing cues for stops excised (i.e. isolated) from spontaneous speech, but the experiments also show that voice timing can be misleading. Thus, in some instances, stops which in context belong to the voiceless category are identified as voiced when the context is removed. The question, then, is to know whether the voicing information is provided by possible acoustic cues, which were suppressed in the second experiment, or by top-down processing. The third experiment shows that, in cases where voice timing is misleading, correct identification of voicing depends on correct understanding of the surrounding words.

Our results as a whole suggest that the available top-down information is always used, either to confirm the information conveyed by the major cue (voice timing) or, in case of contradiction, to restructure the perceptual integration in favour of the alternative cues.

1. Introduction

Il est admis que la compréhension de la parole nécessite l'intégration de deux grandes classes de processus : ascendants (indices acoustiques) d'une part, descendants (issus du contexte) d'autre part(Garnes & Bond, 1976; Ganong, 1980; Connine & Clifton, 1987). L'étude des processus ascendants, la prise en compte des différents indices, pour le trait de voisement des occlusives en français a montré que les relations temporelles entre les vibrations laryngées (la voix) et la détente de l'occlusion orale, ce que nous appellerons le "timing laryngé", jouent un rôle déterminant dans la perception du trait de voisement des occlusives du français (Serniclaes, 1987). En position prévocalique, l'effet perceptif du VOT ("Voice Onset Time"), qui correspond à l'intervalle de temps compris entre la détente de l'occlusion et le départ de la voix (Lisker & Abramson, 1964), l'emporte sur celui de l'ensemble des autres indices. Ce n'est que lorsque le VOT est ambigu que l'identification du voisement dépend de divers autres facteurs acoustiques, notamment la fréquence fondamentale de la voix (F_0) et la durée des transitions formantiques. Mais ces indices n'affectent pas la perception du trait lorsque le VOT atteint des durées comparables à celles qui caractérisent les occlusives voisées ou sourdes dans la parole naturelle. Les résultats obtenus pour la position intervocalique amènent des conclusions semblables. L'occlusive est perçue comme voisée tant qu' il n'y a pas d'arrêt de voix au voisinage de la détente orale et elle est perçue comme sourde dès que l'arrêt de voix atteint une durée comparable à celle qui caractérise les occlusives sourdes dans la production de la parole (Serniclaes, 1975).

Le poids perceptif du timing laryngé s'accorde avec sa fiabilité en tant que critère séparateur entre les occlusives sourdes (ptk) et voisées (bdg) sur la base de mesures acoustiques. Les mesures effectuées sur un large corpus d'occlusives intervocaliques insérées dans des phrases suggèrent que les [bdg] s'accompagnent toujours de vibrations laryngées ininterrompues depuis le début de l'occlusion jusqu'à la fin de la transition d'ouverture vocalique. Quant aux [ptk], ils présentent un arrêt de voix au voisinage de la détente de l'occlusion dont la durée est au moins de 35 ms dans plus de 99% des cas (Serniclaes, 1984). On ne dispose pas de mesures du VOT pour les occlusives prévocaliques enchâssées dans des phrases. D'après les mesures effectuées sur des logatomes et des mots isolés, les [ptk] prévocaliques du français se caractérisent toujours par un délai entre la détente et le départ de la voix (un VOT positif) tandis que pour les [bdg], le départ de la voix intervient avant la détente de l'occlusion (le VOT est négatif) (Caramazza & Yeni-Komshian, 1974; Serniclaes, 1979). Bien que l'on doive s'attendre à des cas de dévoisement de l'occlusion des [bdg] en position prévocalique accentuée (Durand, 1956), ceux-ci sont, d'après des observations informelles, relativement peu fréquents.

L'objectif de notre travail est de confirmer la fiabilité des relations temporelles entre la voix et la détente dans le cas d'occlusives prononcées spontanément au cours d'une conversation. Ceci fera l'objet d'une première expérience. Au cas où ces indices seraient pris en défaut, on peut se demander si la perception du trait de voisement fait alors appel à d'autres indices acoustiques ou, au contraire, à des processus de niveau supérieur (règles phonologiques, informations lexicale, sémantique, etc). L'examen de cette question a fait l'objet d'une seconde et d'une troisième expérience.

2. Expérience I

2.1. Introduction

Lors d'un processus d'identification du voisement basé exclusivement sur la présence/absence de vibrations laryngées au voisinage de la détente de l'occlusion, les risques d'erreur peuvent provenir soit de la persistance des vibrations durant l'occlusion des [ptk] en position médiane, soit de l'absence de vibrations durant l'occlusion des [bdg] en position initiale. L'examen d'échantillons de parole prélevés dans des conditions optimales suggère que de tels risques sont très limités (cf. supra); cependant, le fait de demander aux locuteurs de prononcer un ensemble de logatomes ou de phrases pré-définies, avec un tempo fixe, en chambre anéchoïque, pourrait constituer un facteur de clarification non négligeable. Aussi, avons-nous décidé, afin d'explorer la fiabilité des relations temporelles entre la voix et la détente dans des conditions plus réalistes, de mesurer ces indices sur des échantillons de parole extraits d'une conversation spontanée, enregistrée en chambre anéchoïque.

2.2. Procédure

Locuteurs : 3 francophones bruxellois, dont 2 hommes et 1 femme, âgés de 25 à 40 ans et de formation universitaire.

Procédure d'enregistrement : les locuteurs, qui se connaissent en dehors du cadre de l'expérience, ont été invités à tenir une conversation informelle d'une quinzaine de minutes en chambre anéchoïque. Les sujets, assis, étaient disposés en triangle autour d'un micro Neumann U88. Afin de capter les 3 voix dans des conditions similaires, le micro était réglé en position "omni-directionnelle" et suspendu au centre du triangle formé par les locuteurs. La conversation a été enregistrée à l'aide d'un magnétophone Studder A810 sur une bande Scotch 256 à la vitesse de 19 cm/s. La conversation s'est poursuivie pratiquement sans interruption et a été enregistrée durant 16 minutes et 30 secondes.

Dépouillement : différentes portions de la conversation enregistrée ont été numérisées et transférées sur disque à l'aide d'un logiciel phonétique (Schoentgen & Saerens, 1987) développé sur PDP 11/60. Les portions de signal ont été choisies d'après les critères suivants :

1. Contenir au moins une consonne occlusive, non précédée ou suivie directement d'une autreocclusive ou d'une fricative.
2. Ne pas présenter de recouvrement entre les différentes voix.
3. Atteindre au moins un niveau de 66 db (11 bits), afin d'obtenir un degré de visualisation suffisant lors du traitement par le programme d'analyse.

Après dépouillement de la première minute d'enregistrement, nous avons dénombré 44 portions de signal répondant aux critères 1, 2 et 3. Nous avons alors décidé de modifier le premier critère, et ce pour trois raisons :
Primo : les portions de signal ainsi retenues étaient assez fréquentes (44 portions / 60 s).
Secundo : les analyses acoustiques montrent que les occlusives voisées (qui, à deux exceptions près, se trouvent en position médiane) se caractérisent toutes par des vibrations laryngées continues, donc sans arrêt de voix.
Tertio : l'analyse acoustique des portions prélevées durant la première minute révèle que les occlusives sourdes dont l'occlusion est complètement voisée sont peu fréquentes (cf. 2.3. Résultats).

Aussi, afin d'obtenir un nombre aussi élevé que possible d'occlusives de ce type (Tertio) sans toutefois prélever l'ensemble des cas présents dans les 15 minutes de conversation non explorées, nous avons choisi de ne retenir que les occlusives sourdes qui ont été prononcées de manière *relativement rapide* dans les portions *non accentuées* du discours. Modifié de la sorte, le premier critère de choix nous a conduits à prélever seulement 44 portions supplémentaires. Nous disposons donc, après dépouillement, de deux corpus distincts :
- Un corpus "tout venant" d'occlusives présentes dans la première minute de conversation.
- Un corpus d'occlusives sélectionnées en appliquant le critère modifié aux 15 minutes de conversation restante.

Analyse acoustique : les 88 portions de discours numérisées ont été traitées à l'aide du module "éditeur de signal" du logiciel phonétique. Celui-ci permet notamment de visualiser le signal acoustique, mesurer des durées de segments définis, écouter ces mêmes segments et, éventuellement, les sauver sur disque. Nous avons donc, pour chaque occlusive, mesuré le VOT et l'intervalle de silence (IS), c'est-à-dire le délai entre la fin des vibrations laryngées et la détente de l'occlusion; ceci nécessite le repérage de 3 événements : l'arrêt des vibrations (pré-détente), la détente de l'occlusion et le départ des vibrations laryngées (post-détente). Ces événements ont été localisés en se basant sur les manifestations correspondantes d'énergie apériodique pour la détente, ou périodique pour les vibrations laryngées. Pour tenir compte des évolutions graduelles de l'intensité, nous avons pris un niveau d'amplitude supérieur ou égal au double de l'amplitude du bruit de fond comme limite pour repérer l'arrêt ou le départ des vibrations périodiques, ainsi que le début du bruit de détente.

2.3. Résultats

I. Corpus "tout venant" : parmi les 44 portions extraites de la première minute d'enregistrement, nous avons comptabilisé 52 occlusives en position médiane . Parmi les occlusives médianes, 6 n'ont pas été retenues pour les mesures, l'explosion n'étant pas détectable (cependant chacune de ces 6 occlusives présentait un arrêt de voix de plus de 100 ms). Restent donc 46 occlusives médianes dont 14 voisées et 32 sourdes. Les 14 occlusives voisées présentaient toutes des vibrations laryngées continues. Cette constatation s'est confirmée par la suite; tous les signaux observés ultérieurement et contenant des occlusives voisées en position médiane présentaient les mêmes caractéristiques. Par contre, les caractéristiques temporelles des occlusives sourdes sont moins stables. Cependant, sur les 32 cas observés, nous n'avons rencontré qu'un seul cas de voisement continu.
II. Corpus sélectionné : parmi les 44 portions choisies dans les 15 minutes de conversation restante, nous avons comptabilisé 53 occlusives sourdes, les voisées n'étant plus prises en compte. Dans ce corpus également, une seule occlusive présentait des vibrations laryngées continues.
La comparaison de ces deux corpus montre que le VOT moyen est pratiquement identique et que l'IS ainsi que l'arrêt de voix (IS + VOT) ne diffèrent pas significativement (pour l'IS, $t_{83} = 1.87$; pour l'arrêt de voix, $t_{83} = 1.63$).

Les figures 1 à 3 fournissent les distributions des trois indices temporels en conversation spontanée d'une part, pour des mots en position accentuée et non-accentuée d'autre part (Serniclaes, 1984).

2.4. Discussion

Qu'en est-il de la fiabilité des indices temporels au vu de la première expérience? La prise en considération d'une interruption des vibrations laryngées comme critère de différenciation entre occlusives sourdes et voisées se révèle d'une efficacité tout-à-fait remarquable :

- Toutes les occlusives voisées observées présentent des vibrations laryngées continues,sans interruption de voix.

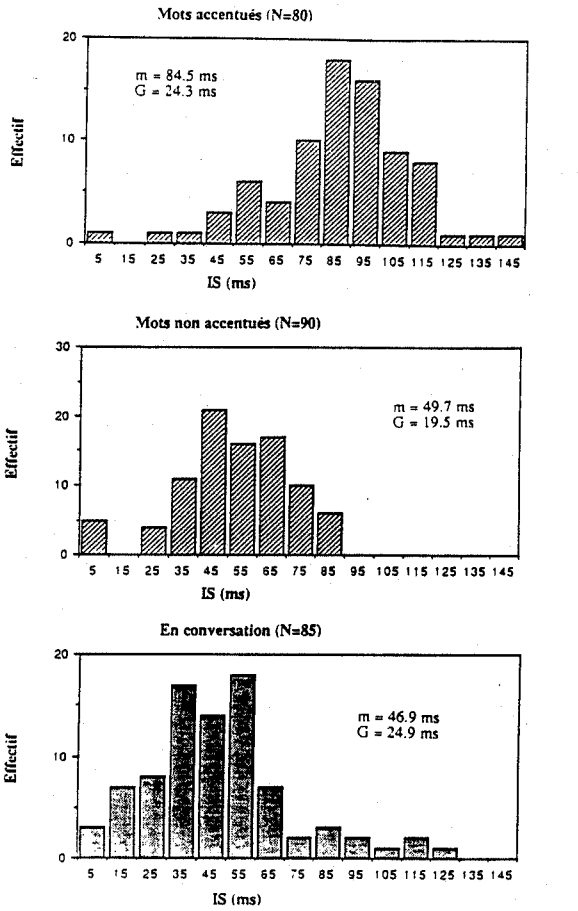


Figure 1 : Comparaison entre les distributions d'IS mesurées sur des occlusives sourdes dans 3 corpus différents : / des mots insérés en phrases, en position accentuée / des mots insérés en phrases, en position non accentuée / en conversation spontanée. Bien que les durées moyennes de l'IS soient équivalentes pour la conversation et les mots non accentués, la variabilité est plus élevée pour les occlusives extraites de la conversation. La différence est significative à 0.05 ($F^{84,89} = 1.63$).

- Parmi les occlusives sourdes observées (32 dans le corpus "tout venant" + 53 dans le corpus sélectionné), deux cas seulement présentent des vibrations laryngées continues.

On a donc sur l'ensemble de l'échantillon un score d'erreur inférieur à 2%, du moins dans les portions de signal répondant aux critères 1 à 3 de la procédure de dépouillement (cf. 2.2. Procédure). Ce score peut être considéré raisonnablement comme une limite supérieure dans la mesure où la recherche dans la sélection de 15 minutes a porté sur les occlusives les plus susceptibles de poser problème quant à la fiabilité des indices temporels. Un score d'erreur aussi faible est exceptionnel par rapport aux résultats usuels (pour le voisement en anglais : Flege & Brown, 1982; Zue, 1985; pour le lieu d'occlusion : Stevens & Blumstein, 1978), ce qui met en évidence la très grande fiabilité des relations temporelles servant à opérer la distinction entre occlusives sourdes et voisées en français. On peut néanmoins se demander d'où provient l'information complémentaire lorsque les indices temporels sont - exceptionnellement - pris en défaut. La question est donc de savoir si l'information qui permet d'identifier correctement le trait de voisement provient d'autres indices acoustiques (F_0, durées de transition, etc) ou, au contraire, d'informations de niveau supérieur (informations lexicales, syntaxiques, sémantiques, etc). Dans l'hypothèse où les indices acoustiques résiduels jouent un rôle critique, l'occlusive qui est perçue comme sourde en contexte de phrase, devrait rester perçue comme telle en dehors de son contexte. Cette question est examinée dans la deuxième expérience.

3. Expérience II

3.1. Introduction

Des expériences antérieures ont montré que si l'on réduit l'IS et le VOT d'occlusives sourdes intervocaliques, elles sont perçues comme voisées lorsque les deux indices temporels sont relativement brefs, sans toutefois être nuls (Serniclaes, 1975).

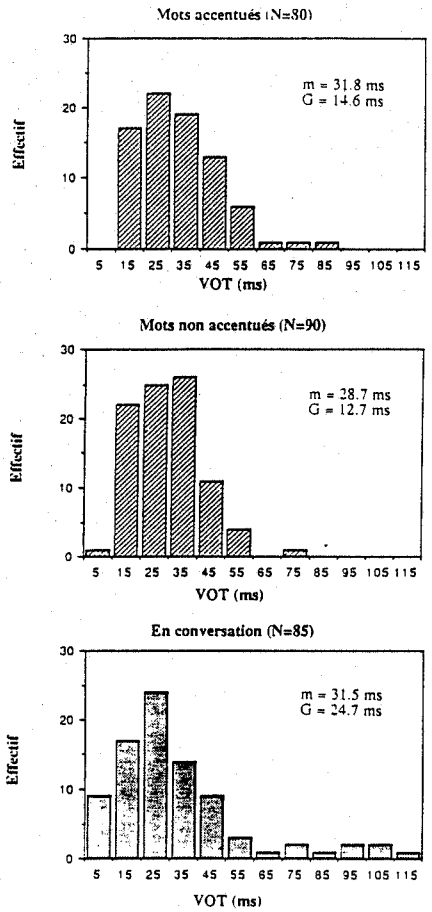


Figure 2 : Comparaison entre les distributions de VOT mesurées sur des occlusives sourdes dans 3 corpus différents : / des mots insérés en phrases, en position accentuée / des mots insérés en phrases, en position non accentuée / en conversation spontanée. Le VOT moyen ne dépend pratiquement pas du corpus envisagé. Mais, à l'instar de ce que l'on observe pour l'IS, la variabilité est plus large pour les items prélevés en conversation. La différence est significative à 0.005 ($F^{84,89} = 3.78$).

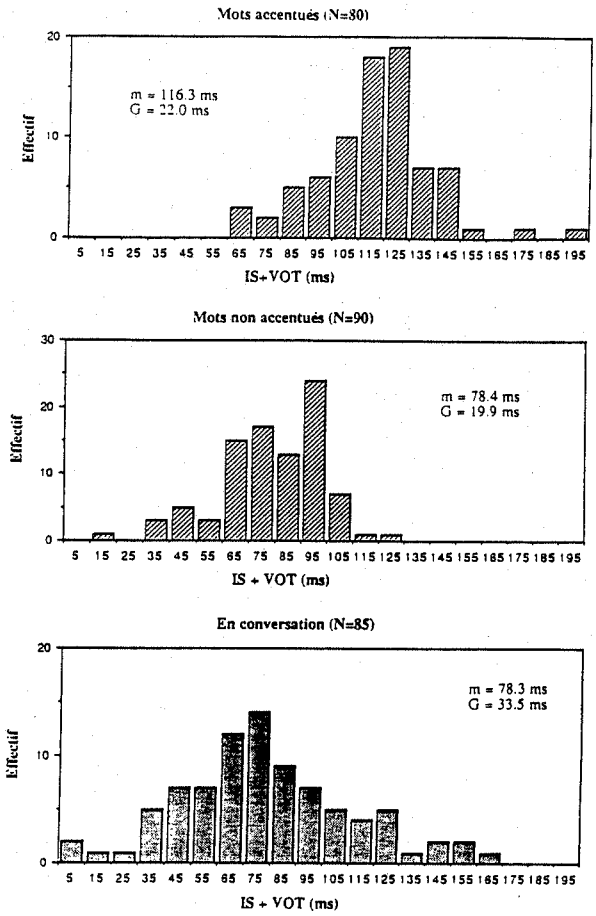


Figure 3 : Comparaison entre les distributions d'arrêts de voix mesurées sur des occlusives sourdes dans 3 corpus différents : / des mots insérés en phrases, en position accentuée / des mots insérés en phrases, en position non accentuée / en conversation spontanée. Comme pour l'IS, les durées moyennes sont pratiquement équivalentes pour la conversation et les mots inaccentués. La variabilité, quant à elle, est plus large pour les occlusives extraites de la conversation. La différence est significative à 0.005 ($F^{84,89} = 2.83$).

On peut ainsi mettre en évidence une fonction de compensation perceptive entre les deux indices. Cette fonction est définie en mettant en relation le VOT et l'IS dans les stimuli pour lesquels il y a équiprobabilité de réponses sourdes et voisées (Stevens & Klatt, 1974). Elle peut être approchée de manière linéaire par l'équation :

$$IS + 2.5\ VOT = 75 \text{ ms} \quad (1)$$

L'examen de cette équation montre que l'IS doit être allongé de 2.5 ms pour compenser une réduction de 1 ms de VOT. Le fait d'obtenir un coefficient de compensation différent de 1 signifie qu'un coefficient AVS (pour "arrêt de voix subjectif") avec :

$$AVS = IS + 2.5\ VOT \quad (2)$$

fournit une meilleure estimation du degré d'ambiguïté perceptive que la durée de l'arrêt de voix (IS + VOT, cf. 2.3. Résultats). Enfin, l'équation 2 suggère également que, en l'absence d'information contextuelle de niveau supérieur, l'occlusive reste perçue comme étant sourde tant que AVS > 75 ms et voisée tant que AVS < 75 ms.

L'objectif de cette seconde expérience est, primo, de voir si une occlusive sourde qui ne présente pas d'arrêt de voix est perçue correctement en dehors du contexte de la phrase, et, secundo, de voir si l'équation 1 permet de prédire les scores d'identification perceptive pour des occlusives sourdes extraites d'échantillons de parole spontanée, et ce en l'absence d'informations contextuelles. Pour ce faire, nous avons excisé des segments comprenant chacun une occlusive et les deux voyelles adjacentes à partir des échantillons de conversation qui ont servi aux mesures acoustiques (cf. 2.2. Procédure). Les occlusives sourdes ont été sélectionnées en fonction du coefficient AVS de manière à obtenir un ensemble approximativement uniforme de valeurs comprises entre 0 et 150 ms.

3.2. Procédure

Stimuli : segments de parole sans signification, extraits des portions de conversation qui ont été numérisées (cf. 2.2. Procédure), et qui contiennent chacun une consonne occlusive parfaitement identifiable et non ambiguë dans son contexte (cf.4.3 Résultats: expérience de contrôle). Les limites de segmentation ont été choisies de manière à inclure la voyelle qui précède l'occlusive ainsi que la voyelle suivante dans le stimulus. Dans chaque cas, le début de la fenêtre de prélèvement correspond au début de la portion stable du segment vocalique précédent tandis que la fin de la fenêtre correspond à la fin du segment vocalique suivant. Les occlusives retenues pour le test perceptif ont été sélectionnées en fonction de la catégorie de voisement, du locuteur, et, pour les occlusives sourdes, de la durée des arrêts de voix. Quatre classes d'occlusives sourdes ont ainsi été constituées en fonction de la valeur de leur coefficient AVS (voir *Tab.1*). Ces 4 classes regroupent un total de 30 occlusives sourdes. De plus, nous avons introduit 18 occlusives voisées dans la série expérimentale. Les voisées présentent toutes des vibrations laryngées continues et peuvent donc être considérées comme non ambiguës.

Procédure : la série expérimentale (48 stimuli) a été constituée à partir de trois blocs de taille approximativement égale. Chaque bloc contient environ le même nombre d'items en provenance des 4 classes d'occlusives sourdes. Les cas ambigus (AVS < 75 ms) ont été placés directement après une occlusive prononcée par le même locuteur et ce en proportion équivalente après des occlusives sourdes et voisées. La série expérimentale est précédée d'une pré-série de 6 stimuli. Chaque stimulus est répété deux fois.

Sujets : 30 francophones belges, sans trouble auditif connu, dont 13 hommes et 17 femmes, âgés de 14 à 50 ans ont participé à l'expérience. L'écoute se faisait à l'aide d'un casque Sennheiser HD222. Les sujets avaient pour consigne d'identifier l'occlusive en choisissant parmi l'une des six réponses : P, B, T, D, K, G.

3.3. Résultats

Les pourcentages d'identification correcte du trait de voisement pour les différents types d'occlusives sont présentés

	Sourdes (N=30)				Voisées (N=18)
AVS (ms)	= 0	0 < ≤ 75	75 < ≤ 100	100 < ≤ 150	(AVS = 0 ms)
N=	2	10	9	9	18
Reconnues significativement comme sourdes (test binomial)	0	6	7	9	0
Reconnues significativement comme voisées (test binomial)	2	3	1	0	18
% de reconnaissance du trait de voisement (N x 30 sujets)	17%	63%	78%	96%	94%
% de reconnaissance du lieu d'occlusion (N x 30 sujets)	60%	83%	84%	98%	78%

Tableau 1 *: Scores d'identification du trait de voisement et du lieu d'occlusion pour les quatre classes d'occlusives sourdes, en fonction du coefficient AVS et pour les 18 occlusives voisées. Chaque score est basé sur les réponses fournies par 30 auditeurs. Toutes les occlusives voisées ont été identifiées comme telles avec un score supérieur au hasard (supérieur à 50 % : $p < 0.05$). Le score d'identification des sourdes dépend de l'AVS. La frontière perceptive se localise approximativement entre 0 et 75 ms. On constate également que 6 occlusives sourdes ont été identifiées comme voisées ($p < 0.05$: label "reconnues significativement comme voisées") lorsqu'elles sont présentées en dehors de leur contexte. Quant au score d'identification du lieu d'occlusion, il augmente en fonction de la durée de l'arrêt de voix, pour les sourdes, et il est plus faible pour les voisées (qui ne présentent pas d'arrêt de voix).*

dans le *Tab.1*. Pour chaque score, l'écart par rapport à l'équiprobabilité a été évalué à l'aide du test binomial. Le même tableau fournit le nombre de cas significatifs ($p < 0.05$) pour chaque type d'occlusive. Le score d'identification des occlusives sourdes augmente en fonction de l' AVS, mais la frontière perceptive (équiprobabilité entre réponses sourdes et voisées) se situe entre 0 et 75 ms, ce qui est inférieur à la frontière de 75 ms mise en évidence précédemment dans le cas de logatomes isolés (cf. 3.1. Introduction; Serniclaes, 1975). Lorsque l'AVS est supérieur à 100 ms, l'occlusive reçoit systématiquement un score de réponses sourdes supérieur au niveau du hasard et, inversément, les deux occlusives sourdes ne présentant pas d'arrêt de voix reçoivent un score de réponses voisées supérieur au hasard. Enfin, les occlusives voisées ont toutes été identifiées comme telles avec des scores supérieurs au hasard.

La *Fig.4* fournit une image détaillée de la relation entre les scores d'identification correcte des différentes occlusives sourdes, prises séparément, et les valeurs d'AVS correspondantes. Il est clair qu'au niveau des items individuels la relation n'est plus monotone. Au-dessous du graphe représentant les pourcentages d'identification en fonction de l'AVS (*Fig.4*), nous avons ajouté la distribution d'AVS pour les 85 occlusives qui ont servi dans la première expérience.

3.4. Discussion

Les résultats confirment la prédominance des relations temporelles entre la voix et la détente dans le processus de perception du trait de voisement des occlusives du français. Nous savions déjà que l'effet perceptif du timing laryngé l'emporte sur celui de l'ensemble des autres indices de voisement connus, c.à.d. ceux qui ont été mis en évidence dans les analyses acoustiques et qui ont en conséquence été manipulés dans les stimuli au cours des recherches antérieures (Serniclaes, 1987). Mais d'autres indices pourraient être présents dans la parole spontanée et, seuls ou en conjonction avec les indices classiques, avoir un poids perceptif supérieur à celui du timing laryngé. Si tel était le cas, la catégorie de voisement pourrait éventuellement être identifiée correctement en dehors du contexte de la phrase lorsque le timing laryngé induit en erreur et, inversément, l'identification correcte ne serait pas garantie lorsque le timing laryngé ne prête pas à confusion. Les résultats obtenus ici vont à l'encontre de cette hypothèse. Ils montrent en effet que les occlusives qui ne comportent pas d'arrêt de voix au voisinage de la détente sont toujours identifiées comme voisées (avec un score supérieur à l'équiprobabilité) en dehors du contexte et, inversément, que l'occlusive est toujours identifiée comme sourde dès que l'arrêt de voix est relativement long.

L'examen de la *Fig.4* révèle qu'une occlusive pour laquelle l'arrêt de voix atteint 91 ms AVS peut encore éventuellement être perçue comme voisée tandis qu'une occlusive dont l'arrêt de voix atteint à peine 29 ms AVS peut encore être perçue comme sourde. Ceci montre que la zone d'ambiguïté du timing laryngé, c.à.d. les valeurs de timing pour lesquelles les

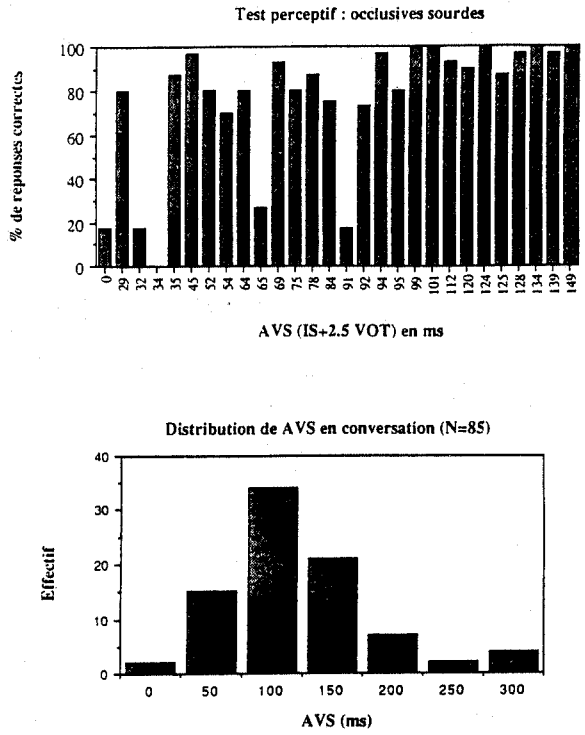


Figure 4 *: La partie supérieure de la figure fournit des scores de réponses correctes pour desocclusives sourdes utilisées comme stimuli dans l'expérience II et ce en fonction du coefficient AVS. Les scores de réponses correspondent généralement à un seul stimulus qui a été présenté à 30 auditeurs, à l'exception du score obtenu pour 0 ms (2 stimuli) et du score obtenu pour 139 ms (2 stimuli). Dans l'hypothèse où les résultats se conformeraient exactement à ceux obtenus précédemment, la frontière perceptive (50 % de réponses sourdes) devrait se localiser à 75 ms et les scores devraient évoluer régulièrement vers 0 ou 100 % de part et d'autre de cette valeur. Ce n'est clairement pas le cas, ce qui laisse supposer que la frontière varie largement en fonction d'autres indices présents dans le segment prélevé. La partie inférieure de la figure donne la distribution de l'AVS pour l'ensemble de l'échantillon d'occlusives sourdes prélevées dans la conversation. On constate que les valeurs d'AVS qui peuvent éventuellement provoquer des erreurs d'identification, la limite se situant approximativement à 100 ms d'après les résultats perceptifs, sont relativement peu fréquentes dans la production.*

indices alternatifs affectent l'identification du trait, atteint au moins 62 ms AVS. En fonction de ces valeurs, le centre de la zone d'ambiguïté se localise grossièrement à 60 ms AVS. Comme la valeur centrale de la zone d'ambiguïté correspond à la frontière perceptive moyenne, celle-ci serait légèrement inférieure à ce que nous avions prévu (cf. 3.1 Introduction : 75 ms) en nous basant sur les données obtenues pour des logatomes VCV. Le désaccord peut s'expliquer par le fait que les données antérieures étaient fondées sur des logatomes isolés prononcés par un seul locuteur.

En convoluant les deux distributions présentées en *Fig.4*, c.à.d. en multipliant les taux d'erreurs d'identification par les fréquences relatives des valeurs d'AVS correspondantes dans la production, nous pouvons grossièrement évaluer le pourcentage d'identification du trait de voisement en conversation pour les occlusives sourdes séparées de leur contexte. On obtient ainsi un score de 85% pour les occlusives sourdes. Quant aux occlusives voisées, nous voyons dans le *Tab.1* que le score peut être évalué à 94%. On peut en conclure que, indépendamment du contexte, les indices acoustiques permettent au système perceptif d'identifier correctement le trait de voisement dans quelque 90% des cas. Une telle performance est relativement faible comparée à la fiabilité du corrélat acoustique; ainsi, les indices fournis par le timing laryngé permettent l'identification correcte du trait dans près de 98% des cas (cf. 2.4 Discussion). Ceci suggère que la fiabilité du timing laryngé n'est pas exploitée de manière optimale par le système perceptif. Cependant un tel résultat pourrait tout simplement provenir du fait que les différences intercatégorielles dans la production sont plus fines que les capacités de discrimination perceptive.

La somme des informations fournies par les indices acoustiques présentes dans le segment prélevé ne permet donc pas d'identifier systématiquement le trait de manière correcte. Rappelons que les occlusives utilisées comme stimuli dans cette expérience peuvent être identifiées correctement lorsqu'elles sont présentées dans leur contexte. En d'autres termes, l'audition des

phrases dont elles ont été extraites prouve que l'on ne détecte pas d'erreurs de prononciation (cf. 4.3 Résultats : Expérience de contrôle). Il est donc clair que des informations extérieures au segment prélevé jouent ici un rôle déterminant. Ces informations peuvent aussi bien provenir d'autres indices acoustiques, présents à l'extérieur du segment, que de processus de niveau supérieur (extraits du contexte). A ce sujet, le cas des six occlusives sourdes perçues comme voisées donne matière à réflexion. Elles sont perçues comme voisées localement, et comme sourdes dans leur contexte. On pourrait en déduire naïvement que l'information contextuelle emporte automatiquement la décision, quelles que soient les valeurs des indices locaux présents dans le segment (AVS et indices résiduels). Mais comment expliquer alors que, dans certains cas, l'on puisse déceler des erreurs de prononciation au cours d'une conversation (production d'un "b" à la place d'un "p"). Cela suggère que les valeurs des indices acoustiques de nos 6 occlusives sourdes perçues comme voisées ne sont néanmoins pas suffisamment voisées pour que le sujet relève une erreur de prononciation .

4. Expérience III

4.1. Introduction

L'expérience II a montré que, dans certains cas, une occlusive perçue comme voisée "localement", c'est-à-dire hors de son contexte, est perçue comme sourde lorsqu'elle réintègre son contexte de phrase. Devons-nous attribuer cette discrépance à des informations situées à l'extérieur des segments et qui induiraient une réponse sourde lorsqu'on les inclut dans le traitement ? Ces informations peuvent être fournies, soit par d'éventuels indices acoustiques situés en dehors du segment prélevé, soit par des informations de niveau supérieur liées au contexte.

L'objectif de cette troisième expérience consistera à évaluer l'effet du contexte par rapport à celui d'éventuels indices acoustiques qui se localiseraient en dehors de la fenêtre de prélèvement utilisée dans l'expérience précédente, sur la perception du trait de voisement des occlusives. Pour ce faire, nous avons récupéré les 6 segments de parole utilisés dans l'expérience II et contenant une occlusive sourde qui n'a pas été identifiée correctement. En plus de ces 6 segments, nous avons retenu 6 autres stimuli, également utilisés dans l'expérience II, et contenant des occlusives sourdes et voisées qui ont été identifiées correctement. Ces 12 segments ont été progressivement allongés de 50 ms de part et d'autre du segment initial afin d'obtenir un total de 60 stimuli. Nous avons donc procédé à 4 allongements progressifs (cf. 4.2. Procédure). De plus, nous avons constitué une série de contrôle de 12 stimuli dans lesquels apparaissent les 12 portions de phrases dont ont été extraits les segments, et ce afin de vérifier si les occlusives extraites sont suffisamment claires dans leur contexte.

4.2. Procédure

Stimuli : les 6 segments de parole, utilisés dans l'expérience II, contenant les occlusives sourdes qui sont identifiées comme étant voisées hors de leur contexte ainsi que 6 segments, également utilisés dans l'expérience II, et contenant des occlusives identifiées correctement en dehors de leur contexte, dont 3 sourdes et 3 voisées. A partir de ces 12 portions originales et des phrases porteuses correspondantes, nous avons construit 48 portions supplémentaires de la manière suivante : chaque portion originale a été systématiquement allongée de 100 ms de façon à inclure le segment de 50 ms situé à gauche de la portion originale dans la phrase porteuse, et le segment de 50 ms situé à droite de la portion originale dans la phrase porteuse. Nous obtenons ainsi un signal centré sur la portion originale, et allongé de 100 ms (50 ms à droite et 50 ms à gauche) par rapport à l'original. La portion originale a ainsi été allongée 4 fois de suite : allongement de 100 , 200 , 300, et 400 ms. Ce travail a été effectué à partir des 12 portions originales, ce qui nous amène à un total de 60 segments, y compris les originaux.

En ce qui concerne l'expérience de contrôle, chaque stimulus comprend une portion de phrase porteuse, d'une durée de 800 ms et centrée sur l'occlusive étudiée. 12 stimuli ont ainsi été constitués à partir des 12 occlusives retenues pour l'expérience III.

Procédure : la série expérimentale (60 stimuli) respecte un ordre pseudo-aléatoire. Nous avons veillé à ce que 2 stimuli issus de la même phrase soient séparés par 5 stimuli différents, ou plus. La série expérimentale est précédée d'une pré-série de 5 stimuli. Chaque stimulus est répété cinq fois.

Sujets : 30 francophones belges, sans trouble auditif connu, dont 14 hommes et 16 femmes, âgés de 14 à 50 ans ont participé à l'expérience. Les sujets avaient pour consigne de reporter par écrit le contenu de chaque segment sur un formulaire ad hoc. Ils ont ensuite été invités à passer l'expérience de contrôle, qui s'est déroulée dans les mêmes conditions. Cette fois-ci, les auditeurs avaient comme consignes de transcrire les phrases et, ensuite, de souligner la ou les portions de phrase qu'ils jugeaient incorrectement prononcées.

Dépouillement : celui-ci a consisté à extraire deux informations des transcriptions des sujets : l'information sur le voisement et sur le contexte. Lors du dépouillement, nous n'avons retenu que les obstruantes; les autres consonnes n'ont pas été prises en compte (les réponses rejetées apparaissent dans le *Tab. 2* sous le label "non-obstruante"). La réponse a été considérée comme voisée lorsque le sujet a perçu une obstruante voisée, et, bien entendu, sourde lorsque le sujet a perçu une obstruante sourde. Quant au contexte, son influence a été mise en évidence comme suit : nous avons considéré

- qu'il y avait intégration du contexte lorsque le sujet transcrivait le segment sous forme de mots correctement orthographiés et séparés
- qu'il n'y avait pas intégration du contexte lorsque le sujet recourait à une écriture pseudo-phonétique.

En cas d'ambiguïté, nous avons questionné le sujet après le test, le dépouillement ayant toujours été effectué immédiatement après la passation.

Les résultats de l'expérience de contrôle ont été dépouillés de la même manière mais nous avons aussi comptabilisé les items pour lesquels le sujet a estimé qu'il y avait erreur de prononciation dans un segment incluant l'occlusive.

	Contexte non intégré			Contexte intégré		Non - obstruante
Durée du segment	Voisée	Sourde	% de réponses sourdes	Voisée	Sourde	
1	100	27	21	15	24	14
2	65	24	27	2	79	10
3	42	12	22	4	116	7
4	26	7	21	1	142	4
5	21	6	22	1	146	6
6	5	0	0	0	175	0

***Tableau 2** : Résultats des transcriptions effectuées par 30 sujets pour des durées de segment de plus en plus élevées (allant de 1 à 6, 6 étant la portion de phrase de 0.8 s utilisée dans l'expérience de contrôle). Le nombre des transcriptions, pour une durée donnée (une ligne), s'élève à 30 sujets x 6 stimuli, c'est-à-dire 180 items. Chacune des transcriptions peut être classée dans une des 5 rubriques : / intervention du contexte (des processus haut-bas : lexique, sémantique, etc) et réponse voisée / intervention du contexte et réponse sourde / pas d'intervention du contexte et réponse voisée / pas d'intervention du contexte et réponse sourde / pas de perception d'une obstruente. Rappelons que les segments de parole servant de stimuli sont centrés sur des occlusives perçues comme voisées "localement" (à partir d'un segment composé de l'occlusive avec les deux voyelles adjacentes) et sourdes dans leur contexte de phrase. On remarque en effet que, pour les segments les plus courts, et sans influence du contexte, le pourcentage de réponses sourdes est peu élevé (21 %) tandis que pour les segments les plus longs, pratiquement toutes les réponses sont sourdes (175 réponses). Comme on peut s'y attendre, le nombre de réponses sourdes ainsi que l'intégration du contexte deviennent de plus en plus élevés, jusqu'à regrouper la quasi-totalité des réponses (175) pour le segment de durée maximale. Parmi les réponses pour lesquelles il n'y a pas eu de contribution du contexte, le pourcentage de réponses sourdes reste très stable (colonne 4). Ceci confirme l'absence d'indices acoustiques en dehors du segment le plus bref. En effet, en excluant l'effet du contexte, dans le cas où des indices interviendraient, on devrait observer une augmentation du pourcentage de réponses sourdes lors de l'allongement de la durée du segment.*

4.3. Résultats

Le *Tab. 2* reprend le nombre de réponses appartenant à chacune des 5 rubriques - contexte intégré/non intégré, réponse voisée/sourde et non-obstruente - en fonction de la durée du segment de parole présenté (allant de 1 à 6, 2 à 5 correspondant aux allongements successifs de 100 ms et 6 étant la portion de phrase de 0.8 sec utilisée dans l'expérience de contrôle). Nous n'avons considéré ici que les résultats concernant les segments de phrase contenant l'une des 6 occlusives sourdes identifiées comme voisées hors de leur contexte. Le total des réponses reprises sur une ligne - une durée de segment donnée - s'élève donc à 6 stimuli x 30 sujets, c'est-à-dire 180 items. De plus, nous avons évalué le score de réponses sourdes pour lesquelles le contexte n'est pas intervenu.

Comme on peut s'y attendre, le score de réponses sourdes ainsi que l'intégration du contexte deviennent de plus en plus élevés pour des durées de segment de plus en plus longues. La quatrième colonne du *Tab.2* reprend le pourcentage de réponses

sourdes en fonction de la durée du segment parmi l'ensemble des réponses pour lesquelles le contexte n'est pas intervenu. Ce score demeure étonnamment stable. Or, dans l'hypothèse où des indices acoustiques complémentaires agiraient à l'extérieur du premier segment (utilisé dans l'expérience II), on devrait observer une augmentation de ce pourcentage de réponses sourdes lors de l'allongement de la durée du segment. Ceci n'est pas le cas (Test chi 2 à 4 dl = 1.6544, effectué à partir du *Tab.2*, tout à fait non significatif).

Quant à l'expérience de contrôle, sur les 180 phrases transcrites, 3 segments comprenant l'occlusive étudiée ont été soulignés et 5 phrases ont été soulignées complètement. Ces occlusives, localement voisées mais perçues comme sourdes dans leur contexte, ne donnent donc pas lieu à des jugements d'erreurs de prononciation.

4.4 Discussion

Les résultats confirment donc le rôle déterminant du contexte (processus haut-bas), et infirment l'hypothèse d'indices acoustiques complémentaires importants situés en dehors des segments prélevés dans l'expérience II. Reste donc à déterminer le rôle joué respectivement par les indices résiduels et le contexte. L'expérience II a confirmé le rôle prépondérant des indices temporels pour la perception du trait de voisement des occlusives extraites de leur contexte (cf. 3.3 Résultats). En effet, les occlusives présentant des vibrations laryngées continues ainsi que quelques autres (6 au total, cf. 3.3 Résultats) ont été perçues comme voisées en dehors de leur contexte (et, rappelons-le, sourdes au sein de la conversation), conformément à la valeur de leurs indices temporels. Or, insérées dans leur contexte, ces occlusives sont clairement perçues comme sourdes, sans détection d'erreur de prononciation, comme l'a montré l'expérience de contrôle (cf. 4.3 Résultats). En outre, dans deux des cas l'indice majeur (AVS) apporte une information "strictement voisée" puisqu'il n'y a pas d'arrêt de voix. Il y a donc clairement dans ces cas une réorganisation du processus d'intégration en faveur d'autres indices acoustiques. Cette réorganisation ne s'opère pas systématiquement, néanmoins (cf.3.3 Résultats), mais *seulement lorsqu'une contradiction surgit* entre l'information apportée par le contexte et les indices majeurs. La détection d'une erreur de prononciation correspondrait dans cette optique, à une confirmation de la contradiction au niveau des autres indices.

5. Discussion générale et conclusions

La première expérience a confirmé la fiabilité des relations temporelles établies à partir de logatomes dans le cas d'occlusives extraites d'une conversation spontanée. Cette fiabilité n'est, bien entendu, pas absolue. Mais, en comparaison avec d'autres travaux (pour le voisement en anglais : Flege & Brown, 1982; Zue, 1985; pour le lieu d'occlusion : Stevens & Blumstein, 1978), les cas contradictoires sont extrêmement rares. La seconde expérience confirme les résultats de perception obtenus pour des logatomes. Lorsqu'une occlusive prononcée dans une conversation est extraite de son contexte, l'identification du voisement dépend essentiellement des indices fournis par le timing laryngé (relations temporelles entre la détente de l'occlusion orale et les vibrations laryngées). L'expérience III montre que, lorsque les indices acoustiques situés dans le segment composé de l'occlusive et des deux voyelles adjacentes ne permettent pas d'identifier correctement l'occlusive, les informations supplétives situées en dehors de ce segment ne proviennent pas d'autres indices acoustiques, mais bien de la prise en compte du contexte.

La relation entre fiabilité et poids perceptif peut être mise en rapport avec les données recueillies à différents stades du développement perceptif. Dans la mesure où la structuration perceptive des indices de voisement résulte du contact avec l'environnement linguistique (Simon & Fourcin, 1978; Bernstein, 1983; voir cependant Eimas & Miller, 1983), la fiabilité des indices pour la distinction entre occlusives sourdes et voisées constitue un critère simple pour sélectionner les indices pertinents et leur attribuer des coefficients de pondération adéquats. Une telle possibilité pouvait sembler invraisemblable au vu de la complexité des relations entre de multiples indices acoustiques individuellement peu fiables et le contexte, sémantique en dernière instance (Bailey & Summerfield, 1980). Mais à partir du moment où un indice, ou du moins un sous-ensemble d'indices apparentés, atteint un degré de fiabilité comparable à celui qui a été mis en évidence dans ce travail pour le timing laryngé, la conjecture devient radicalement différente. Un indice nettement plus fiable que les autres peut, au début du développement perceptif, fournir un indice primaire. Ensuite, ce sont les contradictions entre cet indice et les informations contextuelles qui inciteraient le système à rechercher d'autres indices. Le fait que les indices alternatifs soient peu fiables ne constitue pas un obstacle à leur mise en évidence car l'intervention du contexte modifie complètement la situation. En effet, les indices dont les valeurs dépendent autant du contexte que du timing laryngé fournissent de l'information fiable lorsque le contexte est défini. Au terme du développement, ces indices résiduels sont susceptibles d'être incorporés dans le traitement, et cela en toutes circonstances, c'est-à-dire, même en l'absence d'informations contextuelles (cf. Expérience II et logatomes). On sait que, dans ces conditions, le poids de ces indices est plus faible que celui du timing laryngé. Mais l'intervention du contexte pourrait s'accompagner d'une restructuration des poids perceptifs en accord avec les changements de fiabilité, c'est-à-dire d'un rééquilibrage du traitement en faveur des indices alternatifs.

Bibliographie

Bailey P.J. & Summerfield A.Q. (1980) "Information in speech : observations on the perception of /s/ + stop clusters". J. Exp. Psychol. : Human Perception and Performance 6, pp 536-563.

Bernstein L.E. (1983) "Perceptual development for labeling words varying in voice onset time and fundamental frequency". J. of Phonetics 11, pp 383-393.

Caramazza A. & Yeni-Komshian G.H. (1974) "Voice onset time in two French dialects". J. of Phonetics 2, pp 239-245.

Connine C. & Clifton C. (1987) "Interactive use of lexical information in speech perception". Journal of Experimental Psychology : Human Perception and Performance 13, pp 291-299.

Durand M. (1956) "De la perception des consonnes occlusives : questions de sonorité". Word 12, pp 15-34.

Eimas P.D. & Miller J.L. (1983) "Studies on the categorisation of speech by infants". Cognition 13, pp 135-165.

Flege J.E. (1984) "The detection of French accent by American listeners". J.Acoust.Soc.Am. 76, pp 692-707.

Flege J.E. & Brown W.S.Jr (1982) "The voicing contrast between English /p/ and /b/ as a function of stress and position-in-utterance". J. of Phonetics 10, pp 335-345.

Ganong W.F. (1980) "Phonetic categorization in auditory word perception". Journal of Experimental Psychology : Human Perception and Performance 6, pp 110-125.

Garnes S. & Bond Z.S. (1976) "The relationship between semantic expectation and acoustic information". Phonologica 3, pp 285-293.

Lisker L. & Abramson A.S. (1964) "A cross-language study of voicing in initial stops: acoustical measurements". Word 20, pp 384-422.

Schoentgen J. & Saerens M. (1987) "Un progiciel phonétique". Rapport d'Activités de l'Institut de Phonétique de Bruxelles 21, pp 73-88, 1987.

Serniclaes W. (1975) "Perceptual processing of acoustic correlates of the voicing feature". In Speech Communication, Proc. of the SCS, Stockholm 1974. G. Fant ed., J. Wiley, pp 87-94.

Serniclaes W. (1979) "Sur la dissociation entre bruit, périodicité et fréquence fondamentale en tant qu'indices de voisement des occlusives du français". Rapport d'Activités de l'Institut de Phonétique de Bruxelles 13, pp 71-93.

Serniclaes W. (1984) "Fenêtre de prélèvement temporel des indices d'occlusives". Actes des 13èmes J.E.P., Bruxelles, pp 69-78.

Serniclaes W. (1987) "Etude expérimentale de la perception du trait de voisement des occlusives du français". Thèse de doctorat en Sciences Psychologiques et Pédagogiques. Université Libre de Bruxelles.

Simon C. & Fourcin A.J. (1978) "Cross-language study of speech-pattern learning". J. Acoust. Soc. Am. 63, pp 925-935.

Stevens K.N. & Klatt D.H. (1974) "Role of formant transitions in the voiced-voiceless distinction for stops". J. Acoust. Soc. Am. 55, pp 653-659.

Stevens K.N. & Blumstein S.E. (1978) "Invariant use for place of articulation in stop consonants". J. Acoust. Soc. Am. 64, pp 1358-1368.

Zue V.W. (1985) "The use of speech knowledge in automatic speech recognition". Proc. IEEE 73, pp 1602-1615.

PEUT-ON SUPPRIMER OU MIEUX EXPLOITER LES DISSYMETRIES DES MATRICES DE CONFUSIONS ?

Dominique PASCAL [1] **& Louis-Jean BOË** [2]

[1] CNET LANNION A
TSS/CMC
LANNION

[2] INSTITUT DE PHONETIQUE
ICP, UA 368
GRENOBLE

ABSTRACT

Through three examples of perceptual confusion matrices between speech stimuli (auditory as well as visual ones) and the responses, we consider the information related to the asymmetric responses within these data. We propose the strong form versus week form concept that seems to reflect adequately the fact that a stimulus, in specific conditions, tends to become the class representative. We suggest to process data with the correspondance analysis that can extract the dissymmetry between stimuli and percepts.

INTRODUCTION

Les matrices de confusion inter-stimuli auditifs présentent systématiquement des dissymétries notables. Ce phénomène est encore plus net avec des stimuli visuels (lecture labiale). "Deux stimuli étant dits similaires s'ils tendent à être confondus" [Miller56], les notions perceptives de confusion et de similarité interstimuli ont été considérées comme équivalentes. Pour pouvoir, ainsi, utiliser les matrices de confusions comme données d'entrée d'analyses multidimensionnelles des similarités [Kruskal64], plusieurs procédures de symétrisation ont été avancées et utilisées par Shepard [57, 58], Wagenaard [68] et Houtgast (in [Klein70]) par exemple.

Mais peut-on traiter la dissymétrie comme un biais expérimental qu'il est légitime d'éliminer? Ou au contraire considérer qu'il s'agit là d'une information devant être analysée et exploitée? D'autant que les réductions imposées par les procédures de symétrisation entraînent inévitablement une distorsion des données brutes que seules des contraintes d'exploitation peuvent justifier.

1. LES DISSYMETRIES

Pour illustrer notre propos, nous utiliserons plusieurs données issues des premiers travaux sur les confusions perceptives interstimuli, la célèbre étude sur les consonnes de Miller & Nicely [Miller54], de données publiées par Van Der Kamp & Pols [Van der Kamp71] sur des confusions de voyelles et de tous récents résultats de tests visuels d'identification labiale [Cathiard88].

1.1 Confusions interconsonantiques.

Considérons tout d'abord une matrice de confusions interconsonantiques pour des stimuli réduits aux triplets [p,t,k] et [b,d,g] associés à la voyelle [ɑ]. Les données du tableau 1 correspondent à des stimuli filtrés dans la bande de fréquences 200 - 6500 Hz et présentés dans des conditions de bruit (Rapport S/B = -12 dB) [Miller54]. Les confusions ont été rapportées à un même nombre de stimuli émis (100). Les différences entre les totaux par ligne (stimuli émis) et par colonne (stimuli perçus ou percepts) nous renseignent sur les dissymétries. [d] et [k] ont un score en colonne élevé, ces consonnes ont été perçues plus souvent qu'elles n'ont été émises; [b] et [p] moyen, [t] et surtout [g] faible. Nous dirons que les stimuli [d] et [k] sont des formes fortes auxquelles l'auditeur, à l'intérieur d'un même triplet et dans ces conditions, a tendance à associer son jugement et [g] et [t] des formes faibles attirées, dominées, respectivement par les précédentes. Insistons sur le fait que ces concepts de forme forte et forme faible doivent être associés au contexte des stimuli soumis à l'expérimentation.

Pour le stimulus i, la somme des différences

$$D(i) = \sum_{k=1}^{n} Cki - Cik$$

entre les confusions, pour un même couple percept/stimulus permet ainsi de donner la tendance globale. Elle précise d'autant mieux la notion de forme forte (D(i) >0) et de forme faible (D(i) <0) que la dissymétrie (|D(i)|) est importante : un percept attractif correspondra à une valeur de D(i) très supérieure à 0. La relation plus fine de domination entre deux stimuli i et j est donnée, de la même façon, par la différence

$$Dij = Cij - Cji$$

entre le nombre de perceptions i pour des émissions j (Cij) et le nombre d'émissions de i ayant abouti à une perception j (Cji). Si cette différence est positive, i domine j.

Pour nous en tenir au triplet [p,t,k], nous pouvons symboliser les relations de domination par le schéma suivant qui permet de noter que [k] domine [p] et [t], alors que [p] domine [t]:

1.2. Confusions intervocaliques.

Le tableau 2 présente les confusions obtenues avec des auditeurs hollandais pour des stimuli de type vocalique réduits à une seule période (8 ms), [Van der Kamp71]. Les résultats de la dissymétrie sont les suivants:

- les formes fortes sont les voyelles cardinales extrêmes [i,a,ɑ, u] et une voyelle centrale [oe] qui attire les réponses des auditeurs pour de nombreux stimuli de timbre intermédiaire (vraisemblablement, il s'agit là d'une réponse d'indécision); les stimuli subissant le plus cette attraction vers [oe] sont [e, y,ø];
- les formes les plus faibles sont les voyelles intermédiaires [o, e,ɛ].

Disposant d'un espace bien connu, le quadrilatère vocalique, nous pouvons schématiser les transferts entre stimuli émis et perçepts, c'est-à-dire, entre formes faibles et formes fortes (figure 1, les cercles sont proportionnels à D(i)).

1.3. Stimuli visuels.

Cependant, c'est avec des stimuli visuels [Cathiard88], que les dissymétries sont les plus incontournables, d'autant que les conditions expérimentales n'incluent ici aucune manipulation des stimuli : ni durée écourtée comme dans l'exemple des confusions intervocaliques, ni masquage par un bruit comme dans le cas des confusions de Miller & Nicely. Il s'agit de résultats moyens recuellis pour 21 sujets auxquels ont été présentés des diapositives correspondant à la partie centrale de la consonne ou de la voyelle d'une syllabe CV (Tableaux 3a & 3b). Les visèmes sont extraits d'un film 35 mm et correspondent à une vue de face du locuteur. La procédure de test utilisée était une procédure d'identification avec choix forçé à 4 éventualités. Il était proposé au sujet d'identifier une des 4 syllabes [ʃi, ʃy, si, sy] correspondant au minimum [si] et au maximum [ʃy] de protusion des lèvres ainsi qu'aux réalisations intermédiaires de coarticulation [sy et ʃi] [Abry80]. La diapositive présentée pouvait provenir soit des voyelles [i] et [y] en contexte [s] ou [ʃ], soit des consonnes [s] et [ʃ] en contexte [i] ou [y]. Plus les résultats s'écartent de 25 %, plus ils sont significatifs.

On constate que de nombreux "phonèmes-sons" sont regroupés comme sosies visuels avec de nettes tendances à ce que l'un d'entre eux, que nous appelons la forme forte, soit le représentant de la classe. Avec les visèmes vocaliques comme stimuli, c'est la syllabe [sy] qui est la forme forte, alors qu'avec les visèmes consonantiques c'est [ʃy]. Autrement dit, quelque soit le contexte consonantique, la voyelle [y] est associée à la syllabe [sy]; pour les consonnes [ʃ], quelque soit le contexte vocalique, entraîne la perception [ʃy]. Inversement, avec des stimuli vocaliques, [ʃy] est une forme faible puisque [y] associée à [ʃ] est essentiellement perçue comme [sy], de la même façon que [sy] est une forme faible pour les stimuli consonantiques, puisque [s] associé à [y] est perçu comme [si].

2. LA REDUCTION DES DISSYMETRIES ... ... ET DE L'INFORMATION

Pour réduire les dissymétries, Shepard [57] a, le premier, proposé d'opérer sur des moyennes arithmétiques ou géométriques calculées à partir des quatre éléments de confusions entre les stimuli i et j : Cij, Cji, Cii et Cjj. Le résultat de ces calculs est un indice symétrique Sij de la similarité des deux stimuli.

$$Sij = \frac{C_{ij}+C_{ji}}{C_{ii}+C_{jj}} \qquad Sij = \left[\frac{C_{ij}\times C_{ji}}{C_{ii}\times C_{jj}}\right]^{1/4}$$

Houtgast (cité par [Klein70]) propose la formule suivante, qui tient compte de la distribution complète des confusions possibles d'un stimulus i sur l'ensemble des n items soumis à l'expérimentation :

$$Sij = 0.5\sum_{k=1}^{n}(Cik + Cjk - |Cik - Cjk|)$$

Cette formule revient , en comparant les stimuli deux à deux, à effectuer la somme :

$$\sum_{k=1}^{n} \min |Cik, Cjk|$$

soit, encore, à considérer la somme des confusions communes comme la mesure de la similarité perceptive entre deux stimuli i et j.

Pour notre part, nous proposons d'utiliser la symétrisation suivante qui fournit des dissimilarités DIS entre i et j :

$$DISij = \sum_{k=1}^{n} |Cik - Cjk|$$

En fait, comme on peut le constater sur les exemples que nous avons déjà retenus, ces procédures ont deux effets néfastes :

1. Il n'est évidemment plus question, après leur application, de retrouver formes fortes et formes faibles.
2. Il se peut même que les rapports structurels des données brutes soient bouleversés, et cela d'autant plus que la matrice des confusions est dissymétrique.

A partir des données brutes (Tableau 4), on peut noter que les stimuli [p] sont deux fois plus souvent identifiés comme [k] que comme [t], alors que la présentation de [t] et [k] conduit au même nombre de réponses [p]. En outre, les stimuli [k] ont des probabilités égales de confusions avec [p] et [t], tandis que la présentation de ces deux dernières consonnes conduit au même nombre de réponses [k]; le percept dominant [k] se trouve ainsi également éloigné des stimuli [p] et de [t] de même que le stimulus [k] des percepts [p] et [t]. Après symétrisation par la méthode de Shepard (Tableau 5), le rapport des similarités de [p] vis-à-vis de [k] et [t] tombe à environ 1/3 en faveur de [k]. La symétrisation a ainsi, tout en conservant le rapport des similarités [p,k] et [t,k], artificiellement rapproché [p] et [t], comparativement à [k]. Cette constatation est encore plus vraie avec l'emploi de la formule de Houtgast, le rapport des similarités [p,k] et [p,t] devenant voisin de 1, celui des similarités [p,k] et [t,k] restant encore une fois inchangé (Tableau 6). Avec notre proposition de calcul des dissymétries (Tableau 7), le rapport des similarités [p,k] et [p,t] est analogue à celui

obtenu avec la méthode de Shepard mais celui des similarités [p,k] et [t,k] s'en trouve diminué de moitié.

Faute d'une procédure de symétrisation respectant la matrice des confusions, la meilleure solution consiste à analyser directement les données dissymétriques de façon à ne perdre aucun élément d'information.

3. MISE EN EVIDENCE ET EXPLOITATION DE LA DISSYMETRIE

En utilisant l'analyse factorielle des correspondances [Benzecri73], il est possible de projeter et comparer globalement la proximité de deux ensembles individus et caractéristiques dans un même plan. Comme cette analyse opère sur les profils ligne et colonne, deux stimuli seront proches dans le plan s'ils ont des comportements voisins vis-à-vis de l'ensemble des percepts. Si la matrice de confusion était parfaitement symétrique, on obtiendrait des projections confondues pour les stimuli et les percepts. Les écarts entre les deux projections vont donc nous permettre, à la fois de représenter les dissymétries et d'interpréter les déplacements stimuli ----> percepts par rapport aux formes fortes dégagées, en nous reportant aux matrices de données et aux valeurs de D(i) et Dij calculées. Les écarts les plus importants risquent d'être relatifs aux formes les plus fortes et les plus faibles, puisque ce sont elles qui présentent les plus grandes dissymétries. Cependant, il sera également possible de voir apparaître des déplacements importants pour des couples stimulus/percept qui, tout en n'étant ni forme forte ni forme faible, entretiennent des confusions multiples avec les autres couples.

Par exemple, l'analyse des correspondances que nous avons réalisée sur les données de Miller & Nicely (Tableau 1), concerne les seules confusions internes au triplet [p,t,k] (nous avons, en effet, rassemblé dans un résidu la somme des confusions relatives à [b,d,g]). Deux axes factoriels suffisent à rendre compte de 98 % de la variance des résultats. L'axe 1, expliquant à lui seul 70 % de la variance totale, correspond à l'opposition voisé / non voisé, c'est-à-dire à l'opposition résidu / [p,t,k] et, par ailleurs, les relations de distance à l'intérieur du triplet [p,t,k] sont précisées le long de l'axe 2. C'est, outre le résidu, le stimulus [k] qui contribue le plus à expliquer le premier axe, alors que la contribution du percept [t] est la plus importante sur l'axe 2. Dans le plan factoriel (fig. 2), on constate, que les distances entre [p], [t] et [k] sont conformes aux données brutes du tableau 4. Pour que [t] et [k] soient, très proches en stimuli (points 6 et 7) par rapport au percept [p] (point 1) alors qu'ils sont éloignés en percepts par rapport au stimulus [p] (distance des points 2 et 5 deux fois plus grande que celle des points 3 et 5), il faut que le percept [k], forme forte, se rapproche du stimulus [p] tandis que le percept [t], forme faible, s'éloigne d'autant. De plus, le percept (respectivement le stimulus) [k] se retrouve bien à égale distance des stimuli (resp. des percepts) [p] et [t]; et plus précisèment en position médiane, car l'alignement du triplet [p,t,k] est la conséquence de la domination de [k] qui inhibe l'attraction de [p] sur [t] dans ces conditions de masquage par du bruit.

Avec les confusions intervocaliques de Van der Kamp & Pols, les trois premiers axes factoriels rendent compte respectivement de 31 %, 29 % et 21 % (soit un total de 81 %) de la variance des données. On retrouve des représentations triangulaires, l'axe 1, expliqué par l'opposition [i] / [u], renvoyant à F2 et l'axe 2, avec l'opposition [i] / [a], à F1. Le troisième axe, généralement difficile à interpréter dans ce type d'analyse, renvoie clairement à la voyelle centrale [oe] qui focalise les réponses d'indécision. Dans le premier plan factoriel (fig. 3a), on constate, outre les positions extrêmes des voyelles cardinales [i,a,u], le regroupement des autres voyelles en deux grandes classes [e,y,ø,oe] et [ɑ,ɛ,o,ɔ]. La présence de [e] avec les voyelles frontales arrondies et celle de [ɛ] avec les voyelles postérieures sont vraisemblablement dûes aux nombreuses confusions de ces voyelles avec la catégorie correspondante et aussi, pour [ɛ], avec [a]. La représentation dans le plan (2,3) (fig. 3b) fait de la voyelle [oe] une voyelle extrême, ce qui n'est pas étonnant puisque les confusions la place en position de forme forte. On retrouve, par ailleurs comme dans le plan (1,2), le regroupement de [ɛ] avec les voyelles postérieures.

Les déplacements stimuli ----> percepts indiquent les effets de l'asymétrie, traduite par les distributions différentes des confusions pour les profils stimuli et percepts. Par exemple, pour la voyelle [i], le percept présente un profil plus central que le stimulus : [i] se centralise, alors qu'en revanche [a] s'excentre, et que [ɛ] et [o] se postériorisent. En particulier, les plans (1,2) et (2,3) nous montrent les échanges mutuels entre [ɛ] et [ɔ] et les déplacements respectifs de [o], [ɔ] et [ɑ].

L'analyse des stimuli visuels conduit à une représentation structurelle en deux dimensions pour les voyelles comme pour les consonnes. L'axe 1 explique 69 % de la variance des données pour les voyelles et oppose [i] à [y]. L'axe 2, rendant compte de 27 % de la variance, permet de séparer les contextes [ʃ] et [s]. Pour les consonnes, l'axe 1, avec 84 % de variance expliquée, différencie [s] et [ʃ] et l'axe 2 les contextes [y] et [i] avec un score de 14 %.

Dans le plan factoriel associé aux stimuli vocaliques (fig. 4a), la forme forte [y] en contexte [s] présente un percept dont le profil a une nette composante [y] en contexte [ʃ]. Or les [y] coarticulés avec [ʃ] n'ont plus de trace perçue de [ʃ], ils sont donc entièrement absorbés par la forme forte, celle de [y] avec [s] qui est, en fait, la réalisation vocalique la plus neutre du point de vue articulatoire, peu influencée par la consonne. Ceci se traduit par la confusion des projections du percept [sy] (point 4) et du stimulus [y] en contexte [ʃ] (point 6). Ce sont d'ailleurs ces projections qui participent le plus à la construction de l'axe 1.

Pour les consonnes, quelle que soit la coarticulation avec les voyelles, les sujets reconnaissent une forme "prototypiquement" labialisée, soit [ʃy]. Dans la figure 4b, où le percept [ʃy] participe le plus activement à la construction de l'axe 1, tandis que l'axe 2 est le mieux expliqué par le percept [sy], on observe un regroupement du couple (2,6) stimulus / percept dominant [ʃy] et du stimulus [ʃi] (5). Dans le tableau 3b, en effet, ces trois profils, malgré la dissymétrie des confusions mutuelles de [ʃi] et de [ʃy], sont très proches par opposition aux autres.

C'est ici le couple stimulus / percept [ʃi] qui présente le plus grand déplacement, bien que n'étant ni forme forte ni forme faible. En fait, son apparente symétrie cache des échanges compensatoires avec les percepts [si] et [ʃy], le percept [ʃi] ayant une forte composante [si] qui le rapproche du stimulus [si] (d'où la proximité des points 1 et 7). Ces déplacements traduisent très exactement les données dissymétriques des deux matrices de confusions.

CONCLUSION

Les dissymétries fréquemment constatées dans les matrices de confusions entre stimuli de parole semblent loin d'être fortuites. Il est possible, dans certains cas, de les attibuer à un biais expérimental, le sujet ayant tendance, par exemple, à identifier un segment vocalique de 100 ms comme une voyelle brève plutôt que comme une voyelle longue. Plus généralement, Janson [77] et Lonchamp [78] ont avancé une explication de nature linguistique rendant compte du fait que les stimuli choisis pour l'expérience ne sauraient être idéalement représentatifs des réalisations moyennes dans la langue -le sujet s'attendant, en fonction d'une référence intériorisée pour chaque voyelle, à un timbre différent de celui qui lui est proposé dans l'expérience-. Sans rejeter ces hypothèses, il est facile de remarquer qu'elles ne peuvent s'appliquer dans le cas de dissymétries obtenues avec des stimuli visuels présentés au sujet sans aucun masquage ni aucune autre manipulation. Notre propos est de souligner l'importance de l'information liée à ces dissymétries qui doit être conservée et exploitée lors du traitement des données collectées, car ces dissymétries ont des implications notables sur la compréhension des mécanismes de perception de parole. Nous avons introduit le concept de forme forte et de forme faible qui nous semble bien traduire le fait qu'un stimulus, dans des conditions données, a tendance à représenter une classe. L'analyse factorielle des correspondances, peut permettre de comparer globalement la proximité des deux ensembles stimuli et percepts dans une même représentation et, ainsi, de visualiser, avec les difficultés d'interprétation propres aux analyses multidimensionnelles, les trajets stimuli ----> percepts en cas de dissymétries.

Nous remercions C. Abry pour ses suggestions et M. Cathiard pour nous avoir permis d'utiliser ses résultats de test.

REFERENCES

[ABRY80] ABRY C., BOË L.J., CORSI P., DESCOUT R., GENTIL M. & GRAILLOT P. (1980) : Données fondamentales et études expérimentales sur la géométrie et la motricité labiales. Publication de l'Université des Langues et Lettres de Grenoble.

[BENZECRI73] BENZECRI J.P. (1973) : L'analyse des données. Tome II : L'analyse des correspondances. Dunod, Paris, 2nde Ed. 1976.

[CATHIARD88] CATHIARD M. (1988) : Identification visuelle des cibles des voyelles et des consonnes dans le jeu de la protusion/retraction des lèvres en français. Mémoire de maîtrise de Psychologie. Grenoble II.

[JANSON77] JANSON T. (1977) : Asymmetry in vowel confusion matrices. J. of Phonetics, 5, 93-96

[KLEIN70] KLEIN W., PLOMP R. & POLS L.W. (1970) : Vowel spectra, vowel spaces, and vowel identification. J. Acoust. Soc. Am., 48, 999-1009.

[KRUSKAL64a] KRUSKAL J.D. (1964) a : Multidimensional scaling by optimizing goodness of fit to the nonmetric hypothesis. Psychometrika, 29, 1-27.

[KRUSKAL64b] KRUSKAL J.D. (1964) b : Nonmetric multidimensional scaling : A numerical method. Psychometrika, 29, 115-130.

[LONCHAMP78] LONCHAMP F. (1978) : Recherches sur les indices perceptifs des voyelles orales et nasales. Thèse de IIIè Cycle, Université de Nancy II.

[MILLER54] MILLER G.A. & NICELY P.E. (1954) : An analysis of perceptual confusions among some english consonants. J. Acoust. Soc. Am., 27, 338-352.

[MILLER56] MILLER G.A. (1956) : The perception of speech. In : M. HALLE et al. (Eds) : for Roman Jakobson : essays on the occasion of his sixtieth birthday. The Hague : Mouton, 353-359.

[SHEPARD57] SHEPARD R.N. (1957) : Stimulus and response generalization : a stochastic model relating generalization to distance in psychological space: Psychometrika, 22, 325-345.

[SHEPARD58] SHEPARD R.N. (1958) : Stimulus and response generalization : deduction of the generalization gradient from a trace model. Psychological Review, 65, 242-256.

[VAN der KAMP71] VAN DER KAMP L.J. & POLS L.C.W. (1971) : Perceptual analysis from confusions between vowels. Acta Psychologica, 35, 64-77.

[WAGENAARD68] WAGENAARD W.A. (1968) : Application of Luce's choice axiom to form discrimination. Nederlands Tijdschrift voor psychologie.

	percepts						
stimuli	**p**	**t**	**k**	**b**	**d**	**g**	**somme par stimulus**
p	37	18	37	4	3	1	100
t	25	34	35	3	1	2	100
k	27	29	40	1	2	1	100
b	4	3	3	68	14	8	100
d	1	1	1	13	53	31	100
g	1	4	3	15	47	30	100
somme par percept	95	89	119	104	120	73	
D(i)	-5	-11	+19	+4	+20	-27	

Tableau 1 : Matrice de confusions interconsonantiques d'après MILLER & NICELY (54) (Rapport S/B = -12 dB, bande passante 200-6500 Hz). En ligne les stimuli, en colonne les percepts.

	percepts											
stimuli	i	e	ɛ	a	y	ø	œ	u	ɔ	o	ɑ	somme par stimulus
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	
i 12	56	6	0	2	14	4	3	3	0	0	2	90
e 13	6	9	9	3	10	13	22	3	6	4	5	90
ɛ 14	10	8	15	18	3	2	2	2	8	8	14	90
a 15	12	7	4	35	4	1	2	0	1	2	22	90
y 16	11	13	2	1	26	7	11	12	4	1	2	90
ø 17	6	10	9	1	10	22	26	2	1	2	1	90
œ 18	1	8	6	0	6	17	28	7	2	11	4	90
u 19	0	3	5	1	4	2	4	50	8	8	5	90
ɔ 20	3	0	10	0	0	4	8	10	31	4	10	90
o 21	5	3	6	11	3	6	5	5	18	10	18	90
ɑ 22	6	4	7	21	3	7	1	5	6	7	23	90
somme par percept	116	71	73	93	88	85	112	99	85	67	106	
D(i)	+26	-19	-17	+3	-2	-5	+22	+9	-5	-23	+16	

Tableau 2 : Matrice de confusions pour les voyelles du hollandais d'une durée de 8ms (1 période)
D'après KAMP & POLS (1971)

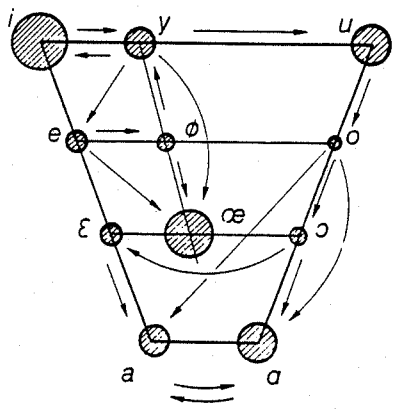


Figure 1 : Schématisation des transferts entre les voyelles: des formes faibles vers les fortes. (D'après la matrice de confusions de Van der Kamp & Pols (71))

	syllabes (percepts)				
voyelles (stimuli)	ʃi	ʃy	si	sy	somme par stimulus
	1	*2*	*3*	*4*	
ʃi 5	57	19.5	12	11.5	100
ʃy 6	8	19.5	10	62.5	100
si 7	41.5	2	51	5.5	100
sy 8	5	8	24	63	100
somme par percept	111.5	49	97	142.5	
D(i)	+11.5	-51	-3	+42.5	

Tableau 3a : Matrice de confusions pour les visèmes vocaliques [i] et [y] en contexte [ʃ] et [s] et identifiés comme syllabes [ʃi,ʃy,si,sy] d'après CATHIARD (88).

	syllabes (percepts)				
consonnes (stimuli)	ʃi	ʃy	si	sy	somme par stimulus
	1	*2*	*3*	*4*	
ʃi 5	24	66.5	0.5	9	100
ʃy 6	9.5	80	1	9.5	100
si 7	33	4	56.5	6.5	100
sy 8	17	8	41	34	100
somme par percept	83.5	158.5	99	59	
D(i)	-16.5	+58.5	-1	-41	

Tableau 3b : Matrice de confusions pour les visèmes consonnantiques [s] et [ʃ] en contexte [i] et [y] et identifiés comme syllabes [ʃi,ʃy,si,sy] d'après CATHIARD (88).

	percepts				
stimuli	p	t	k	t	k
p	37	18	37	61	83 p
t	25	34	35		86 t
k	27	29	40		

Tableau 4 : Confusions, Sous matrice de Tab. 1

t	k	
83	86	p
	92	t

Tableau 6 : Similarités, Résultat réduit au triplet [p,t,k] de la symétrisation de Tab. 1 par la méthode HOUTGAST

t	k	
34	28	p
	16	t

Tableau 7 : Dissimilarités, Résultat réduit au triplet [p,t,k] de la symétrisation de Tab. 1 selon notre proposition

Tableau 5 : Similarités, Résultat de la symétrisation de Tab. 4 par la méthode de SHEPARD (moyenne arithmétique)

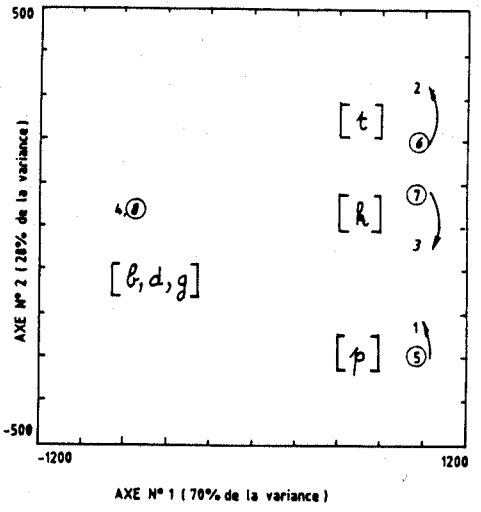


Figure 2 : Analyse des correspondances des données de confusions interconsonantiques reduites (Tab. 1) de Miller & Nicely (54). Plan (1, 2). Les percepts, respectivement les stimuli, [p,t,k] sont repérés par les identificateurs 1 à 3 (resp. 5 à 7), le résidu [b,d,g] est désigné par le couple 4/8.

Figure 3a

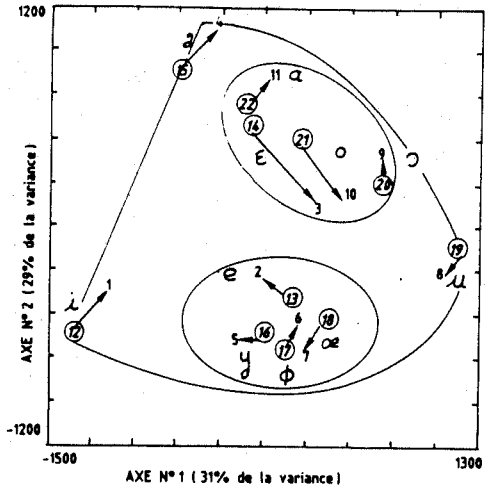
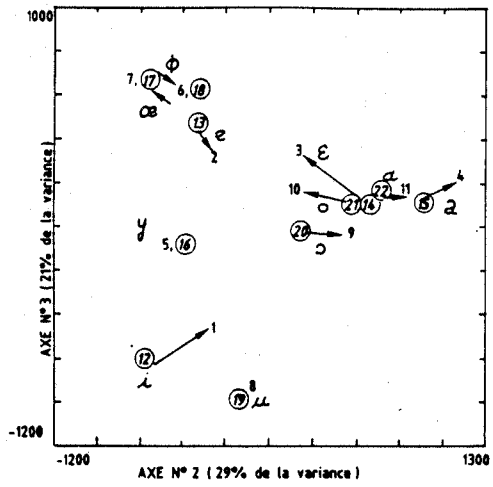


Figure 3b



Figures 3a & 3b : Analyse des correspondances des données de confusions intervocaliques de Kamp & Pols (71). Plan (1, 2) : fig. 3a et plan (2, 3) : fig. 3b. Les percepts sont repérés par les identificateurs 1 à 11, les stimuli par les identificateurs 12 à 22 (voir Tab. 2).

Figure 4a

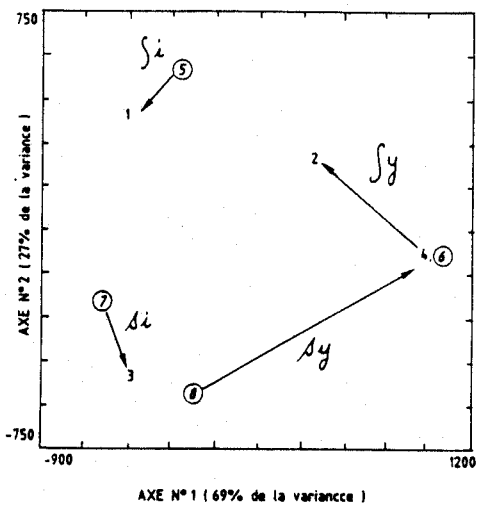
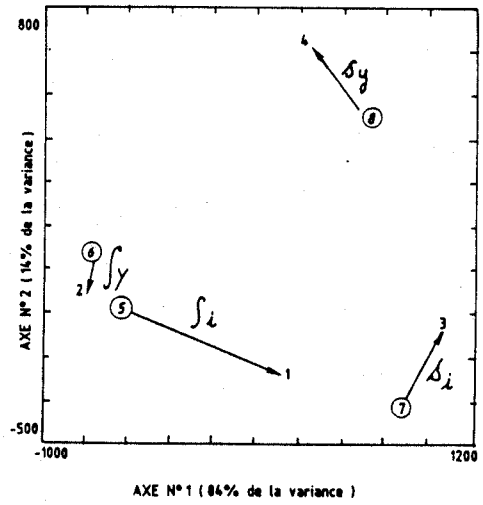


Figure 4b



Figures 4a et 4b : Analyse des correspondances des données d'identification des cibles visuelles des voyelles (fig. 4a) et des consonnes (fig. 4b) de Cathiard (88). Plan des deux premiers facteurs. (Identificateurs : voir Tab. 3).

LE DOUBLE CODAGE DE L'INFORMATION SPECTRALE

F. LONCHAMP

Institut de Phonétique Université de NANCY II
B.P. 33-98 54015 NANCY cedex

Les résultats d'un nombre significatif d'expériences de perception s'interprètent de manière satisfaisante lorsque l'hypothèse d'un double codage des informations spectrales par le système auditif est envisagée. Un premier codage fondé sur le taux de décharge moyen semble bien adapté à la représentation d'informations brèves, non périodiques, et dont le spectre global seul est significatif (barre d'explosion par exemple). Un second codage fondé sur la périodicité des décharges liée à l'évolution temporelle des composantes spectrales paraît plus efficace pour mesurer la composition fréquentielle exacte de sons longs et stables, à structure formantique.

1 - BASES NEUROPHYSIOLOGIQUES DU CODAGE SPECTRAL PAR L'OREILLE.

Ce paragraphe est un rappel succinct d'une partie de nos connaissances actuelles sur la neurophysiologie du nerf auditif. On sait que pour les fréquences inférieures à 4 - 5 kHz une fibre nerveuse auditive répond à une stimulation sonore en manifestant deux caractéristiques. On observe d'une part une augmentation de son taux de décharge moyen d'impulsions nerveuses au delà du taux spontané. Cette augmentation se produit pour des sons d'autant plus faibles qu'ils sont proches de la fréquence de plus grande sensibilité de la fibre (CF : fréquence caractéristique ou centrale). D'autre part, ces impulsions nerveuses sont produites à des instants particuliers, correspondant à une configuration déterminée de la forme de la membrane basilaire en vibration. On parle alors de verrouillage de phase ("phase locking") pour cette analyse du taux de décharge dit instantané, c'est-à-dire à l'échelle d'une période. (cf. par ex. DELGUTTE & KIANG 1984). Le verrrouillage de phase n'est possible qu'à des fréquences inférieures à 3 - 5 KHz en raison de la durée de la période réfractaire et de l'incertitude temporelle sur l'instant de déclanchement.

Un premier aspect capital du taux de décharge moyen est la présence d'une dynamique limitée. Pour chaque fibre, le taux moyen ne varie que pour une gamme limitée d'intensité, de l'ordre de 20 à 30 dB. Au dessous et dessus de deux seuils variables pour chaque fibre, le taux reste constant, présentant donc une saturation non linéaire. DELGUTTE (1987) a néanmoins montré que si l'on tient compte de trois populations de fibres distinguées par leur taux de décharge spontané, qui corrèle avec le seuil de déclanchement et de saturation, on pouvait, sous certaines hypothèses, prévoir les caractéristiques auditives de la discrimation d'intensité. Cette démonstration rend moins probante la conclusion de SACHS & YOUNG (1979) et YOUNG & SACHS (1979) : seule l'utilisation de la réponse verrouillée en phase, ou taux de décharge synchrone, permet au système auditif de disposer d'un spectre interne possédant des maxima spectraux aux fréquences formantiques quelle que soit l'intensité des stimuli sonores. En effet, à l'échelle d'une période, les fibres nerveuses répondent vigoureusement dans une large plage d'intensité à une composante intense de fréquence proche de leur fréquence caractéristique. Il faut néanmoins noter une détérioration pour les intensités les plus fortes. Cette réponse se produit à un moment déterminé du battement de la membrane basilaire. On constate expérimentalement que plus l'intensité d'une composante de fréquence proche de la fréquence caractéristique d'une fibre est grande, plus l'intervalle entre les impulsions est proche d'un multiple entier de la période de la composante. Le degré de synchronisation est donc une mesure indirecte de l'amplitude d'une composante sonore.

Mais le taux de décharge moyen possède une autre caractéristique essentielle. La plage d'intensité capable de moduler ce taux est bien plus grande dans les 10 - 30 ms qui suivent l'apparition d'une composante spectrale (SMITH 1979, SMITH & BRACHMAN 1980, DELGUTTE 1980, par ex.). Le taux présente ensuite une adaptation, qui se traduit par un écrasement des taux de décharge en fonction des différences d'amplitude, et par une saturation réduisant la gamme utile d'analyse de l'intensité. DELGUTTE (1980) montre en particulier que la présence du formant de basse fréquence d'une consonne nasale adapte fortement la réponse des fibres au spectre de la voyelle suivante. Cette adaptation est plus forte pour la zone de fréquence qui a été stimulée par le formant nasal.

2 - CODAGE D'INTENSITE ET CODAGE TEMPOREL.

Intuitivement, les deux mécanismes de codage spectral possèdent des caractéristiques qui les rendent efficaces dans des situations différentes. Le codage par le taux de décharge moyen, que nous appellerons codage d'intensité, semble bien adapté au codage d'informations brèves, non périodiques et qui ne nécessitent pas une grande précision d'analyse. Ce pourrait être le cas des barres d'explosion des occlusives et des bruits fricatifs. Son efficacité est

maximale lorsque le stimulus sonore présente une forte discontinuité d'amplitude, bien que le codage d'intensité soit sensible à la saturation si l'intensité atteinte est trop forte. Au contraire, le codage que nous appelons temporel, fondé sur la capacité de la fibre à répondre en synchronie avec l'évolution temporelle de la composante, paraît plus efficace pour coder la fréquence exacte et l'amplitude des composantes spectrales périodiques, quelle que soit l'intensité globale. Ce codage sera d'autant plus efficace que la durée d'accumulation des informations sera plus longue. Ce mode de codage paraît indispensable pour atteindre la précision de mesure que révèlent les expériences de discrimation de la hauteur tonale et très utile pour le positionnement exact du premier formant auditif à partir du niveau précis des deux harmoniques proéminentes.Il n'est bien sûr pas possible de savoir si ces deux mécanismes peuvent être utilisés simultanément, mais nous faisons l'hypothèse qu'en première approximation, ces traitements sont mutuellement exclusifs.

Notons que l'hypothèse d'un double codage n'est pas nouvelle. Elle joue notamment un rôle important dans les traitements des sorties du modèle d'oreille de CAELEN (1979, 1985). Mais si l'hypothèse de base est la même, de par son fondement physiologique, les visées sont différentes. CAELEN s'intéresse à la structuration détaillée et à très court terme des informations dans l'espace tonotopique de la membrane basilaire et dans le temps, insistant particulièrement sur l'importance de la phase et des formes élémentaires prises par la circulation des "événements" sur ses 'neurogrammes'. Pour nous, en revanche, le double codage est, comme l'a noté un relecteur anonyme, un mécanisme "optimal" (rapport signal / bruit maximal) s'adaptant à deux classes de signaux aux caractéristiques très différentes.

3 - POLS & SCHOUTEN (1978, 1981) : LE ROLE D'UN PREFIXE DE BRUIT DANS LA PERCEPTION DES OCCLUSIVES.

Le modèle que nous proposons a été conçu à l'origine pour rendre compte des résultats de OHDE & SCHARF (1977, 1981) et POLS & SCHOUTEN (1978, 1981). Ces études, ainsi que plusieurs travaux antérieurs, montraient que l'identification des syllabes voyelle + occlusive était meilleure que l'identification des syllabes occlusive + voyelle. Pour rendre les conditions comparables, la barre d'explosion et le bruit de friction étaient supprimés, et la supériorité des groupes VC ne pouvait être due qu'à un traitement plus efficace des transitions vocaliques. Or les transitions sont spectralement au moins aussi marquées dans les groupes CV que dans les groupes VC, ce qui rend difficile une interprétation de cette asymétrie.

POLS & SCHOUTEN (1978), à partir de l'idée intuitive que la brusque transition d'amplitude des groupes CV peut produire une sorte de "clic" auditif, d'ailleurs inaudible objectivement, font précéder le stimulus d'un bruit relativement faible. Les scores d'identification augmentent alors fortement. OHDE & SCHARF (1981) obtiennent également une légère amélioration en utilisant une fenêtre initiale de 6ms. SUMMERFIELD & ASSMANN (1987) rapportent qu'ils ont également observé d'une manière informelle une augmentation de l'intelligibilté de groupes CV dans du bruit blanc lorsque celui-ci débute quelques centaines de ms avant la syllabe. Cette avantage n'existe pas lorque le bruit commence en même temps que le début de la syllabe. SCHEFFERS (1983) a mis en évidence le même avantage pour des voyelles.

On peut tenter de rendre compte de ces résultats dans le cadre d'un modèle à double codage.

Dans le cas d'une brusque discontinuité d'amplitude, le système auditif privilégie, peut-être de façon automatique, le codage d'intensité. Mais les résultats de cette analyse ne sont pas très utiles dans la mesure où le spectre ne possède plus la barre d'explosion caractéristique. Ce sont les transitions seules qui sont soumises à cette analyse dont les résultats peuvent de plus être trop grossiers pour assurer une bonne identification. Si l'on admet que les deux modes de codage ne peuvent être menés en parallèle, on peut faire l'hypothèse que le codage temporel plus précis s'exercera trop tard, sur une partie du stimulus où les déflexions fréquentielles des transitions sont déjà trop réduites. La supériorité des groupes VC s'explique par le fait que le suivi des transitions par le codage temporel peut se faire sans interruption jusqu'à son terme, où réside l'information la plus pertinente. Toute l'information disponible est donc utilisée dans ce cas. La présence d'un bruit précédant la discontinuité d'amplitude ou d'une fenêtre évite peut-être que le codage d'intensité soit utilisé exclusivement, ou décale moins le point de départ de l'analyse temporelle. L'information portée par les transitions est mieux exploitée, et les scores d'identification sont meilleurs.

Nous voulons maintenant présenter quelques études qui confortent, ou précisent, certains aspects de ce modèle très fruste.

CHISTOVICH (1971) a décrit très brièvement une expérience qui s'interprète aisément si l'on fait appel à l'adaptation constatée pour le taux de décharge moyen. Elle concerne une voyelle à deux harmoniques proéminentes dans la zone de F1. Pour maintenir la voyelle sur une frontière phonétique lorsqu'on introduit l'une des harmoniques de F1 avec un retard croissant, il faut diminuer son amplitude. Tout se passe comme si une composante qui apparaît tardivement possédait une sonie supérieure, ou réciproquement, comme si la contribution d'une harmonique présente dès le début de la voyelle diminuait graduellement avec le temps. Comme il n'y a pas d'énergie à la fréquence de l'harmonique manquante, les fibres de fréquence voisine sont dans un état non adapté au moment où celle-ci est introduite, et elles répondent de façon vigoureuse. Les travaux de SUMMERFIELD & ASSMANN (1987) confirment que le système auditif privilégie les variations spectrales sur les zones stables.

COLE & SCOTT (1973) montrent que la perception de l'ordre temporel de groupes de voyelles recyclées est facilité par la présence d'une discontinuité d'énergie entre les voyelles et par la présence de transitions. Ces résultats suggèrent qu'en l'absence de variations brutales d'énergie ou de fréquence, le codage temporel se poursuit sans interruption pendant une longue durée, rendant difficile par la suite la segmentation et l'ordonnancement des informations phonétiques.

SLAWSON (1968) démontre que certains auditeurs ne traitent pas de la même façon une voyelle synthétique selon que l'attaque est douce ou brutale. Comme le contrôle de l'amplitude était placé avant le calcul des résonances, il est probable que les stimuli n'étaient pas physiquement identiques (cf. CARRE &

QUACH-TUAN (1987)), ce qui ne permet pas d'affirmer avec certitude qu'un auditeur analyse l'attaque avec un mécanisme différent de celui utilisé pour la partie stable. Il peut simplement être sensible aux différences spectrales entre les stimuli à attaque douce et brutale quel que soit le mécanisme mis en jeu. Mais CARRE (1988) montre que certains sujets répondent de manière différente à des stimuli physiquement identiques, ce qui suggère fortement la présence de deux mécanismes distints. Cette tendance est plus accusée lorsque la voyelle est brève (50 ms), c'est-à-dire lorsque le codage temporel est moins efficace. Que le timbre des voyelles à attaque brutale soit moins sensible à une variation de la fréquence fondamentale concorde avec l'idée que le codage d'intensité fournit une représentation plus proche du spectre physique, comparable à celle fournie par un filtrage large-bande, ou une analyse LPC. Il est vraisemblablement moins précis dans le cas où un formant est physiquement composé de deux harmoniques d'égale intensité, et il n'est pas, ou guère, utilisable pour des variations spectrales sans discontinuité d'amplitude.

4- UNE FORTE INTENSITE DETERIORE L'IDENTIFICATION DES SONS

De nombreuses études confirment que la précision de l'identification et de la discrimination des sons diminue lorsque le niveau sonore augmente. Mais il n'est pas facile de déduire des résultats lequel des mécanismes est responsable de la détérioration. SYRDAL-LASKY (1978) conclut qu'à 92 dB, l'identification et la discrimination du lieu d'articulation d'occlusives sourdes et sonores en synthèse à 2 formants est de moindre qualité. L'absence de F1 au début des occlusives sourdes ("F1 cutback") permet d'éliminer l'hypothèse d'une extension de masquage important de F1 vers les transitions de F2 ("upward spread of F1 masking"). Il est intéressant de constater que l'addition d'un bruit de fond de 60 dB améliore les performances. Ce résultat doit selon nous être rapproché de celui de POLS & SCHOUTEN. VAN TASSELL & CRUMP (1981) note une réduction générale de l'efficacité de la transition de F3 à séparer [da] de [ga] quand le niveau passe de 60 à 100 dB. La pente d'identification est plus faible et non monotone, et les stimuli extrêmes ne sont plus identifiés qu'à 80%. Les auteurs incriminent l'extension du masquage de F1 et F2 vers les fréquences supérieures, hypothèse rejetée par SYRDAL-LASKY pour ses résultats. DORMAN & DOUGHERTY (1981), pour des syllabes synthétiques à deux formants de forme occlusive sonore+voyelle, confirment que les fonctions d'identification sont très perturbées à 90 dB relativement à celles obtenues à 55 et 70 dB. Les détails des modifications ne correspondent pas à ceux des études précédentes, ce qui est pour nous le signe d'une simple détérioration de la qualité de l'information spectrale conduisant à des réponses aléatoires.

Les études de DORMAN & al. (1986, 1987) confirment que l'intelligibilité de voyelles brèves (50 ms) diminue à partir d'une intensité de 90 - 95 dB SPL. Pour des voyelles plus longue, le réflexe stapédien, d'une latence supérieure à 50 ms, intervient pour réduire au niveau de l'oreille moyenne l'intensité du son transmis, ce qui assure une bonne intelligibilité jusqu'à 110 dB SPL. Les auteurs rappellent que les sujets ayant subi une stapedectomie, c'est-à-dire une section du muscle stapédien qui joue un rôle essentiel dans le réflexe acoustique, montrent également une diminution de l'intelligibilité des voyelles vers 90 dB (BORG & ZAKRISSON 1973). Les voyelles ouvertes sont plus atteintes que les voyelles fermées (DORMAN & al. 1986). Le sens des confusions semble indiquer que F1 décroît et F2 croît avec l'augmentation de l'intensité. CARRE (1988) constate également aux niveaux élevés une distorsion de la courbe d'identification, qu'il attribue à un effet de saturation d'origine neurophysiologique.

Cette diminution de l'intelligibilité à des intensités élevées est compatible avec l'hypothèse que le codage des voyelles est fondé pour une part sur le taux de décharge moyen. Mais il est également probable que le verrouillage de phase devient moins précis aux intensités élevées. S'il est permis de comparer des expériences différentes, la détérioration de l'identification des occlusives à 92 dB est déjà bien établie, alors que celle des voyelles ne fait que commencer, ce qui pourrait pointer vers une pondération différente des mécanismes de codage dans l'identification , ou la discrimination, des voyelles et des occlusives. Mais ces tâches exigent peut-être des degrés différents de précision pour les voyelles et les consonnes, ce qui rendrait la comparaison invalide.

5 - LE CODAGE DE L'ATTAQUE : LES EXPERIENCES DE LACERDA (1987)

Certaines expériences décrites par LACERDA (1987) sont d'un grand intérêt pour tester le modèle proposé. La première compare la discriminabilité d'une variation de F2 dans des stimuli brefs (50 ms) à attaque brutale ou graduelle. Les formants F1, F3 et F4 sont identiques, et F2 est à une fréquence différente, mais stable, pour chaque stimulus. L'hypothèse testée par LACERDA est que l'attaque graduelle adapte les réponses nerveuses et conduit à une représentation auditive moins différenciée. La discrimination est donc plus difficile. Les résultats confirment cette prédiction, mais il nous semble possible d'en donner une autre explication. Tout d'abord, on peut songer à plaider que dans le cas d'une attaque graduelle, un codage temporel intrinsèquement précis est utilisé, mais que la durée du stimulus est trop faible pour que les résultats auditifs soient fiables. La comparaison avec les résultats de FLANAGAN (1955) pour des stimuli de 750 ms, identiques à ceux obtenus par LACERDA pour les stimuli à attaque graduelle, permet de rejeter cette hypothèse. On peut en revanche considérer que le sujet est capable d'utiliser les deux sources d'informations dans les sons à attaque brutale car la structure formantique est stable. Ce surcroît d'informations assure une meilleure discriminabilité.

La deuxième expérience montre qu'une variation de F2 dans la syllabe [ad] est plus discriminable que dans [da]. Elle confirme les résultats présentés plus haut (POLS & SCHOUTEN (1978, 1981)). Nous avons déjà présenté notre interprétation. L'auteur propose deux explications complémentaires et complexes : la première est fondée sur la présence d'une excitation résiduelle d'origine interne dans les syllabes VC. Cette hypothèse est testée en faisant suivre le stimulus d'un appendice formantique constant. La prédiction est que l'effet de cette adjonction doit être négligeable, car cette nouvelle information est déjà disponible ! Le

résultat est conforme aux prédictions, et l'hypothèse est considérée comme validée. On peut démontrer de la même façon que la lune possède une forte personnalité. La prédiction est que la lune refusera de changer de direction, même quand l'expérimentateur se mettra en colère. L'expérience prouve que cette prédiction est vraie, et on "démontre" ainsi que la lune possède effectivement une volonté inflexible. Précisons que notre ironie ne vise ici que le test expérimental, et non l'hypothèse. Aucun modèle ne peut être mis en difficulté par la prédiction d'un effet nul, car il est toujours possible de plaider qu'aucune information nouvelle n'est apportée, ou prise en compte par l'auditeur. La deuxième hypothèse est que la représentation auditive est perturbée par la variation du spectre pendant la durée des transitions car l'adaptation n'est pas instantanée. La discriminabilité devrait augmenter quand la transition de F2 est précédée par un "préfixe" formantique . Les résultats le confirment, ce que notre modèle peut prévoir si on remarque que lorsque le préfixe est soumis au codage d'intensité, une plus grande partie de la transition peut être suivie par le codage temporel. L'expérience suivante examine le rôle de la source sonore. SUMMERFIELD & al. (1985) rapportent que dans des groupes VCV de synthèse à source de bruit, la partie CV est plus discriminable de part et d'autre d'une frontière phonétique. LACERDA, pour les stimuli à spectres constants de la première expérience excités par du bruit, démontre que l'attaque graduelle ou brutale ne joue plus de rôle. Notre modèle rend sans difficulté compte de ce résultat. La possibilité d'un codage temporel étant fortement réduite par la nature non périodique du stimulus, le codage d'intensité est seul pertinent. Il apparaît que la définition spectrale ne dépend pas de la forme spécifique de l'attaque pour ce type de codage. Mais, à l'inverse également des résultats de SIDWELL & SUMMERFIELD (1986), qui portent sur des groupes CVC, la partie VC d'un groupe VCV conserve un léger avantage dans l'expérience décrite ensuite. De toute évidence, les résultats sur ces groupes plus complexes dépendent de nombreux paramètres insuffisamment maîtrisés aujourd'hui. Ils ne peuvent donc fournir un test pour un modèle. LACERDA poursuit en montrant que la direction de la pente n'est pas un paramètre pertinent. L'avantage de la séquence [ab] sur [ba] est claire. Notre modèle ne prévoit pas non plus de différence en fonction de la direction des transitions. La dernière expérience que nous mentionnerons porte sur l'effet d'une barre d'explosion. Son spectre est identique à celui de la première période de la transition. La durée de silence (16 ou 33 ms) entre l'explosion et la partie vocalique n'a presque aucun effet. Mais la présence d'une barre d'explosion améliore la discriminabilité. C'est exactement ce que prévoit notre modèle. La présence de l'explosion devant la partie vocalique permet au codage temporel de se produire plus tôt, et donc un suivi plus complet de la transition formantique.

Il apparaît que le modèle perceptif que nous proposons, fondé sur un double mécanisme de codage, ne se heurte à aucune difficulté majeure. Il permet de donner une interprétation cohérente à un ensemble d'expériences.

6 - LES EFFETS DE L'ADAPTATION : REPP (1987)

Un dernier test est fourni par la série d'expériences décrite par REPP (1987) qui portent sur l'effet du murmure nasal sur l'identification des consonnes [m] et [n]. L'hypothèse testée par REPP, et rejetée en conclusion, trouve son origine dans les résultats, déjà mentionnés, de DELGUTTE (1980) montrant que le premier formant d'une consonne nasale produit un effet d'adaptation sur le spectre neurophysiologique de la voyelle. L'adaptation modifie la partie initiale du spectre de la voyelle et la comparaison avec le spectre de la partie finale de la voyelle fournit, par différence, un indice du lieu d'articulation de la consonne. Une discussion complète de cette hypothèse est hors de propos. Nous noterons cependant avec REPP que la conséquence de l'adaptation due au murmure nasal doit être de réduire l'amplitude spectrale de la voyelle aux fréquences basses. Il peut en résulter une plus grande "audibilité" des fréquences élevées où se concentrent les indices du lieu d'articulation. Concernant notre modèle, notre analyse est la suivante. La présence du murmure nasal, comme le bruit pré-consonantique de POLS & SCHOUTEN (1978) améliore les performances d'identification en permettant un meilleur suivi des transitions vocaliques. Cette prédiction est directement testée par REPP dans la cinquième expérience où il utilise une fenêtre identique à celle de OHDE & SCHARF (1981) sur la partie vocalique de la syllabe. Un effet faible, mais statistiquement significatif, est obtenu. Cette augmentation ne se produit que lorsque la partie vocalique est perçue comme nasale, impliquant la présence d'une consonne nasale éliminée. Elle ne concerne donc que la distiction [m]-[n], et non les stimuli perçus comme étant précédés de [b]-[d]. Nous sommes d'accord avec REPP pour reconnaître que ce fait n'est pas de prime abord en accord avec le rôle attribué à un préfixe par POLS & SCHOUTEN, et surtout par nous même qui prévoyons un suivi de formant plus efficace qui devrait améliorer l'identification du lieu d'articulation dans tous les cas. Mais on peut incriminer un biais de réponse en faveur de [b], comme le fait REPP ailleurs. Ce biais de réponse proviendrait de l'absence de barre d'explosion et non d'une perception différente des transitions. Nous ne somme pas d'accord avec REPP quand il conclut que l'effet de la fenêtre n'explique pas pourquoi des scores d'identification aléatoires d'un murmure et d'une zone vocalique peuvent s'améliorer quand les deux parties sont réunies. Rappelons notre argumentation. Le murmure nasal peut contenir peu d'information phonétique. La partie vocalique seule est mal identifiée car le suivi des formants est rendu difficile par le brusque saut d'intensité qui déclenche un codage d'intensité. La présence du murmure nasal devant la partie vocalique permet un meilleur suivi des formants.

Dans la première expérience de REPP, inverser l'ordre de la partie vocalique et du murmure nasal détériore les scores d'identification comme dans le cas des occlusives non précédées d'un bruit, ce qui est également conforme aux prédictions de notre modèle. La deuxième expérience présente en écoute dichotique , c'est-à-dire dans une oreille différente, le murmure nasal et la partie vocalique. L'adaptation, d'origine périphérique, doit disparaître, et la fiabilité de l'identification diminuer. Les résultats le confirment. Notre modèle fait la même prédiction si on admet que le déclenchement de l'analyse d'intensité est déterminé par une détection unilatérale, donc périphérique, d'une brusque augmentation d'intensité globale. Si l'on sépare le murmure de la partie vocalique par une durée de silence, les résultats d'identification doivent décroître lorsque la durée du silence est augmentée. Le modèle de REPP,

comme le nôtre, prévoit une diminution plus rapide que celle constatée (5 - 7 % en moyenne), qui reste faible même pour une durée de silence de 240 ms. Pour conserver notre hypothèse, il faut admettre que le codage temporel commencé sur le murmure n'est guère perturbé par un silence de plus de 100 ms suivi d'une brusque variation d'énergie. Si ceci est vrai, la barre d'explosion d'une consonne occlusive intervocalique n'a pas d'effet bénéfique sur le suivi de formant subséquent. Elle ne contribue qu'à la mesure de sa propre distinctivité. L'expérience décrite ensuite est difficile à commenter en quelques lignes. La durée du murmure nasal ne joue qu'un rôle mineur de 10 à 60 ms, ce qui est compatible avec notre modèle, qui prévoit un bon suivi des formants par le codage temporel si le codage d'intensité n'est pas déclenché. Il ne fait aucune prédiction sur le rôle d'un murmure nasal extrait d'un contexte différent de celui de la partie vocalique suivante. De manière complexe, cet effet peut être prévu par l'effet d'adaptation. Les expériences 6 et 7 testent directement l'hypothèse d'adaptation en affaiblissant par filtrage la zone de basse fréquence de la partie vocalique. Cette hypothèse prévoit une amélioration des scores d'identification en rendant plus "audibles" les indices de haute fréquence. Or les résultats vont en sens contraire. La détérioration des scores est légèrement plus marquée pour un filtrage de durée brève (10 ms). Notre modèle ne prévoit clairement ni amélioration, ni détérioration de l'identification, et nous ne souhaitons pas proposer une explication ad hoc. Ces expériences conduisent REPP à rejeter l'hypothèse d'un effet bénéfique de l'adaptation sur le mécanisme d'identification phonétique.

7 - CONCLUSIONS

En conclusion, le modèle incorporant un double mécanisme de codage, fondé sur la mesure physiologique du taux de décharge neural moyen et sur la capacité des fibres à répondre en synchronie avec l'évolution temporelle des composantes spectrales, nous paraît compatible avec une série de résultats expérimentaux, le plus important étant l'augmentation des scores d'identification quand une information cruciale portée par une évolution formantique n'est pas présentée juste après une brusque variation d'intensité. Le point le plus délicat demeure le rôle facilitant que l'on doit attribuer à un "préfixe" comme le murmure nasal placé jusqu'à 240 ms devant la partie vocalique à forte variation initiale d'énergie.

***

BORG E. & ZAKRISSON J. (1973) "Stapedius reflex and speech features", JASA, 54, 525 - 527.

CAELEN J. (1979) Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique., Thèse D.E., Toulouse.

CAELEN J. (1985) "Space/time data-information in the ARIAL project ear model", Speech Communication, 4, 163 - 179.

CARRE R. (1988) "Rôle de la fréquence fondamentale dans la perception des voyelles de synthèse", communication personnelle, à paraître.

CARRE R. & QUACH-TUAN (1987) "Effects of non-stationary characteristics on the perception of the vowels", Bull. Lab. Comm. Parlée Grenoble, 1, 307 - 318.

CHISTOVICH L.A. (1971) "Auditory processing of speech stimuli - evidence from psychoacoustics and neurophysiology", Proc. 7th Int. Cong. Acoust., Akademiai Kiado, Budapest, 1, 27 - 42.

COLE R.A & SCOTT B. (1973) " Perception of temporal order in speech : the role of vowel perception", Canadian Journal of Psychology, 27, 441 - 449.

DELGUTTE B. (1980) "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers", JASA, 68, 848 - 857.

DELGUTTE B (1987) "Peripheral auditory processing of speech information : implications from a physiological study of intensity discrimination", in SCHOUTEN M.E.H.(ed.), Psychophysics of speech perception, Martinus Nijhoff, Dordrecht, 333 - 353.

DELGUTTE B. & KIANG N.Y.S. (1984) "Speech coding in the auditory nerve : I. Vowel-like sounds", JASA, 75(3), 866 - 878.

DORMAN M.F., CEDAR I., HANNLEY M.T., LEEK M.R. & LINDHOLM J.M. (1986) "Influence of the acoustic reflex on vowel recognition", JSHR, 29, 420 - 424

DORMAN M.F. & DOUGHERTY K. (1981) "Shifts in phonetic identification with changes in signal presentation level", JASA, 69(5), 1439 - 1440

DORMAN M.F., LINDHOLM J.M., HANNLEY M.T & LEEK M.R (1987) "Vowel intelligibility in the absence of the acoustic reflex : performance-intensity characteristics", JASA, 81(2), 562 - 564.

FLANAGAN J.L. (1955) "A difference limen for vowel formant frequency", JASA, 27, 613 - 617.

LACERDA F. (1987) Effects of peripheral auditory adaptation on the discrimination of speech sounds, Ph.D, U. de Stockholm, PERILUS, 6.

ODHE R.N. & SCHARF D.J. (1977) "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification", JSHR, 20, 543 - 554.

ODHE R.N; & SCHARF D.J. (1981) "Stop identification from vocalic transition plus vowel segments of CV and VC syllables : a follow-up study", JASA, 69(1), 297 - 300.

POLS L.C.W & SCHOUTEN M.E.H. (1978) "Identification of deleted consonants", JASA, 64, 1333 - 1337.

POLS L.C.W. & SCHOUTEN M.E.H. (1981) "Identification of deleted plosives : the effect of adding noise or applying a time window (A reply to OHDE & SCHARF 1981)", JASA, 69(1), 301 - 303.

REPP B.H. (1987) "On the possible role of auditory short-term adaptation in perception of the prevocalic [m]-[n] contrast", JASA, 82(5), 1525 - 1538.

SACHS M.B. & YOUNG E.D. (1979) "Encoding of steady-state vowels in the auditory nerve : representation in terms of discharge rate", JASA, 66, 470 - 479.

SCHEFFERS M.T.M. (1983), I.P.O., manuscript 450/II (cité in SUMMERFIELD & ASSMANN 1987)

SIDWELL A. & SUMMERFIELD Q. (1986) "The auditory representation of symmetrical CVC syllables", Speech Communication, 5, 283 - 297.

SLAWSON A.W. (1968) "Vowel quality and musical timbre as functions of spectral envelope and fundamental frequency", JASA, 43, 87 - 101.

SMITH R.L. (1979) "Adaptation, saturation and physiological masking in single auditory-nerve fibers", JASA, 65, 166 - 178.

SMITH R.L. & BRACHMAN M.L. (1980) "Operating range and maximum response of single auditory-nerve fibers", Brain Research, 184, 499 - 505

SUMMERFIELD Q. & ASSMANN P. (1987) "Auditory enhancement in speech perception", in SCHOUTEN M.E.H. (ed.), Psychophysics of speech perception, Martinus Nijhoff, Dordrecht, 140 - 150.

SUMMERFIELD Q., FOSTER J., TYLER R. & BAILEY J. (1985) "Influence of formant bandwidth and auditory frequency selectivity on the identification of place of articulation in stop consonants", Speech Communication, 4, 213 - 229.

SYRDAL-LASKY A. (1978) "Effects of intensity on the categorical perception of stop consonants and isolated second formant transitions", P&P, 23(5), 420 - 432.

VAN TASSELL D.J. & CRUMP E.S.A. (1981) "Effects of stimulus level on perception of two acoustic cues in speech", JASA, 70(5), 1527 - 1529.

YOUNG E.D. & SACHS M.B. (1979) "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers", JASA, 66, 1381 - 1403.

***

Reconnaissance

LE DECODEUR ACOUSTICO-PHONETIQUE DANS LE PROJET DIRA

J. Caelen, H. Tattegrain

ICP/INPG - UA CNRS
46, Av. F. VIALLET
38031 GRENOBLE CEDEX

ABSTRACT

The DIRA system (Integrated Dialogue for an Advanced Robot) is a multi-expert system for the continuous speech recognition. It is made up of an acoustic-phonetic decoder according to architecture of the whole system, i.e. planning and blackboard architecture. This decoder runs by calling of planner, on the hypothesis/test fashion. However, it is fundamental to clearly distinguish the two acoustic and phonetic levels, by modelling the phonetic macro-structure (by phonetic nets for instance) and the acoustic micro-structure (by acoustic cues adapted to the speaker). Then, the decoding becomes an knowledge independent matching process. Moreover, it is very easy to represent phonetic nets with production rules in Prolog language.

The paper is a discussion about all this problems.

INTRODUCTION

Le rôle du DAP (Décodage Acoustico-Phonétique), bien que fondamental en reconnaissance automatique de la parole, reste encore mal défini à l'heure actuelle. Sa mauvaise formulation tient en partie au fait qu'il s'agit de projeter un sous-espace --celui des observations acoustiques-- dans un espace plus vaste --celui des formes phonétiques et pour une autre partie au fait que son articulation vis-à-vis des niveaux linguistiques est difficile à formuler. Dans un précédent article [Caelen, 87], nous avions mis l'accent sur les ruptures substance/forme et continu/discret que l'on trouve entre les signaux acoustiques et les unités phonétiques. Cette rupture interdit de considérer le DAP comme une suite de transformations d'un domaine de représentation (continu) dans un autre (discret): en fait deux structures pré-existent et le DAP est une mise en correspondance de la micro-structure acoustique du signal et de la macro-structure phonétique (souvent implicite). Ceci n'est rendu possible que si ces structures sont suffisamment explicitées. La mise en correspondance peut alors se faire par "matching" des modèles sous-jacents aux deux structures. Signalons que les méthodes fondées sur les HMM (Hidden Markov Model) ont une vision plus claire sur les modèles de référence que les méthodes dérivées des SE (Systèmes Experts) dans lesquelles les modèles des structures phonétiques sont généralement noyés dans les règles de production dont l'organisation dépend plus de la stratégie d'utilisation que des connaissances elles-mêmes. On peut reprocher en contre-partie aux HMM de ne pas offrir des modèles "explicables" puisqu'ils sont hérités de techniques d'apprentissage aveugles et d'interdire --sauf au prix d'une coûteuse fragmentation-- la prise en compte de plusieurs sources de connaissances, articulatoire et perceptive par exemple. C'est pourtant en prenant en compte toutes les sources de connaissance, que l'on arrivera à plonger le sous-espace des observations acoustiques dans l'espace des formes phonético-phonologiques --ce que fait d'ailleurs un expert en lecture de sonagrammes-- pour restaurer l'information manquante (ou cachée) lors de cette projection.

Ces deux approches (SE et HMM) ne sont pas inconciliables dès lors que sont clarifiées les notions de "modèles phonétiques de référence" et de "matching". Dans une perspective qui utilise des connaissances explicites, plus précisément une technique de SE (ou plus largement d'IA), nous sommes amenés à définir les modèles résultant des macro-structures phonétiques sous forme de réseaux. Les transitions entre les états sont formulées à l'aide de connaissances et de règles idoines et le matching conduit à identifier ces macro-structures sur la micro-structure du signal (ou inversement) c'est-à-dire à parcourir convenablement les réseaux. Il en résulte un autre avantage qui réside dans la nécessaire séparation des connaissances sur ces deux structures: (a) les connaissances acoustiques qui restent très attachées aux procédures de calcul des paramètres (indices, corrélats, etc.) et (b) les connaissances phonétiques qui sont véhiculées essentiellement par les traits (informations symboliques pris au sens le plus large). Le matching peut alors être un processus tout à fait général et indépendant des connaissances mises en oeuvre.

Le DAP revient donc soit à prédire la macro-structure à partir de la micro-structure soit à vérifier l'existence d'une micro-structure dans une macro-structure donnée, ce qui peut être schématisé de la manière suivante (Fig. 1):

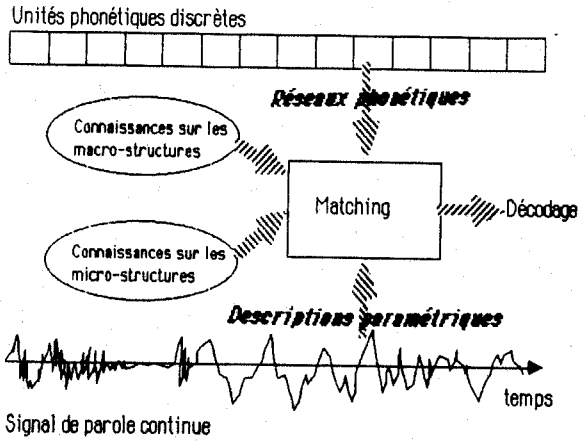


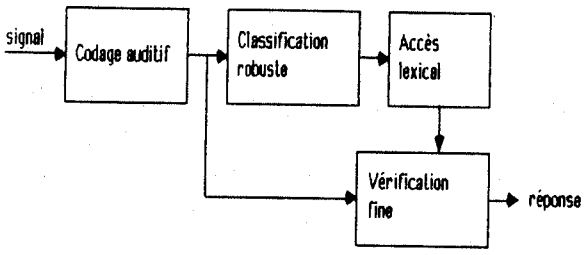
Fig. 1: Le processus du décodage acoustico-phonétique: un modèle phonétique est donné a priori et sa structure est décrite dans les réseaux phonétiques; d'un autre côté le signal a une description paramétrique connue. Des connaissances sur la macro-structure phonétique et la micro-structure acoustique sont rangées dans deux bases qui servent aux processus de "matching" pour en assurer la correspondance. Le résultat produit peut être considéré comme le résultat de décodage.

Dans cette vision du DAP on dépasse la notion de stratégie ascendante ou de stratégie descendante. Certes si les unités phonétiques sont connues a priori par le décodeur le "matching" revient à une sorte de vérification descendante (et une recherche ascendante sinon) mais les mécanismes d'inférence restent les mêmes dans les deux cas: seuls les réseaux phonétiques sont différents pour tenir compte des macro-phénomènes et des micro-phénomènes. Il est donc préférable de parler de stratégie "en proposition" (au lieu d'ascendante) et "en vérification" (au lieu de descendante). C'est sur ce point précisément que diffèrent essentiellement les méthodes IA des methodes stochastiques: les détails qui revètent parfois beaucoup d'importance en parole, peuvent être pris en compte plus facilement par une base de règles que par des coefficients de probabilité qui lissent les phénomènes rares ou de faible amplitude. C'est pourquoi nous avons adopté une vision IA pour le DAP sans se passer des points forts venus des HMM. Le décodeur dans notre projet DIRA (Dialogue Intégré pour un Robot Avancé) opère sur la parole continue avec adaptation au locuteur pour ce qui concerne la micro-structure acoustique. C'est ce décodeur que nous décrivons ci-après.

1. ARCHITECTURES EN DAP

1.1. Tendances actuelles

Les articles contradictoires de Klatt [Klatt, 86b], Stevens [Stevens,86] et Zue [Zue, 86] montrent quelques tendances actuelles en décodage acoustico-phonétique. Zue résume ainsi son point de vue par rapport aux deux autres: "My proposed model of phonetic recognition makes use of broad phonetic analysis and language-specific constraints to reduce the number of lexical hypotheses, and to establish the context for further, detailed phonetic analysis.(...) Like Klatt, I believe that the signal must be transformed into an acoustic, segmental description. However, I do not share his view regarding the feasibility of lexical access from short-time spectra, nor the use of a set of uniform distance metrics to measure phonetic similarities. Like Stevens, I believe in a representation based on distinctive features. However, I am increasingly frustrated by our inhability to find invariance of these features in the acoustic domain, and thus I question the hypothesis that such invariance in fact exists". Ce point de vue, le moins extrême des trois, le conduit au modèle de reconnaissance suivant:



En tous cas cette discussion prouve que le dilemne invariance/variabilité acoustico-phonétique doit être évacué de la reconnaisssance automatique si l'on ne veut pas en rester prisonnier.

Par ailleurs, si un modèle auditif peut paraître satisfaisant au plan épistémologique vis-à-vis de la perception, notre propre expérience dans le domaine [Caelen, 79] nous fait pencher plutôt vers la multiplicité des méthodes d'analyse spectrale qui ont des propriétés complémentaires et qu'il faut harmoniser au sein du DAP. Pour le reste nous sommes très en accord avec la vision générale de Zue, en enlevant toutefois le côté un peu figé que laisse deviner le schéma en ce qui concerne la stratégie.

Du côté des systèmes --et en restant toujours dans le seul cadre des modèles basés-connaissance-- nous examinons ci-après deux approches françaises choisies pour leur représentativité:

- Méloni [Méloni, 86] a une approche ascendante inspirée par la reconaissance des formes structurelle, mise au point par l'observation des courbes calculées sur le signal. Les formes qui jouent le plus grand rôle sont pics et collines (il en existe de plusieurs types définies par les paramètres l=longueur, e.g=émergence gauche, e.d=émergence droite, μv=micro-variation locale, Σ=seuil de détection). A l'aide de ces formes, qui tiennent donc compte du contexte lorsqu'elles sont temporelles, on peut localiser des événements et de là induire des macro-traits puis des phonèmes. Le système de paramètres utilisé est:
  - spectre LPC (et formants),
  - taux de passages par zéro,
  - intensité et Fo.

La notion de durée est implicitement contenue dans une "forme", il n'y a pas de segmentation a priori, les unités phonétiques sont formées à l'issue de la

reconnaissance, par ilôts de confiance. La structure de contrôle du système de reconnaissance est du type blackboard, les règles sont écrites en Prolog II.

- Mercier [Mercier, 88] a une approche différente: la méthode utilisée est ascendante dans l'étiquetage centiseconde hors contexte et descendante depuis les syllabes et noyaux syllabiques. Selon la progression gauche-droite dans le signal une pile de problèmes est créée. La résolution d'un problème donne un résultat de reconnaissance de type phonème (une étape intermédiaire permet de positionner des traits). Chaque problème fait appel à une série de règles qui utilisent des paramètres particuliers (par exemple des indices acoustiques propres aux voyelles) [Bonneau, 86].

La segmentation a pour but ici de
- faire la détection parole/silence
- détecter les frontières de syllabes (sur des critères d'intensité essentiellement)
- détecter les noyaux syllabiques (sauf éventuellement pour les syllabes ayant des /ə/ muets)

Les paramètres utilisés sont:
- spectre du vocodeur à canaux (16 ou 32 canaux),
- taux de passages par zéro,
- intensité et Fo,
- centre de gravité et dérivée spectrale.

La structure de contrôle est du type hiérarchique guidée par les problèmes, les règles sont écrites en Lisp dans SERAC, en C dans KEAL.

Une étude faite par Grenié [Grenié, 1987] montre bien que le champ des recherches sur les stratégies du DAP reste ouvert: il se traduit par la multiplicité des approches qui posent implicitement des questions comme: quelle est la meilleure paramétrisation ? faut-il passer par les traits ? doit-on faire un décodage robuste avant un décodage fin ? quelles est la meilleure stratégie d'utilisation des informations du DAP ? etc.

1.2. Structure de DIRA-DAP

Le décodeur acoustico phonétique dans le projet DIRA (Dialogue Intégré pour un Robot Avancé) est un expert guidé par le superviseur du système de reconnaissance [Caelen, 88]. Sur appel de ce dernier, il peut fonctionner en (a) proposition et en (b) vérification et sur des fenêtres (passé, présent, futur) déterminées par ce même superviseur --qui lui, fonctionne donc en plannificateur de tâches:

(a) en proposition: il délivre des informations aussi robustes que possible de type macro-trait puis, après affinage, une liste de traits en fonction notamment du contexte.

(b) en vérification: (b1) il vérifie par "spotting" la présence ou l'absence d'une information demandée par le superviseur --cette information peut être de diverse nature comme trait, macro-trait, marqueur microprosodique, etc.-- ou (b2) il se rend à un "point de rendez-vous" fixée par le superviseur (ce point de rendez-vous est un point de resynchronisation ou de repli).

Le DAP comprend deux étages:

1- l'analyse acoustique: modèle d'oreille et LPC, paramétrisation, segmentation en phases acoustiques,

2- l'étiquetage (reconnaissance totale ou partielle) en macro-traits et traits acoustiques. Cet étiquetage peut fonctionner sur deux modes:
- en "proposition"
- en "vérification"

et sur des tranches de temps données.

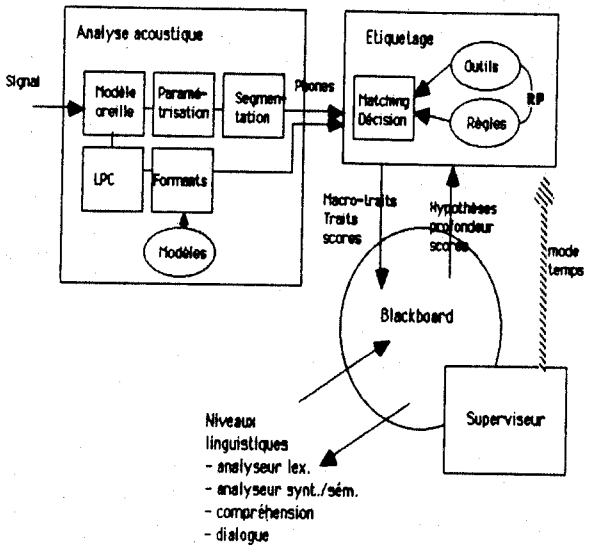


Fig. 3: Structure générale du DAP et situation dans le système de reconnaissance DIRA. Pour être utilisable en DAP, la connaissance sur la macro-structure peut être représentée sous forme de Réseaux Phonétiques (RP) dont l'intérêt sera démontré ci-après. Quant à la connaissance sur la micro-structure elle doit s'exprimer au moyen de règles et de procédures décrivant au mieux les aspects statiques, dynamiques et contextuels de la parole pour les macro-phénomènes aussi bien que pour les micro-phénomènes.

1.2.1. Principe de fonctionnement en mode "proposition"

Dès que le DAP est sollicité par le superviseur il lit les informations déjà contenues dans le blackboard (hypothèses provenant des niveaux linguistiques ou de la base de faits, variables diverses comme débit de parole). La phase d'analyse de la micro-structure consiste (a) à paramétrer le signal et à le segmenter en phases acoustiques, la phase d'étiquetage consiste (b) à assembler ces phases en unités phonétiques en les munissant d'étiquettes de macro-traits et traits. La stratégie d'étiquetage est de type "gauche-droite" par "matching" des Réseaux Phonétiques (RP) sur les phases. Les noeuds des RP (ou états) représentent les phases acoustiques des unités phonétiques. Etiqueter ces unités revient donc à cheminer dans le meilleur RP dans lequel les états et les transitions sont régis par des règles et dans lesquels les traits sont "calculés" par des règles ou des procédures déclenchées à partir des noeuds. Les RP modélisent les macro-classes phonétiques.

La hiérarchie des macro-traits et traits se déduit de l'orde d'appel des règles associées aux réseaux phonétiques; elle est donc implicite à la définition des RP eux-mêmes. Dans la base de connaissance actuelle cette hiérarchie est donnée fig. 4:

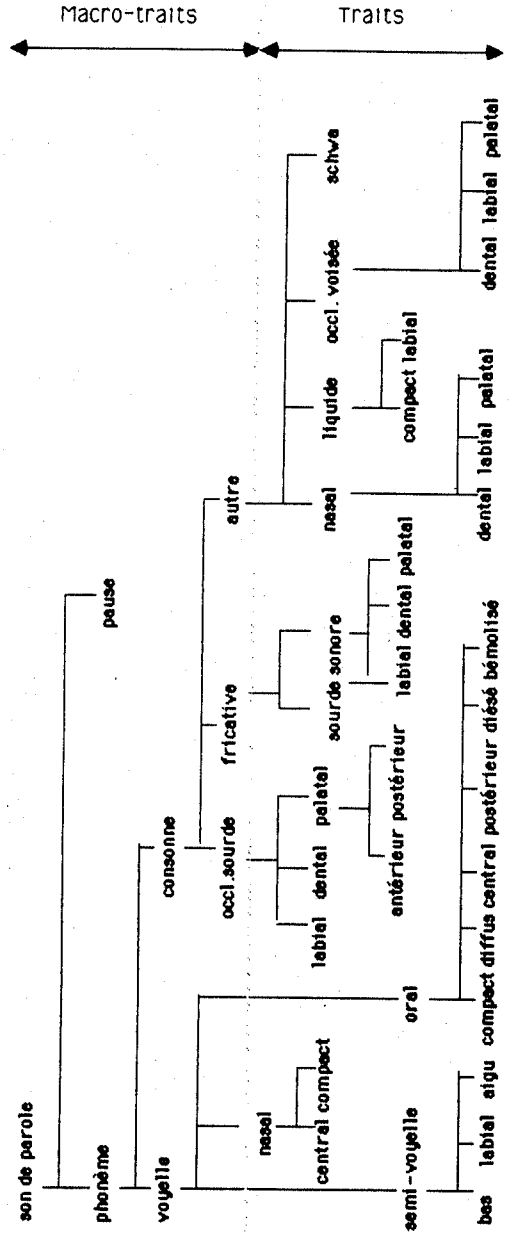


Fig. 4: arbre hiérarchique des macro-traits et traits pour le décodeur.

1.2.2. Principe de fonctionnement en mode "vérification"

La vérification des hypothèses linguistiques se fait en cheminant dans l'arbre hiérarchique de manière descendante. Les contraintes fixées par le superviseur au moment de l'activation de cette tâche sont:

- soit "vérifier H_l avec un coefficient de confiance≥ß et un coefficient de profondeur≥Δ"
- soit "aligner H_l avec le coefficient d'ajustement=μ sur le point de rendez-vous=Ω"

avec:

H_l hypothèses effectuées par les niveaux linguistiques, transcrites en termes d'unités phonétiques

Δ degré de profondeur dans l'arbre hiérarchique des macro-traits et traits (fig. 4)

ß coefficient de confiance homogène à un score de calcul obtenu par "matching"

μ coefficient d'ajustement similaire à ß

Ω point de rendez-vous défini par un ensemble de règles de localisation temporelle (par exemple "précédé_de", "composé_de", "partie_de", etc.) et par des événements quantifiés (par exemple "pause d'au moins 50 ms", "noyau vocalique", etc.).

Exemple: Vérifier "voyelle" pour ß>.75 et Δ="oral"+"compact"
Cela revient à déclencher les règles "voyelle"+"oral"+"compact" soit (a) sur une portion de signal clairement définie, soit (b) sur une portion de signal à localiser. Ces règles sont appelées, au fur et à mesure de la descente dans l'arbre de recherche, à partir des noeuds "voyelle" puis "oral" puis "compact". Le score de "matching" est propagé le long du cheminement. On s'appuie donc toujours sur les connaissances représentées dans les réseaux phonétiques, mis à part le fait que les règles associées à ces réseaux diffèrent de celles des réseaux du mode "proposition".

2. LES OUTILS ET LES CONNAISSANCES MISES EN OEUVRE DANS LE DECODEUR

2.1. L'analyse acoustique

L'analyse acoustique comprend:

- une double analyse spectrale par modèle d'oreille et par LPC

Le modèle d'oreille [Caelen, 79] dont il s'agit ici pour l'obtention du spectre de la parole est un modèle fonctionnel de l'audition périphérique. Il peut se résumer aux deux étages de traitement suivants:

- filtre de préaccentuation
- 24 filtres couplés, bandes critiques de Zwicker --ces filtres sont obtenus en bouclant la sortie de chaque filtre sur les entrées des deux filtres voisins (on obtient ainsi un effet de masque fréquentiel et temporel). Ce modèle répond aux spécifications formulées par Klatt [Klatt, 86a] en ce qui concerne l'étage de prétraitement "ear-like" d'un système de reconnaissance.

Le principal intérêt du modèle en reconnaissance est de fournir un spectre "robuste" c'est-à-dire pour lequel les transitions sont très lisibles et présentant peu de pics. Par contre ce spectre est peu contrasté et du fait des effets de masque, il ne peut permettre de détecter tous les formants. C'est pourquoi il est utile de lui adjoindre une analyse par LPC sur les portions stables du signal ce qui permet de calculer certains paramètres par affinage après le décodage du

macro-trait --nous avons préféré utiliser LPC plutôt que de mettre un étage de type inhibition latérale au modèle d'oreille, dans la mesure où cet étage est mal maitrisé à l'heure actuelle.

- la détection du fondamental

- le calcul d'indices spectraux statiques [Caelen, 81]

Ces indices sont codés de manière non-linéaire sur 5 niveaux. Un étalonnage de ces niveaux est à faire pour les ajuster selon la chaîne analogique utilisée (et le type de voix de certains locuteurs). C'est ici qu'intervient l'adaptation au locuteur. Par convention il existe une relation d'ordre dans ces degrés telle que: "--" < "-" < "=" < "+" < "++"

- les indices spectraux dynamiques

Le codage des variations est capital dans le traitement de la parole: pour cela nous avons opté pour plusieurs types de codages:

- la dérivée de chaque indice statique avec ajout d'un indice de dérivée spectrale normalisée en moyenne,
- la prise en compte (en vérification surtout) de la notion de trajectoire et de cible [Marteau, 87].

- le calcul des formants

Il s'agit ici d'un problème délicat puisqu'on ne peut atteindre par le spectre seul les fréquences de résonance des cavités buccales et nasales. La plupart du temps on détecte les crêtes du spectre (ou de la dérivée seconde) avec ou sans interpolation. Comme pour le fondamental, des procédures de suivi améliorent les transitions et permettent d'éliminer certains pics secondaires. Le risque d'erreur peut être également diminué si l'on contraint la détection dans certaines plages de fréquence: mais ceci ne peut se faire que si l'on a une connaissance a priori de ce que l'on cherche --c'est donc le cas d'une analyse descendante. La méthode retenue pour le moment se fonde sur un modèle markovien guidé par les indices [Bailly, 87].

- la pré-segmentation en phones homogènes

La segmentation s'effectue à partir de tous les indices spectraux dynamiques précédents. Elle est infra-phonémique et vise à produire des phones "homogènes" [Vigouroux, 85] c'est-à-dire des unités homogènes dans leur déroulement temporel --soit stables soit variables continument. Son principal intérêt est de réduire le débit d'information (sans perte notable) en regroupant des trames voisines. Les phones obtenus correspondent aux phases acoustiques des phonèmes (période de changement) comme: établissement, tenue, coda d'une voyelle ou implosion, occlusion, burst, détente d'une occlusive, etc. Par cette méthode, on obtient toujours au moins un phone dans un phonème (même pour des semi-voyelles brèves ou un débit d'élocution rapide). V. Zue [Zue, 1986] aboutit à une segmentation très voisine en rattachant la trame courante à sa voisine la plus proche.

2.2. Les réseaux phonétiques

Il y a actuellement 5 modèles de réseaux phonétiques: les occlusives sourdes, les fricatives, les voyelles, les autres consonnes, les pauses et 5 types de règles associées à ces réseaux:

a) sur le noeud courant
b) sur les transitions avec les noeuds précédents
c) sur les contextes
   - antécédant
   - subséquent
d) sur les informations venant du blackboard
e) sur les actions et mécanismes de décision

La fig. 5 donne un exemple de réseau pour les voyelles:

Les règles sont écrites ci-après dans un langage externe facilement compréhensible. On distingue les règles:

- non-actives par défaut: elles sont activables par la clause: action<-R, où R est le nom de la règle
- actives par défaut: elles sont de deux types
  + non-récursives: elles ne sont exécutées qu'une fois
  + récursives: elles sont exécutées autant de fois consécutivement que cela est possible. Les actions qu'elles déclenchent sont alors soit exécutées après le dernier appel, soit exécutées à chaque appel (dans ce cas elles sont notées à l'aide de la facette $appel: action$appel<-R). Les variables qu'elles manipulent sont cumulées (par exemple la durée) ou moyennées si aucun autre calcul spécifique n'est indiqué.

Pour le réseau de la fig. 5 les règles s'écrivent:

VO-Règle Début_voyelle
  Action <- Lecture_blackboard (contextes, variables)
  Action<- (V1 OU V2 OU V11)

V1-Règle Etablissement_voyelle_ou_semi-voyelle
  ! montée d'énergie sauf éventuellement pour les fermées en contexte vocalique,
  ! transition lente pour les semi-voyelles
  si ((pente(intensité)>p_V ET Etat_précédent='néant')
    OU (pente(intensité)≥0
      ET contexte_précédent='vocalique'
      ET état_précédent='néant')
    OU (Etat_précédent='Etabl_voyelle_et_semi-voyelle' ET indice(CD)≤'-'))
    ET intensité≥B+3*(S/B)/4 ! S/B rapport signal (S) à bruit (B)
    ET Nb(formants)≥2
  alors Etat<-'Etabl_voyelle_et_semi-voyelle';note<-1
    ET frontière_début$défaut<-@phone_courant
    ET action<- J1
    ET ACTION <- (V1 OU V2 OU V11)

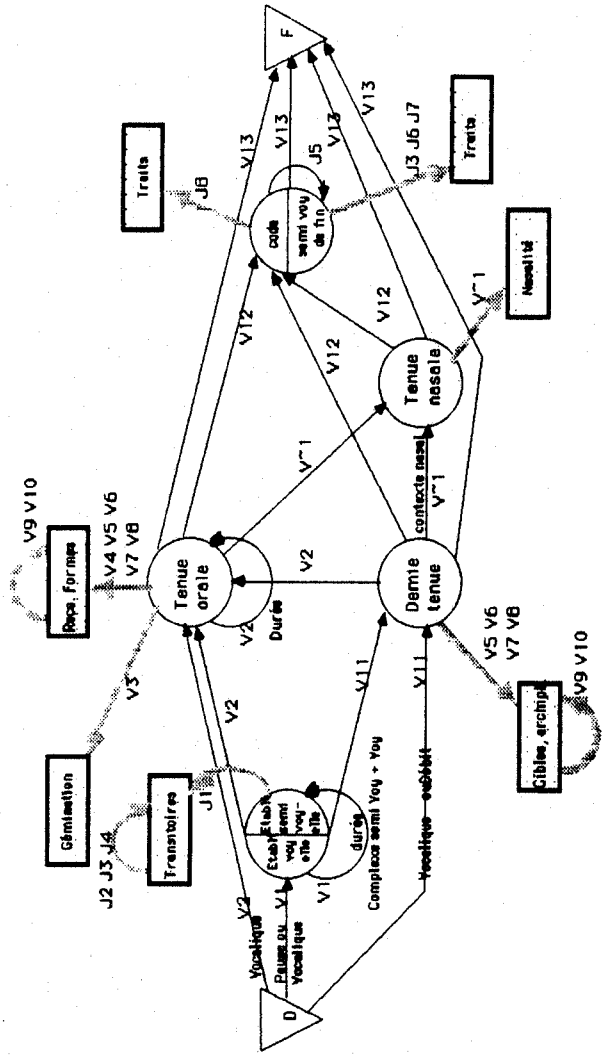


Fig. 5: Réseau phonétique pour le décodage en proposition des voyelles et du complexe semi-voyelle/voyelle, voyelle/liquide. Les noeuds (cercles) correspondent aux phases acoustiques de ces unités. Les procédures d'affinage (rectangles) sont déclenchées à partir des noeuds et des règles (identificateurs sur les arcs) conditionnent les transitions d'un noeud au suivant. Les deux états "début" et "fin" (triangles) sont les entrées et sorties du réseau, contraintes éventuellement par des règles contextuelles. Un tel réseau représente la combinatoire des phases acoustiques (axe paradigmatique) d'une unité phonétique relativement à sa structure temporelle (axe syntagmatique).

J1-Règle Semi_voyelle
! séquence semi-voyelle+voyelle (JV)
si Etat='Etabl_voyelle_et_semi-voyelle'
ET trait='semi_voyelle'
ET durée(Etat)≥D_J*débit ! durée(Etat)=durée cumulée, D_J durée moyenne d'une semi-voyelle
ET (cible('f',F_1)-cible('b',F_1)≥200
OU |cible('f',F_2)-cible('b',F_2)|≥400)
alors trait='semi-voyelle'
ET action<-J2, J3, J4

! 'cible' est une procédure qui ajuste la trajectoire formantique par régression linéaire et calcule la cible par rapport aux frontières du segment telle que: 'c'=centre, 'b'=backward, 'f'=forward

J2-Règle Semi_voyelle+bas
! /w/ est parfois fusionné avec /a/ donc non détectable sinon son F_2 est montant d'une valeur initiale ≈ 800 Hz
si Etat='Etabl_voyelle_et_semi-voyelle'
ET trait='semi-voyelle'
ET trait='bas'
ET cible('b',F_2)≤1200 ET cible('f',F_2)-cible('b',F_2) ≥ 200
alors Trait<-'bas'

J3-Règle Semi_voyelle+labial
! /μ/ est parfois sourd (séquence tμit dans 28) - non traité ici
si (Etat='Etabl_voyelle_et_semi-voyelle'
OU Etat='Coda_voyelle_ou_semi-voyelle')
ET trait='labial'
ET trait='semi-voyelle'
ET 2000≥cible('b',F_2)≥1500
alors Trait<-'labial'

etc.

3. DISCUSSION ET CONCLUSION

Ce système est en cours de test sur des corpus. Les premiers résultats sont très encourageants mais méritent d'être discutés très en détail car notre système de reconnaissance est très intégré [Caelen, 88] ce qui nous empêche de parler de performance au seul niveau du décodage acoustico-phonétique. Le but étant de comprendre la parole pour un robot, c'est-à-dire d'exécuter correctement des opérations commandées, le pourcentage de détection d'une plosive par exemple, n'a pas beaucoup de sens. L'appel du décodeur par le superviseur étant fait de manière opportuniste (donc régi après analyse de la situation) il est évident que les cas de décodage difficiles seront évités ou traités "en vérification" après émission d'hypothèses préalables. Si bien que se pose le problème du comptage des succés et des échecs (au sens classique du terme), de leur degré de gravité et de leur répercussion dans l'ensemble du système. Des méthodologies spécifiques doivent être envisagées --ce qui déborde largement le cadre de cet article-- dans un cadre de réflexion plus général tel celui de SAM (Speech Assessment Methods). Ce n'est qu'après une telle démarche que les résultats pourront être discutés valablement.

En conclusion, la structuration des connaissances sous forme de réseaux phonétiques permet de séparer clairement les connaissances sur la macro-structure phonétique et sur la micro-structure du signal de parole. La notion de réseau phonétique permet quand à elle de jeter un pont entre les modèles stochastiques tels que HMM et les systèmes de règles plus classiques dans lesquels la macro-structure (sur les axes paradigmatiques et syntagmatiques) de la parole n'émerge pas suffisamment. Cette notion permet également d'adapter et de passer facilement d'un mode de fonctionnement en "proposition" à un mode de fonctionnement en "vérification", sans modifier les procédures de "mise en correspondance" et éventuellement, sur des calculateurs spécialisés, de mener plusieurs stratégies en parallèle.

Le formalisme des règles s'adapte bien au langage Prolog II dans lequel est écrit ce décodeur, ainsi qu'à la représentation des connaissances disponibles dans le domaine phonétique. Une des difficultés demeure dans la propagation des scores lors du cheminement dans les réseaux: comment ajuster les seuils pour ne pas bloquer trop tôt un cheminement admissible (système trop robuste mais pauvre en information) sans toutefois autoriser de fausses solutions (problème général du "treillis phonétique ascendant") ? ou en d'autres termes comment ajuster la pente de coupure du filtre phonétique ?

BIBLIOGRAPHIE

[Bailly, 87] G. Bailly, D. Liu, 1987
Détection d'indices par quantification vectorielle et réseaux markoviens. 16° JEP, Hammamet, pp. 60-63.

[Bonneau, 86] A. Bonneau, M. Rossi, G. Mercier, 1986
Hierarchical representation of French vowels by Expert System. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 20-21.

[Caelen, 79] J. Caelen, 1979

Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique. Thèse d'Etat Sciences, Toulouse.

[Caelen, 87] J. Caelen, 1987
Le décodage acoustico-phonétique: état de l'art, Bulletin de l'Institut de phonétique de Grenoble, Vol. 16, pp. 177-221.

[Caelen, 88] J. Caelen, 1988
Meta-stratégie en reconnaissance dans le système DIRA-RAP. 17° JEP Nancy.

[Grenié, 87] M. Grenié, 1987
Nature et hiérarchie d'indices acoustiques indépendants du locuteur: application à la reconnaissance automatique des voyelles du Français.
Thèse de 3° cycle, Aix-en-Provence.

[Halle, 85] M. Halle, 1985
Speculations about the representation of words in memory. in V.A. Fromkin ed., Phonetic Linguistics, Academic Press, New-York, pp. 101-114.

[Huttenlocher, 86] D. Huttenlocher, M. Withgott, 1986
On acoustic versus abstract units of representation. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 61-62.

[Klatt, 86a] D.H. Klatt, 1986
The problem of variability in speech recognition and in models of speech perception. in J. Perkell and D. Klatt eds., "Variability and Invariance in Speech Processes", Erlbaum.

[Klatt, 86b] D.H. Klatt, 1986
Models of phonetic recognition I: Issues that arise in attempting to specify a feature-based strategy for speech recognition. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 63-66.

[Marteau, 87] P.F. Marteau, J. Caelen, M.T. Janot-Giorgetti
Extraction automatique de caractéristiques dynamiques du signal de parole. Application à l'analyse des voyelles nasales. 16° JEP, Hammamet, pp. 88-91.

[Méloni, 86] H. Méloni, R. Bulot, 1986
Un système de traitement de connaissances pour le

décodage acoustico-phonétique. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 26-27.

[Mercier, 88] G. Mercier, A. Cozannet et J. Vaissière, 1988
Recognition od speaker-dependent continuous speech with Keal-Nevezh. In Recent Advances in Speech Understanding and Dialig System, NATO ASI Series, Nieman, Lang & Sagerer, ed., Springer Verlag.

[Stevens, 86] K.N. Stevens, 1986
Models of phonetic recognition II: An approach to feature-based recognition. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 67-68.

[Vigouroux, 85] N. Vigouroux et J. Caelen, 85
Segmentations en vue de l'organisation d'une base de données acoustiques et phonétiques. 14èmes JEP, SFA, Paris, pp. 152-155.

[Withgott, 86] M. Withgott, M.A. Bush, 1986
On the robustness of phonetic information in short-time speech spectra. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 101-102.

[Zue, 86] W. Zue, 1986
Models of phonetic recognition III: The role of analysis by synthesis in phonetic recognition. Proc. of Montreal Symposium on Speech Recognition, McGill University, pp. 69-70.

INTERPRETATION DE SPECTROGRAMMES DE PAROLE.

Claudie FAURE - Xuan Shi WANG

ENST - Département SIG - UA CNRS 820
46 rue Barrault
75634 Paris cedex 13

ABSTRACT

Knowledge Based Pattern Recognition is used for speech spectrogram interpretation. This paper focuses on the segmentation problem which is described as a hierarchy of sub-problems. The leaves of this hierarchy correspond to the recognition of basic features in the image and the other levels to morphological reasoning which combine the basic features. Reasonning is achieved by rules, their left part is structured in order to reveal the abstract category to which the specific reasoning belongs. A memory, as a relationnal attributed graph, is used to collect the results of each sub-problem resolution. The nodes represent homogeneous parts of the image or locations which are candidates for segmentation. A confidence degree is assigned to each node. Resolution of sub-problems leads to change the structure of the memory graph by adding or merging nodes or to update the confidence degree of nodes.

1. PRESENTATION DU PROJET.

Ce travail se situe dans le cadre général de la conception de systèmes de perception et d'interprétation d'images, qui incorporent un savoir spécifique. Dans le cas présent, ce savoir est l'expertise du lecteur de spectrogrammes de parole.

Les données sont constituées de spectrogrammes de parole que l'expert peut "traduire" en suite de phonèmes. Cette image est la représentation visuelle de l'évolution spectrale des sons prononcés par un locuteur.

Ces dernières années de nombreux travaux s'intéressent au problème de l'expertise des spectrogrammes de parole. Leur objectif principal est de formaliser et d'évaluer les connaissances de l'expert humain, en vue de leur utilisation ultérieure dans des systèmes de reconnaissance automatique de la parole, plus particulièrement au niveau du décodage acoustique-phonétique.

Il est rare que ces systèmes opèrent sur le spectrogramme brut pour simuler complètement une tâche de lecture en se plaçant dans un environnement identique à celui de l'expert du point de vue des informations traitées. Nous définissons un double objectif pour notre système :

- d'une part explorer la problématique de la perception visuelle et des processus d'interprétation de formes mettant en jeu une connaissance spécifique qui, dans ce cas, se révèle complexe;
- d'autre part fournir à un système expert de lecture de spectrogrammes des données obtenues automatiquement à partir de l'image, ce qui permettra d'expliciter l'expertise intervenant dans le processus d'identification phonétique effectuée par l'expert.

Les références concernant l'analyse des spectrogrammes et de la connaissance experte s'y rapportant sont cités : [CAR,87], [GRE,87], [MIZ,86], [MOR,85], [ZUE,86], [CON,86], [STE,86a], [JOH,83].

On distingue deux étapes dans l'analyse de l'image : la segmentation en phonèmes et l'identification des phonèmes. Cette décomposition est imposée par le fait que le système de RdF doit servir d'entrée à un système expert (SE) en lecture de spectrogrammes de parole qui travaille à partir de descriptions de segments phonémiques. Ce système expert a été conçu et réalisé au LIMSI, [STE,86b]. Le rôle de la reconnaissance de formes (RdF) est donc de segmenter puis de décrire les segments dans des termes compatibles avec le vocabulaire du SE. La figure 1 présente le système dans son ensemble.

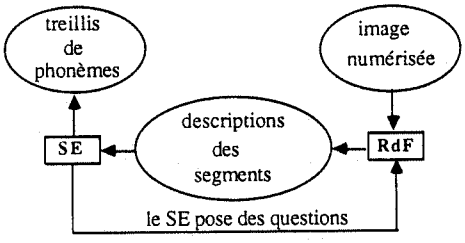


Figure 1

Le SE poursuit un raisonnement à partir des descriptions grossières des segments acoustiques. Ces descriptions peuvent être complétées ou affinées à sa demande en cours de raisonnement. Le SE interprète l'image à un niveau **purement symbolique.**

Le système de RdF a pour but de segmenter le spectrogramme de parole en événements acoustiques et de décrire spontanément et/ou à la demande du SE ces événements. Les descriptions sont données par le système de RdF dans des termes directement utilisables par le SE. Le système de RdF doit **calculer sur l'image** les traits pertinents pour les tâches de segmentation et de description des événements. Le choix des procédures de calcul est conditionné par la nature de ces traits.

Si la segmentation et l'identification des segments diffèrent du point de vue des objectifs, les décisions de l'expert sont argumentées verbalement par des descriptions de l'image qui se réfèrent quasiment aux mêmes traits pour les deux étapes. De sorte qu'à la fin de la segmentation, le système aura presque résolu l'étape de description des segments.

2. LA SEGMENTATION : ARCHITECTURE ET MECANISME DE RAISONNEMENT.

On s'intéresse d'abord au problème de la segmentation qui est la première tâche a réaliser par le système de RdF. La segmentation visuelle d'une image suppose la perception de **ruptures** dans les données et/ou de **zones** présentant une certaine régularité. Les fonctions actives pour la détection de ruptures et de régularités sont largement coopérantes. La segmentation de l'image du spectrogramme de parole est un cas particulier d'application où les changements et les régularités ne sont pas seulement évaluées d'un point de vue perceptif mais avec l'intervention d'une connaissance liée au domaine de la parole.

Le problème à résoudre : on veut obtenir à partir d'une image digitalisée de spectrogramme de parole un ensemble d'objets de description de cette image. Ces objets sont de deux types, les "ruptures" et les "zones". La segmentation est un problème complexe au sens où différentes informations se combinent pour sa résolution. Comme tout problème complexe, il faut le décomposer. Un arbre de résolution du problème sera adopté, une vue partielle en est donnée sur la figure 2. Les feuilles sont les sous-problèmes (SP1, SP2, SP4 sur la figure 2) liés aux **informations primitives** calculées sur l'image et les sous-problèmes d'ordre supérieur (SP3, SP5 sur la figure 2) concernent la **coopération d'informations.** Dans la version actuelle du segmenteur automatique, la résolution du problème se fait de manière purement ascendante. On définit une **situation** comme étant l'état de résolution du problème au niveau de chaque nœud de l'arbre. La situation finale correspond à la résolution du nœud racine, elle contient le résultat obtenu pour la tâche de segmentation. La situation est en fait un ensemble d'objets de description (ruptures et zones) munis d'attributs, elle est localisée dans une Mémoire Commune (MC) de travail accessible par les nœuds de l'arbre.

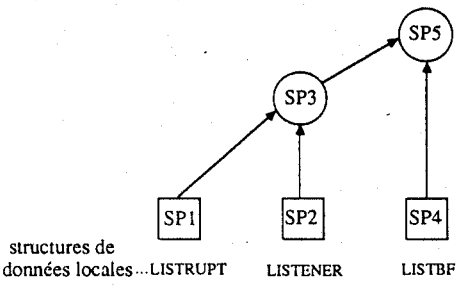


Figure 2

Les informations primitives sont liées de manière très étroites aux traits pertinents de l'image décrits par l'expert. Le calcul de ces informations se fait au niveau des sous-problèmes feuilles, elles ont pour fonction de faire évoluer les situations soit directement en créant et caractérisant des objets de description, soit indirectement en se combinant aux informations déjà contenues dans les situations.

Quand un sous-problème feuille a été résolu, on progresse vers le haut dans l'arbre de décomposition du problème de segmentation pour résoudre le noeud père. Un problème intermédiaire est résolu par activation de règles. Ces règles supportent différents types de connaissances : celles qui concernent la connaissance interne au système, c'est à dire qui portent sur des comportements connus des traitements numériques; celles qui sont liées à la perception d'image et celles qui sont spécifiques à l'application. La vérification de la **prémisse** d'une règle déclanche des **actions** qui ont pour but de faire évoluer la situation de la MC. On tend vers une définition la plus abstraite possible des prémisses pour différencier leur description, que l'on souhaite assez générale, de leur instanciation par les informations effectivement calculées sur les données numériques. Ceci permet de réduire le nombre de prémisses et par suite le temps de calcul nécessaire à leur recherche, et de les utiliser à plusieurs étapes de l'analyse. Cet a-priori conduit à une structuration des prémisses en **"description"** et **"contexte".** La description étant le niveau le plus abstrait. Le contexte concerne les informations calculées sur l'image disponibles lors de la résolution d'un sous-problème intermédiaire et définit les conditions d'appel aux actions. Par cette structure à deux niveaux, on peut faire apparaître l'interaction des raisonnements typiques de l'interprétation d'image et de ceux qui sont typiques de l'interprétation de spectrogrammes. Les règles sont alors de la forme :

$(P_i, C_j) \dashrightarrow A_{ij}$

La recherche de prémisses se fait dans la MC qui évolue sous l'effet des résultats des actions.

On donne les descriptions des prémisses des premières règles qui seront introduites pour la résolution des sous-problèmes intermédiaires.

P1 : (accord), plusieurs informations s'accordent pour localiser dans un même voisinage une rupture dans l'image.
P2 : (accord), plusieurs informations s'accordent pour définir une zone homogène.
P3 : (conflit) un ensemble d'informations {x} s'accordent pour localiser dans un même voisinage une rupture de l'image, un ensemble d'information {y} s'accordent pour signaler ce voisinage comme inclus dans des zones homogènes, $\{x\} \cap \{y\} = 0$.

3. LA MAQUETTE DU SEGMENTEUR.

Une maquette du segmenteur a été réalisé afin de valider l'approche décrite au paragraphe précédent et d'estimer ses performances. On a choisi pour cette maquette 3 informations primitives qui sont associées à 3 sous-problèmes feuilles, SP1, SP2, SP4. L'arbre de décomposition pour cette maquette est donné figure 2.

On pourra se référer à [STE,86b] pour la description de la méthode d'obtention des spectrogrammes numériques. Les images qui sont utilisées ici correspondent à des mots prononcés en moins de 1 s. Le signal échantillonné à 10 kHz est préaccentué puis traité par un banc de filtres pour réaliser l'analyse spectrale. La sortie de ces traitements constitue l'image initiale du spectrogramme pour notre système. Elle est constituée d'une matrice de 100 lignes et 1000 colonnes. Les 100 lignes correspondent à 5000 Hz, et les 1000 colonnes à 1 s. de parole. L'unité de temps sera l'intervalle entre deux colonnes successives de cette matrice initiale, soit 1ut = 1ms.

Par la suite on utilisera une matrice réduite pour certains calculs. On appelle x_{ij} les éléments de la matrice initiale et y_{kl} ceux de la matrice réduite. Chaque valeur y_{kl} résume un pavé de la matrice initiale de 10 lignes et 5 colonnes. Les x_{ij} sont moyennés sur 10 éléments d'une colonne et sommés sur 5 éléments d'une ligne.

Simultanément au calcul de la matrice réduite, la valeur maximale de y_{kl} est mémorisée, elle dépend du spectrogramme traité.

3.1. Les Sous-Problèmes feuilles.

On ne décrira que trés brièvement le calcul des informations primitives (IPs) qui se font au niveau des sous-problèmes feuilles, pour plus de précision voir [FAU,88]. Sur la figure 2, les mémoires locales qui contiennent les résultats de SP1, SP2 et SP4 sont indiquées par LISTRUPT, LISTENER et LISTBF.

L'information primitive associée à **SP1** est issue du calcul des ruptures verticales dans l'image qui existent indépendamment d'une connaissance a-priori sur l'application. Les résultats servent de marqueurs sur l'image, les endroits ainsi détectés signalent des changements de structure de l'image. Ils ont un rôle d'indicateur et non de trait pertinent pour la segmentation au même titre que ceux définis par l'expert.

Chaque verticale de l'image réduite est représentée par une suite de symboles qui codent l'intensité de chaque y_{ki}. L'algorithme de recherche de ruptures donne un coût de comparaison pour chaque paire de verticales adjacentes. Les ruptures sont choisies sur la séquence des valeurs des coûts de comparaison.

L'information primitive associée à **SP2** est liée à l'énergie du signal. Sur la matrice réduite on recherche les zones de silence par l'analyse de la projection de l'image sur l'axe temporel. La courbe obtenue aprés projection est assimilable à une courbe d'énergie.

Dans le sous-problème **SP4** l'information primitive porte sur la zone des basses fréquences, de 0 à 500 Hz. Son analyse permet de définir deux types de régulariiés qui correspondent pour la première à une absence dénergie en basse fréquence et pour la seconde à une forte énergie en basse fréquence, que l'on appellera le voisement fort.

3.2. Les sous-problèmes intermédiaires.

Ils ont pour fonction de réaliser une combinaison des IPs qui sont calculées indépendament les unes des autres. Chaque sous-problème feuille a pour tâche d'engendrer des hypothèses sur la segmentation. La fonction d'un sous-problème intermédiaire étant de gérer cet ensemble d'hypothèse en vue d'une décision. La MC est un graphe orienté dont chaque nœud représente une hypothèse. Il existe deux types d'hypothèses dans la MC : Les zones homogènes (ZSEG) et les frontières (ZFTR). Chaque hypothèse est caractérisée (entre autre) par sa nature, sous forme d'étiquettes qui sont héritées des mémoires locales attachées aux feuilles de l'arbre de résolution. Chaque hypothèse a un facteur de confiance qui est nul au moment de la création de l'hypothèse et va évoluer lors de l'application des règles de combinaison d'informations, il prendra des valeurs positives ou négatives. Le succés d'une segmentation se traduit par le fait que chaque ZSEG de confiance strictement positive correspond à un segment phonémique identifié par l'expert dont les frontières sont donnés par les ZFTR dont la confiance est strictement positive.

Σ_1	Σ_2
R : rupture de structure dans l'image	S : silence
DM : position du début	E : énergie
FM : position de la fin	VHV : voisement fort
DS : début de silence	SIL : silence en basses fréquences
FS : fin de silence	
DVHV : début de voisement fort	
FVHV : fin de voisement fort	
DSIL : fin d'énergie en basses fréquences	
FSIL : début d'énergie en basses fréquences	

Figure 3

L'ensemble des étiquettes utilisées dans la maquette est donné par Σ_1 pour les ZFTR et Σ_2 pour les ZSEG, voir la figure 3.

La combinaisons des informations se fait par application de règles. La figure 4 donne le shéma d'évolution des situations et les règles appliquées. Une règle Ri correspond à la prémisse Pi décrite plus haut.

On donne les règles R1, R2a, R2b, R3 dans leur contexte d'utilisation.

R1: la partie description de la prémisse correspond à la proximité de deux ruptures. Les conditions posent une contrainte sur la durée de la zone frontière (≤15 ut) et sur la non-contradiction d'étiquettes (par exemple on ne peut pas avoir {DS, FSIL} ou {FS, DS} pour caractériser la nature d'un ZFTR). La règle s'applique en fusionnant les deux hypothèses de rupture dans un seul ZFTR qui voit son facteur de confiance augmenter.

R2 : la partie description correspond à deux configurations du graphe illustrées figure 5a pour la règle R2a et 5b pour la règle R2b. Il n'y a pas de contrainte de contexte et les effets de la règle sont illustrés sur cette même figure, il s'agit de supprimer les arcs inutiles (5.a) ou de regrouper, dans un seul nœud, les informations sur une zone réparties sur plusieurs nœuds du graphe (5.b).

R3 : la règle est instanciée par le contexte {x}={R} et {y}={SIL,E}. Ce qui conduit à : si une rupture a été détectée sur l'image dans une zone d'énergie non nulle et d'énergie nulle en basse fréquence alors faire décroitre la confiance de ZFTR. Cette règle est liée à la connaissance que l'on a sur le comportement du module de recherche des ruptures qui sur-segmente dans les bruits h.f. L'évolution de la confiance permet de conserver la rupture comme marqueur de changement de structure dans l'image et de l'éliminer des résultats tant qu'aucune information n'est venue la confirmer.

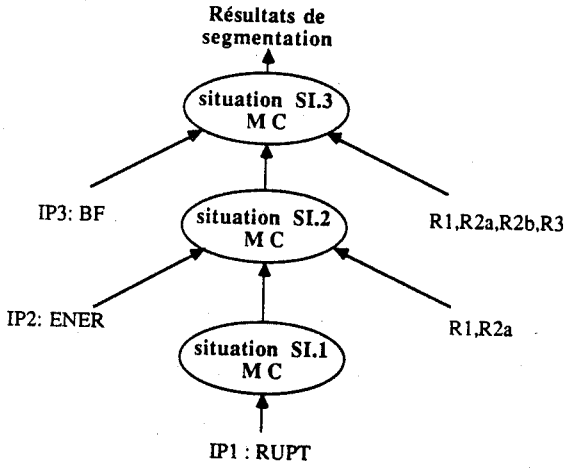


Figure 4
MC : Mémoire Commune
IP : Information Primitive
SI : Situation
R : Règle

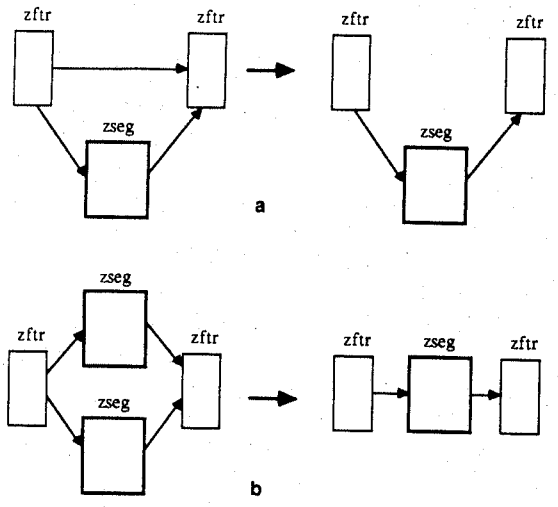


Figure 5

4. RESULTATS ET DISCUSSION.

La maquette a été testée sur 56 images qui contiennent 387 frontières réparties en six classes. On note ces classes <V-C> (<voyelle-consonne>), <C-V> (<consonne-voyelle>), <C-C> (<consonne-consonne>), <V-V> (<voyelle-voyelle>), <S-DMOT> (<silence-premier segment>), <FMOT-S> (<dernier segment-silence>). La figure 6 indique le nombre de frontières correctement détectées (notées O) et le nombre de frontières entre segments non détectées (notées M). Les notations sont les suivantes : R pour /r/, N pour consonne nasale, FN pour fricative non voisée, FV pour fricative voisée, PN pour plosive non voisée, PV pour plosive voisée, V pour voyelle, DMOT pour le premier segment du mot et FMOT pour le dernier. Le sous ensemble de phonèmes utilisés est : /a/, /i/, /w/, /e/, /ã/, /r/, /f/, /s/, /Z/, /t/, /k/, /d/. Dans ce test on a considéré les frontières dont le facteur de confiance est strictement supérieur à zéro.

Les meilleures segmentations correspondent à <FN-V>, <PN-V>, <V-FN>, <V-PN>, c'est à dire à des frontières où la variation en basse fréquence est un des traits pertinents que relève l'expert. Pour <N-V>, <V-N> et <V-V> aucune frontière n'est détectée, trés souvent SP1 a marqué cette rupture mais elle n'a pas été confimée par la suite (le facteur de confiance reste donc égal à zéro) dans la mesure où les traits pertinents qui permettent à l'expert de décider d'un changement de phonème sont évalués dans des régions fréquentielles qui ne sont pas prises en compte dans cette maquette. Les sursegmentations sont essentiellement localisées dans le dernier segment, on en compte 48. La disparition lente du signal provoque des ruptures non significatives. Un premier segment correspondant à une explosion entraîne une localisation erronée du début de ce segment, 16 erreurs de ce type ont été faites. Des règles spécifiques pour le premier et le dernier segment seront nécessaires. Il existe aussi des segmentation avant ou aprés le signal de parole, on en compte 22, ceci étant du à l'ambiance bruité dans laquelle l'enregistrement a été effectué. On donne, figure 7, un exemple de segmentation obtenue sur un spectrogramme. On indique pour chaque hypothèse confirmée la valeur du facteur de confiance et les étiquettes des informations qui ont engendrées cette hypothèse. Certaines segmentations sont localisées à un instant précis indiqué par une ligne verticale sur la figure 7, d'autres correspondent à une plage indiquée par un rectangle.

<C-V>	(O)	(M)
<FN-V>	24	1
<PN-V>	21	4
<PV-V>	13	4
<R-V>	19	13
<FV-V>	5	12
<N-V>	0	19
<V-C>	(O)	(M)
<V-PN>	19	0
<V-FN>	19	4
<V-PV>	4	2
<V-FV>	9	8
<V-R>	9	16
<V-N>	0	17
<V-V>	(O)	(M)
	0	2
<S-DMOT>	(O)	(M)
<S-V>	14	0
<S-FN>	13	0
<S-FV>	1	0
<S-R>	3	0
<S-N>	2	0
<S-PV>	4	1
<S-PN>	3	15
<FMOT-S>	(O)	(M)
<V-S>	42	0
<FN-S>	6	0
<FV-S>	1	0
<PN-S>	5	0
<R-S>	2	0
<C-C>	(O)	(M)
<PN-R>	3	7
<FN-R>	1	3
<R-PN>	1	0
<R-FN>	0	2
<R-PV>	4	0
<R-N>	0	1
<PN-FN>	1	0
<FN-PN>	2	2
<PN-PN>	0	1
<PN-N>	1	0
<N-PV>	0	2

Figure 6

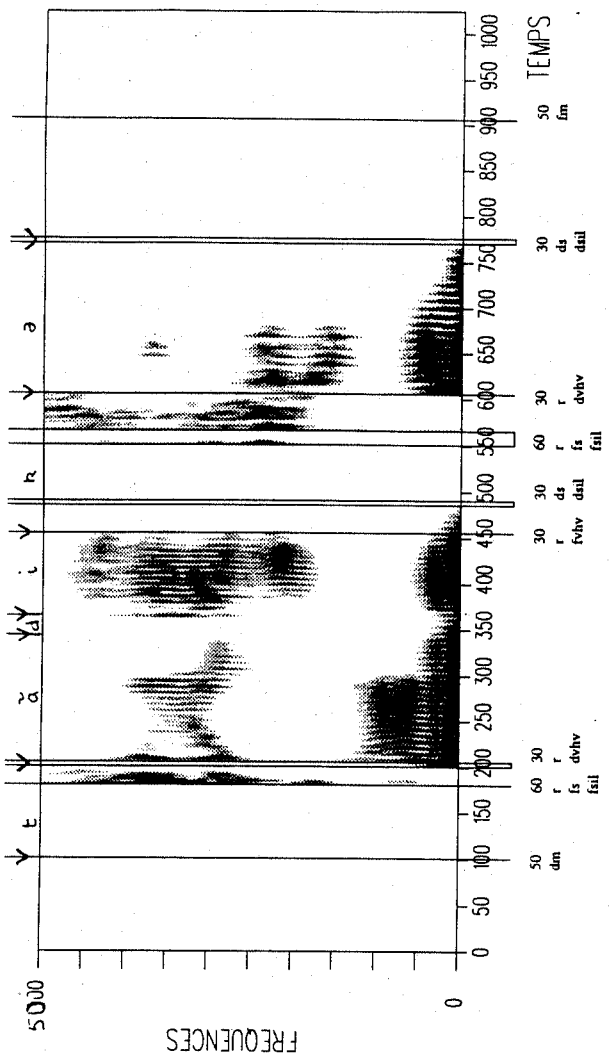


Figure 7
Chaque segmentation est donnée avec son facteur de confiance et les étiquettes des événements détectés (cf la liste de la figure 3).
↓ : segmentations de l'expert.

L'expertise doit se rattacher aux sous-problèmes intermédiaires. Elle se présente sous formes de règles qui doivent s'appliquer sur les hypothèses de la MC. L'apprentissage symbolique dont il n'a pas été question ici mais qui a fait l'objet d'une étude dans le cadre de l'acquisition des connaissances [SAI,88], nous a permis de dégager quelques règles à partir d'un corpus d'apprentissage de 58 frontières décrites par l'expert. L'application de ces règles ne peut se faire qu'aprés avoir réalisé les procédures de calcul des traits qu'elles utilisent. On remarque cependant qu'une des règles obtenues est la suivante : il existe une frontière si il y a fin ou début de voisement. Il est a noter que cette règle peut s'appliquer dans le contexte actuel de la maquette mais qu'elle est la seule. Or la stratégie de décision adoptée est basée sur une concertation de règles, toutes les règles de la base de connaissances qui peuvent s'appliquer sur une hypothèse sont déclanchées, la décision étant prise sur l'ensemble de ces règles. Il est donc dangereux de donner à la seule règle applicable un pouvoir de décision dans la mesure où le système est privé du contexte global de la connaissance.

5. CONCLUSIONS.

Un système d'interprétation automatique de spectrogramme de parole est un système complexe dont la conception fait intervenir de nombreuses compétences : connaissances phonétiques, traitement et analyse d'image, raisonnement automatique, représentation et acquisition des connaissances, apprentissage automatique. On a pu définir des tâches relativement autonomes liées à chacune de ces compétences : le SE, l'apprentissage symbolique et l'analyse de l'image à laquelle le travail présenté se rattache. Le fond commun de tous ces travaux est la connaissance experte spécique à l'application traitée qui a fait l'objet de nombreuses discussions avec M. Eskenazi. Sa compétence et sa disponibilité sont à la base des résultats les plus encourageants.

REFERENCES :

[CAR,87] N. CARBONELL, D. FOHR, J.P. HATON, " APHODEX, An Acoustic-Phonetic Decoding Expertt System", Int. J. of PR and IA, Vol. 1, No 2, pp. 31-46, 1987.

[CON,86] J.H. CONNOLLY, E.A. EDMONDS, J.J. GUZY, S.R. JOHNSON, A. WOODCOCK, "Automatic speech recognition based on spectrogram reading", Int. J. Man-Machine Studies, Vol. 24, pp. 611-621, 1986.

[FAU,88] C.FAURE, "Interprétation de spectrogrammes de parole", Rapport interne ENST, 88DOO1, 1988.

[GRE,87] P.D. GREEN, M.P. COOKE, H.H. LAFERTY, J.H. SIMON, "A Speech Recognition Strategy Based on Making Acoustic Evidence and Phonetic knowledge explicit", NATO-ASI, Bad Windsheim, 1987.

[JOH,83] J. JOHANNEN, J. MACALLISTER, T. MICHALEK, S. ROSS, " A Speech Spectrogram Expert", Proc. ICASSP, Boston, pp. 746-749, 1983.

[MIZ,86] R. MIZOGUCHI, K. TSUJINO, O. KAKUSHO, "A continuous recognition system based on knowledge engineering techniques", Proc. ICASSP, Tokyo, pp 1221-1224, 1986.

[MOR,85] R. DE MORI, M. PALAKAL, "On the use of computer vision techniques for automatic speech recognition", IEEE Conf. Computer Vision and Pattern Recognition, pp. 691-693, 1985.

[SAI,88] L. SAITTA, "Spectrogram Segmentation. Acquisition of the Expert's Knowledge", Rapport Interne ENST, 1988.

[STE,86a] P.E. STERN, M. ESKENAZI, D. MEMMI, "An expert system for speech spectrogram reading", Proc. ICASSP, Tokyo, pp. , 1986.

[STE,86b] P.E.STERN, "Un système expert en lecture de spectrogrammes", Thèse de Docteur-Ingénieur, Orsay, 1986.

[ZUE,86] V.W. ZUE, L.F. LAMEL, "An expert spectrogram reader: approach to speech recognition", Proc. ICASSP, Tokyo, pp. 1197-1200, 1986.

RECONNAISSANCE DE MOTS ISOLES EN UTILISANT UN RESEAU DE NEURONES

Fan YANG Li WU Jean-Paul HATON

C.R.I.N / BP 239
54506 Vandœuvre-les-Nancy, France

Abstact

A fundamental problem in speech recognition is the representation of the acoustic vectors characterizing each sppeech time frame into a neural networks. In this paper, we first present a method of speech processing for improving isolated word recognition. Afterwords we show the results of experience running on following parameters: LPC order, number of hidden units and number of input units. At the end, we give illustrating results of performance comparison between RNA and TDW.

Résumé

Un des problèmes fondamentaux est de relier aux réseaux les vecteurs acoustiques qui caractérisent chaque fenêtre du signal. Dans ce papier, nous proposons d'abord une méthode pour améliorer la performance de reconnaissance de mots isolés. Ensuite nous montrons les résultats de l'expérience en faisant varier les différents paramètres qui influencent la reconnaissance: l'ordre de LPC, le nombre d'unités de la couche cachée et le nombre d'unités de la couche d'entrée, et à la fin, nous illustrons le résultat par la comparaison de performance entre le réseau RNA et système classique DTW.

I. INTRODUCTION

Aprés avoir été abandonnés il y a vingt ans, les réseaux de neurones artificiels réapparaissent dans le monde de la recherche en Intelligence Artificielle suite à leurs récents succès. Le réseau de neurones et le DTW (Dynamic Time Warping) appartiennent tous deux aux systèmes dynamiques. Mais le réseau de neurones est mieux adapté pour représenter le cerveau humain[6]. Dans ce dernier modèle, le temps nécessaire pour la reconnaissance est indépendant du nombre de références utilisées, ce qui n'est pas le cas de la DTW. Dans cet article, nous traiterons le cas de la reconnaissance de mots isolés (dix chiffres 0-9 en chinois) en utilisant un réseau de neurones.

Le problème le plus important est de représenter les mots sous forme de signal temporel à l'entrée du réseau. En se referant à la méthode de H.Bourlard et C.J. Wellekens[1] et aux caractéristiques de la parole naturelle, nous proposons dans cet article une méthode pour améliorer le taux de reconnaissance.

Nous avons enregistré deux cents chiffres isolés en chinois, chaque chiffre étant prononcé vingt fois. La moitié sert pour l'apprentissage du réseau, tandis que le reste est pris comme ensemble de test. Nous faisons ici une étude comparative des capacités de la reconnaissance, des différents paramètres et modèles utilisés:

* ordre de LPC pour la phase de la quantification
* mode d'entrée
* nombre des unités de neurones

Dans la dernière partie, nous avons comparé la performance du réseau de neurones avec celle de la programmtion dynamique.

II. PRESENTATION GENERALE DU RESEAU

Les réseaux de neurones artificiels sont fondés sur des modèles théoriques qui tentent d'expliquer comment les cellules du cerveau et leur interconnexion parviennent à exécuter des calculs complexes. Bien sûr, dupliquer sur un circuit l'architecture massivement parallèle et complexe du cerveau est impossible, mais des modèles simplifiés qui ne s'intéressent qu'aux transmissions entre neurones ont déja démontré les capacités de tel réseaux à apprendre, mémoriser et effectuer des calculs en temps réel.

Les réseaux de neurones artificiels sont formés de simples processeurs interconnectés qui communiquent entre eux en se transmettant des signaux d'activition ou d'inhibition. Chaque neurone additionne les signaux qu'il reçoit en entrée et produit un signal de sortie si cette somme dépasse un seuil fixé.

Dans un réseau composé de plusieurs couches, le signal introduit dans la couche d'entrée se propagera entre les couches en subissant à chaque étape un traitement parallèle. A chaque connexion entre deux neurones est associé un coefficient (le poids) qui pondère la transmission du signal: le signal reçu est égal au signal appliqué en entrée, multiplié par ce poids. Le neurone est modélisé par unités qui additionnent les N entrées pondérées et transmet le résultat par l'intermèdiaire d'une fonction de seuil non linéaire (Fig.1).

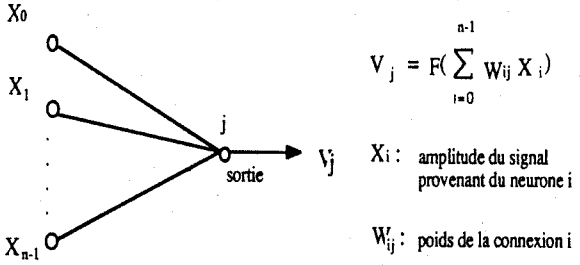


Figure 1: le neurone est modélisé par une unité qui additionne les N entrées pondérées et transmet le résultat à travers une fonction non-linéaire.

Dans un réseau de neurones artificiels, les coefficients de pondération sont déterminés par un apprentissage qui consiste à présenter au réseau une série d'entrée, et à modifier les connexions du réseau pour que chacune de ces entrées corresponde à la sortie souhaitée. Il existe beaucoup d'algorithmes d'appentissage, le plus connu est la rétro-propagation du gradient [2].

III. INTERFACE ENTRE LES MOTS ISOLES ET LE RESEAU

Dans l'utilisation d'un réseau de neurones pour la reconnaissance de mots isolés, le premier problème est de relier le signal de parole au modèle. Au niveau du mots isolés, la façon la plus directe est de représenter un mot par un tableau de dimention de M*N [3], M étant le nombre de valeurs de chaque vecteur acoustique, N étant le nombre maximum de vecteurs composant un mot. Une autre méthode plus proche du processus naturel du décodage de la parole est d'utiliser un réseau de neurones prenant en compte le facteur de déformation temporelle. Le TDNN (Time Delay Neural Network) est un tel réseau qui est capable d'apprendre les décisions non-linéaires automatiquement en calculant l'erreur de rétro-propagation, et également de retrouver les caractéristiques acoustiques et leurs relations temporelles indépendantes de la position du

temps[4][5]. En considérant qu'il n'est pas essentiel de conserver la contrainte temporelle dans la production et la perception de la parole, H.Bourlard et C.J Wellekens [1] ont employé une méthode plus simple et plus efficace. Ici, nous avons choisi cette dernière méthode pour représenter (Fig. 2) un mot par une suite de codes obtenus par analyse LPC du signal et quantification vectorielle (QV) [7].

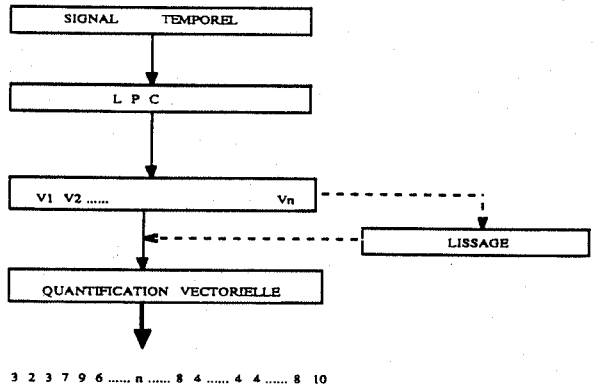


Figure 2: le signal temporel est d'abord transformé en une série de codes de QV qui seront ensuite utilisés par le réseau de neurones

La suite de codes de quantification vectorielle est rentrée parallèlement dans le réseau des neurones Ne-Nh-Ns, Ne étant le nombre de cellules d'entrée qui est égal au nombre des prototypes QV, Nh étant le nombre de cellules de la couche cachée, Ns étant le nombre des cellules de sortie qui indiquent la reconnaissance d'un des dix chiffres (Fig. 3).

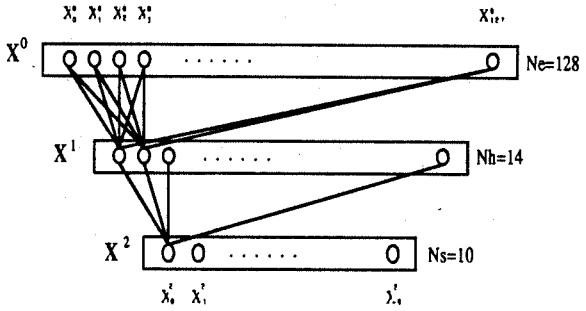


Figure 3: le réseau de neurones à trois couches ayant une couche de sortie de 10 cellules, une couche cachée de 14 cellules et une couche d'entrée de 128 cellules

Une cellule d'entrée déjà activée le reste si elle apparaît à nouveau dans la forme à reconnaître.

Une suite de codes de QV d'un mot isolé est toujours un sous-ensemble des codes de la couche d'entrée, donc la machine examine les codes de QV, si elle rencontre un code N, elle active la cellule du $N^{ème}$ neurone.

L'ordre d'apparition des cellules d'entrée n'est pas pris en compte; de même on néglige totalement la fréquence d'apparition d'un code d'entrée. L'ignorance de ces deux facteurs simplifie donc l'entrée d'un signal dans le réseau. Mais nous devons tenir compte des conséquences amenées par ces deux simplifications.

Dans la production de la parole, les phonèmes fondamentaux sont intégrés à une séquence. Chaque phonème porte ses propres caractéristiques mais aussi les effets de coarticulation que lui imposent ses voisins.

Un phonème apparaissant à plusieurs reprises dans différents contextes correspond donc à autant de réalisations acoustiques différentes, surtout les phonèmes ayant une durée très courte. Le contenu acoustique est suffisant pour la description de chaque mot[1].

La suppression de l'importance de la fréquence d'apparition d'un mot amènera pourtant un réel problème. Dans ce cas, un code qui apparaît même une seule fois joue aussi un rôle aussi important que celui qui apparaît plusieurs fois. Ceci est équivant à affaiblir la résistance du système aux parasites. Si nous ne considérons pas les bruits extérieurs au système, il reste deux sortes de bruits qui peuvent influencer les résultats. Premièrement, le bruit dû à la position de la fenêtre d'analyse de LPC qui peut faire varier le vecteur des coefficients LPC. Deuxièmement, le bruit de la Quantification Vectorielle. Nous proposons donc un lissage pour réduire ces deux bruits avant d'entrer les codes dans le réseau (Fig.2).

Il n'y a pas de raison de faire un lissage directement sur les codes de QV de la série, parce que les codes ne sont pas une description acoustique, mais seulement des symboles. Alors nous le faisons sur la suite des vecteurs de LPC. Ce lissage est un simple calcul de moyenne sur les coefficients. Dans la partie suivante, le rôle du lissage est évident.

IV. EXPERIENCE

Nous avons fait des essais dans un contexte monolocuteur pour la reconnaissance de dix chiffres isolés en utilisant un réseau à trois couches (Fig.3).

Nos expériences consistent à examiner la performance du réseau en faisant varier différents paramètres, tel le nombre de neurones de la couche d'entrée et la couche cachée, l'ordre de le LPC.

Le réseau de neurones de base se compose de la façon suivante: le nombre de neurones de couche d'entrée est de 128, 14 pour la couche cachée, et 10 pour la couche de sortie.

A l'aide de ce réseau de base, nous avons obtenu les courbes comparatives des performances de la reconnaissance données plus loin.

Chaque chiffre a été prononcé vingt fois, soit au total deux cents chiffres, la moitié servant à l'apprentissage, le reste pour le test.

Notre travail est réalisé sur MASSCOMP, la fréquence d'échantillonnage est 16KHZ avec une précision de 12bits. L'analyse acoustique LPC determine, pour chaque fenêtre de 12.8 ms de Hamming avec 0.95 de préaccentation, un vecteur de coefficients.

Notre expérimentation est de mesurer la performance du système en fonction du nombre de point de lissage. La courbe ci-dessous (Fig.4) montre l'évolution du taux de reconnaissance. Prenant la courbe de Nh16 par exemple, nous voyons que le taux de reconnaissance sans lissage est égal à 96% et qu'il arrive à son meilleur score lorsque Nbl=14 ou 15. En fixant Nbl, Nbl=4 par exemple, nous obtenons le meilleur résultat quand Nh=14, le pire résultat se produit quand Nh=5. Ces courbes montrent une variation irrégulière. Nous pouvons constater qu'après lissage, les resultats de reconnaissance peuvent être bien améliorés et d'autre part qu'ils dépendent du choix du nombre d'unités de la couche cachée.

Nous avons aussi fait varier le nombre d'unités de la couche d'entrée (Ne) en fixant le nombre d'unités de la couche cachée à Nh=14 (Fig.5). Nous constatons une reconnaissance optimale quand Ne=128. Si le nombre d'unités de la couche d'entrée est fixé à Ne=64, le codebook de quantification vectorielle n'est pas suffisant. En revanche, le résultat avec le codebook de Ne=256 est aussi moins bon, cela est probablement dû à l'insuffisance du nombre de références d'apprentissage.

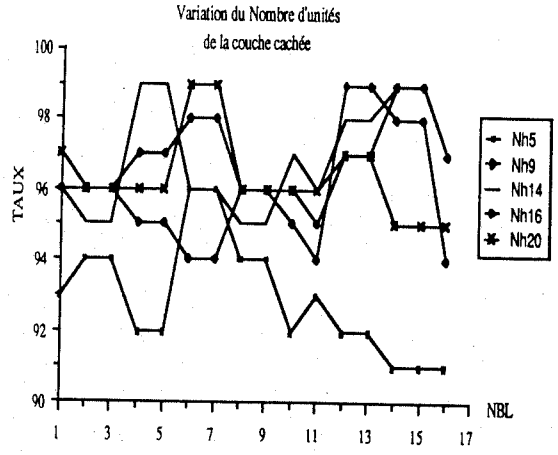


Figure 4: les courbes de l'évolution du taux de reconnaissance en fonction du nombre de points de lissage (NBL) selon le nombre d'unités de la couche cachée (Nh).

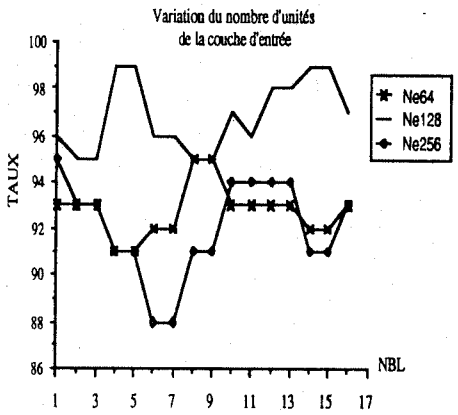


Figure 5: les courbes du taux de reconnaissance en fonction du points de lissage (NBL) selon le nombre d'unités de la couche d'entrée (Ne).

La Fig.6 ci-dessous montre l'évolution du taux de reconnaissance en fonction de l'ordre de LPC. Les résultats de l'ordre de 20 et de 24 sont similaires, alors, nous prenons 20.

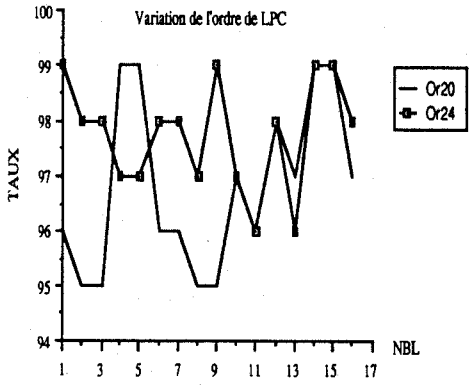


Figure 6: les courbes du taux de reconnaissance en fonction du points de lissage selon l'ordre de LPC

V. COMPARAISON ENTRE RNA et DTW

Nous avons comparé le réseau de neurones artificiel avec la programmation dynamique.

Les deux méthodes ont été testées en prenant trois ensembles de formes de références:

--- 10 formes par chacun des 10 chiffres, soit 100 formes
--- 5 formes par chacun des 10 chiffres, soit 50 formes
--- 1 formes par chacun des 10 chiffres, soit 10 formes.

Dans les trois cas, l'ensemble de test est formé de 100 chiffres n'ayant pas servi à l'apprentissage. Les temps indiquant sont des temps CPU sur Masscomp, y compris l'analyse LPC et la quantification vectorielle.

On constate que dans la méthode de RNA, le temps de reconnaissance et l'espace nécessaire sont indépendant du nombre de formes de référence et que l'intérêt de RNA dépend beaucoup du nombre de références.

Ensemble de références: 10 * 10 = 100 chiffres		
	DTW	RNA
TAUX	99%	99%
TEMPS (minutes)	41:23.6	5:38.3
ESPACE (octet)	477396	31308

Ensemble de références: 5 * 10 = 50 chiffres		
	DTW	RNA
TAUX	97%	91%
TEMPS (minutes)	15:22.4	5:38.3
ESPACE (octet)	236844	31308

Ensemble de références: 1 * 10 = 10 chiffres		
	DTW	RNA
TAUX	59%	41%
TEMPS (minutes)	7:21.9	5:38.3
ESPACE (octet)	47276	31308

VI. CONCLUSION et PERSPECTIVES

Le système expérimental réalisé presente les caratéristiques suivantes:

* stockages des connaissances dans les connexions du réseau sous une forme plus proche du cerveau humain
* vitesse de reconnaissance et espace mémoire indépendants du nombre de références
* taux de reconnaissance maximal de 99% pour les chiffres en mono-locuteur

Cette expérience préliminaire montre l'intérêt des modèles connexionnistes pour la reconnaissance de la parole. Néanmoins de nombreux problèmes restent à résoudre, par exemple, la structure interne du réseau de neurones; le contrôle automatique de l'apprentissage. Nous travaillons actuellement à améliorer notre modèle.

REFERENCES

[1] H.Bourlard & C.J.Wellekens, "Speech Pattern Discrimination and Multilayer Perceptrons", Manuscript M.211, Philips Research Laboratory. Sept. 1987.

[2] F.Fogelman Soulie, "Le Connexionnisme". Support de cours MARI 87-COGNTIVA, Ecole des Hautes Etudes en Informatique Université de Paris 5.

[3] S.M Peeling & R.M. More and M.J. Tomlinson, "The Multi-Layer Perception as a tool for Speech Pattern Processing", Proc.IoA Autumn on Speech and Hearing (1986)

[4] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, "Phoneme Recognition: Neural Networks vs. Hidden Markov Models", IEEE-ICASSP, 1988.

[5] Raymond L. Watrous and Lokendra Shastri, "Learning phonetic Features Using Connectionist Networks: An Experiment in Speech Recognition", MS-CIS-86-87 LINC LAB44 Oct.1986.

[6] Bart Kosko, "Constructing an Associative Memory", BYTE, Sep.1987

[7] Yoseph Linde, A.Buzo, R.M. GRAY, "An algorithm for Vector Quantizer Design", IEEE trans. on communication. Vol. Com-28. No.1. Jan.1980

REMERCIEMENTS

Nous remercions A. Boyer, Y. Gong et A. Gourinda pour les programmes de QV et DTW.

EXPERIMENTATION SUR DES INDICES DE CONVEXITE POUR DES MODELES MARKOVIENS

Christine Dours - Guy Pérennou

Laboratoire CERFIA UA-CNRS N°824 Université Paul Sabatier,
118 route de Narbonne - 31062 TOULOUSE

ABSTRACT

We present some experiments of speech recognition in order to evaluate the efficiency of different approaches to hidden Markov models (HMM):
- centisecond models compared with infra-phonemic segment approaches,
- efficiency of the convexity of spectral cues.

1 INTRODUCTION

L'étude que nous présentons est centrée sur la présélection de sous-vocabulaires dans les systèmes de reconnaissance de la parole. Cette étape est intéressante surtout en présence de grands vocabulaires. (voir par exemple les travaux de [Jelinek 85] et [Mérialdo 87]).

On peut aussi dériver de ces approches des modèles de reconnaissance indépendants du locuteur pour des petits vocabulaires.

Le travail exposé utilise une modélisation par HMM continus dont Levinson a fait une présentation générale [Levinson 85]. Concernant la reconnaissance utilisant des HMM continus, on pourra se référer à [Levinson 83], [Liporace 82], [Bourlard 85] et à [Jouvet 87], ce dernier présentant une étude comparative des différents modèles.

Notre travail vise avant tout à démontrer:

1) La faisabilité d'une approche par classes de phonèmes utilisant peu d'indices spectraux (ici 4); de ce point de vue, deux modèles sont comparés, l'un faisant intervenir les segments infra-phonémiques stables [Vigouroux 85] et l'autre les centisecondes.

2) La possibilité de renforcer cette approche en adjoignant des indices robustes déduits des quatre premiers.

2 CADRE DE L'EXPERIMENTATION

2.1 LE CORPUS UTILISE

Nos expériences ont été développées sur un corpus comportant dix fichiers de cinq locuteurs, chaque fichier contenant dix séries de quatre nombres connectés. (numéros de téléphone) [Vigouroux 88]. Ce corpus est constitué de 3513 phonèmes répartis comme décrit dans le paragraphe suivant. Il a été étiqueté selon la méthode d'"étiquetage fréquentiel fin" [Barrera 87] utilisée aussi pour l'étiquetage de la BDSON [Descout 86]; le codage comporte des champs pour caractériser au mieux le signal aux niveaux acoustique, phonétique et syntactico-lexical.

Pour notre expérience, seul le champ contenant les phases acoustiques (établissement, tenue et coda du phonème) est utilisé. Une analyse spectrale préalable nous a fourni les indices Aigu/Grave (A/G), Fermé/Ouvert (F/O), Doux/Strident (D/S) et l'énergie totale E [Caelen 81].Ces indices ont été normalisés et sont à valeur continue dans l'intervalle [-5,+5].

Pour l'évaluation d'un système de reconnaissance complet, ce corpus est trop réduit. Mais il est suffisant pour les buts que nous nous sommes fixés dans ce travail.

2.2 LES CLASSES DE PHONEMES

Tous les phonèmes du Français ne sont pas attestés dans notre corpus. Le /l/ et le /p/ par exemple, n'apparaissent pas dans les nombres compris entre 0 et 100. Ceux qui y figurent ont été regroupés en classes conformément au tableau de la fig.1. Le critère de regroupement est la ressemblance phonétique par rapport aux indices utilisés.

CLASSES	N	S	Z	R	Q	D	A	IN	E	O	I	U
PHONEMES	n	s f	z v	r	k t	d	a ɑ̃ ɔ̃	ɛ ɛ̃ œ̃	ə ø œ	o ɔ ω	i y e ɥ	u
effectif de la classe dans le corpus	319	158	308	280	404	107	664	237	118	118	528	272

figure 1: Les classes phonémiques.

2.3 LES MODELES

Plusieurs types de modèles ont déjà été testés dans différentes applications, modèles d'unités phonétiques à deux états de Jelinek [Jelinek 76] etc. Pour notre application, nous avons développé des HMM continus à trois états correspondant aux phases acoustiques du phonème (voir fig.2).

Ces modèles sont régis par:
- la matrice de probabilité de transition des états; la probabilité de transition de l'état i à l'état j sera notée a_{ij}.
- les matrices des moyennes et des covariances qui permettent de calculer la densité de probabilité multivariable gaussienne $b_j(y_t)$ pour l'état j et le vecteur acoustique de l'observation au temps t.

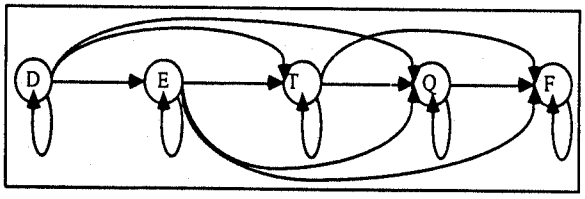


figure 2: Le modèle de l'application ; D et F sont les états fictifs de début et de fin du phonème alors que E,T et Q sont les états correspondants à l'établissement, la tenue et la coda du phonème.

3 LES EXPERIENCES SUR LES CLASSES DE PHONEMES

3.1 L'APPRENTISSAGE

3.1.1 L'INITIALISATION DES MODELES

Les matrices du modèle ont été initialisées à partir de statistiques effectuées sur le corpus.

3.1.2 LA REESTIMATION

Les modèles ont été réestimés par l'algorithme de Baum-Welch qui permet de calculer une probabilité totale par récurrence. Dans la suite, nous utiliserons le terme trajectoire pour désigner une suite de transitions d'états.

Cette probabilité met en œuvre (on peut se référer aux travaux de [Baker 74]):
- soit le calcul de α_t (i) qui est la probabilité pour que l'observation $Y_t=(y_1,y_2,...,y_t)$ ait été émise conditionnellement à une trajectoire quelconque issue de l'état 1 et aboutissant à l'état i au temps t.

- soit le calcul de β_t (j) qui est la probabilité pour que l'observation $Y^t=(y_t,...,y_T)$ ait été émise conditionnellement à une trajectoire commençant en j au temps t et aboutissant à F.
La probabilité totale de production d'une observation Y s'écrit: $P(Y)=\alpha_{T+1}(F)=\beta_0(D)$, (T étant la longueur de l'observation, D et F les états fictifs de début et fin de phonème).
Le calcul se heurte à des difficultés d'ordre numérique; les probabilités qui entrent dans le calcul de α_t et de β_t étant faibles, on aboutit très rapidement à un underflow que la présence de sommes ne permet pas de contrôler par le passage aux logarithmes.
Plusieurs approches ont été proposées pour remédier à cet inconvénient:
- calculs à partir d'approximations de logarithmes de sommes.
- introduction d'un facteur d'échelle K>1 dans le calcul de α_t et de β_t, le résultat étant alors multiplié par K^{T-1}. Le choix du facteur K est assez délicat en raison des variations notables de certains paramètres estimés.
- en ce qui nous concerne, une approche similaire à celle développée par Rabiner a été utilisée [Rabiner 83].

3.2 LA RECONNAISSANCE

Nous avons utilisé l'algorithme de Viterbi qui consiste à trouver la probabilité conjointe maximale d'une observation Y donnée de longueur T et d'une trajectoire. Elle s'écrit [Baker 74]:

$P^*(Y) = \gamma(T,F)$ où γ est défini récursivement par:

$$\gamma(t,j)=\max_i(\gamma(t-1,i).a_{ij}).b_j(y_t) \qquad (1)$$

On évite les problèmes d'underflow en passant aux logarithmes. La formule (1) devient alors:

$$-\log\gamma(t,j)=\min_i(-\log\gamma(t-1,i)-\log(a_{ij}))-\log(b_j(y_t))$$

3.3 LES RESULTATS

Les modèles de classes ont été estimés par apprentissage à partir du tiers du corpus et la reconnaissance a été effectuée sur la totalité.

Deux expériences ont été menées afin de montrer d'une part la supériorité des centisecondes sur les segments infra-phonémiques et d'autre part l'intérêt d'introduire la convexité d'indices spectraux.

— Centisecondes ou segments infra-phonémiques ?

L'utilisation des segments infra-phonémiques [Caelen 83] a pour avantage de réduire les données à traiter. Il était donc intéressant de comparer les taux de reconnaissance des classes par l'algorithme de Viterbi, obtenus d'une part pour les segments et d'autre part pour les centisecondes du corpus décrit en 2.1. La fig.3 donne les résultats et les gains obtenus pour chaque classe, seuls les indices E, A/G, F/O, et D/S sont utilisés.

On peut constater que la reconnaissance sur le corpus "centisecondes" est supérieure de 10% en moyenne à celle effectuée sur le corpus "segments". On peut noter de très grands écarts de gain entre les diverses classes. Cette différence peut se justifier pour les classes S et Z par le fait que la classe S est constituée de phonèmes relativement stables, ce qui n'est pas le cas de ceux de la classe Z. La segmentation utilisant la notion de moyenne, il est normal que la reconnaissance de la classe S soit moins perturbée que celle de la classe Z. Les mauvais résultats de certaines classes, (notamment pour la classe E), peuvent également s'expliquer par la segmentation automatique qui a tendance à sur-segmenter. L'étiquetage ne pouvant intervenir que sur les marques de segmentation, il arrive donc que la phase "établissement du phonème" ne soit pas notée au profit de la phase "coda" du phonème précédent.

Ceci nous a conduits à rejeter les segments infra-phonémiques pour la suite des expériences.

On remarquera au passage que l'apprentissage permet une amélioration de 8% en moyenne sur le corpus "segments".

CLASSES	SEGMENTS		CENTISECONDES	GAIN
	av. appr.	ap. appr.	ap. appr.	
I	25	34	51,87	17,87
A	25	45	46,77	1,77
O	45	50	78,18	28,18
U	50	60	90,91	30,91
E	12	28	28,21	0,21
Q	70	75	87,59	12,59
D	60	67	70,79	3,79
S	70	89	89,24	0,24
Z	20	22	28	6
R	45	45	50,8	5,8
N	40	50	55,32	5,32
IN	48	50	75,1	25,1

figure 3: Comparaison du taux de reconnaissance des classes sur les segments et sur les centisecondes (résultats en % pour 4 indices, Viterbi et 1 candidat); 'av.', 'ap.' et 'appr.' sont les abréviations respectives de 'avant', 'après' et 'apprentissage'.

—Amélioration de la reconnaissance par les indices de convexité

Notre expérience visant essentiellement à améliorer les performances précédentes, nous avons introduit en tant que paramètres supplémentaires les convexités des indices. En effet, les phonèmes sont non seulement caractérisés par la valeur absolue d'indices spectraux mais aussi par leur variation relative, ce que traduit partiellement la convexité de la courbe d'évolution de ces indices en fonction du temps. On pourra, par exemple, souvent opposer phonétiquement voyelles et consonnes par la convexité de l'énergie. De même, la convexité de l'indice A/G pemet de distinguer les fricatives des autres classes. De plus, les indices de convexité sont robustes pour une approche indépendante du locuteur contrairement aux indices absolus. En revanche, ils sont sensibles au contexte du phonème.

Le calcul d'une convexité possible pour un indice i à l'instant t est le suivant:

$c(i,t) = v(i,t-2)+2v(i,t-1) -6v(i,t)+2v(i,t+1)+v(i,t+2)$

avec c(i,t) convexité de l'indice i à l'instant t et v(i,t) valeur de l'indice i à l'instant t.

Nous donnons en fig.4 les résultats (pour le 1er et les trois 1ers candidats) obtenus par l'algorithme de Viterbi pour les quatre indices initiaux, puis pour ces mêmes indices et leur convexité.

classes	1 CANDIDAT			3 CANDIDATS		
	4 indices	4 indices + 4 convexités	gain	4 indices	4 indices + 4 convexités	gain
I	51,87	57,01	5,14	77,57	82,24	4,67
A	46,77	72,81	26,04	95,85	99,31	3,46
O	78,18	80	1,82	89,02	92,73	3,71
U	90,91	90,91	0	90,91	100	9,09
E	28,21	30,26	2,05	67,18	77,95	10,77
Q	87,59	93,07	5,48	98,36	99,27	0,91
D	70,79	86,52	5,73	93,26	97,75	4,49
S	89,24	91,77	2,53	96,2	96,2	0
Z	28	39,56	11,56	67,11	75,11	8
R	50,8	55,6	4,8	75,6	74	-1,6
N	55,32	57,45	2,13	74,47	82,98	8,51
IN	75,1	75,1	0	98,85	98,85	0

figure 4: Comparaison du taux de reconnaissance (en %) des classes par Viterbi pour des vecteurs spectraux à 4 indices et pour des vecteurs spectraux faisant intervenir ces mêmes indices et leurs convexités respectives.

A l'examen du tableau de la fig.4, il apparait une amélioration significative des résultats lorsque l'on tient compte de la forme d'évolution des indices par l'intermédiaire de leur convexité.

Les types d'erreurs les plus fréquents qui apparaissent dans la matrice de confusion sont décrits ci-dessous. On fait remarquer au passage que lorsqu'un phonème est dit être confondu avec une classe, la mauvaise reconnaissance ne porte que sur un sous-ensemble de ses occurences.

-***pour la classe I*** : le phonème /e/ est confondu avec la classe E (16,36% d'erreurs pour 4 indices et 14,49% pour 8 indices), il est également confondu avec la classe IN (9,81% d'erreurs pour 4 indices et 8,88% pour 8 indices) quand il est placé devant les phonèmes /œ̃/ ou /ɛ̃/.

- ***pour la classe A*** : le phonème /ɑ̃/ est confondu avec la classe IN (31,8% d'erreurs pour 4 indices et 11,52% pour 8). Des confusions apparaissent également avec la classe O (9,91% pour 4 indices et 9,68% pour 8); il s'agit d'erreurs portant sur l'ensemble des phonèmes de la classe A.

- ***pour la classe R*** : le phonème /r/ est confondu avec la classe A (13,2% d'erreurs pour 4 indices et 9,2% pour 8) quand il est suivi des phonèmes /ɑ/ ou /ɑ̃/. Il est également confondu avec la classe O (13,6% pour 4 indices et 12% pour 8) quand il est suivi du phonème /o/.

- ***pour la classe E*** : le phonème /œ/ est confondu avec la classe IN (13,33% d'erreurs pour 4 et 8 indices) quand il est précédé du phonème /n/. Le phonème /ø/ est confondu avec la classe A (16,92% pour 4 indices et 32,82% pour 8) quand il est précédé de la séquence de classes Q-R, et avec la classe R (15,38% pour 4 indices et 5,13% pour 8) quand il est précédé du phonème /r/.

- ***pour la classe IN*** : les phonèmes /ɛ̃/ et /ɛ/ sont confondus avec la classe A (8,81% et 9,2% d'erreurs) et avec la classe E (15,33% et 13,79% d'erreurs).

- ***pour la classe Z*** : le phonème /z/ est confondu avec la classe S (12,44% d'erreurs pour 4 et 8 indices) quand il précédé d'un /r/ ou suivi d'un phonème des classes S ou Q. Les phonèmes /z/ et /v/ sont également confondus avec la classe E (12,44% et 6,67% d'erreurs pour 4 et 8 indices) quand ils sont précédés d'un phonème de cette même classe.

- ***pour la classe D*** : le phonème /d/ est confondu avec la classe N (22,47% et 12,36% d'erreurs pour 4 et 8 indices) quand il est précédé d'un phonème des classes A ou IN.

- ***pour la classe Q*** : les phonèmes /k/ et /t/ sont confondus avec la classe D (4,74% et 4,56% d'erreurs) quand ils sont précédés d'une voyelle ou d'un /n/.

- ***pour la classe O*** : le phonème /o/ est confondu avec la classe A (18,18% et 16,36 % d'erreurs) quand il est précédé ou suivi d'un /r/.

- ***pour la classe N*** : le phonème /n/ est confondu avec la classe D (8,51% et 10,64 % d'erreurs) quand il est précédé d'un phonème des classes A ou IN.

-***pour les classes U et S*** : aucune grosse erreur n'est détectée.

4 LES EXPERIENCES SUR LES MOTS

Le but de ce paragraphe est uniquement de montrer la répercussion des indices de convexité au niveau de la reconnaissance des mots d'un lexique.

4.1 LA COMPILATION DES MODELES DE MOTS

Pour cette application, les modèles des mots du corpus sont créés par compilation des modèles des classes de phonèmes. On entend par "compilation" non seulement la concaténation des états (en ayant soin d'ôter les états fictifs de début et de fin de phonème) mais aussi le réajustement des probabilités de transition entre phonèmes. On trouvera une description de la méthode chez Jelinek [Jelinek 76].

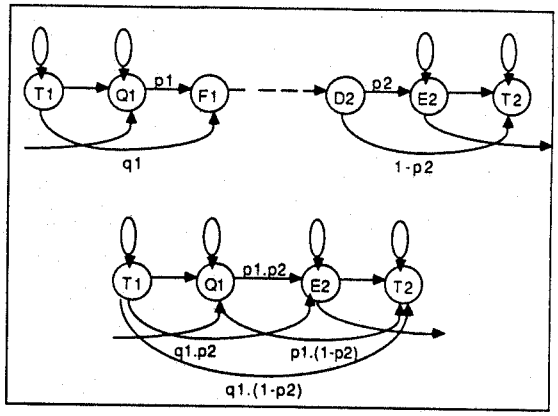


figure 5: La compilation des modèles de mots selon Jelinek.

4.2 LES RESULTATS

Un lexique comportant dix modèles créés suivant le principe précédent a permis de faire la reconnaissance de mots appartenant au corpus multilocuteur de base. On donne en fig.6 les résultats obtenus sur les modèles élaborés d'une part pour des vecteurs spectraux à 4 indices, d'autre part pour des vecteurs spectraux faisant intervenir ces mêmes indices et leurs convexités respectives. Les tests ont été effectués sur vingt occurences par modèle.

mots	transcription en classes	erreurs avec 4 indices	erreurs avec 4 indices + 4 convexités
DOUZE	DUZ	5%	0%
TRENTE	QRAQ QRAQE	10%	0%
QUARANTE	QARAQ QARAQE	50%	20%
CINQUANTE	SINQAQ SINQAQE	0%	0%
QUATRE-VINGT	QAQREZIN QAQRZIN	50%	15%
ZERO	ZIRO	0%	0%

figure 6: Mise en évidence de l'intérêt de la convexité au niveau de la reconnaissance des mots d'un sous-lexique du corpus. On notera au passage qu'un mot peut avoir plusieurs transcriptions en classes si l'on tient compte des différentes prononciations possibles.

Le corpus de taille beaucoup trop faible pour évaluer un système de reconnaissance permet cependant de montrer l'apport très significatif de la convexité au niveau de la reconnaissance des mots.

5 CONCLUSION

Nous avons pu vérifier qu'une meilleure prise en compte de l'évolution temporelle des indices acoustiques permet des améliorations importantes en reconnaissance de la parole.
Dans une communication [Dours 88], nous avons montré que l'introduction de contrôles d'évolution de scores dans le déroulement de l'algorithme de Viterbi permettait d'éviter de fausses détections.
Un contrôle local peut être fait par les indices de convexité qui ont par ailleurs l'avantage d'être robustes. Ceci nous a permis d'obtenir un gain de performance de plus de 10% en reconnaissance de mots connectés.

Nous avons également pu montrer que des entrées acoustiques sous forme de segments acoustiques infra-phonémiques donnent des résultats inférieurs à ceux obtenus à partir de représentations centisecondes. Sans doute faut-il voir là une conséquence d'un moins bon contrôle temporel.

6 REFERENCES

[Baker 74] J.K. Baker : "Stochastic modeling for Automatic Speech Understanding", Speech Recognition, R. Reddy (ed), p521-542, 1975.

[Barrera 87] C. Barrera, J.F. Malet, N. Vigouroux, G. Caelen-Haumont, J. Caelen 1987 : "Towards an automatic labelling system", 11eme ICPH, Tallinn Août 1987.

[Bourlard 85] H.Bourlard, Y.Kamp,C.J. Wellekens: "Speaker dependant connected speech recognition via phonemic Markov models", proc.IEEE Int.Conf. on Acoustic, Speech and Signal Processing, Tampa, FL, p31.5.1-4, 1985.

[Caelen 81] J. Caelen, G. Caelen-Haumont, 1981 : "Indices et propriétés dans le projet ARIAL II", actes du séminaire Encodage et Décodage Phonétiques, GALF-CNRS, Toulouse , p128-143, 1981.

[Caelen 83] J. Caelen, N. Vigouroux, G. Pérennou, 1983 : "Structure des informations acoustiques dans le projet ARIAL", Speech Communication n° spécial 2,3, p219-222, 1983.

[Descout 86] R. Descout, J.F. Serignat, O. Cervantes, R. Carré, 1986 : "Une base de données des sons du Français", 12th ICA, july 1986.

[Dours 88] C. Dours, G. Pérennou : "Role of intermediary scores in word spotting", congrès FASE, Edinburgh Août 1988 (à paraître).

[Jelinek 76] F. Jelinek : "Continuous speech recognition by statistical methods", Proc.IEEE vol. 64,n°4, p532-556, 1976.

[Jelinek 85] F. Jelinek : "The development of an experimental discrete dictation recognizer", Proc.IEEE vol.73, n°11, p1616-1623, nov.1985.

[Jouvet 87] D. Jouvet : "Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques", Congrès AFCET-IRIA Antibes, p65-72,1987.

[Levinson 83] S.E. Levinson, L.R. Rabiner,M.M. Sondhi: "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", Bell system tech. vol.62, n°4, p1035-1074,April 1983.

[Levinson 85] S.E. Levinson : "Structurals methods in automatic speech recognition", Proc.IEEE vol.73, n°11, p1625-1647, nov.1985.

[Liporace 82] L.A. Liporace : "Maximum likelihood estimation for multivariate observations of Markov sources", IEEE Trans. Inform. Theory, vol.IT-28, p129-136, 1982.

[Mérialdo 87] B. Mérialdo, A.M. Dérouault, M. Elbeze, S. Soudoplatoff : "Reconnaissance de parole avec un très grand vocabulaire", 16ème JEP Société Française d'Acoustique, p 161-164, Hammamet 5-9 oct 1987.

[Rabiner 83] L.R. Rabiner, S.E. Levinson, M.M. Sondhi: "On the application of vector quantization and hidden markov models to speaker-independant, isolated word recognition", Bell System Technology, J.62 n° 4 , p1100-1105, 1983.

[Vigouroux 85] N. Vigouroux, J. Caelen, 1985 : "Segmentations en vue de l'organisation d'une base de données acoustique et phonétique", 14ème JEP, GALF-CNRS PARIS, p152-155, 1985.

[Vigouroux 88] N. Vigouroux: Rapport partiel convention CNET n° 85 7B 00 790 92 45, Février 1988.

Apprentissage des modèles markoviens par maximum d'information mutuelle

Bernard MERIALDO

Centre Scientifique IBM France
5, place Vendôme, 75021 Paris, FRANCE

ABSTRACT

In this paper, we present a new method for training Hidden Markov models. This method tries to maximize the Mutual Information between the text that has been pronounced and the acoustic observation that corresponds to this utterance. While the usual training, based on Maximum Likelihood, take only the pronounced text into account, the Maximum Mutual Information training also considers all possible texts. This leads to a better discrimination between the pronounced text and the others. Experimentaly, the acoustic models trained using Maximum Mutual Information training are more accurate than those trained using Maximum Likelihood training.

INTRODUCTION

Les Modèles de Markov Cachés sont maintenant largement utilisés en Reconnaissance de la Parole. L'intérêt de ces modèles provient de leur capacité d'apprentissage automatique, et plus précisément du compromis qu'ils réalisent entre l'information fournie explicitement par le "concepteur" du modèle et celle extraite automatiquement des données d'apprentissage. Dans les applications actuelles:

- la structure du modèle (états et transitions) est fournie explicitement par le "concepteur", en utilisant ses propres connaissances sur le phénomène à modéliser (par exemple la décomposition phonétique),
- les paramètres du modèle (distributions de probabilités de transitions) sont appris automatiquement sur un certain volume de données d'apprentissage.

Historiquement, l'utilisation intensive des modèles de Markov a été provoquée par la découverte d'une inégalité par Baum [1]. Cette inégalité a permis de construire une procédure itérative qui calcule les valeurs optimales des paramètres selon un critère de maximum de vraisemblance (en fait la procédure ne garantit qu'un maximum local). Cette procédure est efficace d'un point de vue informatique, et permet de réaliser l'apprentissage des modèles complexes dont on a besoin en reconnaissance de la parole (plusieurs dizaines de milliers de paramètres, plusieurs centaines de phrases d'apprentissage...). Cette procédure est connue sous la dénomination de "algorithme de Baum-Welch" ou "algorithme Forward-Backward".

Une fois que l'on a choisi d'utiliser des modèles de Markov, on est en droit de se demander quelle est la meilleure procédure d'apprentissage à mettre en oeuvre, c'est-à-dire quelle est la procédure qui, à partir d'un corpus d'apprentissage donné, calculera les valeurs des paramètres qui donneront le meilleur taux de reconnaissance. Malheureusement, on ne connaît pas encore la réponse à cette question fondamentale. Les seuls résultats théoriques existants sur ce problème sont des résultats de convergence des estimées lorsque la taille de l'apprentissage croît infiniment, et ils supposent que certaines conditions de régularité soient vérifiées. Par exemple, Arthur Nadas [6] a prouvé que, sous certaines conditions, l'apprentissage par maximum de vraisemblance était le meilleur possible (au sens qu'il converge vers le modèle optimal avec la plus faible variance). Les conditions nécessaires sont que:

- le modèle optimal appartienne à la famille de modèles considérée,
- le vrai modèle de langage soit connu,
- le corpus d'apprentissage soit suffisamment "grand" (tende vers l'infini),
- les performances du système s'améliorent lorsqu'on approche du modèle optimal.

De ces conditions, les trois premières ne sont certainement pas vérifiées, tant il est vrai que personne n'ose espérer que la parole soit le résultat d'un processus markovien (au niveau acoustique ou linguistique), et que les corpus d'apprentissage sont toujours limités. Seule la quatrième condition peut laisser quelques espoirs, bien que sa validité mathématique semble difficile à démontrer.

Il reste donc qu'au delà des considérations théoriques en faveur de telle ou telle méthode d'apprentissage, seule l'expérimentation permet de comparer réellement deux méthodes.

Modèles de Markov et apprentissage

Dans ce paragraphe, nous rappelons quelques généralités sur les modèles de Markov et explicitons les apprentissages par maximum de vraisemblance (Maximum Likelihood ML) et par maximum d'information mutuelle (Maximum Mutual Information MMI).

Si l'on suit la présentation classique du problème de la Reconnaissance de la Parole comme un problème de Théorie de l'Information [2], on sait que le système de reconnaissance optimal est celui qui, à partir de l'observation acoustique A, produit le texte $\hat{W}$ vérifiant:

$$p(\hat{W} \mid A) = \max_{W} \frac{p(A \mid W) . p(W)}{p(A)}$$

Dans cet article, nous supposons que le modèle linguistique qui permet de calculer $p(W)$ a été construit, et nous nous intéressons à la construction du modèle acoustique $p(A \mid W)$.

Nous supposons qu'à partir du texte W, on sait construire un modèle de Markov (par exemple en concaténant des modèles phonétiques), et qu'alors $p(A \mid W)$ est la probabilité d'émission de l'observation A par ce modèle. Appelons θ

le vecteur des paramètres du modèle et notons $p_\theta(A \mid W)$ pour expliciter la dépendance du modèle par rapport à θ.

L'apprentissage du modèle consiste, étant donné un texte d'apprentissage W_T et un enregistrement A_T correspondant, à trouver la "meilleure" valeur pour θ.

L'apprentissage par maximum de vraisemblance ML définit cette valeur comme:

$$\underset{\theta}{\text{Argmax}}\ p_\theta(A_T \mid W_T)$$

L'apprentissage par maximum de l'information mutuelle MMI (proposé par L. Bahl et al [4] et P. Brown [5]) définit cette valeur comme:

$$\underset{\theta}{\text{Argmax}}\ p_\theta(W_T \mid A_T)$$

Explicitons cette seconde formulation:

$$\begin{aligned}\underset{\theta}{\text{Argmax}}\ p_\theta(W_T \mid A_T) &= \underset{\theta}{\text{Argmax}}\ \frac{p_\theta(A_T, W_T)}{p_\theta(A_T)} \\ &= \underset{\theta}{\text{Argmax}}\ \frac{p_\theta(A_T, W_T)}{p_\theta(A_T) \cdot p(W_T)} \\ &= \underset{\theta}{\text{Argmax}}\ Log\left(\frac{p_\theta(A_T, W_T)}{p_\theta(A_T) \cdot p(W_T)}\right) \\ &= \underset{\theta}{\text{Argmax}}\ I_\theta(A_T, W_T)\end{aligned}$$

(la seconde égalité utilise le fait que la probabilité linguistique $p(W_T)$ ne dépend pas de θ). $I_\theta(A_T, W_T)$ est l'information mutuelle entre l'acoustique et le texte, ce qui explique le nom donné à cette méthode d'apprentissage.

Un raisonnement simple permet de comprendre pourquoi l'apprentissage MMI est plus discriminant que l'apprentissage ML. Rappelons que le but de l'apprentissage est de trouver la valeur des paramètres qui donnera le système de reconnaissance le plus précis. Comme les seules données de Parole que l'on peut utiliser sont celles de l'apprentissage, il est cohérent de demander que le système appris reconnaisse parfaitement le texte d'apprentissage. C'est-à-dire, d'après la formulation du décodeur optimal donnée précédemment, que:

$$p_\theta(W_T \mid A_T) \geq p_\theta(W \mid A_T) \quad \forall\ W$$

En fait, il n'est pas certain qu'il existe une valeur de θ vérifiant la condition précédente, mais on peut remarquer que les deux apprentissages ML et MMI "essayent" de vérifier cette condition. En effet, cette condition peut se réécrire:

$$p_\theta(A_T \mid W_T) \cdot p(W_T) \geq p_\theta(A_T \mid W) \cdot p(W) \quad \forall\ W$$

L'apprentissage ML maximise le terme de gauche (indépendamment des termes de droite) en maximisant le facteur:

$$p_\theta(A_T \mid W_T)$$

alors que l'apprentissage MMI maximise directement le terme de gauche, en minimisant par là-même ceux de droite puisque l'on a la contrainte:

$$\sum_W p_\theta(W \mid A_T) = 1$$

Le fait que l'apprentissage MMI n'essaye pas seulement d'améliorer la probabilité du texte correct, mais essaye en même temps de minimiser la probabilité des autres textes possibles, conduit à une meilleure discrimination entre le texte correct et les autres textes lors de la reconnaissance.

L'apprentissage ML

Dans ce paragraphe, on rappelle la formulation mathématique de l'algorithme de Baum-Welch pour l'apprentissage ML, ce qui servira ensuite pour réaliser l'apprentissage MMI.

Prenons les notations suivantes:
- l'observation acoustique est une suite de symboles élémentaires $A = y_1^T = y_1 y_2 \ldots y_T$,
- a_{ij} est la probabilité de transition de l'état i à l'état j dans le modèle correspondant au texte W,
- b_{ijy} est la probabilité d'émission du symbole y pendant la transition i-j,
- c_i est la probabilité que le modèle soit au départ dans l'état i,
- le vecteur de paramètres est $\theta = (a_{ij}, b_{ijy}, c_i)$,

alors la probabilité d'émission peut être calculée par:

$$p(A \mid W) = \sum_{i_1} \ldots \sum_{i_T} c_{i_1} \prod_{j=1}^{T} a_{i_j i_{j+1}} \cdot b_{i_j i_{j+1} y_j}$$

Introduisons les variables α et β définies par:

$$\begin{cases}\alpha_i(t) = \text{probabilité que le modèle soit dans l'état i et ait émit } y_1^t \\ \beta_i(t) = \text{probabilité que le modèle émette } y_{t+1}^T \text{ en partant de l'état i}\end{cases}$$

Les valeurs de α et β sont calculées facilement en utilisant des relations de récurrence [1].

Alors le compte moyen de la transition i-j pendant l'émission de $A = y_1^T$ vaut:

$$c_{ij} = \frac{\sum_{t=1}^{T} \alpha_i(t-1) \cdot a_{ij} \cdot b_{ijy_t} \cdot \beta_j(t)}{p_\theta(A_T \mid W_T)}$$

La valeur réestimée $\bar{a}_{ij}$ de a_{ij} est proportionnelle à c_{ij}:

$$\bar{a}_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}$$

On définit de même les réestimées $\bar{b}_{ijy}$ et $\bar{c}_i$ de b_{ijy} et c_i, ce qui donne un vecteur réestimé $\bar{\theta} = (\bar{a}_{ij}, \bar{b}_{ijy}, \bar{c}_i)$. L'inégalité de Baum assure que:

$$p_{\bar{\theta}}(A_T \mid W_T) \geq p_\theta(A_T \mid W_T)$$

et que l'itération de ces formules de réestimation conduira à un maximum local de $p_\theta(A_T \mid W_T)$.

Remarquons que:

$$\sum_{ij} c_{ij} = T$$

et que l'on peut prouver que [1]:

$$\frac{\partial p_\theta(A_T|W_T)}{\partial a_{ij}} = \frac{1}{a_{ij}} \cdot p_\theta(A_T|W_T) \cdot c_{ij}$$

L'apprentissage MMI

Pour l'apprentissage MMI, on ne connaît pas de formule de réestimation, on applique donc une technique classique de maximisation par gradient avec contraintes, la fonction à maximiser étant:

$$F(\theta) = \log p_\theta(A_T|W_T) \cdot p(W_T) - \log \sum_W p_\theta(A_T|W) \cdot p(W)$$

avec les contraintes:

$$\sum_j a_{ij} = 1 \quad \forall i; \qquad \sum_y b_{ijy} = 1 \quad \forall i,j; \qquad \sum_l c_l = 1$$

La composante du gradient correspondant à a_{ij} est:

$$\frac{\partial F(\theta)}{\partial a_{ij}} = \frac{\dfrac{\partial p_\theta(A_T|W_T)}{\partial a_{ij}}}{p_\theta(A_T|W_T)} - \frac{\displaystyle\sum_W \frac{\partial p_\theta(A_T|W)}{\partial a_{ij}} \cdot p(W)}{\displaystyle\sum_W p_\theta(A_T|W) \cdot p(W)}$$

$$= \frac{1}{a_{ij}} \cdot (c_{ij} - c'_{ij})$$

où c_{ij} est le compte moyen de la transition i-j tel qu'il a été défini dans les formules de réestimation du paragraphe précédent, et c'_{ij} ressemble à une moyenne des comptes moyens des réestimées sur les textes possibles W.

Le second terme c'_{ij} est à priori difficile à calculer, puisqu'il suppose une sommation sur tous les textes possibles W. La solution proposée par L. Bahl et al. [4], pour un apprentissage de modèles de mots est d'approximer le second terme en réduisant la sommation aux seuls mots acoustiquement proches (d'après leur système de reconnaissance) du mot appris. P. Brown [5] effectue des expériences sur le "E-set", ce qui constitue un vocabulaire suffisamment réduit pour que la sommation ne pose pas de problème.

Dans cet article, nous proposons un autre moyen pour appliquer facilement l'apprentissage MMI à des modèles phonétiques, en utilisant le modèle phonétique rebouclé (Looped Phonetic Model LPM). Le LPM est réalisé en plaçant toutes les machines phonétiques en parallèle, et en ajoutant des transitions vides depuis l'état final de chaque machine phonétique ϕ vers l'état initial de chaque autre machine phonétique ϕ' avec une probabilité égale à la probabilité "biphonème" $p(\phi' \mid \phi)$ que le phonème ϕ' suive le phonème ϕ.

Le LPM permet d'effectuer une reconnaissance phonétique où la contrainte linguistique sur les suites de phonèmes est une contrainte biphonème. Un texte W est alors une suite de phonèmes ϕ_1^n et sa probabilité est:

$$p(W) = p(\phi_1^n) = \prod_{i=1}^{n} p(\phi_i | \phi_{i-1})$$

La reconnaissance d'une suite de phonèmes s'effectue simplement en appliquant l'algorithme de Viterbi au LPM.

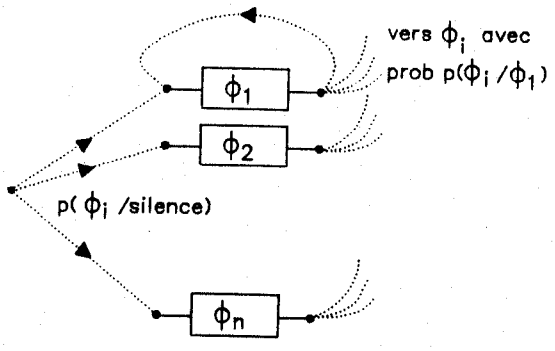


Figure 1: Modèle phonétique rebouclé LPM

L'intérêt du LPM pour l'apprentissage MMI est que l'on a l'égalité suivante:

$$\sum_W p_\theta(A_T|W) \cdot p(W) = p_\theta(A_T|LPM)$$

si bien que le LPM permet de calculer d'un seul coup le second terme de l'expression du gradient MMI, bien que ce terme corresponde à une sommation sur toutes les suites de phonèmes possibles. Il s'ensuit également que le terme c'_{ij} n'est autre que le compte moyen pour la réestimation sur le modèle LPM, et est donc très facilement calculé en appliquant l'algorithme de Baum-Welch sur ce modèle.

Les remarques précédentes permettent de réaliser le calcul du gradient MMI à partir de deux réestimations selon l'algorithme de Baum-Welch, une sur le modèle correspondant au texte d'apprentissage W_T, l'autre sur le LPM.

En pratique, le calcul du vrai gradient pose quelques problèmes d'erreurs d'estimation et de rapidité de convergence. Dans l'expression:

$$\frac{\partial F(\theta)}{\partial a_{ij}} = \frac{1}{a_{ij}} \cdot (c_{ij} - c'_{ij})$$

une valeur faible de a_{ij} peut conduire à concentrer l'effet du gradient sur des composantes correspondant à des transitions peu fréquentes, et à laisser les probabilités des transitions fréquentes pratiquement inchangées. Nous avons été amenés à remplacer ce vrai gradient par un vecteur "heuristique" qui supprime ces problèmes, et nous avons trouvé que le vecteur de coordonnées:

$$\left(\frac{c_{ij}}{\sum_k c_{ik}} - \frac{c'_{ij}}{\sum_k c'_{ik}}\right)$$

(c'est-à-dire en normalisant les comptes, et en ne divisant pas par a_{ij}) donnait les meilleurs résultats. Expérimentalement, ce vecteur a toujours conduit à une augmentation de l'information mutuelle. C'est ce vecteur que nous avons donc utilisé dans les expériences décrites au prochain paragraphe.

Résultats expérimentaux

Nous utilisons un système phonétique à 40 machines [8]. Les enregistrements sont réalisés dans une cabine insonorisée, à l'aide du "Yorktown acoustic front-end" [7]. Celui-ci consiste en:

- un micro-table PZM,
- un convertisseur A/D qui échantillonne le signal de parole à 20kHz,
- un processeur de signal qui effectue une analyse spectrale, suivie de l'application d'un modèle d'oreille et d'une quantification vectorielle (200 classes) pour produire une observation centiseconde.

5 locuteurs masculins, non professionnels (c'est-à-dire n'appartenant pas à un groupe de reconnaissance de la parole), ont enregistré deux textes d'apprentissage: T_{200} composé de 200 phrases phonétiquement équilibrées, et T_{250} composé de 250 phrases contenant des phonèmes en contexte [11], plus un texte de reconnaissance T_{79} composé de 79 phrases extraites de lettres. Tous les locuteurs ont prononcé ces textes en mode Syllabes Isolées, c'est-à-dire en laissant une courte pause entre chaque syllabe.

Pour chaque locuteur, on a appris un modèle acoustique par apprentissage ML (sur T_{200} et T_{250}). Ce modèle sert de point de départ pour l'apprentissage MMI (sur T_{200}). L'apprentissage MMI consiste en une série de déplacements à pas variables le long du vecteur "heuristique". A chaque pas, on calcule l'information mutuelle, et on effectue une reconnaissance phonétique sur T_{79} pour étudier la qualité du modèle (mais ce résultat n'est pas utilisé par la procédure d'apprentissage). L'évolution de l'information mutuelle et du taux d'erreurs phonétiques en fonction du nombre d'itérations est présenté pour chaque locuteur dans les figures suivantes.

Nous indiquons également dans une table les valeurs suivantes:

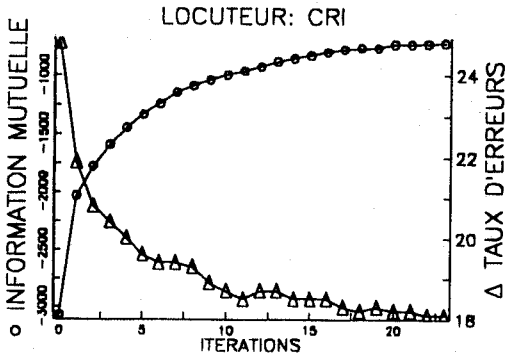
- le taux de reconnaissance phonétique:

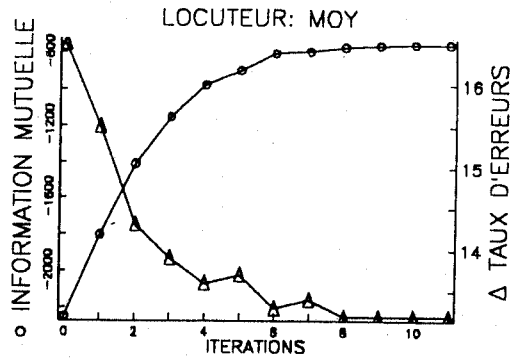
$$\frac{\text{nb. phonèmes reconnus}}{\text{nb. phonèmes}}$$

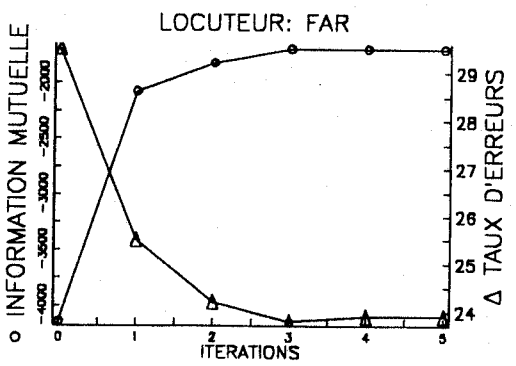
- le taux d'insertions: $\dfrac{\text{nb. insertions}}{\text{nb. phonèmes}}$
- le taux d'erreurs global, calculé selon la formule [10]:

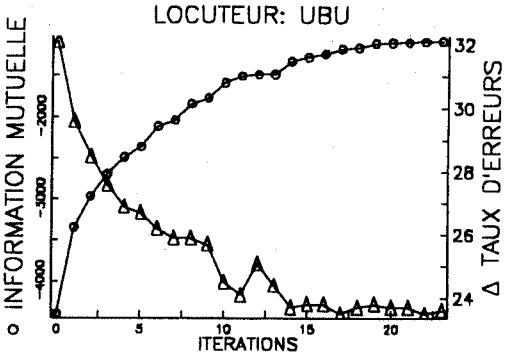
$$\frac{\text{substitutions + délétions + insertions}}{\text{nb. phonèmes + insertions}}$$

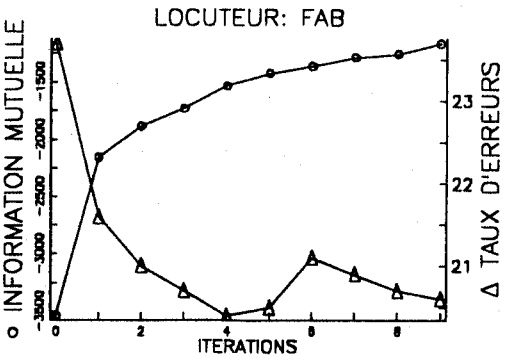
à la fois pour l'apprentissage ML et pour l'apprentissage MMI.











locuteur		rec.	ins.	erreurs
CRI	ML	84.1	11.8	24.8
	MMI	85.2	4.1	18.1
MOY	ML	88.2	5.7	16.5
	MMI	89.7	3.3	13.2
FAR	ML	79.4	12.6	29.5
	MMI	80.4	5.7	23.9
UBU	ML	79.2	16.7	32.1
	MMI	81.1	6.1	23.6
FAB	ML	80.9	6.1	23.7
	MMI	81.4	2.6	20.6
moyenne	ML	82.36	10.58	25.32
	MMI	83.56	4.36	19.88

Ces résultats montrent que, dans tous les cas, l'apprentissage MMI améliore fortement les performances du modèle. La plus grande partie de ce gain vient de la réduction des insertions, en moyenne 6%. Le taux de reconnaissance augmente lui d'un peu plus de 1% en moyenne. Le taux d'erreurs global moyen diminue de 5,5%.

CONCLUSION

Nous avons présenté une méthode d'apprentissage fondée sur un principe de maximisation de l'information mutuelle (MMI) qui conduit à une meilleure discrimination que l'apprentissage par maximum de vraisemblance ML classique. Nous avons proposé une façon originale d'application de cette méthode aux modèles phonétiques, et une variante "heuristique" qui conduit à une meilleure convergence.

Cette méthode a été expérimentée sur 5 locuteurs et a montré une diminution du taux d'erreurs phonétiques de 5,5% en moyenne, par rapport à l'apprentissage ML.

Bibliographie

[1] L. Baum, *An inequality and association Maximization technique in Statistical Estimation for Probabilistic Function of Markov Processes*, Inequality, Vol III, 1972, pp 1-8.

[2] L. Bahl, F. Jelinek, R. Mercer, *A Maximum likelihood Approach to Continuous Speech Recognition*, IEEE Trans on PAMI, PAMI-5 No 2, March 83.

[3] L. Bahl, P. Brown, P. de Souza, R. Mercer, *Maximum Mutual Information of Hidden Markov Model Parameters*, ICASSP 86, Tokyo, vol 1 pp 49.

[4] L. Bahl, P. Brown, P. de Souza, R. Mercer, *Estimating HMM parameters so as to maximize Speech Recognition accuracy*, Research Report RC-13121, 9/10/87, IBM TJ Watson Research Center, PO Box 218, Yorktown Heights, NY10598.

[5] P. Brown, *The Acoustic-Modeling Problem in Automatic Speech Recognition*, PhD Dissertation, Carnegie Mellon University, May 87.

[6] A. Nadas, *A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood*, IEEE trans. on ASSP, ASSP-31 No 4, August 1983, pp 814-817.

[7] A. Averbuch et al., *An IBM PC Based large-vocabulary isolated-utterance speech recognizer*, ICASSP 86, Tokyo, vol 1 pp 53.

[8] H. Cerf-Danon, A-M Derouault, M. El-Beze, B. Merialdo, S. Soudoplatoff, *Speech Recognition experiment with 10,000 word vocabulary*, NATO Advanced Institute on Pattern Recognition, June 18-20, 1986, Brussels..

[9] B. Merialdo, *Speech Recognition using Very Large Size dictionary*, ICASSP 87, Dallas.

[10] R. Schwartz, Y. Chow, F. Kubala, *Rapid speaker adaptation using a probabilistic spectral mapping*, ICASSP 87, Dallas, vol 2, pp 633-636.

[11] A-M. Derouault, *Context-dependent phonetic Markov models for Large Vocabulary speech recognition*, Proc. of NATO Advanced study Institute on Speech Understanding, 6-18 July 1987, Ed. Springer-Verlag.

Adaptation en Cours de Reconnaissance d'un Dictionnaire de Références Phonétiques, à un Nouveau Locuteur[1]

H. Y. SU
R. ANDRE-OBRECHT

IRISA, Campus de Beaulieu, 35042 Rennes cedex

ABSTRACT

Vector quantization has been employed to realize an acoustic-phonetic recognition system with success for continuous speech. The main idea is to build a dictionary by building first the sub-dictionaries corresponding to the phonetic classes, from a training set, using vector quantization; the recognition is then just a comparison of a vector with the elements of this dictionary. The accuracy of this system in the mono-speaker case is about 70% and 90% using the nearest neighbor rule and the 3 nearest neighbors rule respectively, but it is getting unacceptable (about 40% and 75%) for a speaker-independent recognition even with a pluri-speaker dictionary built from a 3-speaker training set. Adaptive vector quantization is then used to improve the quality of the pluri-dictionary **in the process of recognition** (the accuracy has been improved from 39% to 67% for the nearest neighbor rule and from 80% to 90% for the 3 nearest neighbors rule).

I - INTRODUCTION

L'indépendance d'un système de reconnaissance de parole vis-à-vis du locuteur est un problème difficile. De nombreuses études ont été menées ces dernières années, plus particulièrement dans le cadre de systèmes faisant intervenir des ***dictionnaires de références***.

Dans le cas spécifique de l'emploi de la quantification vectorielle, deux approches ont été jusqu'ici étudiées:

- construction d'un dictionnaire multi-locuteur [2,5,7]; l'ensemble d'apprentissage est suffisamment conséquent (nombre de locuteurs, volume de données par locuteur) pour que le dictionnaire soit représentatif de tout locuteur. Il faut cependant noter que ce dictionnaire n'est employé qu'à des fins de codage et non d'identification: c'est le prétraitement avant l'utilisation d'algorithmes de reconnaissance de type DTW, HMM,...
- adaptation d'un dictionnaire standard à un locuteur [3,8]; l'idée est de trouver un couple de projecteurs vers un espace commun où la variabilité inter-locuteur sera réduite. Il est nécessaire de considérer une phase d'apprentissage pour chaque locuteur, même si elle peut être très réduite.

Notre approche diffère des précédentes dans la mesure où nous nous sommes fixés comme but une adaptation au locuteur, ***dynamique sans nouvelle phase d'apprentissage***, en cours de reconnaissance.

[1]. Etude réalisée dans le cadre de la convention CNET/INRIA N°86 7B029007909245 LAA/TSS/DAP

Ce travail fait suite à une étude sur les algorithmes de quantification vectorielle dans un cadre mono-locuteur [10], en vue d'une reconnaissance phonétique. Nos expérimentations nous ont permis de constater que l'emploi d'un dictionnaire pluri-locuteur ne pouvait être envisagé dans le cadre d'une identification phonétique (paragraphe **III**) et qu'il n'était pas possible d'adapter un dictionnaire mono-locuteur à un nouveau locuteur, sans une "certaine" phase d'apprentissage. Ces réflexions nous ont conduits à la démarche suivante:

-construire un dictionnaire pluri-locuteur,
-adapter ce dictionnaire au locuteur en cours de reconnaissance.

Cette adaptation peut être réalisée à l'aide d'un algorithme de gradient stochastique ou un algorithme de Lloyd généralisé. L'évaluation des deux algorithmes a été réalisée dans le cadre d'un système de reconnaissance de mot isolés. Bien que le vocabulaire soit restreint à 21 mots et petites phrases, le taux de reconnaissance phonétique pour un nouveau locuteur passe de 49% à 68% pour le plus proche voisin et de 80% à 90% pour les trois plus proches voisins, après une adaptation en cours de reconnaissance de 105 prononciations.

Après un rappel de l'algorithme de quantification vectorielle utilisé et des performances du dictionnaire en terme de taux de reconnaissance phonétique (paragraphe **II**), nous décrirons les deux algorithmes d'adaptation (paragraphe **III**) et leur mise en œuvre dans le cas d'un système de reconnaissance de mots isolés (paragraphe **IV**).

II - DESCRIPTION DE L'ALGORITHME DE QUANTIFICATION VECTORIELLE

II.1 - Construction du dictionnaire

Nous nous proposons d'utiliser la quantification vectorielle dans le cadre d'une reconnaissance phonétique des parties homogènes (ou quasi-homogènes) du signal de parole; ces zones ne sont pas des zones stationnaires, mais elles correspondent à des zones de signal "spectralement stables" [1]. Nous construisons un dictionnaire à partir de sous-dictionnaires, chacun d'eux correspondant à une classe phonétique. Nous utilisons présentement 21 classes phonétiques:

[a]=/ a,ɑ/	[i]=/ i /	[j]=/ j /
[y]=/ y,ɥ /	[u]=/ u /	[w]=/ w /
[E]=/ e,ɛ /	[&]= / ø, ɔ ,œ /	[o]= / o, ɔ /

[on]= / ɔ̃ /	[an]=/ ɑ̃ /	[un]=/ œ̃, ɛ̃ /
[m]=/ m /	[n]=/ n /	[bv]=/ b,d, ʒ ,v /
[l]=/ l /	[ge]=/ ʒ /	[r]=/ r /
[s]=/ s /	[ch]=/ ʃ /	[f]=/ f /

A partir d'un ensemble de données d'apprentissage (3 listes de 10 phrases phonétiquement équilibrées par locuteur[2]), nous utilisons la segmentation manuelle: elle consiste à sélectionner manuellement les parties "homogènes" du signal et à les répartir en 21 sous-ensembles d'apprentissage. Chaque zone est décomposée en blocs (de 1 à 3 par zone) de 40 ms, éventuellement non disjoints; chacun est paramétré après une analyse de Fourier (24 canaux, échelle Mel)[3]. L'algorithme de Lloyd couplé à la méthode de "splitting" avec arrêt sur un seuil (**LBG** [6],[9]) est utilisé sur chaque sous-ensemble d'apprentissage pour obtenir les 21 sous-dictionnaires comprenant chacun en moyenne 7 éléments (figure 1). Le dictionnaire global est la simple réunion de ces sous-dictionnaires.

Pour un dictionnaire mono-locuteur, la taille du dictionnaire global est de l'ordre de 150. Le dictionnaire pluri-locuteur, obtenu par apprentissage sur 2 hommes et 1 femme (4058 vecteurs) comporte 512 éléments (une étude détaillée de ces dictionnaires est décrite dans [9]).

L'avantage de cette construction du dictionnaire de références est que chaque élément du dictionnaire porte une seule étiquette, ce qui facilite considérablement la procédure de reconnaissance (nous n'avons qu'à trouver les plus proches voisins d'un vecteur d'entrée dans les sous-dictionnaires différents) et l'adaptation du système au locuteur (comme nous le verrons).

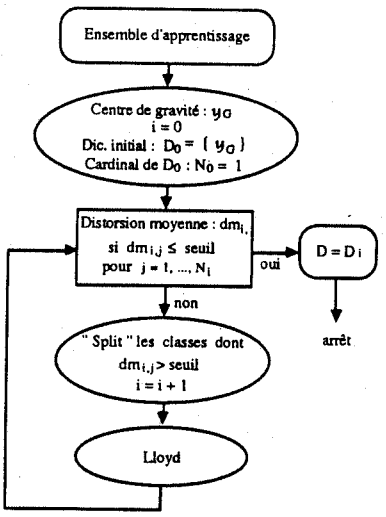


Figure 1. La méthode de "splitting" avec l'arrêt sur seuil

II.2 - Expérimentation en reconnaissance

En phase de reconnaissance, nous utilisons un pré-traitement automatique afin d'extraire du signal les parties "homogènes". Ce traitement se compose d'une segmentation automatique et d'un étiquetage événementiel et transitoire [1,11]. Sur les segments homogènes, est centrée et paramétrée une fenêtre de 20 ou 40 ms (selon la longueur du segment).

[2] Données fournies par le CNET, Lannion

[3] Les fréquences de coupure ont été données par le CNET-Lannion

Les tests de reconnaissance sont faits à partir de 5 locuteurs, E_i, i = 1,...,5, dont trois ont participé à l'apprentissage E_1, E_2, E_3. Pour chaque locuteur, un ensemble-test est réalisé à partir de 2 listes de 10 phrases phonétiquement équilibrées. Ces listes sont différentes de celles utilisées en apprentissage, elles correspondent à 450 phonèmes traités, le cardinal de chaque ensemble-test est de l'ordre de 570 (pour E_5, locuteur féminin, cinq listes sont traitées, pour obtenir 1280 éléments).

Règle de reconnaissance \ Ensemble-test	E_1	E_2	E_3	E_4	E_5
Plus proche voisin	$70_{\pm 4}$	$65_{\pm 4}$	$67_{\pm 4}$	$43_{\pm 4}$	$58_{\pm 3}$
2 plus proches v.	$83_{\pm 3}$	$85_{\pm 3}$	$82_{\pm 3}$	$61_{\pm 4}$	$78_{\pm 2}$
3 plus proches v.	$91_{\pm 2}$	$91_{\pm 2}$	$89_{\pm 3}$	$74_{\pm 4}$	$87_{\pm 2}$

Figure 2. Taux de reconnaissance phonétique pondérés par l'intervalle de confiance à 95%, sur les ensembles-test à partir d'un dictionnaire pluri-locuteur

Les expériences de cette reconnaissance phonétique (figure 2) montrent de bons résultats lors de l'utilisation pluri-locuteur (E_1, E_2, E_3). Ils se dégradent lorsque le locuteur est inconnu du système; la reconnaissance d'un nouveau locuteur par un dictionnaire mono-locuteur est très laborieuse (figure 3).

Cette remarque nous conduit naturellement à la possibilité d'adapter un dictionnaire pluri-locuteur à un nouveau locuteur pour en faire un mono-locuteur.

Ensemble test \ Dictionnaire	D_1 (H)	D_2 (H)	D_3 (F)	D_4 (H)
E_1 (H)	75 ±3	52 ±4	32 ±4	38±4
E_2 (H)	55 ±4	70 ±4	44 ±4	51±4
E_3 (F)	38 ±4	43 ±4	65 ±4	42±4
E_4 (H)	40 ±4	38 ±4	40 ±4	54±4

(a)

Ensemble test \ Dictionnaire	D_1 (H)	D_2 (H)	D_3 (F)	D_4 (H)
E_1 (H)	94±2	80±3	64±4	73±4
E_2 (H)	84±3	94±2	71±4	79±3
E_3 (F)	63±4	67±4	88±3	67±4
E_4 (H)	66±4	65±4	65±4	82±3

(b)

Figure 3. Taux de reconnaissance phonétique pondérés à partir de dictionnaires mono-locuteur en fonction des locuteurs

(a). règle du plus proche voisin

(b). règle des trois plus proches voisins

III - ALGORITHMES D'ADAPTATION

III.1 - Adaptation du dictionnaire par algorithme de gradient stochastique

Dans sa thèse de 3ème cycle, B.Delyon [3] a décrit un algorithme de gradient stochastique pour "rafraîchir" un dictionnaire par rapport à un nouvel ensemble d'apprentissage dans le domaine de la quantification vectorielle. Le but de cette opération est de modifier le dictionnaire pour minimiser la moyenne de la dégradation:

$$E\{ f(x-\hat{x}) \}$$

où $f(x-\hat{x})$ est une fonction dépendante de la métrique, dans l'espace gaussien et représentative d'un coût, nous pouvons avoir

$$f_1(x-\hat{x}) = \frac{1}{2}\|x-\hat{x}\|^2 + C_1$$

ou

$$f_2(x-\hat{x}) = \frac{1}{2}\|x-\hat{x}\|^2 1_{\|x-\hat{x}\|\le R} + \frac{1}{2}R\|x-\hat{x}\| 1_{\|x-\hat{x}\|>R} + C_2$$

où R est un rayon à choisir.

Pour un dictionnaire donné par

$$D = \{ y_i , i = 1, 2, \ldots, N \},$$

nous avons:

$$f(x-\hat{x}) = \inf_i f(x-y_i)$$

$$= f(x-y_j) \wedge \inf_{i\neq j} f(x-y_i)$$

et les dérivations:

$$\frac{\partial}{\partial y_j} f_1(x-\hat{x}) = (y_j - x) 1_{\hat{x}=y_j}$$

$$\frac{\partial}{\partial y_j} f_2(x-\hat{x}) = \left[(y_j-x)\times 1_{\|x-y_j\|\le R} + \frac{y_j-x}{\|y_j-x\|}\times R \times 1_{\|x-y_j\|>R} \right] \times 1_{\hat{x}=y_j}$$

L'algorithme de gradient stochastique s'écrit alors : en supposant que x_n soit reconnu comme le j^{ème} élément du dictionnaire D_n,

$$\hat{x}_n = y_j^n ,$$

nous avons:

$$y_j^{n+1} = y_j^n + \lambda (x_n - y_j^n)$$

ou

$$y_j^{n+1} = y_j^n + \lambda (x_n - y_j^n) \times 1_{\|x_n - y_j^n\|\le R} + \lambda R \times 1_{\|x_n - y_j^n\|>R}$$

y_j^{n+1} remplace y_j^n dans le nouveau dictionnaire D_{n+1}, les autres éléments restent inchangés. λ est un pas à choisir : il peut être soit une constante (pas constant), soit une variable décroissante (pas décroissant).

Cet algorithme de mise à jour du dictionnaire est très rapide mais la performance dépend malheureusement beaucoup de la fiabilité de la reconnaissance du système. De plus, nous devons choisir le pas λ, et l'éventuel rayon R.

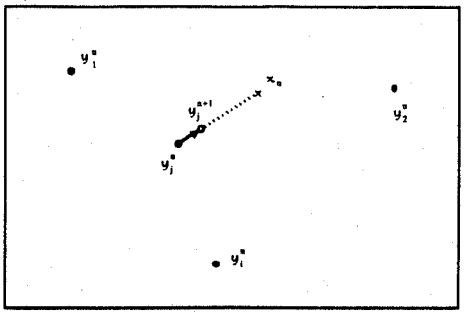


Figure 4. n^{ième} adaptation du dictionnaire: $\hat{x}_n = y_j^n$, y_j^n est remplacé par le nouvel élément y_j^{n+1}, $y_i^{n+1} = y_i^n$, $i \neq j$, reste inchangé

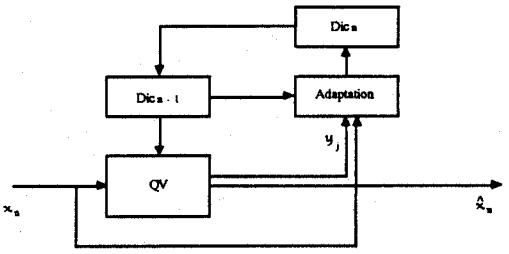


Figure 5. Adaptation du dictionnaire par algorithme de gradient stochastique en cours de reconnaissance

III.2 - Adaptation du dictionnaire par algorithme de Lloyd généralisé

C'est un sous-produit de la méthode de construction du dictionnaire par classe phonétique: la reconnaissance d'un ensemble de phrases prononcées par le locuteur étranger à l'aide du dictionnaire initial D, nous permet de construire un ensemble d'"apprentissage automatique"; nous appliquons l'algorithme de Lloyd à cet ensemble et le dictionnaire D, afin d'obtenir un nouveau dictionnaire D' plus adapté au nouveau locuteur.

Ce processus se déroule de la manière suivante: au cours de la reconnaissance, chaque vecteur x_n, reconnu comme appartenant à la j^{ème} classe phonétique, est étiqueté de cette classe et mémorisé, d'où l'ensemble d'"apprentissage automatique".

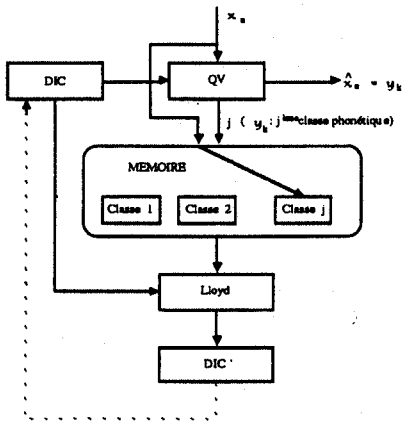


Figure 6. Système d'adaptation du dictionnaire par algorithme de Lloyd

Il est évident que cet ensemble peut être plus ou moins grand, et plus il sera grand, mieux le dictionnaire sera adapté au locuteur. Mais nous devons choisir un compromis entre place mémoire et taille de l'ensemble. La performance de cet algorithme dépend de la qualité de l'ensemble d'"apprentissage automatique", qui dépend elle du taux de reconnaissance à partir du dictionnaire initial.

IV - MISE EN ŒUVRE DE L'ADAPTATION DANS UN SYSTEME DE RECONNAISSANCE

IV.1 - Introduction d'un module de calage

Les deux algorithmes que nous avons décrits précédemment, utilisés brutalement en prenant comme valeur de λ_n, le plus proche voisin montre que le taux de reconnaissance phonétique n'est pas assez élevé pour espérer une réelle adaptation; le dictionnaire ne se dégrade cependant pas [9]. Il est donc nécessaire pour évaluer ce type d'adaptation de l'intégrer dans un système de reconnaissance réel, ou du moins d'utiliser un module de calage lexical.

Le module de calage a pour but de choisir le bon candidat parmi les trois candidats proposés par la reconnaissance acoustico-phonétique et même éventuellement de corriger cette proposition (figure 5). Nous utilisons un module lexical, développé à l'IRISA par P.Frison, consistant à comparer la suite des étiquetages d'un mot (ou d'une petite phrase) avec les éléments d'un ensemble de références (le vocabulaire phonétique du système), et à trouver la plus vraisemblable (modèle markovien). Par exemple, la phrase "connecteur visible /kɔnɛktœ:r vizibl/" a été étiquetée de la manière suivante par le module de la reconnaissance acoustico-phonétique, et calée avec la référence trouvée [..-k-o-n-E-..-k-..-t-&-r-bv-i-z-i-bv-l-&]:

numéro de segment	plus proche voisin	2ème plus proche voisin	3ème plus proche voisin	référence trouvée	sous-dic. à adapter
1	..	..	..	..	--
2	**	**	**	k	--
3	o	on	w	o	[o]
4	m	n	l	n	[n]
5	y	&	E	E	[E]
6	..	..	..	..(k	--
7	..	..	..	..	--
8	**	**	**	t	--
9	&	E	r	&	[&]
10	r	s	l	r	[r]
11	bv	m	n	bv	[bv]
12	E	bv	j	i	[i]
13	s	bv	&	z	--
14	i	j	l	i	[i]
15	bv	&	w	bv	[bv]
16	&	E	l	RR	[l]
17	l	bv	&	l	[l]
18	bv	l	E	II	--
19	bv	i	&	&	[&]
20	..	..	..	..	--
21	..	..	..	-/	--

où

(k = omission de /k/
** = phonème explosion
RR = répétition
I I = insertion
.. = silence
-/ = fin d'une suite

Ce calage permet d'éviter les fausses adaptations, par exemple:

- le 19ème vecteur de "connecteur visible" ne servira qu'à modifier le sous-dictionnaire correspondant à la classe phonétique de [&] (et non pas celui de [bv]);
- le 13ème vecteur ne servira pas car [z] n'existe pas dans le dictionnaire;
- le 18ème non plus car considéré comme inséré par le module du calage, il est sans doute transitoire;
- le 12ème servira à modifier le sous-dictionnaire correspondant à la classe phonétique de [i] malgré son absence parmi les trois candidats;
- le 16ème est considéré comme répétition de [l] et servira évidemment à modifier le sous-dictionnaire de [l].

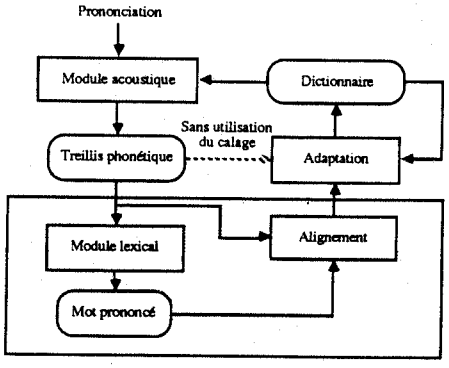


Figure 7. Utilisation du module de calage pour l'adaptation du dictionnaire

Ce module lexical est limité à un vocabulaire de 100 mots, mais il est évident que tout autre type de calage, fournissant le même type de résultat, peut être utilisé. L'adaptation par l'algorithme de gradient stochastique n'aura lieu qu'à la fin d'une reconnaissance complète d'un mot, après la validation de la reconnaissance acoustico-phonétique.

IV.2 - Expérimentations et résultats

Une première expérience est faite en utilisant une base de données de 210 occurrences: 21 mots et petites phrases, prononcés chacun 10 fois par un locuteur masculin différent des cinq locuteurs mentionnés précédemment. Pour les besoins de l'étude, ils ont été regroupés en deux bases (105 mots chacune) qui nous donnent l'ensemble pour l'adaptation et l'ensemble-test (787 vecteurs et 888 vecteurs respectivement). En fonctionnement normal, l'adaptation se fait de manière continue.

Le tableau suivant donne les résultats de la reconnaissance sur l'ensemble-test avec le dictionnaire pluri-locuteur D et les dictionnaires adaptés par les deux méthodes proposées.

Méthode d'adaptation / Règle de reconnaissance	Algorithme de Lloyd	Algorithme de gradient stochastique		Reconnaissance indépendante du locuteur (avec D)
		f_1	f_2	
plus proche voisin	67 ±3	57 ±3	51 ±3	39 ±3
2 plus proches v.	83 ±3	74 ±3	72 ±3	65 ±3
3 plus proches v.	90 ±2	85 ±2	84 ±2	80 ±3

Figure 8. Taux de reconnaissance acoustico-phonétique pondérés par l'intervalle de confiance à 95%, sur l'ensemble-test avant et après l'adaptation automatique du dictionnaire en cours de reconnaissance

Il est à remarquer que

-les dictionnaires adaptés sont nettement meilleurs que le dictionnaire initial;

-bien que l'algorithme de Lloyd ne soit pas séquentiel (car il nécessite un ensemble d'"apprentissage automatique"), c'est une excellente méthode d'adaptation, puisque le taux de reconnaissance passe de 39% à 67% pour le plus proche voisin;

-l'algorithme de gradient stochastique donne de bons résultats, et l'adaptation avec la fonction f_1 est plus efficace que celle avec f_2. Ce phénomène s'explique sans doute par le fait qu'après la correction des erreurs (calage), les vecteurs qui participent à l'adaptation sont des vecteurs justes, il est normal de ne pas limiter leur action.

Pour apprécier la valeur des résultats précédents, cette expérience a été renouvelée à partir du même vocabulaire et du même dictionnaire initial, auquel a été ajouté un sous-dictionnaire de [z] = /z/, obtenu par un apprentissage sur les nombres de 0 à 99 prononcés par trois locuteurs (2 hommes et 1 femme, base de données du GRECO Communication Parlée). Pour chaque nouveau locuteur (un homme CL et une femme VL), un enregistrement de 20×21 prononciations a été réalisé sans précaution dans un environnement de type salle-machine (avec un système OROS-AI). L'adaptation du dictionnaire a été faite par l'algorithme de Lloyd en cours de reconnaissance de 105 prononciations, et a confirmé les résultats précédents (figure 9, ensemble-test CL=2289 éléments et VL=2419 éléments).

Dictionnaire utilisé / ensemble-test / Règle de reconnaissance	Dictionnaire initial (D)		Dictionnaire adapté par Lloyd	
	CL (homme)	VL (femme)	CL	VL
plus proche voisin	29 ±2	32 ±2	64 ±2	56 ±2
2 plus proches v.	48 ±2	48 ±2	82 ±2	73 ±2
3 plus proches v.	63 ±2	58 ±2	88 ±1	82 ±2

Figure 9. Taux de reconnaissance acoustico-phonétique pondérés par l'intervalle de confiance à 95%, sur les ensembles-test avant et après l'adaptation du dictionnaire par Lloyd

V - CONCLUSION

Deux méthodes de quantification vectorielle adaptative sont proposées, justifiées par la façon de construire le dictionnaire de références par classe phonétique. L'intérêt de cette approche est de réaliser un système de reconnaissance de parole continue capable de s'adapter automatiquement au futur locuteur en cours de reconnaissance.

La comparaison de ces méthodes montre que l'algorithme de Lloyd généralisé est un excellent algorithme d'adaptation, rapide et efficace, et qu'à son avantage, l'algorithme de gradient stochastique permet une adaptation **permanente** sans interruption du système de reconnaissance et sans coût réel.

Dans notre système d'adaptation présenté, nous avons utilisé un module de calage pour augmenter la fiabilité de l'étiquetage automatique à partir du dictionnaire initial. Il est évident que c'est un module indispensable pour rendre l'adaptation possible, mais qu'il peut être remplacé par beaucoup d'autres types de module lexical.

VI - REFERENCES

[1] R. ANDRE-OBRECHT, H.Y. SU et B. DELYON : ***"Expériences en vue du Décodage Acoustico-Phonétique à Partir d'une Recherche Statistique d'Evénements Articulatoires et d'un Codage Vectoriel"***. 16e JEP p.64-67, 1987, à paraître dans la Revue d'Acoustique

[2] D.K. BURTON, J.E. SHORE, & J.T. BUCK : ***"Isolated-word speech recognition using multisection VQ codebooks"***. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 n°4, August 1985

[3] K. CHOUKRI and G. CHOLLET : ***"Adaptation of Automatic Speech Recognizers to new speakers using cononical correlation analysis techniques"***. Computer Speech and Language, Vol.1 n°2, December 1986

[4] B. DELYON : ***"Un théorème de limite centrale pour certaines équations différentielles aléatoires"***. Thèse de 3ème cycle, Université Pierre et Marie CURIE, Juillet 1986

[5] E. KOPEC, & M.A. BUSH : ***"Network-based isolated digit recognition using vector quantization"***. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 n°4 pp. 850-867, August 1985

[6] Y. LINDE, A. BUZO, & R.M. GRAY : ***"An algorithm for vector quantizer design"***. IEEE Transactions on Communications COM-28 pp. 84-95, January 1980

[7] L.R. RABINER, & S.E. LEVINSON : ***"A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building"***. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 , June 1985

[8] K. SHIKANO, K.F. LEE and R. REDDY : ***"Speaker Adaptation through Vector Quantization"***. Rapport interne, Carnegie Mellon University, December 1986

[9] SU Huan-yu : ***"Reconnaissance Acoustico-Phonétique en Parole Continue par Quantification Vectorielle; Adaptation du Dictionnaire au Locuteur"***. Thèse de l'Université de Rennes I, 1987

[10] SU Huan-yu : ***"Utilisation de la Quantification Vectorielle en reconnaissance de la Parole Continue"***. 15e JEP p.247-249, 1986

[11] H.Y. SU, R. ANDRE-OBRECHT, B. DELYON, V. LE MAIRE, A. MORIN : ***"Etude de Méthodes statiques pour le Décodage Acoustico-Phonétique de Parole Continue"***. Rapport de convention CNET/INRIA N° 86 7B029007909245 LAA/TSS/DAP, Janvier 1988

COMMANDE VOCALE D'UN ROBOT MANIPULATEUR

Jean-Marie CONDOM André LOZES

Laboratoire CERFIA, UA-CNRS N°824

Université Paul Sabatier, 118 route de Narbonne - 31062 TOULOUSE CEDEX

ABSTRACT

We present here a voice-controlled robot able to move blocks. This application, for teaching goals, uses two types of commands : on one hand, analogic orders telling the robot in which direction to move ("go to_the_left" for example), and on the other hand, discrete commands specifying the goal to be reached ("approach the_block X", for example) and necessitating for the robot to know its environment and to reason about the tasks that are given to him.

Recognizing connected words allows creating multiple analogic commands leading to a variety of motions. The slowness of the speech rate, however, results in a longer response time when compared with a system operated by remote control which is immediate. Spoken language is better adapted to giving discrete commands because the response time does not have to be immediate. However, spoken language necessitates a dialog management module in order to attain an efficient and user-friendly interactive system.

1- Introduction

Le Salon Internationnal des Techniques et des Energies du Futur (SITEF) organisé à Toulouse en septembre 1987 nous a donné l'occasion d'exposer la première version d'une interface vocale pour un bras manipulateur. Ce prototype constitue un support expérimental pour l'enseignement de Robotique, Reconnaissance des Formes et Intelligence Artificielle (IRR) ayant déjà mis à contribution trois groupes d'étudiants en bureau d'étude : Luc-Henri PAMPAGNIN, Jean-Marc MICHELIN, Frédéric PLAISANCE, Laurence NEGRELLO, Bruno SUAREZ et Christian SCHEPENS.

Cette application réalise l'intégration de matériel industriel de série : un robot manipulateur cinq axes à vocation pédagogique, une carte de reconnaissance de la parole, une carte de synthèse et un mini-ordinateur. Ce choix est motivé par la facilité de mise en œuvre et le parti pris de présenter aux étudiants des produits disponibles et performants.

Le logiciel comporte trois processus concourants qui assurent les trois tâches fondamentales de perception, de décision et d'action. Cette décomposition naturelle accroît la modularité des programmes et la flexibilité de l'application (choix du robot, modification du langage de commande).

Le langage de commande regroupe deux types d'ordres :

1 - Les ordres explicites encore appelés commandes discrètes [MARIANI 87] qui spécifient totalement le but et peuvent être triviales comme dans le cas de STOP ou nécessiter le calcul d'une trajectoire complexe comme dans le cas d'APPROCHER LE_CUBE Y.

2 - Les ordres incomplètement spécifiés encore appelés commandes analogiques [MARIANI 87] tels que REPLIER, MONTER A_DROITE qui indiquent seulement au robot de suivre une trajectoire.

Ce support expérimental permet d'illustrer divers aspects de la communication homme-machine :

- spécification d'un langage de commande perçu comme naturel,
- différenciation entre mots isolés et enchaînés au plan syntaxique,
- amélioration de l'intéractivité par la synthèse vocale,
- ergonomie et fiabilité,
- domaines d'application.

Par rapport à des expériences antérieures [WINOGRAD 72], nous devons prendre en compte ici les contraintes du dialogue oral, ce qui implique l'utilisation de commandes simples avec une syntaxe simple.

C'est dans ce domaine que le LIMSI a déjà travaillé en présentant à l'exposition "Communication" du CNRS en 1984, la commande vocale d'un bras articulé capable de déplacer des cubes.[MARIANI 87] [BEROULE 84]

2 - Présentation de l'application [CONDOM 87]

L'application met en œuvre quatre modules :

- La carte de reconnaissance RME186 développée par VECSYS à partir des recherches du LIMSI qui permet de reconnaitre des mots enchaînés (mots prononcés les uns à la suite des autres sans silence). Cette carte utilise exclusivement des paramètres acoustiques (approche globale), par conséquent chacun des mots intervenant dans le projet doit être appris. Un apprentissage de mots enchaînés améliore la reconnaissance. Le système est monolocuteur et peut traiter jusqu'à 200 références. [VECSYS 86]
- La carte de synthèse TELEVOX, développée par ELAN INFORMATIQUE sous licence CNET qui permet de synthétiser du texte : celui-ci est dans un premier temps traduit en diphones pré-enregistrés lesquels sont envoyés au synthétiseur qui émet les sons correspondants [ELAN INFORMATIQUE 87]. La carte est supportée par un micro-ordinateur compatible PC qui joue le rôle d'un terminal de commande et de visualisation.
- Le robot ERICC de la société BARRAS PROVENCE. Ce bras manipulateur destiné à l'enseignement, comporte cinq axes rotoïdes ainsi qu'une pince permettant de déplacer des objets dans l'espace opérationnel.
- Un VAX 750 auquel sont reliés par liaison série les trois modules précédents et qui synchronise la reconnaissance des ordres et le déplacement du robot.

3 - Commande vocale du robot

Les primitives du langage de commande appartiennent au deux types : ordres incomplètement spécifiés et ordres explicites.

Les **ordres incomplètement spécifiés** sont destinés à modifier la situation (position et orientation) du robot de manière régulière sous le contrôle de l'opérateur : la réception d'un ordre à un instant donné met fin au mouvement en cours du robot et lance un mouvement uniforme selon une nouvelle trajectoire. Le temps intervient directement dans la valeur du déplacement pour ce type de commande analogique.

Les **ordres explicites** précisent l'état final du robot et ne tiennent pas compte du temps : afin d'éviter l'énumération de valeurs numériques pour désigner les situations remarquables, on mémorise par apprentissage, un nombre fini de points et de cubes.

3.1 - Ordres incomplètement spécifiés (ou ordres analogiques)

Le repérage spatial est obtenu par référence à un repère fixe lié au bâti du robot. Les déplacements se font en itérant sur un incrément (déplacement élémentaire).

Les mouvements de base correspondent à des déplacements élémentaires selon trois directions x, y et z d'un repère orthonormé lié à la base du robot (fig.1) ou selon des portions d'hélices définies en coordonnées cylindriques μ, θ, z (fig.2).

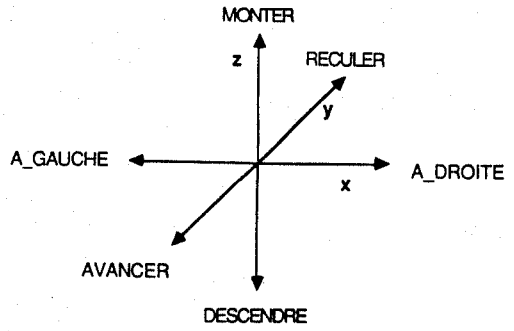


Figure 1 - déplacements linéaires

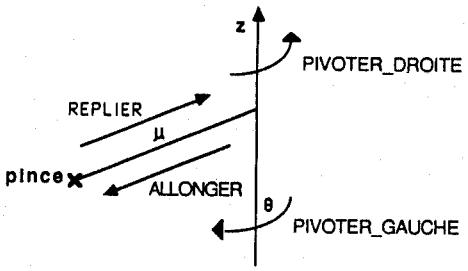


Figure 2 - déplacements circulaires

Les mouvements de base correspondent à l'énoncé de mots isolés c'est à dire à des mots délimités par un temps de silence significatif (supérieur à 200 ms) [VECSYS 86].

Si l'on enchaîne les mots on obtient des mouvements composés correspondant aux différentes combinaisons possibles:

Exemples : - Séquence de deux mots enchaînés. (fig.3)

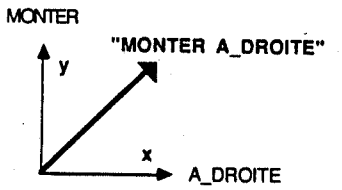


Figure 3 - la direction prise par la pince est donnée par la droite y = x

- Séquence de trois mots enchaînés. (fig.4)

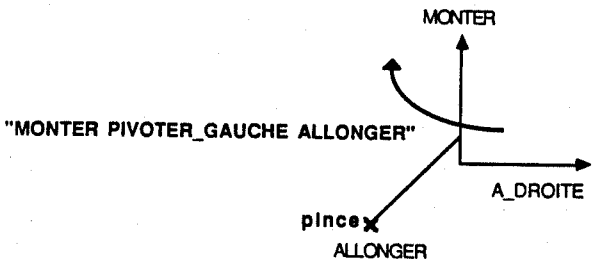


Figure 4 - la trajectoire de la pince décrit un arc d'hélice

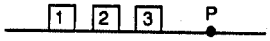
Trois vitesses de déplacement RAPIDEMENT, NORMALEMENT, LENTEMENT permettent une approche rapide puis un déplacement lent pour la saisie de l'objet. Cette vitesse est proportionnelle à la taille de l'incrément. En combinant tous les mots relatifs au mouvement et à la vitesse, on obtient 321 vecteurs de déplacement élémentaire. La pince de son côté tourne sur elle-même avec VISSER et DEVISSER. L'opérateur est averti des incidents détectés par le robot (arrivée en butée, configuration inaccessible) à l'aide de la synthèse vocale .

3.2 - Ordres complètement spécifiés (ou ordres discrets)

Il s'agit tout d'abord des ordres élémentaires, comme STOP (renforcé par ARRET) qui interrompt un mouvement en cours, le système restant toujours en attente d'une commande, ou PAUSE qui désactive la reconnaissance et permet à l'utilisateur de quitter l'application ou de modifier l'une des trois vitesses. La pince peut SAISIR, RELACHER se mettre en position verticale ou horizontale.

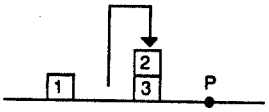
Si l'on désire maintenant donner à la machine une faculté de raisonnement lui permettant de déplacer des objets sans que l'opérateur soit contraint de guider la trajectoire, le système doit passer par une phase d'apprentissage qui consiste à mémoriser les positions de chacun des objets

(I) Apprentissage des trois cubes et du point P



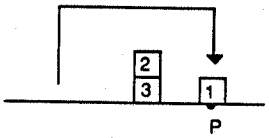
- Approche de chacun des éléments suivie de "APPRENDRE LE_CUBE (resp. LE_POINT)"
- Le système donne alors 4 noms :
  "Cube i appris" (i=1,2,3)
  "point P appris"

(II) Saisie de 2 et pose de 2 sur 3



- "APPROCHER LE_CUBE 2"
- saisie du 2 guidée par l'opérateur
- "APPROCHER LE_CUBE 3"(amène 2 au-dessus de 3)

(III) Pose de 1 en P



- "APPROCHER LE_CUBE 1"
- saisie du 1 guidée par l'opérateur
- "APPROCHER LE_POINT P" (amène 1 en P)

Figure 5 - manipulation de 3 cubes et d'un point par le robot manipulateur

intervenant dans l'univers opérationnel du bras manipulateur. L'opérateur fait saisir le cube et commande au système la mémorisation des coordonnées spatiales de la pince ("apprendre le_cube"). Le robot signale le nom du cube par le message de synthèse "cube vert appris". Les noms des objets sont prédéfinis dans le programme de l'application. Une fois l'apprentissage terminé, l'opérateur peut donner l'ordre d'aller chercher le cube ("approcher le_cube vert") : la pince ne va pas le saisir mais se placer à

proximité pour permettre la saisie du cube sous le contrôle de l'opérateur. Le message "approche terminée", avertit l'opérateur de la fin de l'opération. Chaque fois qu'un cube est relâché, ses coordonnées spatiales sont mises à jour.

La trajectoire effectuée pendant l'approche tient compte uniquement des obstacles que constituent les cubes empilés. Les ordres erronnés sont bien entendu traités par le système qui en avertit l'utilisateur au moyen de la synthèse: Par exemple "impossible cube non appris" à la suite d'un ordre d'approche sur un cube ignoré du système ou encore "impossible d'apprendre plus de n cubes" n étant le nombre maximum de noms connus du système.

Application : Considérons un univers composé de trois cubes 1, 2, 3 et d'un point P. L'utilisateur se fixe pour but de poser le 2 sur le 3 et de placer le 1 en P. La figure 5 illustre les trois principales étapes de l'application.

3.3 - Organisation de l'application

L'application s'articule autour d'une organisation multitâches mettant en œuvre trois processus s'exécutant en parallèle (fig.6) :

- une interface de reconnaissance permettant :
  - le chargement des références des mots appris dans la RME186,
  - le lancement de la reconnaissance.
  - la récupération des mots reconnus par la carte.
- un programme superviseur qui fait l'analyse syntaxique et sémantique de la phrase récupérée par le programme précédent.
- un programme de pilotage qui commande le déplacement du robot à la position demandée et contrôle son exécution.

Le premier avantage du parallélisme réside dans le découpage modulaire des tâches, bénéfique lors de l'analyse, du développement et de l'adaptation (extention du vocabulaire, adaptation à un autre robot...) mais sa justification tient aux contraintes de rapidité imposées par le mode de fonctionnement choisi; si l'on veut interrompre une action en cours "immédiatement" il est nécessaire que le système soit continuellement réceptif. De ce fait nous devons considérer les tâches de décision et de pilotage comme concourantes et non pas seulement séquentielles.

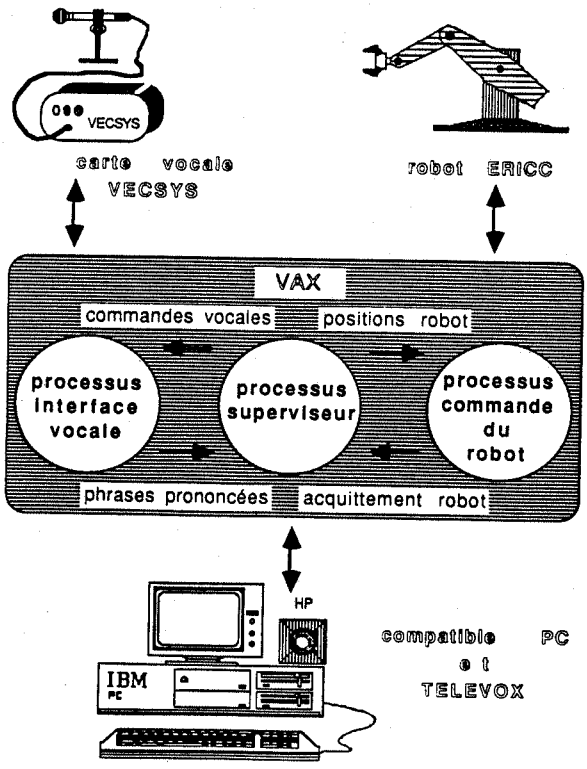


Figure 6 - synchronisation des processus

4 - Ergonomie du système

L'utilisation pratique de ce support expérimental nous a permis de faire quelques considérations de nature ergonomique en considérant principalement trois critères :

4.1 - Les taux de reconnaissance

Trois facteurs ont joué un rôle important dans l'amélioration des taux de reconnaissance :

- l'utilisation de mots, d'une part différents sur le plan phonétique pour réduire les confusions possibles, et d'autre part constitués d'au moins deux syllabes afin de conserver une zone acoustique stable minimale lors des enchaînements de mots où les liaisons modifient les extrémités des mots,
- l'apprentissage de mots enchaînés qui diminuent les erreurs dues aux liaisons.
- la prise en compte d'une syntaxe réduisant le nombre de références intervenant dans la reconnaissance à un instant donné et par conséquent le taux de confusion.

Afin d'évaluer la faisabilité du système nous avons procédé à une mesure expérimentale des taux de reconnaissance dans les conditions suivantes :

- L'apprentissage a été réalisé sur l'ensemble du vocabulaire (33 mots dont 25 pour les ordres analogiques et 8 pour les ordres discrets) dans une ambiance non bruitée.
- La reconnaissance a lieu dans les conditions normales de pilotage.

Les résultats obtenus sont résumés dans le tableau suivant (fig.7) :

	Ordres Analogiques	Ordres Discrets
mots isolés	95 %	97 %
mots enchaînés	91,5 %	94 %

Figure 7 - taux de reconnaissance

Pour les commandes analogiques on constate quelques confusions inhérentes au choix du vocabulaire (principalement "visser" et "dévisser", "replier" et "reculer") et surtout des mots non reconnus qui ont pour cause les changements de rythme et d'intonation du locuteur dont l'attention est essentiellement captée par le pilotage du robot.

Pour les commandes discrètes les performances sont meilleures. Cela tient d'une part aux images acoustiques des mots, distinctes entre elles et de l'ensemble du vocabulaire, et d'autre part à ce que l'utilisateur n'est plus préoccupé par le pilotage et peut donc émettre naturellement ses ordres.

Les performances de reconnaissance sont donc fonction de la vitesse d'élocution.

4.2 - Le temps de réponse du système

Ce paramètre a été minimisé à 3 niveaux :

Parole : - par une restriction du vocabulaire à 33 mots.
- avec l'analyse syntaxique signalée précédemment.
- en réduisant le plus possible le nombre de références (60).

Informatique par le parallélisme des processus. (fig.6)

Robotique par des mouvements très variés permettant de joindre quasi-directement deux points quelconques de l'espace opérationnel.

Si le temps de réponse de la carte de reconnaissance est en moyenne de 0.5 sec. et la durée de l'ensemble décision - action immédiate, le système présente une certaine inertie à l'exécution due à la lenteur du débit vocal qui ne permet pas, en moyenne, de prononcer 4 ou même 3 mots enchaînés en moins de 2 sec. Il s'ensuit que l'opérateur doit anticiper l'émission des ordres en vue des changements de direction.

De fait, la commande vocale est mieux adaptée aux ordres discrets n'exigeant pas une réaction immédiate, qu'aux ordres analogiques plus rapidement et facilement exécutées avec une télécommande [MARIANI 87].

4.3 - le pilotage du bras manipulateur

Le langage de commande a été choisi de manière à exprimer le plus naturellement et le plus rapidement possible le mouvement désiré.

En ce qui concerne les ordres discrets, l'expérience montre qu'il est plus convivial de saisir directement les cubes ce qui permet d'enchaîner des actions de plus haut niveau du type "prendre le cube bleu et poser sur le cube vert".

En résumé, si les ordres discrets s'intègrent mieux dans le dialogue oral homme - machine que les ordres analogiques, leur cœxistence est souhaitée pour la prise en compte d'évènements imprévus perçus uniquement par l'opérateur (par exemple un obstacle non appris).

5 - Conclusion.

Le pilotage vocal, dans l'esprit où nous l'avons réalisé, ne prétend pas surclasser le pilotage manuel du fait de la lenteur du débit de parole. Les applications de notre projet - aide aux handicapés, troisième main, télémanipulation en milieu hostile - ramènent toutes à l'idée de prothèse. Le temps de réponse d'un tel système dépassera toujours l'imprécision du 4ième top de l'horloge parlante.

Rappelons que l'un des buts poursuivis est d'ordre pédagogique. Le système actuel est, de ce point de vue, une base permettant de développer des applications avec un robot pouvant actualiser la connaissance de son environnement et raisonner par rapport à des tâches qui lui sont commandées. Il pourra par exemple trouver la solution optimale pour modifier la configuration d'un empilement de cubes, détecter un obstacle sur sa trajectoire et en prévenir l'utilisateur ou encore mémoriser un nouveau plan d'actions...

Une telle application, pour être efficace et conviviale, doit s'articuler autour d'un système de compréhension et de gestion de dialogues qui devra par exemple prendre en compte des messages ambigus, incomplets ou incohérents émis par l'opérateur [PIERREL 87].

Références :

[BEROULE 84] Béroule D.
"Communication parlée avec un système robotisé simulé"
Rapport d'activité 1983 du LIMSI, février 1984.

[CONDOM 87] Condom J.M. et Lozes A.
"Commande vocale d'un robot manipulateur dans l'espace opérationnel"
Rapport d'activité du CERFIA, 1987.

[ELAN INFORMATIQUE 87]
Manuel de présentation de la carte de synthèse TELEVOX, 1987.

[MARIANI 87] Mariani J.J.
"Commande vocale d'un bras articulé ou la construction d'une pyramide... de cubes grâce à la voix humaine"
Le Courrier du CNRS, juillet - décembre 1987.

[PIERREL 87] Pierrel J.M.
"Dialogue oral homme-machine", pp 159 - 162
Editions Hermès PARIS, 1987.

[VECSYS 86] :
Manuel de présentation de la carte de reconnaissance RME186, 1986.

[WINOGRAD 72] Winograd T.
"Understanding Natural Language".
N. Y. : Academic Press, 1972.

Dialogue
et
système à base de connaissances

LE DECODAGE ACOUSTICO-PHONETIQUE AU GIA
IDENTIFICATION ASCENDANTE DES VOYELLES

BULOT Rémy

GIA, Faculté de Luminy
70 r. Léon Lachamp, Marseille

MELONI Henri

Faculté des Sciences
33 rue Louis Pasteur , Avignon

Abstract :

In the bottom-up phase of the Acoustico-Phonetic Decoding, we first localize the acoustic and phonetic events by means of forms appearing in the time-based-evolution of certain speech parameters.
These events are grouped together by contextual rules to constitute labelled units (vocalic and consonantic cores).
The vowels are identified by means of a beam of hierarchized features. These are calculated from the characteristic zones of the vocalic cores and the transitions. The attribution of probability factor to each of the rules describing these features makes it possible to produce a lattice of valued phonetic hypotheses.

Introduction

Dans le cadre d'un système de Reconnaissance Automatique de la Parole Continue, nous envisageons le Décodage Acoustico-Phonétique comme un traitement qui permet de passer d'un signal concret et continu à une représentation en unités phonétiques discrètes ou de projeter celles-ci sur la matière sonore d'un énoncé. Dans le premier cas, il s'agit de proposer de manière ascendante des accès lexicaux à des cohortes de mots au moyen de symboles acoustico-phonétiques divers (formes, associations de formes, événements, phonèmes, syllabes, diphones, propriétés, indices, traits, etc.) [8], [7], [5], [3], [1]. L'autre aspect consiste à vérifier de manière descendante, dans un contexte connu, des hypothèses phonétiques fournies au cours du traitement par des connaissances linguistiques (phonologie, lexique, syntaxe, etc.) [2], [5].

Pour chacune de ces phases de la reconnaissance, le DAP suppose, dans un premier temps, que l'on effectue une mise en correspondance temporelle des unités discrètes avec les zones de signal correspondantes ([8], [2], [4]) puis, dans une seconde étape, que les événements acoustico-phonétiques soient caractérisés au moyen d'attributs coïncidants avec la description formelle de ces unités. Nous présentons ici la méthodologie utilisée pour réaliser l'identification ascendante des voyelles après leur localisation [1].

Les segments représentant des noyaux vocaliques issues du processus de localisation sont caractérisés plus finement au moyen de traits pseudo-phonétiques valués. Ces connaissances sont définies au moyen de règles représentées dans un formalisme proche du langage Prolog [9]. La décision brutale mais nécessaire d'affecter positivement ou négativement un trait à une macro-classe vocalique est valuée en fonction de la "qualité" des règles employées pour déterminer cet attribut ; cette mesure permet de déduire un taux de vraisemblance pour chaque hypothèse phonémique proposée.

1 - Localisation ascendante des voyelles

Les noyaux vocaliques sont localisés au moyen de formes apparaissant sur l'évolution temporelle de certains paramètres [1] [2]. Toutefois, les unités ainsi repérées peuvent correspondre à une voyelle, une portion de voyelle (essentiellement pour les nasales) ou une séquence de phonèmes comportant une voyelle et certaines consonnes vocaliques adjacentes (liquides et semi-consonnes). Nous définissons, à l'intérieur du noyau vocalique, des zones caractéristiques sur lesquelles les attributs de la voyelle sont calculés. La détermination de ces intervalles est liée au mode de production supposé du phonème (oral ou nasal), elle s'effectue donc après l'évaluation grossière du trait correspondant (les 2 hypothèses sont fréquemment envisagées conjointement).

Pour les noyaux vocaliques étiquetés *nasal* avec un taux de vraisemblance suffisant, la portion significative du phonème est située avant la zone finale (partie nasale) et sur un intervalle déterminé au moyen des phénomènes suivants (figure 1) :

- colline importante pour la fréquence du premier pic spectral (ouverture),
- maximum d'énergie (intervalle de 50 ms centré sur cette trame)

```
zone-voyelle-nasale(z0, z-voy) ->
        ou(forme-sur(z0, colline1-fp1(z, t)),
           position-limite-sup(z0,Er0,t))
        etendre(<t, t>, 2, 2, z')
        intersection(z', z0, z-voy) / ;
```

Pour les noyaux vocaliques étiquetés *oral* le segment sur lequel les traits sont évalués est défini par les phénomènes suivants:

- zone stable sur le second formant (figure 2),
- positions des maxima des énergies basse, moyenne et haute,
- maximum d'amplitude des 2 premiers formants,
- maximum fréquentiel du premier formant.

La zone retenue correspond à la coïncidence contextuelle maximale de ces phénomènes.

zone-voyelle-orale(z0, z-voy) ->
milieu(z0, o)
forme-sur(z-f2, Formant2(z0, o), plat-ff2(z))
inferieur-eg(5, longueur(z))
milieu(z, m)
étendre(<m, m>, 2, 2, z-voy) / ;
zone-voyelle-orale(z0, z-voy) ->
centre-des-maxima-d-energie(z0, m)
etendre(<m, m>, 2, 2, z)
intersection(z0, z, z-voy) ;

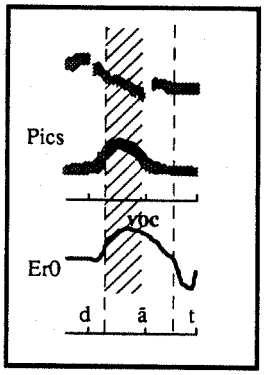


Figure 1
zone caractéristique pour une voyelle nasale

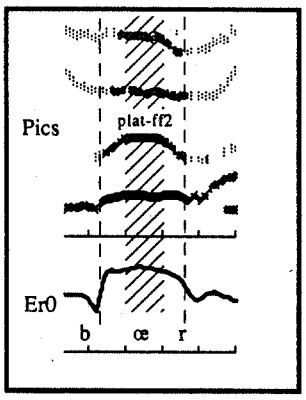


Figure 2
zone caractéristique pour une voyelle orale

2 - Identification des traits pseudo-phonétiques

Les phonèmes sont projetés sur un faisceau de traits non redondants de nature binaire ou ternaire (figures 3 et 4). Cette représentation dérivée des systèmes couramment utilisés [12] permet de proposer des accès lexicaux robustes plus fins dans les cas où la valuation globale des phonèmes candidats serait moins discriminante. La décision prise à un niveau conditionne directement les calculs des étapes suivantes ; toutefois, plusieurs hypothèses peuvent être simultanément considérées en fonction de la valuation accordée aux traits dominants. En ce qui concerne les voyelles, la séparation s'effectue dans un premier temps sur l'opposition des traits *nasal* et *oral* ; les traits de niveaux inférieurs sont ensuite évalués sur les zones caractéristiques correspondantes.

Un point de choix dans l'arbre des traits est nommé par un identificateur unique donnant accès à un ensemble de règles qui décrivent des critères d'appartenance à deux ou trois classes

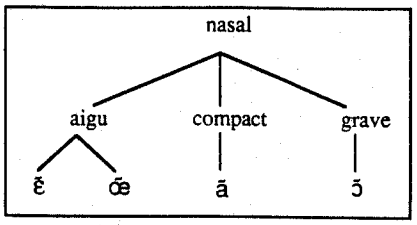


Figure 3
Traits discriminants pour les voyelles nasales

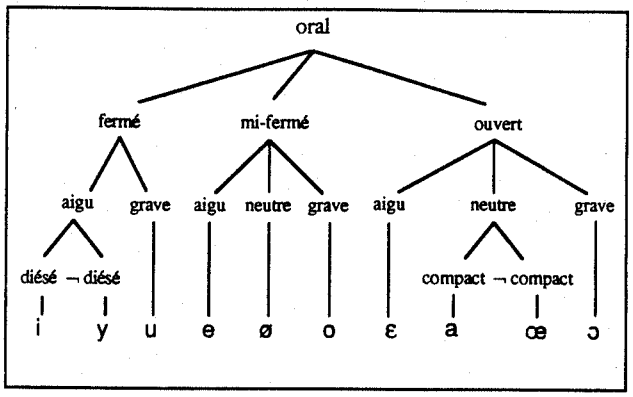


Figure 4
Traits discriminants pour les voyelles orales

phonétiques et qui définissent ainsi des traits distinctifs de nature binaire ou ternaire.Les choix de type ternaire sont utilisés lorsqu'une classe semble s'insérer entre deux autres sur l'axe de représentation. Chaque trait est défini par une disjonction de règles visant à décrire les phénomènes à travers des paramètres différents et dans des contextes distincts.

ouverture-voy-orale(z0, z1, <fermé, 1, 100>) ->
inférieur(cible-ff1(z0), 300) ;
ouverture-voy-orale(z0, z1, <ouvert, 3, 100>) ->
inférieur(400, cible-ff1(z0))
inférieur(680, cgb(z1)) ;
ouverture-voy-orale(z0, z1, <fermé, 2, 98>) ->
inférieur(cible-ff1(z0), 400)
peu-d-energie-sur-le-noyau(z0) ;

L'application d'une règle du type ci-dessus provoque l'insertion dans le treillis de résultats d'un triplet constitué d'une étiquette représentant le trait, du numéro de la règle qui a attribué ce trait ainsi que d'une valeur comprise entre 0 et 100 caractérisant la vraisemblance de la déduction.

3 - Valuation des traits

Les systèmes de traits utilisés couramment par les phonéticiens sont généralement redondants. Or, il est difficile, en démonstration automatique, d'évaluer des combinaisons de résultats partiels, surtout lorsque ceux-ci sont ambigus et parfois même contradictoires. C'est la raison pour laquelle nous avons préféré employer un système d'attributs minimal pour la discrimination des phonèmes.Toutefois, les règles qui décrivent un même trait ont été définies indépendamment les unes des autres et

sont partiellement redondantes. Aussi, ont-elles été évaluées séparément sur la base de sons afin de pouvoir leur attribuer un score respectif. Celui-ci est calculé sous la forme d'un pourcentage par la formule :

$$\frac{\text{nbre de fois où la règle a identifié le bon trait}}{\text{nbre de fois où la règle s'est appliquée}} \times 100$$

Cette valeur n'est absolument pas représentative d'un taux de reconnaissance mais mesure le degré confiance que l'on peut attribuer à chaque règle de façon indépendante.

La difficulté pour évaluer la vraisemblance d'une décision sur un point de choix réside dans le fait que plusieurs règles peuvent être démontrées simultanément avec des scores différents, et même avec des résultats différents (il n'est pas toujours possible de vérifier la cohérence d'un ensemble de règles lorsque celles-ci mettent en jeu des paramètres distincts). De plus, il n'est pas évident que l'application de plusieurs règles attribuant le même trait conforte la valuation de celui-ci (contextes différents, règles plus générales mais moins fiables), et si oui, dans quelle proportion ?

Pour résoudre ce problème, nous avons décidé d'effectuer l'approximation suivante : nous prenons en considération uniquement le résultat affecté du score le plus élevé. Compte tenu de ce principe, les règles attribuant un trait sur un point de choix sont rangées dans la base de connaissances suivant leur capacité décroissante de discrimination ; nous conservons le résultat et le score de la première règle démontrée.

Dans le cas de traits binaires, l'identification d'un trait avec un score v signifie aussi que le trait opposé est affecté du score complémentaire, c'est-à-dire $(100-v)$.

Dans le cas de traits ternaires t1/t2/t3, les règles qui décrivent les deux classes les plus éloignées (t1 et t3) sont suffisamment restrictives pour qu'un phonème mal classifié par celles-ci appartienne nécessairement à la classe centrale t2. De là, on déduit que : Si t1 (t3) est démontré avec une vraisemblance v, alors t2 et t3 (t1) sont affectés respectivement des vraisemblances $(100-v)$ et 0. Par contre, la démonstration de t2 avec un score v ne permet pas de décider entre t1 et t3 et ceux-ci sont affectés d'un score égal à $(100-v)/2$.

Ce système de valuation vérifie les propriétés suivantes :

- la somme des vraisemblances attribuées aux traits associés à un point de choix est égale à 100 %,
- si un chemin dans l'arbre des traits est valué par le produit des vraisemblances de ses composants, alors la somme des scores des chemins ayant le même trait d'origine est égale à 100 %.

Dans le cas où l'on souhaite attribuer un score à des hypothèses phonémiques sur un segment de parole, sa valeur est donnée par le produit des valuations des traits qui les caractérisent.

4 - Identification des traits nasal/oral

Afin d'illustrer les techniques que nous employons, nous décrivons les principales règles utilisées pour l'identification des traits *nasal* et *oral*. Les connaissances codées sont une transposition des informations robustes proposées par différents auteurs ([11], [6]) et qui ont été validées dans notre système de représentation.

4.1 Trait *nasal*

L'indice le plus fiable pour la nasalité est la détection du formant nasal grave sur la fin du noyau vocalique. On observe alors la présence de deux formants dans les basses fréquences :

- un formant à moins de 350 Hz (formant nasal),
- un formant entre 400 et 900 Hz (formant oral).

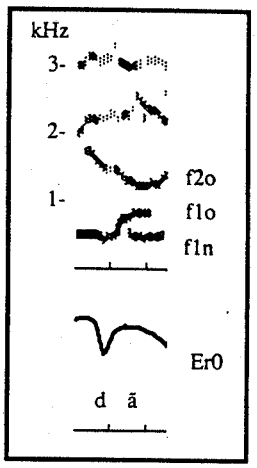


Figure 2
Présence de 3 formants en dessous de 1900 Hz pour une nasale

En plus des voyelles nasales, seule la voyelle /u/ peut vérifier cette propriété spectrale et l'existance de pics émergeant suffisamment dans le haut du spectre suffit à lever l'ambiguïté. Ces pics peuvent-être généralement associés au deuxième formant oral ou au deuxième formant nasal. Une des règles qui décrit ce phénomène est donnée ci-dessous :

```
nasalite-voy(z0, <nasal, 8, 92>) ->
    zone-continue(z0, Fp1, <100, 350>, 50, <i, j>)
    inferieur-eg(4, longueur(<i, j>))
    appartient(moyenne(z-f1, Formant(<i, j>, i), <400, 900>)
    inferieur(ap1(z0), plus(20, afsup(z0, <900, 5000>)) ;
```

Cette règle ne concerne que 30% des voyelles nasales présentes dans la base de sons (le modèle LPC avec 14 coefficients ne sépare pas toujours très bien les deux formants graves nasal et oral) et son taux de réussite est de 92%. Les quelques erreurs d'identification apparaissent sur des voyelles ouvertes dans des contextes nasals ; on observe parfois le prolongement d'un formant nasal (résiduel ou anticipé) dans le noyau vocalique oral.

Dans le cas où l'on ne détecte pas le formant nasal grave, un critère de nasalité sera évalué en fonction de l'affaissement du premier pic sur la fin de la voyelle. On effectue alors une paramétrisation du signal avec 21 coefficients de LPC et la présence de deux pics dans les basses fréquences augmente la vraisemblance du trait nasal. Dans l'exemple ci-dessous, le terme *<21, "sans", +1.>* spécifie que les paramètres de parole concernés sont calculées avec 21 coefficients, pas de préemphase et un rayon égal à 1 ;

lorsque ce terme n'est pas précisé, les conditions d'analyse du signal sont implicites et correspondent à *<14, "sans", +1.>*.

nasalite-voy(z0, <nasal, 7, 77>) ->
inferieur(400, limite-sup(z1, Formant1(z0, m)))
inferieur(ap1(z0), plus(20, afsup(z0, <900, 5000>)))
ecrasement-f1(z0, k)
inferieur(fp1(k, <21, "sans", +1.>), 350)
inferieur(fp2(k, <21, "sans", +1.>), 900)) ;

L'étude de l'émergence et de la largeur du premier pic permet de mettre en évidence sa faiblesse dans le spectre. Les critères retenus sont :

- une faible émergence du premier pic (< 10 db),
- une chute de 10 db sur l'émergence du premier pic avant la fin du noyau,
- une largeur importante du premier pic (> 250 Hz),
- une augmentation importante de cette largeur (plus de 100 Hz) entre le début et la fin du noyau.

La plupart des erreurs dans l'application de ces règles concernent la voyelle /a/ ainsi que les voyelles mi-ouvertes lorsqu'elles sont suivies d'un /r/ appartenant à la syllabe et faisant partie du même noyau vocalique.

4 . 2 Trait *oral*

les règles qui attribuent ce trait doivent être considérées plutôt comme la description d'événements peu probables sur des voyelles nasales que comme la définition de l'oralité d'une voyelle. Par exemple :

- la bonne tenue du premier formant :
nasalite-voy(z, <oral, 4, 97>) -> non(ecrasement-f1(z, k)) ;

- la localisation du deuxième formant au delà de 1900 Hz :
nasalite-voy(z, <oral, 9, 100>) ->
inferieur(cgb(z), 700)
inferieur(1900, moyenne(z-f2, Formant2(z, m))) ;

- l'existance d'une forte résonnance ailleurs que dans les basses fréquences :
nasalite-voy(z, <oral, 10, 100>) ->
inferieur(1000, fem-lissee(z)) ;

- la détection d'un noyau non ouvert :
nasalite-voy(z, <oral, 2, 96>) ->
restreindre(z, 2, 2, z')
inferieur(cgb(z'), 600) ;

sont autant de propriétés spectrales propres aux voyelles orales.

Les deux tableaux ci-dessous résument les performances de chaque règles évaluées séparément. La *vraisemblance* a été calculée d'après la formule donnée au § 3 ; la colonne *Reconnus* donne le taux de voyelles reconnues par chaque règles dans la classe concernée . On constate que les règles utilisées sont largement redondantes mais cela ne gêne en rien l'identification puisque le système retient la plus fiable parmi celles qui s'appliquent.

Phonèmes : voyelles nasales

n°	Vraisemblance	Reconnus
5	94 %	21 %
6	94 %	31 %
7	77 %	62 %
8	92 %	30 %
10 % non reconnus		

Phonèmes : voyelles orales

n°	Vraisemblance	Reconnus
2	96 %	42 %
3	99 %	51 %
4	97 %	39 %
9	100 %	18 %
10	100 %	15 %
11	99 %	25 %
13	92 %	74 %
11 % non reconnus		

5 - Résultats sur la reconnaissance des voyelles

L'identification remontante de certaines unités phonétiques (voyelles, consonnes constrictives, consonnes occlusives sourdes) à partir d'un ensemble de traits discriminants a donné des résultats satisfaisants pour un locuteur particulier . De plus, les règles concernant la détection des macro-classes ainsi que l'identification de certains traits (sourd/sonore pour les fricatives et les occlusives, oral/nasal pour les voyelles) se sont révélées aussi efficaces pour d'autres locuteurs (masculin). La généralisation du processus à d'autres locuteurs est relativement simple dans la mesure où les phénomènes caractérisés restent qualitativement semblables malgré des variations quantitatives au niveau de certains paramètres. Les règles qui décrivent ces traits intègrent des informations contextuelles sur la nature acoustique des sons environnants et modélisent partiellement les effets "visibles" de la coarticulation. L'attribution d'un coefficient de vraisemblance pour chacune d'elles permet de produire un treillis valué d'hypothèses phonétiques ; une matrice de confusion pour les voyelles donne une idée des caractéristiques de notre module de DAP dans une analyse remontante.

	i	y	u	e	ø	o	ɛ	œ	a	ɔ	ɛ̃	ã	ɔ̃	dans 3 premiers
i	81	3		16										**98**
y	6	65		14	10						5			**80**
u			88		3	3					6			**92**
e	1	4		83			10				2			**96**
ø				30	50	7	3		3		7			**82**
o			29		3	56			3				9	**85**
ɛ				38			43				19			**65**
œ				11			18	12	4		44	11		**55**
a					1	1	18	1	44	7	22	5	1	**90**
ɔ						7		3		51		25	14	**93**
ɛ̃				3	1		11	1	5		76	3		**92**
ã						2	7	2	10	8	7	54	10	**83**
ɔ̃			3			3				10		5	79	**92**

base de sons: un locuteur, vitesse d'élocution "normale", salle non-insonorisée, 3 mn de parole continue et 700 voyelles.

Dans la table ci-dessus, une ligne est composée du phonème analysé suivi des taux de répartition pour les différents candidats proposés en premier choix. On notera les deux résultats médiocres obtenus sur les voyelles /œ/ et /ɛ/ liés à la fois au locuteur et au corpus d'évaluation. En effet, notre base de sons contient peu d'exemples du phonème /œ/ ; il en résulte donc un déséquilibre dans la valuation des règles qui concernent ce dernier puisque sa probabilité d'apparition est déjà très faible. De plus, cette voyelle apparaît fréquemment suivi de /r/ qui compacifie les spectres et rend la séparation difficile avec /a/ ou les voyelles nasales. De plus, l'étiquetage des sons a été fait en fonction de la prononciation "standard" et non pas en fonction de la réalisation effective des phonèmes. Ainsi, l'énonciation de /ɛ/ par notre locuteur est souvent plus proche de /e/ et ceci explique la forte confusion entre ces deux phonèmes.

Bien que cette table permet d'avoir une idée des performances, celles-ci ne pourront être évaluées de façon plus précises qu'avec un système complet de RAPC. D'une part, ces statistiques ne tiennent pas compte de l'écart entre les valuations des diverses solutions et, d'autre part, des stratégies de reconnaissance de type "plus forte densité" peuvent choisir comme premier candidat une unité qui n'est pas nécessairement la plus vraisemblable.

6 - Conclusion

La phase ascendante du DAP dans un système de RAPC ne permet pas d'identifier avec une égale vraisemblance des traits acoustico-phonétiques discriminants les phonèmes. Certains phénomènes mieux marqués et relativement indépendants du contexte fournissent cependant des indications précieuses pour accéder à des classes de mots limitées et justifient l'utilisation d'un système de traits robustes. L'identification phonémique est renvoyée à une phase descendante de vérification dans laquelle seraient prises en compte les déformations contextuelles pertinentes pour un ensemble de mots donnés.

Les techniques employées pour la représentation et le traitement des connaissances acoustiques et phonétiques permettent d'enrichir la base de règles et d'en améliorer constamment les performances (la qualité des résultats ne peut que croître en raison de la stratégie employée). L'association de coefficients de vraisemblance à chaque règle permet de proposer plusieurs solutions pour l'identification des phénomènes et d'assouplir ainsi le mécanisme abrupt d'une classification.

Bibliographie

[1] BULOT R., *Techniques d'Intelligence Artificielle pour la Reconnaissance de la Parole : Application au Décodage Acoustico-Phonétique ;* Thèse de l'Université d'Aix-Marseille II, sept. 87

[2] BULOT R., MELONI H., *Reconnaissance des formes et segmentation ;* 16èmes J.E.P., Hammamet 1987

[3] CAELEN J., Cervantes O., Fernandez Y., *Mecanisme de consultation dans la base de données et de connaissances parole (BDC Parole)* ; 16èmes JEP, Hammamet septembre 1987

[4] GIBELLI M., SOUSSI N., *Segmentation automatique d'un énoncé connu*; Mémoire de DEA, GIA Fac. de Luminy, sept. 86

[5] GISPERT J., *Les accés : interface entre le DAP et le lexique* ; 16èmes JEP, Hammamet septembre 1987

[6] LONCHAMP F., *Analyse acoustique des voyelles nasales françaises* ; Verbum tome 2, Fascicule 1, 1979

[7] MELONI H., *Etude et réalisation d'un système de reconnaissance automatique de la parole* ; Thèse d'état, Faculté des Sciences de Luminy, Marseille, fevrier 1982

[8] MELONI H., GISPERT J., GUIZOL J., *Traitement de connaissances pour l'identification analytique de mots dans le discours continu* ; Congrès AFCET Informatique 5ème Génération, Paris 5-7 mars 1985

[9] MELONI H., BULOT R., *A knowledge based system for acoustic and phonetic decoding of continuous speech.*; Congrès International d'Intelligence Artificielle de Marseille, décembre 1986

[10] MELONI H., BULOT R., *Paramétrisation du signal et reconnaissance des formes pour le décodage acoustico-phonétique en Prolog* ; Congrès AFCET R.F.I.A., Antibes 87

[11] MRAYATI M., *Etude des voyelles nasales françaises* ; Bulletin de l'Institut de Phonétique de Grenoble, n°4, pp 1-26

[12] ROSSI M., *Les traits acoustiques* ; La Linguistique, 13, pp 63-82, 1977

SEANCE D'EXPERTISE POUR L'ACQUISITION DE CONNAISSANCES ACOUSTICO-PHONETIQUES DANS LE SIDOC-Parole

Y.FERNANDEZ (*,**,@), O.CERVANTES (*,**), J. CAELEN (*), J.F. SERIGNAT (*)

(*) Laboratoire de la Communication Parlée
I.C.P., Unité Associée au C.N.R.S., No. 368
INPG/ENSERG. 46, Av. Félix Viallet
38031 Grenoble Cédex.

(**) Laboratoire de Génie Informatique
I.M.A.G. - Université de Grenoble
B.P. 68 38402 St. Martin d'Hères

ABSTRACT

The "Système d'Intégration de DOnnées et de Connaissances Parole" (SIDOC-Parole) is a knowledge based system providing a flexible environment for speech research and knowledge acquisition. It is based on an object-oriented **knowledge model** which provides features for defining and manipulating speech related information. The model allows the definition of a knowledge-base schema where **Speech-Objects** are related through **Semantic-Links**. The schema is complemented by associated **Rules**, which are special kinds of objects. SIDOC-Parole runs on several modes. In **knowledge acquisition mode** and through an **experimentation session**, an expert can explore and validate knowledge over large volumes of speech data. The **Observation** is another type of object which serves as the vehicle in the evolution from data to knowledge. In this article we present the basic representation notions of SIDOC-Parole, as well as its role and that of the expert during the different phases of an experimentation session.

1. INTRODUCTION.

Le Système d'Intégration de DOnnées et de Connaissances Parole (SIDOC-Parole) est un système d'aide à la recherche et à l'acquisition de connaissances du domaine parole [1]. Son but est de permettre aux utilisateurs de mener des séances interactives d'expertise sur des objets-parole. Le SIDOC-Parole offre des fonctionnalités variées pour la représentation et la manipulation de connaissances, obtenues et validées par des analyses sur une large base de données parole. Ce système est composé de quatre modules principaux : un Gestionnaire d'Objets (GO), un Gestionnaire de Connaissances (GC), un Gestionnaire de Traitements (GT) et une Interface Utilisateur (IU). Le GT exécute des calculs numériques (e.g. traitement du signal, application de procédures statistiques, etc), tandis que le GC obtient des réponses par des inférences. Le GO fournit les moyens pour l'accès aux objets concernés par ces deux grands types de processeurs. Les détails de l'architecture du système ont été présentés par ailleurs [1].

Le SIDOC-Parole est construit sur un **modèle de connaissances** permettant la représentation et l'organisation conceptuelle des entités du domaine. Il s'agit d'un modèle hybride intégrant les notions d'**OBJET** et de **REGLE**. La définition des entités, de leurs propriétés et des relations existant entre les entités constituent le schéma. Le système offre un ensemble d'opérateurs pour la manipulation de ces informations permettant le stockage, la consultation, la modification et l'inférence des informations. Ce dernier type d'opération nécessite la mise en oeuvre de mécanismes de raisonnement pour l'obtention de nouvelles informations.

@ Avec le support financier de CONACYT, México.

Le modèle permet ainsi de définir et de manipuler les **Objets-Parole** sous une forme déclarative comme des objets complexes multimedia, et les **Règles-du-Domaine** qui encapsulent des connaissances procédurales du domaine parole [2]. Un ensemble d'informations organisées par ce modèle constitue une Base de Connaissances (BC).

Dans la section 2, nous présentons les notions de base, nécessaires à la représentation et à la manipulation de connaissances : les objets-parole, les observations, les règles. Nous énonçons, dans la section 3, les étapes à suivre, du point de vue de l'expert, au cours d'une séance d'expérimentation. Dans la section 4, nous présentons en détail un exemple de séance permettant la validation et l'acquisition de connaissances. Ensuite nous présentons d'une manière succincte un deuxième type d'exemple. Enfin, dans la section 5, nous présentons un bilan de l'état d'avancement de nos travaux ainsi que les perspectives pour nos développements futurs.

2. ELEMENTS POUR LA REPRESENTATION DE CONNAISSANCES-PAROLE.

L'univers du SIDOC-Parole est composé d'un ensemble d'objets complexes auxquels on associe un ensemble de propriétés. Les objets peuvent être GENERIQUES ou INSTANCES de ces objets génériques. Les objets qui partagent les mêmes propriétés et qui ont le même comportement sont groupés dans des CLASSES. Les relations existant entre les objets sont exprimées à travers les LIENS SEMANTIQUES. Nous présentons ici brièvement, la description des notions de base utilisées pour la représentation des connaissances dans le système.

2.1 LES OBJETS.

Tout objet dans le SIDOC-Parole peut être défini par 4 sections :
. une IDENTIFICATION : lui donnant une existence unique dans la base
. des ATTRIBUTS : permettant la description de ses caractéristiques structurelles,
. des LIENS SEMANTIQUES : exprimant ses relations avec les autres objets de la base,
. des CONTRAINTES D'INTEGRITE : spécifiant des restrictions à vérifier sur plusieurs attributs de l'objet ou portant sur lui en tant qu'unité indépendante.

Nous remarquons l'importance des liens sémantiques pour organiser les informations. Ils définissent de façon implicite un réseau d'associations [3] spéciales. Les liens sémantiques sont variés [4]. Parmi les plus importants :
- liens de **COMPOSITION** : exprimant la notion d'ordonnancement pour la création des objets nouveaux (ou dérivés),
- liens d'**EQUIVALENCE** : exprimant une équivalence "conceptuelle" entre deux objets-parole,
- liens de **CLASSIFICATION** : permettant la structuration d'un treillis **sorte-de**, **est-un** d'objets et l'héritage de propriétés entre eux.

2.2 LES OBSERVATIONS.

Une OBSERVATION est définie comme un type spécial d'objet. Pour aider aux expérimentations menant à la mise en forme des OBSERVATIONS, le GT fournit plusieurs traitements (signal, statistiques, analyse de données). L'ensemble de traitements disponible dans le SIDOC-Parole a été inspiré du système ARCANE [5] en ce qui concerne l'analyse de données. Le SIDOC-Parole permet ainsi à l'utilisateur de spécifier les objets-parole sur lesquels il souhaite vérifier des propriétés à priori (par exemple, valider des prédicats qui les concernent). Il peut ensuite ajouter à postériori des modulateurs de vraisemblance aux éléments observés.

L'OBSERVATION est un objet qui concrétise le passage des données vers les connaissances après une phase de réduction et une phase d'analyse suivies de l'interprétation de son contenu. Dans chaque phase, les opérations restent sous le contrôle de l'expert. La définition et le remplissage de ces observations sont effectués avec les spécifications obtenues lors d'un dialogue dirigé entre l'expert et le système. En général, une OBSERVATION est définie par un "**item**" (à partir duquel sont précisés les individus à étudier) et par la liste des propriétés à analyser sur chacun de ces individus. Deux exemples sont illustrés dans la section 4.

2.3 LES REGLES.

Il s'agit d'un autre type spécial d'objet qui complète la représentation des connaissances. Toute REGLE possède les composants classiques d'une règle de production [6] : **SI** {conditions} **ALORS** {conclusion}. Chaque condition est écrite comme une Expression Logique Bien Formée (ELBF) où interviennent des **prédicats**. Lorsque toutes les conditions sont satisfaites la règle est applicable, ce qui entraine le déclenchement d'un ensemble d'actions variées constituant la **conclusion**. Les actions entrainent souvent la mise à jour d'objets ou l'activation d'autres règles.

Afin d'organiser les règles, le système propose deux critères : les CLASSES et les PAQUETS de REGLES. (La notion de CLASSE est la même que pour les autres objets du système). Les deux types d'organisation permettent l'accès convenable aux règles par groupes en plus de l'accès individuel:

- Les **CLASSES** organisent les règles en accord avec différentes caractéristiques: leur **degré de généralité** [7], leur **type** (règles pondérées, règles déterministes), ou leur **forme d'intervention** dans des raisonnements. Notamment il est possible de distinguer les **REGLES-DU- DOMAINE** (règles-d) et les **META-REGLES** (règles-m). Les règles-d expriment des connaissances parole provenant de l'expertise tandis que les règles-m indiquent COMMENT UTILISER les autres connaissances de la base. Il existe différents types de règles-d, notamment les **règles-d pondérées** qui représentent des connaissances pragmatiques de l'expert. Une règle pondérée associe un coefficient de confiance [8] à ses conclusions.

- Les **PAQUETS** de règles sont constitués selon différents objectifs d'utilisation. Ces objectifs peuvent être fixés soit par l'utilisateur, soit par le système lui-même. Les PAQUETS facilitent l'établissement de STRATEGIES pour l'exploitation ultérieure des règles par un moteur d'inférences. Par exemple, pour l'apprentissage automatique [9] [10], pour expliquer le contenu de la BC [11], ou pour une application précise comme le décodage acoustico-phonétique [12].

En tant qu'objet de SIDOC-Parole [4], toute règle doit être complétée avec :
a) un identificateur, b) la liste des objets concernés par la règle, c) un coefficient de confiance associé à la conclusion, d) l'emplacement de la règle dans la base générale de connaissances. (Voir figure 1). Le coefficient de confiance est choisi par l'expert, qui au cours d'une séance d'expérimentation a analysé, pondéré et sélectionné les prédicats intervenant comme conditions dans une règle. Ce coefficient peut résulter d'une combinaison algébrique des coefficients associés aux prédicats (fréquences, moyennes, etc.) selon une algèbre dépendante de l'objectif d'un paquet particulier de règles (par exemple, pour des systèmes experts en reconnaissance de la parole).

```
IDENTIFICATION
    %nom: <nom-règle>
EMPLACEMENT
    %est-un: <type-de-règle>
    %dans:   <nom-paquet>
OBJETS_CONCERNES
    %liste-de: OBJET_PAROLE
CONDITIONS
    %liste-de: PREDICAT
CONCLUSION
    %liste-de: PAIRE_CONCLUSION
    [ (<action-1>, <coefficient-1>)
      (<action-2>, <coefficient-2>)
                .....
      (<action-3>, <coefficient-3>) ]
```

Figure 1.
Structure schématisée d'une Règle du SIDOC_Parole

3. ETAPES D'UNE SEANCE D'EXPERIMENTATION.

L'acquisition de connaissances est un processus long et complexe. Le SIDOC-Parole offre à l'expert les moyens pour analyser une large base de données parole et en déduire des connaissances qui peuvent être intégrées dans des systèmes intelligents. L'expert utilise le système en mode d'acquisition de connaissances au cours d'une **séance d'expérimentation.** Une des possibilités actuelles du système est l'expérimentation des hypothèses de l'expert visant à la formulation de règles du domaine. Très souvent l'expert a l'intuition de la forme de règles qu'il veut expérimenter mais il n'a pas une connaissance précise sur les prédicats de cette règle et notamment de la valeur exacte des constantes attachées à ces prédicats. Pour cela, l'expert recherche les éléments qui lui permettent de formaliser puis de valider une règle. Une fois validée, la REGLE représente une encapsulation des connaissances parole constatée sur un ensemble important de données-parole.

Pour la formulation d'une règle-d, la séance d'expérimentation se déroule en plusieurs étapes dont l'enchaînement est guidé par un dialogue entre l'expert et le système. Ces étapes sont :

1. CREATION D'UN ESPACE D'OBSERVATION :
   - 1.a consultation du schéma conceptuel défini,
   - 1.b sélection des objets concernés par l'expérimentation.

2. CREATION D'OBSERVATIONS :
   - 2.a spécification du contenu de l'observation: les propriétés à étudier et les individus choisis,
   - 2.b manipulation du contenu (par application de traitements d'analyse de données).

3. VALIDATION DE PREDICATS :
   - 3.a validation de prédicats (sur le contenu de l'observation) entrant dans la composition des conditions de la règle,
   - 3.c pondération des prédicats validés,
   - 3.d choix de prédicats à intervenir dans les conditions.

4. INTEGRATION DE LA REGLE DANS LA BASE :
   4.a formalisation de la règle (en utilisant des conditions tirées de l'analyse de résultats de l'étape précédente),
   4.b insertion dans la BC suivant les spécifications de l'expert.

Chacune de ces étapes entraine le déclenchement de plusieurs actions internes dans le système, soit pour préciser les entités concernées, soit pour exécuter des opérations sur ces entités. Lorsque le système n'est plus capable d'obtenir les informations nécessaires par ses mécanismes internes, il sollicite de l'expert son intervention à travers un dialogue (de type graphique) qui permet de préciser la suite des actions à dérouler. Le système présente à l'écran les fonctionnalités disponibles à chaque étape et l'expert fait un choix selon ses besoins.

4. LA SEANCE D'EXPERIMENTATION.

EXEMPLE A:

Supposons que l'expert a constaté dans certains cas la validité d'un ensemble de propositions caractérisant la tenue voisée d'une consonne. Il souhaite vérifier si la règle qu'il définit de manière intuitive est généralisable sur d'autres données. Nous remarquons ici que la forme d'une règle, vue par l'expert, n'est pas la même que celle de la REGLE en tant qu'objet spécial dans la base du SIDOC-Parole.

Soit par exemple la règle (1) de l'expert exprimée sous la forme classique
SI [conditions] ALORS [conclusion] :

```
    SI [      (ETIQUETTE_COURANTE = |C|)
      ET    (ETIQUETTE_PRECEDENTE = |C + E|
            OU ETIQUETTE_PRECEDENTE = |V + Q|
            OU ETIQUETTE_PRECEDENTE = |C + T + NV|)
(1)   ET     (BRUIT + SIGNAL/BRUIT >=INTENSITE
             >= BRUIT + 2(SIGNAL/BRUIT)/4 )
      ET    (NOMBRE_DE (FORMANTS) >= 1)
      ET    (DUREE (PHONE) > 0.75 * DUREE_CONSONNE)
      ET    (Fo > 0   OU  INDICE_AIGU/GRAVE >= 0) ]
    ALORS    [ PHASE_CONSONNE = TENUE_VOISEE ]
```

Où la signification des étiquettes fréquentielles fines qui apparaissent est: |C+E|, établissement de la consonne; |V+Q|, coda de la voyelle; |C+T+NV|, tenue de la consonne, non voisée. DUREE_CONSONNE est la durée moyenne d'une consonne, BRUIT est l'intensité moyenne du bruit ambiant, et le rapport SIGNAL/BRUIT est celui du corpus en cours d'analyse.

La partie-conditions de la règle peut contenir des propositions logiques simples -comme dans le système expert MYCIN [8] -ou des prédicats avec variables ou encore des relations quelconques.

Chacune des conditions concerne une ou plusieurs propriétés des objets-parole. La liste des prédicats intervenant dans ces conditions pour la règle de l'expert est :

P1 : INTENSITE (x) <= C1 ET INTENSITE (x) >= C2
P2 : Nombre (FORMANTS (x)) >= C3
P3 : Durée (x) > C4 * DUREE_CONSONNE
P4 : (Fo (x) > C5 OU INDICE_AIGU_GRAVE >= C6)

Il faut maintenant déterminer l'item x (portion du signal) sur lequel on va valider ces prédicats et optimiser les constantes Ci, i=1 à 6. Les étapes à suivre pour cette validation sont décrites ci-dessous

4.1 CREATION D'UN ESPACE D'OBSERVATION.

Le premier pas consiste à déterminer les objets et les instances (stockées comme des données) sur lesquels l'expert effectuera son expérimentation. Cette tâche est effectuée en deux étapes :

4.1.1 Consultation du schéma de la base

Un premier schéma conceptuel de la base d'objets du SIDOC-Parole a été défini [1]. Ce schéma contient les caractéristiques de plusieurs objets-parole génériques, ainsi que les contraintes à respecter pour la création de leurs instances.

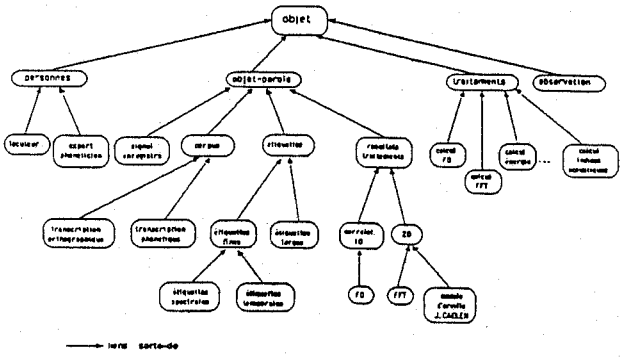


Figure 2.a
Exemple de Schéma Conceptuel

```
def-obj : SIGNAL_ENREGISTRE
    %sorte-de : OBJET-PAROLE
  #caract-enregistrement
    %un: CARACT_ENREG
  #no-élément
    %un : entier
  #longueur-blocs
    %un : entier

  %lien (équivalence,
   SIGNAL_ENREGISTRE, TRANSCRIPTION_PHONETIQUE)
  %lien (équivalence,
   SIGNAL_ENREGISTRE, TRANSCRIPT._ORTOGRAPHIQUE)
  %lien (composition-directe,
   SIGNAL_ENREGISTRE, TRANSCRIPT._ORTOGRAPHIQUE)
fin-def
```

Définition de l'objet SIGNAL_ENREGISTRE

```
def-obj : TRAITEMENT
    %sorte-de : objet
  #nom-traitement
    %un : chr(15)
  #objets-entrée
    %liste-de : OBJETS-PAROLE
  #objets-sortie
    %liste-de: OBJETS-PAROLE
  #programme-exécutable
    %un : chr(15)
  #taille-fenêtre-analyse
    %un : entier
fin-def

CALCUL_SPECTRE
    %est-un TRAITEMENT
  #nom-traitement : SPECTRE-PUISSANCE
  #objets-entrée  : SIGNAL-ENREGISTRE
  #objets-sortie  : ELEM_SPECTRE
  #progr.-exécutable : SPECTRE.EXE
fin-inst
```

Définition de l'OBJET TRAITEMENT
et exemple d'une de ses instances

```
def-obj : ELEM-SPECTRE
    %sorte-de : RESULTAT-TRAITEMENT
  #élément-traité
    %un : SIGNAL-ENREGISTRE
  #vecteur-spectre-par-fenêtre
  #bloc-début
    %un : entier

  %lien (composition-directe,
   ELEM-SPECTRE, SIGNAL-ENREGISTRE)
fin-def
```

Définition de l'objet ELEM-SPECTRE, résultat de l'application d'un traitement

Figure 2.b
Définitions obtenues par la primitive
VOIR-OBJ ‹ obj-gen›

L'expert peut interroger la base sur son contenu général et obtenir la liste des objets actuellement définis (figure 2.a). Avec cette liste, l'expert pourra déterminer ceux qui sont indispensables à son expérimentation. S'il souhaite obtenir des détails sur la définition d'un objet générique en particulier, il peut demander par exemple : VOIR_OBJ ‹ Signal_Enregistré›. (Voir figure 2.b).

4.1.2 Sélection d'un sous-ensemble d'objets et de leurs instances associées.

Il s'agit maintenant de créer l'espace d'observation avec les objets concernés par son expérimentation. Pour effectuer cette sélection, l'expert dispose des fonctionnalités offertes par le GO. Il fixe la liste des objets qui l'intéressent ainsi que les conditions pour sélectionner un sous-ensemble de leurs instances.

Pour valider la règle (1), l'expert pourrait par exemple sélectionner: ⌊ -tous les enregistrements du corpus "CVC07", prononcés par des locuteurs masculins ayant un accent parisien, les étiquettes fines spectrales associées ainsi que d'autres objets résultant de traitements et gérés par le GO -⌋. L'expert demande alors l'accès dans son espace d'observation:

- aux enregistrements sélectionnés avec les conditions: corpus.code = "CVC07", locuteur.sexe = "M", locuteur.accent = "Paris" (voir figure 3a)
- aux objets associés par des liens sémantiques, à ceux qu'il vient de sélectionner (on peut en effet avoir besoin ultérieurement de ces objets pour calculer certains paramètres), (voir figure 3b).

a)

```
SELECT    LOCUTEUR, CORPUS, SIGNAL_ENREGISTRE
  POUR    LOCUTEUR.sexe    = "M"
    ET    LOCUTEUR.ACCENT  = "Paris"
    ET    CORPUS.code      = "CVC07"
```

Requête qui copie dans l'espace de l'expert,
les objets génériques LOCUTEUR, CORPUS et SIGNAL_ENREGISTRE
ainsi que les instances satisfaisant les conditions spécifiées.

b)

* Par les liens d'EQUIVALENCE, il est possible de récupérer les instances des objets TRANSCRIPTION-ORTHOGRAPHIQUE et TRANSCRIPTION-PHONETIQUE liés à l'instance CVC071JA.SL (mot numéro 13 du corpus CVC07 prononcé par le locuteur SL) de l'objet SIGNAL_ENREGISTRE :

```
COPIER-OBJS-LIES-A  CVC071JA.SL  PAR-LIEN  EQUIVALENCE
```

* Par les liens de COMPOSITION, il est possible d'accéder à tous les objets dérivés (calculés) à partir de l'instance CVC071JA.SL (F0, INTENSITE, ENERGIE, etc). Par exemple, pour un premier niveau de dérivation l'expert peu copier les instances correspondantes avec :

```
COPIER-OBJS-LIES-A  CVC071JA.SL  PAR-LIEN  COMPOSITION  1
```

Figure 3.
Exemple de la Création de l'Espace d'Observation

Dans cet exemple, il a besoin d'analyser plusieurs paramètres. Ceux qui existent déjà dans la base d'objets, seront recopiés dans son espace d'observation : VALEUR_Fo, FORMANTS, INTENSITE, rapport SIGNAL/BRUIT. Les INDICES ACOUSTIQUES et la DUREE PHONEMIQUE seront obtenus au cours de la création de l'OBSERVATION.

4.2. CREATION D'OBSERVATIONS.

Pour la création d'une instance de l'objet OBSERVATION, il est nécessaire que l'expert détermine les entités concernées (ou individus) et les propriétés qu'il souhaite étudier. Les OBSERVATIONS sont définies de manière interactive entre l'expert et le système. Il s'agit donc de préciser, à partir des objets existant dans l'espace d'observation, quelles sont les propriétés à étudier et pour quels individus (au sens statistique du terme).

Les individus sont des instances d'un item défini comme la portion de signal que l'expert désire analyser. Il est possible de déterminer les frontières de l'item à considérer de deux manières : soit par des étiquettes (larges ou fines), soit par des repères numériques sur le signal (bloc,échantillon).

Pour notre exemple, les individus sont déterminés par les portions de signal repérés par les étiquettes fréquentielles fines: ETIQUETTE_COURANTE = C, ETIQUETTE_PRECEDENTE=(C + E), ou =(V + Q), ou =(C +T + NV). La liste des propriétés qui doivent apparaitre dans l'OBSERVATION correspond à celles qui apparaissent dans la règle de l'expert. La figure 4 présente l'OBSERVATION correspondant à l'exemple donné.

Du point de vue système, le premier pas consiste alors à déterminer la nature des propriétés,à savoir: un attribut simple d'un objet, un attribut composé, le résultat de l'application d'une fonction sur un attribut (simple ou composé), ou bien une notion qu'il est nécessaire de rendre explicite (soit par une simple définition, soit par le déclenchement d'un processus de calcul numérique ou de raisonnement [13]). Le GO produit ainsi une première sélection des instances des individus à analyser. Remarquons par exemple que l'indice_aigu_grave est un attribut de l'objet indices_acoustiques. Il n'est pas disponible directement dans l'espace d'observation de l'expert, mais il est obtenu pour la création de l'OBSERVATION.

		INTENSITE	NB_FORMANTS	DUREE	Fo	INDICE_AI/GR
INSTANCES	I1	valeurs				...
	I2					
			numériques ...			
	In					

Les Instances correspondent à l'ITEM sélectionné par une paire d'étiquettes

a) Avant les manipulations de l'expert

PREDICATS	P1	P2	P3	P4
poids / fréquence	.7	.3	.6	.8
inclure	*	no	*	*

b) Après les manipulations de l'expert

Figure 4.
Exemple du contenu de l'OBSERVATION

4.3 VALIDATION DE PREDICATS.

Il existe un ensemble d'opérations que l'expert peut utiliser pour manipuler les OBSERVATIONS puisque la nature de celles-ci est de type objet. Lorsque l'expert est satisfait du contenu de son OBSERVATION, il pourra l'interpréter et en tirer des conclusions à leur tour pondérées avec des poids (ou fréquences). L'expert doit choisir ensuite, quels sont les prédicats qui feront partie de la règle (ceux marqués avec *, dans la figure 4.b). Ceci est nécessaire, car le nombre de prédicats qu'il est possible de vérifier est très large et ils ne sont pas tous du ressort de la règle en cours de formulation.

4.4 FORMULATION DE LA REGLE ET SON INTEGRATION DANS LA BASE

Le GC du SIDOC-Parole offre un cadre complet pour la définition, la validation et la manipulation de REGLES. L'expert a alors la possibilité de

valider des règles nouvelles avant de les intégrer dans la base de REGLES mais il peut aussi les intégrer directement. A chaque ajout d'une règle, le GC est chargé de vérifier la cohérence avec le contenu précédant de la base de REGLES. L'expert peut demander au système l'exploitation d'une base de REGLES en particulier, en chaînage avant ou arrière sur un ensemble précis de la BC générale. Ce type de séance n'est pas présenté dans cet article.

La figure 5 montre l'écran proposé à l'expert pour l'intégration de sa règle dans la base de règles du SIDOC-Parole. Comme on l'a déjà signalé, avant d'intégrer la nouvelle règle, le système doit vérifier qu'elle ne remet pas en cause la cohérence et l'intégrité de la base déjà existante.

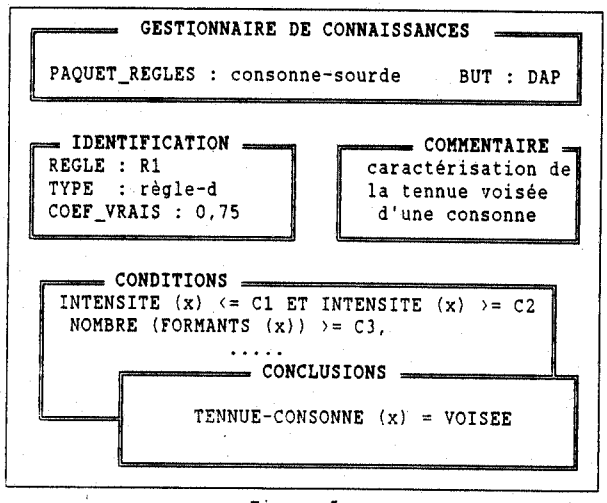


Figure 5.
Structure pour l'insertion d'une règle-d dans la BC

EXEMPLE B:

Cet exemple montre d'autres possibilités du SIDOC-Parole concernant le stockage de nouvelles connaissances. Supposons que l'expert désire observer s'il existe ou non une corrélation importante (>= 0,9) entre la valeur de l'**indice_F/O** (Fermé_Ouvert) et l'**aperture** des voyelles. Les activités dans les étapes de la séance d'expérimentation sont:

Création de l'espace d'observation:

L'expert regarde la définition de l'objet "voyelle" (voir figures 6 et 7) et il décide d'utiliser les liens sémantiques pour accéder aux objets de son intérêt. En donnant les **étiquettes** phonémiques des voyelles "V" (par exemple: /a/) précédées et suivies de [P, T ou K] il demandera certains **corpus** pour obtenir les segments de **signal enregistrés** qui deviendront les individus de son OBSERVATION. (Voir la figure 8).

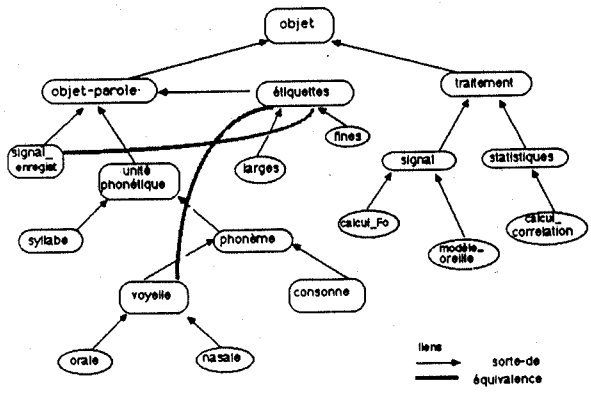


Figure 6.
Une partie du Schéma Conceptuel

```
def-obj : VOYELLE
    %sorte-de : OBJET-PAROLE
  #indices_spéctraux_acoustiques
    %liste-de : ELEM-INDICES-ACOUSTIQUES
  #indices-articulatoires
    %liste-de : ELEM-INDICES-ARTICULATOIRES
  %lien (équivalence,
   (ETIQUETTES_FINES, ETIQUETTES_LARGES)
fin-def

def-obj : ELEM-INDICES-ACOUSTIQUES
    %sorte-de : RESULTAT_TRAITEMENT
  #élement-traité
    %un : ELEMENT_SIGNAL_ENREGISTRE
  #indices-élem-par-fenêtre
    %liste-de : INDICES_ACOUSTIQUES
  %lien (composition,
   (ELEM_ENERGIE, ELEM_Fo, ELEM_SPECTRE)
fin-def

def-obj : INDICES_ACOUSTIQUES
    %sorte-de : OBJET_PAROLE
  #aigu_grave          %un : réel
  #fermé_ouvert        %un : réel
  #bémolisé_diésé      %un : réel
  #compact_écarté      %un : réel
  #doux_strident       %un : réel
fin-def

def-obj : ELEM-INDICES-ARTICULATOIRES
    %sorte-de : OBJET_PAROLE
  #description_articulatoire
    %liste-de : INDICES_ARTICULATOIRES
  %lien (équivalence,ELEM_SIGNAL_ENREGISTRE
fin-def

def-obj : INDICES_ARTICULATOIRES
    %sorte-de : OBJET_PAROLE
  #aperture            %un : réel
  #protrusion          %un : réel
  #position_machoire   %un : réel
  #position_langue     %un : réel
fin-def
```

Figure 7.
Définitions de quelques objets liés à l'objet VOYELLE

```
SELECT  CORPUS, SIGNAL_ENREGISTRE
  POUR  CORPUS.code     =  "PVPV"
    ET  CORPUS.code     =  "TVTV"
    ET  CORPUS.code     =  "KVKV"

Requête d'inclusion dans l'espace d'observation
de l'expert des objets génériques CORPUS et
SIGNAL_ENREGISTRE ainsi que leurs instances
satisfaisant les conditions spécifiés.

* par les liens d'EQUIVALENCE l'expert peut
  utiliser les ETIQUETTES associées aux CORPUS
  contenant des VOYELLES :

  COPIER-OBJS-LIES-A "PVPV" PAR_LIEN  EQUIVALENCE
  COPIER-OBJS-LIES-A "TVTV" PAR_LIEN  EQUIVALENCE
  COPIER-OBJS-LIES-A "KVKV" PAR_LIEN  EQUIVALENCE

* par les liens de COMPOSITION l'expert peut
  accéder à l'objet INDICE_F/O correspondant
  aux portions de signal sélectionnées

  COPIER-OBJS-LIES-A "PVPV" PAR_LIEN  COMPOSITION
  COPIER-OBJS-LIES-A "TVTV" PAR_LIEN  COMPOSITION
  COPIER-OBJS-LIES-A "KVKV" PAR_LIEN  COMPOSITION

* le système copie les objets dérivés nécessaires
```

Figure 8.
Requête schématisée de la création d'un Espace d'Observation pour étudier les VOYELLES

Création d'une observation:
L'expert, guidé toujours par le système, définit ensuite le contenu de sa première observation pour les voyelles "a" dans la base (voir la figure 9a).

Validation de prédicats:
Pour cette expérience la validation d'un prédicat unique est demandée (voir la figure 9b) :

	INDICE_F/O	APERTURE
instances de /a/	valeurs numériques	....
	....	....

a) Avant les manipulations de l'expert

prédicat PU	CORRELATION(INDICE_F/O, APERTURE)
résultat	0,95
inclure	*

b) Après la validation du prédicat

Figure 9.
Contenu d'une OBSERVATION pour les VOYELLES

L'expert effectue ensuite la même expérience en construisant des observations pour les autres voyelles.

Intégration de la connaissance dans la base:
Après avoir constaté que la corrélation est significative dans plusieurs cas de voyelles, l'expert décide de stocker cette connaissance sous deux formes: encapsulée dans une règle pour des buts de reconnaissance (voir figure 10), ainsi que comme une contrainte dans la définition de l'objet VOYELLE. Pour ceci il utilise la primitive d'ajout de contrainte aux objets [1]:

AJOUT_CONTR
< VOYELLE, CORRELATION(IND_F/O, APERTURE) > 0,85 >

Nous remarquons que l'utilisation de certaines primitives de manipulation d'objets est restreinte. La modification du schéma (comme cette modification à l'objet VOYELLE) est une opération privilégiée pour conserver la cohérence de la base [2].

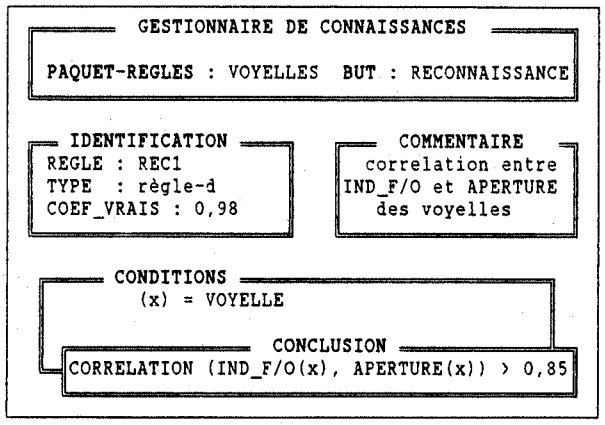


Figure 10.
Structure pour l'insertion d'une Règle-d dans la BC

5. ETAT D'AVANCEMENT ET PERSPECTIVES

A l'heure actuelle, le premier noyau des fonctionnalités du GO est disponible. L'expert peut consulter la structure et le contenu de la base d'objets du SIDOC-Parole. Le GC est en cours de développement pour permettre le déroulement des séances d'expertise. Les spécifications pour la mise en oeuvre des processus d'échange et de partage d'informations sont définies et partiellement réalisées. Le SIDOC-Parole vise à devenir un outil performant pour assister les experts dans leurs travaux de recherche sur la parole. Il reste pourtant plusieurs aspects à approfondir, notamment ceux concernant l'apprentissage automatique et les explications sur les raisonnements.

6. BIBLIOGRAPHIE

[1] CERVANTES O.
"Bases de Données et d'Objets Complexes Multimedia pour la recherche sur la parole". Thèse Docteur en Informatique de l'INPG. Janv.1988.

[2] FERNANDEZ Y., CERVANTES O.
"Eléments pour la gestion de connaissances dans le SIDOC-Parole"
Bulletin ICP, NO. 2, INPG, (à paraitre).

[3] FINDLER N. (ed)
"Associative Networks"
Academic Press, 1979.

[4] CERVANTES O., SERIGNAT J.F.
"Représentation Centré Objet dans la BDC-Parole"
16èmes JEP-SFA, pp.331-334,
Hammamet, Tunisie, Octobre, 1987.

[5] CAELEN J., CAELEN-HAUMONT G., et al.
"ARCANE: Acquisition et Recherche de Connaissances Acoustico-Phonétiques"
15èmes JEP-SFA, pp.207-211,
Aix-en-Provence, 1986.

[6] NILSSON N.
"Principles of Artificial Intelligence"
Tioga Pub. Co., 1980.

[7] CLANCEY, W.J.
"The Epistemology of a Rule-based Expert System : a Framework for Explanation".
Artificial Intelligence Vol. 20, 1983.

[8] LAURIERE J.L.
"Représentation et Utilisation de Connaissances"
Techniques et Sciences Informatiques, Vol. 1, No. 1 et 2, 1982.

[9] DIETTERICH T.G., MICHALSKY R.S.
"Inductive Learning of Structural Descriptions: Evaluation Criteria and Comparative Review of Selected Methods".
Proceedings 6th Int.Joint Conference on Artificial Intelligence, Tokyo, Août 1979

[10] GUIZOL J.
"Apprentissage Inductif de Règles pour le Décodage Acoustico- Phonétique".
15èmes JEP-SFA, pp. 227-230,
Aix-en-Provence, Mai 1986.

[11] HASLING, D.W.
"Abstract Explanations of Strategy on a Diagnostic Consultation System"
Proc.National Conf. on Artificial Intelligence, Washington, D.C.,1983.

[12] CAELEN J., TATTEGRAIN H.
"Le Décodeur Acoustico-Phonétique DIRA-DAP"
Proposé aux 17èmes JEP-SFA, Nancy, 1988.

[13] CAELEN J., CERVANTES O., FERNANDEZ Y.
"Mécanismes de Consultation dans la Base de Données et de Connaissances Parole".
16èmes JEP-SFA, pp.327-330,
Hammamet, Tunisie, 1987.

DETECTION DE FRONTIERES SYNTAGMATIQUES EN PAROLE CONTINUE : UTILISATION DE LA FREQUENCE FONDAMENTALE.

N. CARBONELL, J. J. BONIN

CRIN-INRIA
BP 239 - 54536 Vandoeuvre lès Nancy Cedex

ABSTRACT

We are studying how prosodic cues could facilitate continuous speech recognition and understanding. In this paper, results are discussed concerning the detection of syntagm boundaries from the analysis of fundamental frequency variations. Three corpus have been tested corresponding to various situations : reading, memorized sentences (short term memory), quasi-spontaneous task-oriented dialogues. Our present rates for syntagm boundary detection are roughly : 95% correct detections and 7% insertions.

1 CADRE ET OBJECTIFS DE L'ETUDE.

Il semble qu'**en reconnaissance/compréhension de la parole continue (RCPC)**, les variations temporelles des paramètres prosodiques, en particulier l'évolution de la fréquence fondamentale, soient porteuses d'informations linguistiques et que la prise en compte de ces informations dans les systèmes de compréhension de la parole continue (SCPC) puisse accroître sensiblement les performances qualitatives de ces systèmes.

Si, actuellement, dans les approches classiques, on n'envisage plus d'utiliser les informations prosodiques pour guider le processus de reconnaissance/compréhension, à la différence de LEA (LEA 75), les systèmes qui reposent sur la coopération de plusieurs sources de connaissances (acoustico-phonétique, lexicale, syntaxico-sémantique et pragmatique principalement) incluent souvent une composante prosodique (MELONI 82, MARTIN 79, PERENNOU 82, VAISSIERE 82, WAIBEL 86).Toutefois, dans ces systèmes, le rôle d'une telle composante est relativement limité, en raison de la relative pauvreté et du manque de fiabilité des informations disponibles ; en outre, la segmentation automatique du signal de parole introduit souvent un nombre d'erreurs supllémentaires non négligeable. Signalons enfin que rartes sont les systèmes qui comme ceux de Vaissière et Martin prennent en compte l'ensemble des paramètres prosodiques.

Le travail que nous relatons ici a pour but d'**intégrer informations et connaissances prosodiques à un système de compréhension** de dialogues oraux en langue naturelle. La Figure 1 (d'après PIERREL 87) décrit l'architecture du système à sources de connaissances multiples en cours de développement au CRIN. Nous envisageons d'y intégrer les résultats de la présente étude sous forme d'une composante qui, à partir d'une analyse du signal et de résultats ponctuels fournis par le module de décodage acoustico-phonétique (APHON) du système, fournirait des informations utilisables par les composantes lexicale (LEX), syntaxico-sémantique (SYN-SEM) et même pragmatique (DIAL).

Dans un premier temps, nous avons **limité nos objectifs** :

- pour l'instant, les résultats que nous présentons concernent uniquement les variations temporelles de F0,

- notre démarche et nos hypothèses de base nous ont conduits à une restriction plus fondamentale : nous n'envisageons pas actuellement de construire une représentation prosodique complète d'un énoncé car, pour y parvenir, il nous faudrait nous appuyer largement sur les analyses et les conclusions des autres modules du système, en particulier sur le treillis phonétique produit par la composante acoustico-phonétique.

Or, cela nous paraît dangereux, dans la mesure où les résultats prosodiques et acoustico- phonétiques servent de base à l'élaboration des représentations lexicale et syntaxico-sémantique d'un énoncé (cf. le fonctionnement de notre système décrit dans CARBONELL 86). Pour assurer la robustesse de ces représentations, il est donc impératif, d'une part, que les modules prosodique et acoustico-phonétique évitent d'utiliser les conclusions des autres modules et, d'autre part, qu'ils travaillent de manière aussi indépendante que possible l'un de l'autre. En particulier, il est nécessaire qu'ils s'appuient sur des paramètres acoustiques différents et des analyses acoustiques indépendantes, de façon à éviter la production par le module prosodique d'hypothèses déduites d'interprétations phonétiques erronées, hypothèses qui renforceraient ces interprétations (et réciproquement) ; l'un ou l'autre de ces deux modules pourrait même, dans ces conditions, détecter les erreurs de l'autre et contribuer à les corriger.
C'est pourquoi le seul résultat du module acoustico-phonétique que nous comptons utiliser, mis à part la détection des noyaux vocaliques, est la durée vocalique moyenne, notion globale (car relative à l'énoncé entier) et donc robuste ; cette notion est définie dans le paragraphe 3.2.

2 DIFFICULTES D'UTILISATION EN RCPC DE LA PROSODIE.

La difficulté majeure est de concevoir une composante prosodique indépendante, au sens où nous l'avons défini en 1, des autres composantes.
En effet, les variations de la fréquence fondamentale, du rythme d'énonciation et de l'intensité ont de multiples fonctions (VAISSIERE 88) et résultent de facteurs hétérogènes, entre autres :

- la nature des phonèmes et les phénomènes de coarticulation très marqués en parole continue,

- la syntaxe de l'énoncé,

- le contenu sémantique de l'énoncé et l'environnement pragmatique, plus généralement, la manière dont le locuteur se situe par rapport à ce qu'il dit (cf. par exemple, les accents emphatiques),

- la variabilité inter et intra locuteur (paramètres physiologiques et psychologiques essentiellement).

Or les effets de ces paramètres sur le signal sont souvent semblables entre eux, au moins en ce qui concerne la parole non contrainte. A titre d'illustration nous indiquons quelques observations résultant de l'analyse d'un corpus multi-locuteur de parole quasi-spontanée décrit plus loin (cf. paragraphe 3.1) :

- des réalisations de /a/ au sein d'un syntagme et en position non accentuée peuvent avoir une durée égale à celle de certains /a/ en fin de syntagme (1,5 fois la durée vocalique moyenne) ;

- un accent d'emphase peut engendrer sur la première syllabe d'un mot comme "maxima" une variation de F0 égale ou supérieure à celle observée sur la dernière voyelle du syntagme contenant ce mot (dans le cas d'une continuation majeure) ;

- chez certains locuteurs, on n'observe aucune variation significative de la durée des phonèmes au cours d'un énoncé ; chez d'autres on constate des variations importantes pouvant aller jusqu'au triple de la durée vocalique moyenne.

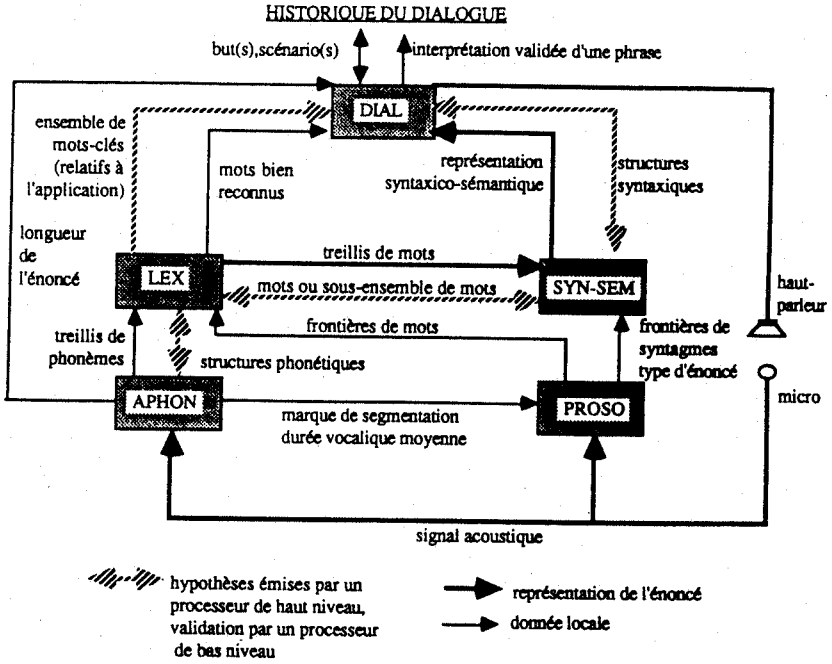


Figure 1 : Architecture générale du système de dialogue

(D'après PIERREL 87)

En ce qui concerne l'évolution temporelle de F0, nous avons ignoré les variations intravocaliques et pris en compte uniquement les zones centrales des voyelles, car ces variations sont très liées aux phénomènes de coarticulation ; leur interprétaion nécessite donc l'identification phonétique préalable du contexte. Or, dans un premier temps, nous recherchons des algorithmes qui soient aussi indépendants que possible des résultats du décodage acoustico-phonétique.
Mais cette solution conduit à ignorer des informations très riches sur le plan linguistique, par exemple celles que fournissent les "glissandos" (ROSSI 81, pp 55 et 203) et les continuations majeures/mineures (DELATTRE 66a), sur la structure syntaxico-sémantique de l'énoncé.
Nous envisageons donc, dans une phase ultérieure de notre étude, de remettre en cause cette stratégie et de prendre en compte les variations temporelles de F0 au sein de certains noyaux vocaliques ; mais la sélection de ces noyaux pose problème si on refuse de s'appuyer sur les conclusions (treillis phonétique) du décodage acoustico-phonétique.

Quant aux variations du rythme, que nous sommes en train d'aborder, nous ne les utilisons pour le moment, qu'à titre de confirmation des informations fournies par l'analyse de F0.

3 NOTRE ETUDE.

3.1 Corpus utilisés.

Notre travail s'appuie sur trois corpus recueillis respectivement dans trois situations de prise de parole différentes. Ces situations correspondent aux principales familles d'applications potentielles des SCPC à moyen terme :

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

Ce choix résulte de notre volonté de tester l'hypothèse intuitive suivante :
la nature et la forme des manifestations macroprosodiques linguistiquement significatives varient d'une situation de prise de parole à l'autre ; dictée et lecture d'une part, expression spontanée d'autre part, diffèrent sensiblement au niveau prosodique.

Caractéristiques des trois corpus :
Fréquence d'échantillonnage : 16 kHz.
Segmentation et étiquetage phonétique "manuels" par des experts phonéticiens, à partir de spectrogrammes numériques de qualité comparable à celle des spectrogrammes analogiques (LAPRIE 88); trois experts, un par corpus. L'homogénéïté est donc assurée pour chaque corpus.

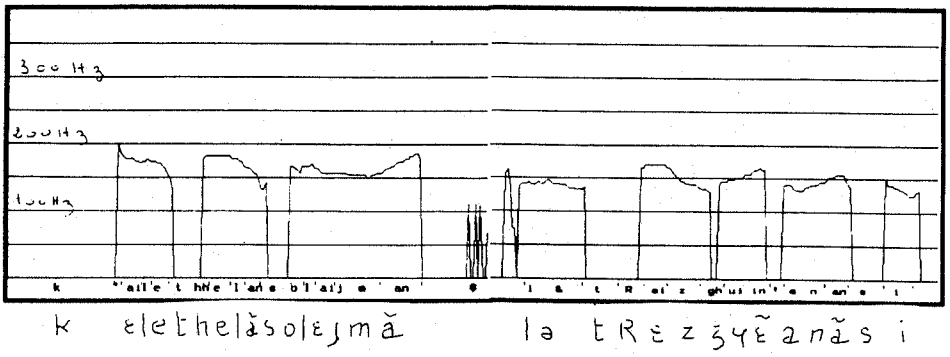
Corpus CMB :
50 phrases différentes prononcées par 5 locuteurs masculins non professionnels.
Protocole : le texte d'une phrase du corpus de Combescure (COMBESCURE 81) est présenté au locuteur qui, après l'avoir lu et mémorisé, le dit aussi naturellement que possible.
L'acquisition a été effectuée dans l'atmosphère assez calme d'une salle non insonorisée.

Corpus LABISE :
Il s'agit de la lecture d'un texte relativement simple "La bise et le soleil". Le corpus (BDSONS) a été constitué sous la direction et avec le soutien du GRECO Communication parlée ; il a été enregistré dans d'excellentes conditions (chambre sourde / dynamique importante).
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 quasi-spontanée en situation de dialogue oral finalisé : consultation d'un centre de renseignements météorologiques (CARBONELL 82).
10 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 nous nous sommes efforcés de n'induire ni la forme des questions ni, plus généralement, la structure du dialogue.

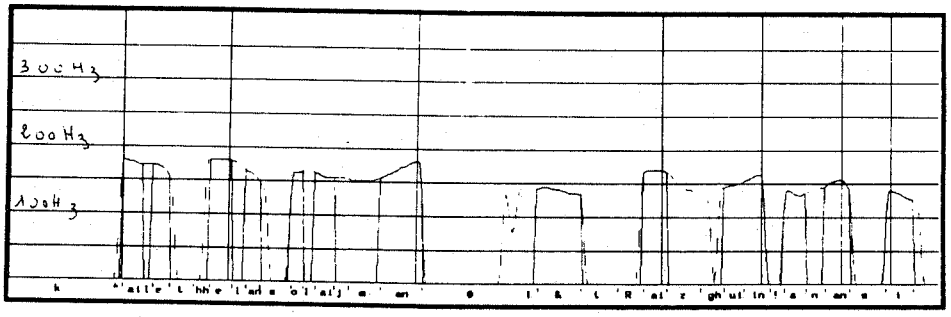
exemple :
continuation du froid (jusqu'à quand)

Conditions d'enregistrement : atmosphère relativement calme, salle légèrement insonorisée. Actuellement, segmentation et étiquetage n'ont été réalisés que pour 3 locuteurs.



"Quel était l'ensoleillement ... le treize Juin à Nancy ?"

Figure 2 : Calcul de la fréquence fondamentale (corpus METEO)



"Quel était l'ensoleillement... le treize Juin à Nancy ?"

Figure 3 : Positionnement des marques de frontières syntagmatiques par notre algorithme.

3.2 Outils logiciels d'analyse.

Calcul de la fréquence fondamentale :
Notre choix s'est porté sur l'algorithme d'autocorrélation de Sondhi (SONDHI 68) repris dans (RABINER 78), en raison de sa relative résistance au bruit et de sa rapidité de calcul ; nous avons adopté, en l'affinant, l'algorithme de correction de Bristow (BRISTOW 82).
Ces algorithmes ont été implantés sur MASSCOMP (BONIN 87) et utilisent un processeur vectoriel.
La figure 2 fournit un exemple pris dans le corpus METEO.

Détection des noyaux vocaliques et estimation de la "durée vocalique moyenne" :
La détection des noyaux vocaliques repose sur le repérage des maxima d'énergie dans une bande d'énergie adaptée ; la durée vocalique moyenne est la valeur médiane des durées des noyaux vocaliques de l'énoncé. Ces algorithmes sont décrits dans (FOHR 86). Actuellement, nous n'avons utilisé cet algorithme que ponctuellement pour valider notre détection de frontières syntagmatiques ; lorsque cette dernière sera au point, nous comptons l'utiliser à l'exclusion de la segmentation manuelle.

Détection des pics linguistiquement significatifs de la fréquence fondamentale :
L'analyse porte uniquement sur les valeurs de F0 correspondant aux noyaux vocaliques. Pour l'instant, nous nous limitons à une seule valeur de F0 par noyau, mesurée au maximum énergétique du noyau (BONIN 87). Après lissage de la courbe des variations de F0, l'algorithme sélectionne les pics de la courbe les mieux marqués à l'aide de seuils relatifs ; il place ensuite les marques de frontières syntagmatiques au sommet des pics retenus.
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 (LAPRIE 88).

3.3 Approche et résultats.

Nos résultats actuels portent uniquement sur la détection de frontières syntagmatiques à partir de l'analyse des variations de la fréquence fondamentale, car l'intonation constitue le paramètre prosodique actuellement le mieux connu des phonéticiens, au moins pour le français.
Parmi les marqueurs intervenant dans les modèles intonatifs proposés par Rossi (ROSSI 81, pp 179 à 233) aussi bien que par Martin (MARTIN 81, pp 234 à 271) et Hirst (HIRST 80), les plus accessibles à une détection automatique sont certainement les contours mélodiques indiquant les continuations majeures/ mineures ou , plus généralement, la structuration hiérarchique de l'énoncé.
Nous avons donc mis au point un algorithme de recherche et d'analyse des pics significatifs de F0 au sein d'un énoncé, susceptible de détecter certaines fins (dernier noyau vocalique) de syntagmes, donc certaines frontières entre unités lexicales (ou fins de mots).
Cet algorithme a été testé sur les trois corpus dont nous disposons et fournit des résultats qui sont détaillés dans le tableau de la figure 4.

A noter que, dans ce tableau, "marques trouvées" désigne pour nous les frontières syntagmatiques détectées par l'algorithme ; "marques correctes" indique, parmi les frontières détectées, celles qui correspondent à une frontière syntagmatique réelle et "marques attendues", toutes les frontières syntagmatiques présentes dans les différents corpus.

Pour déterminer, sur les trois corpus, les marques attendues, nous avons appliqué les règles proposées par Rossi (cf. ROSSI 81, le tableau page 203) pour rendre compte de l'augmentation sensible de F0 sur le dernier noyau vocalique de certains syntagmes au sein d'un énoncé. Nous avons ajouté à ces marques les fins d'énoncés, frontières lexicales aisément détectables automatiquement. Notre algorithme détecte donc, outre les frontières syntagmatiques à l'intérieur d'un énoncé, les fins d'énoncés.

Corpus	marques trouvées	marques correctes	marques attendues
CMB	161	153	160
LABISE	457	425	450
METEO	461	425	-

Figure 4 : Résultats.

Rappelons que ces résultats ont été obtenus à partir de l'étude des noyaux vocaliques fournis par l'étiquetage manuel. L'utilisation de NOVOCA pour la détection des noyaux vocaliques ne modifierait vraisemblablement pas beaucoup les conclusions, en raison de la qualité des résultats que cet algorithme fournit (voisins de 95%). Nous préférons améliorer et mettre au point nos divers algorithmes de détection d'indices prosodiques en nous appuyant sur une segmentation manuelle plutôt qu'automatique pour des raisons de commodité essentiellement, les essais ponctuels effectués avec NOVOCA fournissant des résultats très semblables à ceux obtenus à partir de la segmentation manuelle.
Sont comptées comme erreurs, uniquement les marques qui ne correspondent ni à une fin de syntagme ni à une fin de mot, par

Ces résultats appellent les commentaires suivants :

Les frontières syntagmatiques internes à un énoncé sont détectées avec une bonne fiabilité, de l'ordre de 95%. Nous n'avons pas tenu compte des "e" dits muets, estimant que c'est à la composante lexicale de traiter les problèmes posés par le présence facultative de cette voyelle. Notre objectif étant de fournir des informations sûres aux différents modules d'un SRCPC, nous nous sommes intéressés aux insertions plutôt qu'aux omissions, tout au moins dans un premier temps.

Nous avons constaté que les **"erreurs" de l'algorithme** provenaient essentiellement de la détection d'accents que l'algorithme confond avec des frontières syntagmatiques ; il s'agit principalement :

- d'une part, d'accents d'insistance ou d'emphase ; par exemple :

"Le mal s'envenime faute de soins." (CMB)
"J'aimerais connaître le temps prévu le douze juin mille neuf cent quatre-vint deux." (METEO)
"Alors la bise se mit à souffler de toutes ses forces..." (LABISE)

Dans ces exemples une frontière lexicale est bien détectée mais elle ne correspond pas à une frontière syntagmatique.

- d'autre part, d'accents affectant les mots longs, chez certains locuteurs uniquement :

"On entend les gazouillis d'un oiseau dans le jardin." (CMB)
"Quelle sera la température maxima à Remiremont..." (METEO)
"...un voyageur qui s'avançait..." (LABISE, 6 locuteurs sur 9)

De même les mots commençant par le préfixe "re" ou "ré" reçoivent parfois une marque sur le préfixe, par exemple : "réchauffé" et "reconnaître", dans LABISE.

Signalons enfin qu'une **étude perceptive** menée sur le corpus METEO (CARBONELL 82) suggère que la perception humaine est capable de distinguer les accents des variations de F0 en fin de syntagme ; mais on peut s'interroger sur la nature de ce filtrage : n'est -il pas induit par la prise en compte d'informations syntaxico-sémantiques plutôt que par l'interprétation d'informations acoustiques ?

Pour conclure, précisons que : si, dans le cadre de la RCPC il n'est pas nécessaire d'obtenir des marqueurs prosodiques pour toutes les fins de syntagmes, un tel marquage perd beaucoup de son intérêt si 10% des marques coïncident, non pas avec le dernier noyau vocalique d'un mot mais avec le premier ou le deuxième (cf. les mots longs).
On pourrait peut-être accroître la robustesse de cet algorithme en tenant compte également de la durée des noyaux vocaliques, Delattre (DELATTRE 66b), Rossi (ROSSI 81) et Di Cristo (DI CRISTO 81) signalant un accroissement de la durée des noyaux vocaliques en fin de syntagme.

4 CONCLUSION.

Les connaissances acquises par les phonéticiens sur l'intonation paraissent utilisables et utiles en RCPC. En particulier, nous avons montré qu'il était possible de construire un algorithme robuste et fiable de détection des frontières syntagmatiques.
En outre, les indications fournies par cet algorithme, même lorsqu'elles ne correspondent pas à des frontières syntagmatiques mettent en évidence des frontières lexicales, sauf dans les cas relativement peu fréquents d'accentuation de mots longs, l'accent plus ou moins marqué pouvant apparaître sur la première syllabe ou une syllabe intermédiaire du mot.et être confondu par l'algorithme avec une frontière syntagmatique. Notre algorithme détecte donc avec une bonne fiabilité les frontières lexicales

En ce qui concerne le rythme, cette conclusion doit être nuancée, dans l'état actuel de nos connaissances. Nous venons de commencer l'étude de ce paramètre en utilisant la notion de durée vocalique moyenne (cf. paragraphe 3.2) pour, d'une part, distinguer les pauses de la partie silence des plosives sourdes et d'autre part contribuer à éliminer les marques correspondant à des accents.
Pour l'instant, nous avons effectué une étude purement descriptive des trois corpus par rapport à la notion de durée vocalique moyenne.
Cette étude fournit des résultats qui diffèrent selon les corpus. Si, dans la lecture (cf. LABISE) les allongements vocaliques coïncident dans l'ensemble avec des fins de syntagmes, et les pauses, fréquentes, se produisent toujours en fin de syntagme, ces régularités sont moins fréquentes dans le corpus CMB. Enfin, une certaine anarchie caractérise le corpus METEO sur le plan du rythme. C'est du moins ce que suggèrent les résultats bruts ; nous envisageons d'affiner l'analyse pour déterminer s'il n'existe pas d'autres régularités, spécifiques de la parole spontanée, et si l'on peut définir un modèle adéquat de la distribution temporelle en parole continue spontanée.
Il semble donc que la prise en compte d'informations relatives au rythme puisse faciliter et améliorer recherche lexicale et analyse syntaxico-sémantique lorsqu'il s'agit de lecture, mais pas dans le cas de la parole spontanée. Des études fines sont à faire dans ce domaine.

L'intensité quant à elle, constitue un paramètre encore plus difficile à utiliser en RCPC, l'énergie des différentes voyelles présentant une grande variabilité ; par exemple, l'énergie intrinsèque de /a/ est nettement supérieure à celle de /i/, /y/, ou /u/.

Par ailleurs, il paraît difficile de mettre en oeuvre toute l'expertise des phonéticiens en ce qui concerne l'intonation : les imperfections des algorithmes de segmentation actuels, la difficulté à isoler les phénomènes macroprosodiques sans faire appel aux résultats du décodage acoustico-phonétique, rendent hasardeuse la mise en oeuvre des connaissances acquises sur l'évolution de F0 au sein des noyaux vocaliques en fin de syntagme (cf. par exemple, la notion de continuation majeure ou mineure). Il paraît donc impossible pour l'instant d'utiliser les informations syntaxiques les plus riches que véhicule la macroprosodie.

Enfin, la question reste ouverte du rôle à donner aux informations prosodiques dans un SRCPC : validation (cf. VAISSIERE 82) ou émission d'hypothèses? composante à part entière interagissant avec les autres au cours de la reconnaissance/compréhension d'un énoncé, ou bien composante exclusivement productrice d'informations, fournissant aux autres modules des informations ponctuelles susceptibles de compléter celles fournies par le décodage acoustico-phonétique? Il est nécessaire d'approfondir les études avant de pouvoir conclure et être en mesure d'élargir le domaine des connaissances prosodiques implantables dans un SRCPC.

REFERENCES.

[BONIN 87] Bonin J. J., "Détection d'indices prosodiques linguistiquement significatifs",
Mémoire de DEA Informatique, Université de Nancy I, 1987.

[BRISTOW 82] Bristow G. J., Fallside F., "An autocorrelation pitch detector with error correction",
IEEE, pp 184-187, 1982.

[CARBONELL 82] Carbonell N., Haton J. P., Lonchamp F., Pierrel J. M., "Elaboration expérimentale d'indices prosodiques pour la reconnaissance, application à l'analyse syntaxico-sémantique dans le système MYRTILLE II",
in Di Cristo A., Haton J. P., Rossi M., Vaissière J. (éd.) "Prosodie et reconnaissance automatique de la parole", GALF Groupe de la Communication parlée, pp 59-91,1982.

[CARBONELL 86] Carbonell N., Pierrel J. M., "Architecture and knowledge sources in a human- computer oral dialogue system",
Workshop NATO on multimodal dialogues including voice, Corse, 1986.

[COMBESCURE 81] Combescure P., "Vingt listes de dix phrases phonétiquement équilibrées",
Revue d'Acoustique 14, 1981.

[DELATTRE 66a] Delattre P., "Les dix intonations de base du français",
French Review, 40 (1), pp. 1-14, 1966.

[DELATTRE 66b] Delattre P., "A comparison of syllable length conditioning among languages",
Applied Linguistics, vol. 4, n° 3, 1966.

[DI CRISTO 81] Di Cristo A., "De la microprosodie à l'intonosyntaxe",
Thèse pour le Doctorat d'Etat, Université de Provence, 1981.

[FOHR 86] Fohr D., "APHODEX : un système expert en décodage acoustico-phonétique de la parole continue",
Thèse de Doctorat d'Université, Université de Nancy I, 1986.

[HIRST 80] Hirst D., "Un modèle de production de l'intonation.
Travaux de l'Institut de Phonétique d'Aix en Provence, Vol 7, 1980.

[LAPRIE 88] Laprie Y., Snorri : un système interactif d'étude de la parole",
Article soumis aux XVII^{ème} JEP pour acceptation , Nancy, 1988.

[LEA 75] Lea W., Medress M. F., Skinner T. E., "A prosodically guided speech understanding strategy",
IEEE Trans. Vol. ASSP-23, pp 30-38, 1975.

[MARTIN 79] Martin P., "Automatic location of stressed syllable in French",
Current Issues in Linguistic Theory, vol. 9, pp. 1091-1094, 1979.

[MARTIN 81] Martin P., "Pour une théorie de l'intonation",
in "L'intonation de l'acoustique à la sémantique", Etudes linguistiques XXV, Klingsieck, 1981.

[MELONI 82] Meloni H., Guizol J., "Utilisation des paramètres prosodiques dans un système de reconnaissance automatique de la parole continue",
in Di Cristo A., Haton J. P., Rossi M., Vaissière J. (éd), "Prosodie et reconnaissance automatique de la parole", GALF Groupe de la Communication parlée, pp 93-120, 1982.

[PERENNOU 82] Perennou G., Caelen G., "Utilisation de la prosodie pour la reconnaissance de la parole dictée",
in Di Cristo A., Haton J. P., Rossi M., Vaissière J., "Prosodie et reconnaissance automatique de la parole", GALF Groupe de la Communication parlée, pp 25-57, 1982.

[PIERREL 87] Pierrel J. M., "Dialogue oral homme-machine (connaisances linguistiques, stratégies et architecture des systèmes)",
Hermès, 1987.

[RABINER 78] Rabiner L. R., Schafer R. W., "Digital processing of speech signals",
by Bell Laboratories, Prentice Hall, 1978.

[ROSSI 81] Rossi M., "L'interprétation perceptive", "Intonation, énonciation, syntaxe",
in "L'intonation de l'acoustique à la sémantique", Etudes linguistiques XXV, pp 54-63, 184-233, Klingsieck 1981.

[SONDHI 68] Sondhi M. M., "New methods of pitch extraction", IEEE Trans. AU-16, 262-266, 1968.

[VAISSIERE 82] Vaissière J., "A suprasegmental component in a french speech recognition system : reducing the number of lexical hypotheses and detecting the main boundary",
Recherches Acoustiques, Centre National d'Etudes des Télécommunications, Vol VII, 109-125, 1982.

[VAISSIERE 88] Vaissière J., "The use of prosodic parameters in automatic speech recognition",
à paraître in Nieman, Lang, Sagerer (ed), "Recent advances in speech understanding and dialog system", NATO ASI Series, Springer Verlag.

[WAIBEL 86] Waibel A., "Prosody and speech recognition",
PhD Dissertation, Computer Science Departement, Carnegie Mellon University, 1986.

Prédiction et vérification lexicale dans le cadre d'un dialogue oral homme-machine

Laurent ROMARY - ESE/CRIN/INRIA
Bernard MANGEOL - CRIN/INRIA

BP 239, 54506 Vandœuvre.
(romary@crin.UUCP)

Abstract

As a central part of the man-machine oral dialogue system under development at the CRIN at Nancy, we present here our conception of its lexical component. We first show the importance of linguistic knowledge as a guide to word recognition, and particularly, how contextual information can be obtained at any step of a dialogue in a task oriented application. Then, we present some techniques of combining hypotheses thanks to a specific lexical representation and the use of the Dempster-Shafer theory for combining word evidence. At last, we describe the prediction and verification of word presence along the speech signal through the use of macro-classes of phonemes and dynamic programming algorithms.

1. Introduction.

Malgré la relative continuité des recherches en intelligence artificielle ces dernières années, de nouvelles façons d'analyser les problèmes qui touchent ce domaine tendent à faire évoluer celui-ci de manière profonde, grâce aux efforts conjoints de nombreuses disciplines regroupées sous le terme général de sciences cognitives. En particulier, un système dit intelligent est de moins en moins vu comme un ensemble clos où se trouvent centralisées toutes les informations et les décisions, mais plutôt comme un univers où plusieurs entités coopèrent pour permettre la réalisation d'une fonction particulière vis-à vis du monde extérieur. Ce paradigme permet, par exemple, la considération d'environnements multi-experts, ou la gestion de l'interaction entre plusieurs agents autonomes (des robots) dans un univers physique particulier.

Cette évolution se retrouve dans les travaux touchant la reconnaissance de la parole où l'on parle maintenant de dialogue oral homme-machine en considérant, non plus un système de reconnaissance indépendant du monde extérieur qui tente (parfois vainement) de comprendre un signal qui lui est présenté en entrée, mais un agent intelligent, destiné à converser avec d'autres usagers et pour cela, convié à intégrer sa composante de reconnaissance dans une description plus générale de l'univers qui l'entoure. Cela nécessite, au niveau d'un tel système, la mise en place d'un double mécanisme, à savoir la prise en compte de nouveaux éléments à chaque interaction avec l'extérieur et l'utilisation de ces objets au niveau du module de compréhension pour optimiser son analyse des énoncés à venir.

C'est dans cette optique que nous avons conçu l'architecture du système de dialogue oral homme-machine en cours de développement au CRIN à Nancy [Carbonell 87][Pierrel 87], dont l'organisation interne est elle-même particulièrement modulaire. Nous n'allons détailler ici qu'une certaine vision du lexique, en regardant comment celui-ci échange des informations avec le reste du système, ainsi que les choix de techniques que nous avons été amenés à faire pour intégrer ces informations et réaliser des prédictions sur des mots, à partir de la représentation phonétique d'un énoncé.

2. La place d'une composante lexicale dans un système de dialogue oral homme-machine.

Les informations lexicales sont les premiers éléments linguistiques, manipulés par le système, qui ne soient pas propres à la communication orale. En effet, tous les niveaux inférieurs, filtrage, transformées, et décodage acoustico-phonétique travaillent essentiellement sur des informations acoustiques et ceci, d'une manière purement ascendante, du signal jusqu'à une représentation sous forme de phonèmes ou de syllabes. L'expérimentation relative à Hearsay II avait déjà montré ce résultat en adoptant une structure des sources de connaissance dans le Blackboard où le premier grand pôle d'échanges se situait au niveau du mot [Erman 80].

Une entité lexicale peut en effet être vue comme un point de convergence d'informations de types divers, comme un patron phonétique, une classe (ou une structure) grammaticale, et une représentation sémantique. Chacune d'entre elles peut servir de point d'accès au lexique tout entier ou, dans le cas de la reconnaissance d'un énoncé, d'espace de décision particulier en fonction des informations disponibles dans le système.

On considère communément que l'impossibilité d'obtenir un décodage acoustico-phonétique d'excellente qualité, impose aux niveaux supérieurs de limiter progressivement les cas d'ambiguité entre les différentes interprétations possibles du signal. Cependant, nous allons voir que les niveaux linguistiques peuvent posséder un rôle analogue à l'étape de décodage, au sens où eux aussi peuvent proposer certains éléments du lexique qui s'avèrent pertinents à chaque stade de la reconnaissance. Afin de comprendre l'origine de ce processus, il est nécessaire de décrire les connaissances qui vont être utilisées à différents niveaux d'analyse et que nous nommerons de manière générale le *contexte*.

Le contexte n'apparaît pas ici comme une entité externe aux intervenants d'un dialogue, comme parfois il peut être défini en pragmatique [Latraverse 87], entité qui représenterait toutes les conditions préalables dans lesquelles s'insère une suite d'énoncés. Cette vue du contexte n'a aucun sens si l'on désire modéliser le comportement d'un agent intervenant dans le dialogue. En effet, ce qui importe est l'état "mental" d'un des protagonistes et l'influence qui en résulte au niveau de la reconnaissance. Le contexte est donc l'ensemble des connaissances du système à un instant donné de son existence, qui forme son espace de représentation du monde extérieur.

Nous allons raisonner dans le cadre précis de l'application envisagée qui est l'interrogation par un utilisateur d'un centre de renseignements administratifs simulé par la machine (correspondant sensiblement aux informations disponibles dans les pages roses d'un annuaire). L'architecture envisagée (cf fig.1) fait ressortir quatre processeurs indépendants qui interagissent avec le lexique (LEXIQUE), du décodeur acoustico-phonétique (APHON) au module de dialogue (DIALOGUE), en passant par le détecteur d'indices prosodiques (PROSODIE) et les analyseurs syntaxico-sémantiques (ANALYSEURS). A chaque niveau, des informations de plus en plus précises sont disponibles qui toutes entrent dans la définition du contexte.

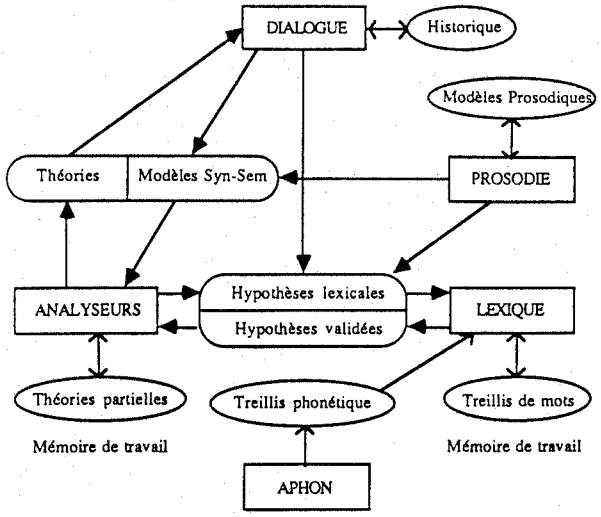


Figure 1.

Le module de dialogue gère en continu les échanges entre le locuteur et la machine. En particulier, il dirige partiellement le dialogue et déclenche les phases d'acquisition d'un nouvel énoncé. En début de dialogue, il va ainsi pouvoir prédire un énoncé de salutation en réponse à une introduction de la part de la machine du type "*Centre de renseignements administratifs, Bonjour!*". Puis un sous-lexique particulier relatif à une demande d'informations va pouvoir être proposé (typiquement : "*Je voudrais...*"), correspondant à l'expression de la requête de l'usager. Enfin, après satisfaction de celui ci, on peut s'attendre à deux types d'énoncés auxquels correspondent deux sous-lexiques particuliers, à savoir un énoncé de relance ou de fin de dialogue ("*Je vous remercie, au revoir.*").

Ces connaissances générales au niveau d'un dialogue peuvent s'affiner à chaque échange puisque progressivement le thème général du discours se définit, comme par exemple une demande d'information relative au renouvellement d'une carte d'identité, ou l'accession à la nationalité française. Localement, seul un sous-ensemble restreint du lexique relatif à ce thème peut être mis en exergue, tout en autorisant éventuellement l'usage de mots moins probables, en cas de rupture de séquence, ou d'énoncés concernant la gestion du canal de communication si l'usager désire préciser une réponse de la machine ("*Pouvez vous répéter, s'il vous plaît*"). Plus précisément, l'usage de certaines structures ou expressions par le locuteur peut être fortement conditionné par un énoncé particulier de la machine comme une question du type :

- *Où habitez vous ?*

où la structure standard de réponse est du type :

- *J'habite à [Nancy]*

ou de façon plus elliptique :

- *à [Nancy]*

Cette grande corrélation entre deux énoncés successifs permet de restreindre de manière importante l'espace d'analyse au niveau du lexique, et bien sûr, de façon plus générale au niveau des structures de la langue.

Les prédictions plus précises au niveau de la structure détaillée d'un énoncé ne concernent plus le module de dialogue, mais plutôt les deux modules d'analyse syntaxico-sémantique et de prosodie. Ce dernier processeur fournit à l'analyseur lexical des frontières possibles de mots, sans être en mesure de préciser la nature exacte de ces mots. Les hypothèses fournies sont donc purement temporelles. Les analyseurs peuvent, quant à eux utiliser des résultats relativement sûrs, obtenus à partir de la reconnaissance d'une partie d'énoncé pour induire des hypothèses sur les éléments restants. Si par exemple le système a reconnu la portion d'énoncé :

J'habite ...

Une hypothèse sémantique va pouvoir être générée concernant un sous-lexique de lieu. Ceci peut se faire dans la pratique grâce à des contraintes sémantiques sur les constructions possibles autour du verbe "*habiter*". Ce phénomène de déclenchement sémantique est d'ailleurs connu en psychologie dont les expériences apportent beaucoup à notre approche [Heyer 85].

La phrase suivante montre un cas plus complexe d'analyse faisant intervenir la représentation du monde qu'a la machine en cours d'analyse :

"*Je suis Marocain et je désirerais renouveler ma carte de séjour.*"

L'analyse du début de l'énoncé permet de qualifier l'élément en mémoire qui représente le locuteur avec l'attribut "Marocain", ce qui restreint l'ensemble des demandes possibles de papiers administratifs susceptibles d'être émises par celui-ci. Ce type d'analyse impose que l'énoncé soit interprété dès le début de sa reconnaissance, ce qui n'est pas encore réalisé au stade actuel de développement du système.

Le problème se pose maintenant d'intégrer ces informations dans le cadre d'une analyse particulière. En effet, la relative généralité du domaine donne une taille importante au lexique non-grammatical, aussi est-il nécessaire de posséder un mécanisme de sélection assez efficace pour que le temps et la qualité de la reconnaissance restent raisonnables.

3. Intégration d'une hypothèse descendante.

Afin de conserver une certaine souplesse à la reconnaissance, nous avons choisi de préserver la disponibilité de tout le lexique à chaque instant, de sorte qu'un apport d'information relatif à un sous-lexique particulier ne fasse que renforcer celui ci. Nous supposons ici qu'une hypothèse lexicale est par définition précise et sélectionne un sous-ensemble vrai de l'ensemble total des mots. Pour cela le lexique est structuré par avance sous forme d'une arborescence multiple correspondant à une organisation syntaxique ou sémantique particulière. Au niveau sémantique nous avons adopté une grammaire de cas étudiée par Guy Deville et Hans Paulussen [Deville 87], qui structure les mots prédicatifs (verbes, noms ou adjectifs) en fonction de primitives définies à partir de traits sémantiques de base. Pour illustrer ceci, nous pouvons donner une représentation partielle des informations relatives à "*signer*" et de certains de ses voisins.

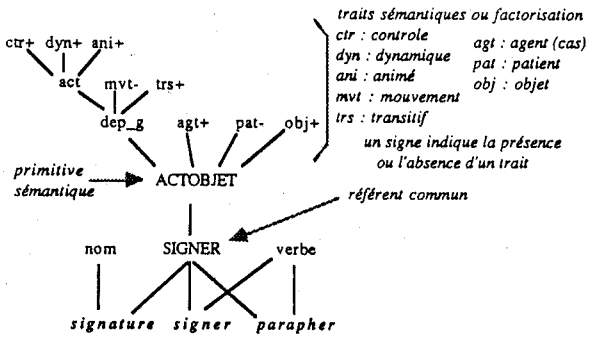


Figure 2.

Au sens défini ci-dessus chaque nœud de cette structure permet de désigner un sous-lexique particulier, et grâce à des opérations ensemblistes élémentaires, il est possible d'exprimer la plupart des contraintes qui nous semblent utiles au niveau du système. Ainsi, une expression du type :

(inter (non verbe) (union ctr+ ben+))

désigne tous les éléments du lexique autres que verbes qui expriment une action contrôlée ou acceptant un bénéficiaire.

Une hypothèse lexicale va donc pouvoir être décrite sous le format suivant :

(*Sous-lexique Score Identifiant Plage-phonétique*)

où *Sous_lexique* est une expression ensembliste signalée ci-dessus, *Score* une évaluation de la qualité de l'hypothèse, *Identifiant* une marque propre à l'émetteur de l'hypothèse et qui permet à celui-ci de la retrouver en cas de validation. Enfin, *Plage-phonétique* est un intervalle au sens large, indiquant la zone temporelle où l'hypothèse est effective.

L'intégration de telles hypothèses se fait relativement simplement grâce à la théorie de Dempster Shafer adaptée ici à un type particulier de distribution de vraisemblance (on peut se référer à [Barnett 83] pour une bonne introduction à cette théorie). Chaque hypothèse accompagnée d'un score va être combinée à la distribution initiale existant au niveau du lexique

en considérant celui-ci comme deux sous-ensembles, à savoir le sous-lexique désigné et son complémentaire.

Avant de détailler le mécanisme de combinaison, il est nécessaire de signaler la provenance de la distribution sur le lexique a priori. Nous sommes partis d'un histogramme de présence des entrées lexicales dans un corpus relatif au même domaine d'application en situation réelle. L'information fréquentielle résultante nous donne une distribution probabiliste vraie au niveau de chaque mot. Cette information peut ensuite être remontée par sommation le long de l'arborescence, pour obtenir ainsi la masse totale d'incertitude que représente un sous-lexique particulier désigné par une hypothèse.

Le problème d'agrégation d'une hypothèse sur la base lexicale peut alors se formaliser ainsi :

Une distribution initiale sur le lexique, qui peut se réduire à une répartition de masse m1 sur les éléments focaux X et $\neg X$ telle que :

$$m1(X) = p \; ; \; m1(\neg X) = 1-p$$

Une hypothèse sur le sous-lexique X avec le score q : (X q), ce qui correspond à une distribution d'incertitude m2 sur les éléments focaux X et Θ (le lexique tout entier) telle que :

$$m2(X) = q \; ; \; m2(\Theta) = 1-q$$

L'application de la règle de combinaison de Dempster-Shafer appliquée aux distributions m1 et m2 schématisée par la figure 3 donne une nouvelle distribution m dont les éléments focaux sont X et $\neg X$ telle que :

$$m(X) = K*p$$
$$m(\neg X) = K*(1-p)*(1-q)$$

où $K = 1/(1-q+p*q)$ est un facteur de normation tel que :

$$m(X) + m(\neg X) = 1$$

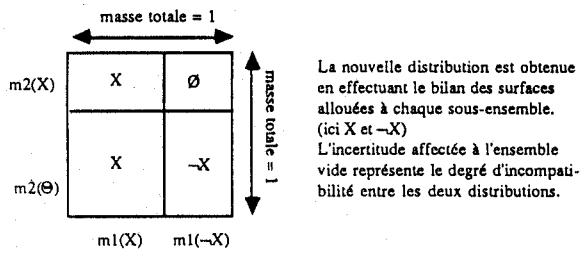


Figure 3.

Au niveau de chaque élément du lexique la distribution finale est obtenue en répartissant m(X) et m($\neg$X) sur X et $\neg$X respectivement, en respectant les proportions correspondant à la distribution initiale m1. L'incertitude obtenue correspond à un certain état de connaissance au niveau du lexique, qui va se traduire de manière effective lors d'une demande de validation de la part des analyseurs syntaxico-sémantiques par exemple. En effet, pour tout sous-lexique dans lequel les analyseurs cherchent à retrouver un mot particulier sur le signal, la distribution relative de certitude pour chaque élément va déterminer l'ordre dans lequel ils seront vérifiés sur le signal de manière à privilégier les plus probables, tout en éliminant plus ou moins les éléments particulièrement improbables, suivant le degré d'avancement de la reconnaissance.

Un tel mode de combinaison peut sembler en première analyse relativement irrévocable, puisqu'il modifie continuellement la base de certitude sur le lexique. Cependant, la règle de Dempster Shafer appliquée à ce type de répartitions bien particulier possède des propriétés intéressantes qui permettent au système de ne pas être contraint par des erreurs antérieures éventuelles.

En dehors des propriétés classiques de commutativité et d'associativité propre au mode de combinaison, qui rendent les effets d'un ensemble d'hypothèses indépendantes de l'ordre de leur application, la structure particulière des distributions de certitude employés ici fait que toute hypothèse (X q) est rétractable, au sens qu'il est possible d'émettre une nouvelle hypothèse ($\neg$X q) pour la réduire complètement. Ainsi, un module particulier peut revenir sur une de ses analyses et de l'extérieur, intervenir sur le lexique pour redonner une cohérence interne à ses informations.

L'exemple fourni en annexe montre l'effet de l'application d'une hypothèse sémantique sur un ensemble de verbes, dans le cadre d'un lexique réduit pour les besoins de l'exemple. Les mots ainsi prédits doivent alors être vérifiés sur le signal pour être ensuite renvoyé au module émetteur. Nous allons voir maintenant comment une telle vérification peut s'opérer et comment des mots peuvent être prédits de manière ascendante, directement à partir du signal.

4. Prédiction et vérification au niveau phonétique.

La phase de prédiction doit être rapide et l'algorithme doit tenir compte des propriétés des résultats du décodage acoustico-phonétique. Le système APHODEX [Fohr 87] détecte très bien les noyaux vocaliques (environ 95% de noyaux bien détectés, et 5% d'insertions), les fricatives et les plosives (détection de l'ordre de 90%, peu d'insertions pour les fricatives et environ 10% pour les plosives). La détection des liquides (l et R) et des nasales (m,n et η) est nettement moins bonne, les efforts de recherche s'étant concentrés sur les trois premières classes de phonèmes dans un premier temps. La détection des semi-voyelles (w et μ) n'est pas encore opérationnelle. De plus, si la reconnaissance des trois grandes classes (voyelles, fricatives et plosives) est bonne, l'étiquetage en terme de phonèmes reste perfectible. Un programme de recherche a priori de mots, ne disposant que des seules informations fournies par le décodage, devra donc disposer d'une représentation des mots adaptée à ce niveau de décodage.

Deux solutions principales s'offrent au concepteur du système:

a) Une première approche consiste à regrouper les phonèmes d'un mot de façon contextuelle en blocs ou macro-phonèmes, qui s'avèrent être moins facilement omis par le décodage. Par exemple, le mot "*extension*" sera codé :

	GV	GP&F	GV	GF	GV
avec	(ɛ)	(kst)	(â)	(s)	(iô)
ou	(ɛ)	(kst)	(â)	(sj)	(ô)

Aucun des macro-phonèmes retenus ne pourra être élidé complètement, une au moins de ses composantes devra être fournie par le décodage, pour qu'un mot puisse être candidat. Une recherche exacte de cette chaîne de macro-phonèmes sera faite sur le treillis fourni par les niveaux bas, après l'avoir transformé en regroupant la liste des phonèmes fournis en macro-phonèmes. Une deuxième étape avec recherche exhaustive de tous les phonèmes donnera un score de reconnaissance plus fin au mot, ou même le rejettera si ce score devient trop faible.

b) Une deuxième méthode de codage et de recherche très simple fournit des résultats comparables, voire meilleurs. Le patron phonétique d'un mot se réduit à la liste des grandes classes qui le composent, parmi plosives, voyelles ou fricatives: par exemple, le patron de "*chapeau*" sera F V P V.

Un affinage consisterait à prendre en compte des traits supplémentaires tels que le voisement des consonnes ou la nasalisation des voyelles, s'ils sont obtenus avec des degrés de confiance suffisants. Cela permettrait d'augmenter le nombre de classes, et rendrait un patron plus sélectif, mais une seule erreur sur ces indices empêche le repérage d'un mot présent. Dans un premier temps, nous avons donc implanté la version la plus simple, avec trois classes seulement. Le patron d'un mot est alors équivalent à une écriture de ce mot sous une forme numérique, en base n, où n est le nombre de classes retenues. ("*chapeau*" peut ainsi être associé à 1202 en base 3 si l'on associe 0 à une plosive, 1 à une fricative et 2 aux voyelles)

Si ces quatre classes sont correctement trouvées par la phase de prétraitement du décodage acoustico-phonétique, repérer chapeau sur le signal revient à rechercher ce nombre sur quatre positions contiguës. Il suffit de promener un masque de largeur quatre sur la segmentation fournie par le prétraitement. Chaque valeur associée au masque courant se déduit de la précédente en ajoutant la contribution entrante à droite puis en ôtant la contribution sortante à gauche, pour un balayage gauche droite. On peut ainsi examiner tous les mots de longueur L et (L + 1)

en un seul passage. Bien sûr un même patron correspond à plusieurs mots, mais reste assez sélectif. Par exemple, le patron F V P V ne correspond qu'à une dizaine de mots parmi 1300 entrées lexicales environ, tels que *"chaque"*, *"chacun"*, *"chapeau"*, *"Jean-Paul"*, *"Jacques"*, *"jusque"*, *"jusqu'à"*, *"jeudi"*... Une deuxième phase de vérification de chacun de ces mots permet d'en éliminer plus ou moins, selon la dissemblance autorisée.

Ainsi, sur la phrase *"Ils ont de beaux chapeaux tyroliens"*, seuls *"chapeaux"* et *"Jean-paul"* sont retenus, ce qui semble très satisfaisant. De récents progrès dans le calcul du fondamental et l'évaluation du critère voisé vs non-voisé d'une plage de signal devrait faciliter notre tâche.

Le nombre de mots significatifs retenus pour un énoncé court tel que celui ci-dessus varie de 0 à 10 maximum, dont environ 50 % de valides. Les mots erronés seront de toute façon retenus en cas de demande sur le signal, leur décodage apparaissant parfois de façon parfaite dans la chaîne phonémique fournie. Le taux de mots détectés en fonction de tous les mots effectivement présents, donc à reconnaître, varie de 30% à plus de 50 % selon les locuteurs. Supposons que la probabilité globale de détecter une classe phonétique parmi les trois que nous avons retenues soit p, et que la probabilité d'insertion d'une de ces classes soit q.

La probabilité de trouver un mot de longueur L est alors $p^{L}(1-q)^{(L-1)}$ (il faut trouver L classes consécutives, et ne rien insérer dans les L - 1 intervalles), par exemple pour L=6 , p=0.95 et q=0.05 on devrait repérer 56% des mots présents et pour L=6 , p=0.9 et q=0.1 on doit détecter environ 31% de mots.

On remarque que cette méthode est très sensible au niveau de décodage, et que des progrès modestes en segmentation du signal peuvent la rendre très performante. Les résultats pratiques et théoriques sont assez proches, et la différence est liée au niveau de la deuxième phase du décodage, un très mauvais étiquetage pouvant conduire au rejet d'un mot correctement détecté, mais trop mal étiqueté.

Dans une deuxième phase, le niveau lexical ne travaille plus qu'à la demande des niveaux supérieurs. Si la phase de prédiction doit éviter de produire des mots erronés, le but à ce moment de la reconnaissance sera inversé : il faudra éviter d'omettre des mots candidats existants. Nous avons déjà précisé qu'un mot candidat avait le format suivant :

(*Sous-lexique Score Identifiant Plage-phonétique*)

Selon les informations de voisinage déjà prise en compte, la plage phonétique est plus au moins stricte : une hypothèse pourra apparaître au milieu de la plage proposée, sans cadrer ni à gauche ni à droite par exemple, être contrainte d'un côté (le mot attendu par les modules d'analyse est un voisin immédiat d'un autre mot déjà validé), ou même couvrir toute la plage. Le score sera aussi dépendant du nombre d'informations déjà prises en compte, ainsi que de la confiance que le système place dans ces informations. Il ne tient encore compte d'aucune information phonétique. Il peut servir à fixer un seuil maximum de pénalité avant rejet pour la comparaison, ce seuil étant proportionnel au score a priori du mot.

La vérification est faite sur le signal avec un algorithme de programmation dynamique classique, qui fait trois hypothèses à chaque pas : mise en correspondance des deux phonèmes, élision du phonème attendu, insertion du phonème présent, chacun de ces choix ajoutant les pénalités induites au total déjà trouvé. Si ce total excède le seuil de rejet, on abandonne le chemin correspondant. Si aucun chemin n'aboutit, le mot sera rejeté, sinon la pénalité trouvée sera combinée au score à priori pour établir le score final du mot. Si la plage était floue, elle est remplacée par la plage de signal qui a fourni le meilleur chemin.

Les pénalités à associer à chacun des choix de progression de l'algorithme sont tirées de trois bases de connaissances : La première indique le coût d'une substitution d'un phonème par un autre. Elle n'est pas complètement symétrique, car certains phonèmes ne sont jamais étiquetés par le système, qui n'en a pas la description (w, μ) et certaines confusions sont orientées. Cette base de connaissances est une expertise a posteriori sur les confusions faites réellement par le système de décodage, et prend en compte à la fois les raisons phonétiques (proximité de deux phonèmes), phonologiques (altération d'un phonème par co-articulation) et imperfections du système (erreurs systématiques ou fréquentes). Une deuxième base de connaissances indique la gravité d'une élision. Elle prend surtout en compte les résultats du système, mais aussi des informations contextuelles : si l'élision d'un noyau vocalique est grave, car c'est une erreur rare, l'absence de détection d'un "i" entre s et ʃ l'est beaucoup moins. La troisième base de connaissance traite les insertions. Elle sera associée à la longueur du phonème "coupable" pour produire une pénalité la plus juste possible. Toutefois, un segment plosif très long, et placé en début de mot, pourra être une pause.

Ce module de vérification est déjà opérationnel pour de petits systèmes développés en parallèle avec le système de dialogue oral homme-machine. Le niveau de décodage est suffisant pour un locuteur entraîné pour permettre la reconnaissance de phrases, avec un vocabulaire limité (une centaine de mots). Le nombre de mots à retenir pour ne pas perdre des mots réellement présents est très variable en fonction des locuteurs, les extrêmes vont de 3 à 10 pour 1. La figure donnée en annexe 2 nous montre des copies d'écran d'un test ayant abouti à une reconnaissance parfaite de l'ordre < Copie core dans essai > pour une application "commandes vocales à un système informatique".

5. Conclusion et perspectives.

Les éléments de réflexions et de techniques présentés ici s'insèrent en réalité dans une analyse plus complète menée sur la mise en place d'un système de dialogue oral homme-machine sur un domaine relativement complexe. Les derniers éléments dont nous disposons au stade actuel de nos travaux montrent qu'il est difficile d'envisager la composante lexicale de manière totalement indépendante du reste du système. Il est préférable de spécifier son statut exact au regard des phases d'analyse structurale, mais surtout vis à vis de l'espace cognitif de la machine, pour espérer "comprendre" effectivement un dialogue.

Le point important est ici de considérer la machine comme un réel interlocuteur quand il s'agit de mettre en œuvre ce type de dialogue. De récentes expériences [Amalberti 88] ont permis d'ailleurs de montrer l'utilité de l'étude du dialogue homme-homme comme référence en la matière. Cette approche permet d'établir des modèles plus complets d'une organisation de haut niveau , où les techniques de base que nous avons montrées s'insèrent de façon satisfaisante.

Références.

[Amalberti 88] R. Amalberti, N. Carbonel, P. Falzon, "Dialogue Homme-Homme, Dialogue Homme-Machine : Un même modèle ?", *Actes du 3eme Colloque International de l'ARC*, Toulouse 9-11 mars 1988.

[Barnett 83] J.A. Barnett, "Computational methods for a mathematical theory of evidence". *Proc. IJCAI 83*, pp.868-875.

[Carbonell 87] N.Carbonell and J.M.Pierrel, "Architecture of knowledge sources in a human-computer oral dialogue system". in: M.M.Taylor, F.Neel and D.G.Bouwhuis, eds., *structure of multimodal dialogues*, North-Holland, Amsterdam, 1987 .

[Deville 87] G.Deville, H.Paulussen et J.M. Pierrel, "Une grammaire de cas comme modèle de représentation sémantique d'énoncés de dialogues oraux homme-machine finalisés", *Proc. AFCET-INRIA 6th Cong. RFIA*, Antibes, nov. 1987.

[Erman 80] Lee D. Erman et Victor R. Lesser, "The Hearsay-II Speech Understanding System: A Tutorial". in : W.A. Lea, *Trends in Speech Recognition*, Prentice-Hall,1980

[Fohr 87] D. Fohr, N. Carbonel, J.P.Haton, "APHODEX, an acoustic-phonetic decoding expert system". Proc. IEEE Workshop on Expert Systems and Pattern Analysis. in *International Journal of Pattern Recognition and Artificial Intelligence*. C.H.Chen ed.,V.1 N.2 1987,pp. 207-222.

[Heyer 85] K. den Heyer, A. Goring et G.L. Dannenbring, "Semantic priming and word repetition , the two effects are additive", *Journal of memory and langage*, v24 1985,pp.699-716.

[Latraverse 87] François Latraverse, *La pragmatique : histoire et critique*, P.Margaga,Bruxelles.

[Pierrel 87] J.M.Pierrel, *Dialogue Oral Homme-Machine*, Hermes, Paris, 1987.

Annexe I : Effet d'une hypothèse.

lexique (la description complète de l'arborescence n'est pas fournie ici)

NOEUD	NOMBRE	SCORE	ANCETRES	
document	5	0,0123	anime-	nom
je	249	0,6118	anime+	pronom
mesurer	2	0,0049	mesure1	verbe
compliquer	1	0,0025	process2	verbe
important	1	0,0025	statut2	adjectif
importance	1	0,0025	statut2	nom
importer	4	0,0098	statut2	verbe
habitation	1	0,0025	statut1	nom
habiter	5	0,0123	statut1	verbe
appartenir	4	0,0098	location1	verbe
remise	1	0,0025	exchprod	nom
remettre	2	0,0049	exchprod	verbe
remercier	2	0,0516	actanime	verbe
changement	1	0,0025	process1	nom
changer	8	0,0197	process1	verbe
couter	1	0,0025	mesure2	verbe
savoir	29	0,0712	extension	verbe
garder	2	0,0049	location2	verbe
vote	1	0,0025	atrans	nom
obtention	2	0,0049	echobt	nom
obtenir	24	0,0590	echobt	verbe
donner	21	0,0516	exchprod	verbe
signer	1	0,0025	actobjet	verbe
apporter	2	0,0049	mvmt2	verbe
aller	18	0,0442	mvmt1	verbe

La sélection d'un sous-lexique est seuillée à **niveau = 0,02**

l'hypothèse : **(verbe ())** {est une demande de validation d'un verbe sur le signal.}

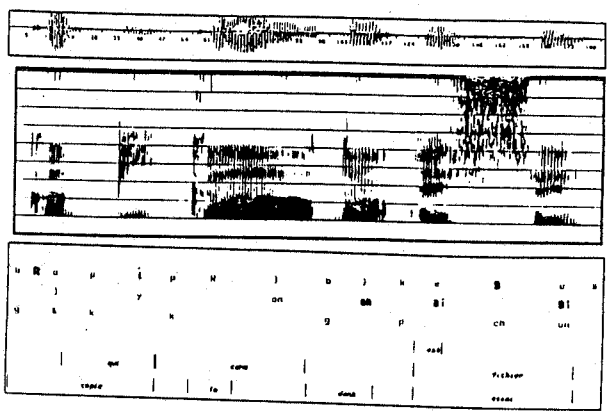
savoir -> score = 0,0712 {premier élément à être vérifié}
obtenir -> score = 0,0590 {deuxième...}
donner -> score = 0,0516
remercier -> score = 0,0516
aller -> score = 0,0442

l'hypothèse : **((inter mvt+ ctr+) 0,9)** {renforce l'ensemble des verbes exprimant un mouvement et un contrôle.}

l'hypothèse : **(verbe ())** {donne alors :}

aller -> score = 0,3066 {"aller" à été mis en évidence
savoir -> score = 0,0494 par l'hypothèse précédente et
obtenir -> score = 0,0409 sera donc vérifié en priorité.}
remercier -> score = 0,0358
donner -> score = 0,0358
apporter -> score = 0,0341

Annexe II : Exemple de reconnaissance de la phrase "*copie core dans essai*".



- Signal temporel
- Spectrogramme
- Treillis phonétique fourni par APHON.
- Mots reconnus en phase ascendante (les barres verticales indiquent les limites temporelles de ceux-ci).

META-STRATEGIE EN RECONNAISSANCE DANS LE PROJET "DIRA"

J. Caelen

ICP/INPG - UA CNRS
46, AV. F. VIALLET
38031 GRENOBLE CEDEX

ABSTRACT

This paper describes an architecture of multi-expert system for the continuous speech recognition The DIRA system (Integrated Dialogue for Advanced Robot) is organized around a blackboard and supervisor which manages the blackboard and plans the tasks of various experts. The experts are: acoustic-phonetic decoder, lexical analyser, syntactic-semantic decoder and dialogue manager. The main advantage of this architecture relies upon the role devoted to supervisor which is able to reasoning on knowledges stored in the blackboard. The type of reasoning is opportunistic, i.e. the supervisor (a) examines the evolution of events during the time, (b) builts a decision tree according to goals to reach and hypothesis to prove, and (c) performs a confidence values on the branchs of the decision tree. Then, the plan consists in selecting nodes which have the best values in the decision tree. Therefore, the supervisor is a planner which has an knowledge base enclosing the meta-strategy written in terms of production rules.

The paper provides an example in order to show the system operating cycle.

INTRODUCTION

Le concept de système multi-experts s'est affirmé depuis les années 80, à la suite du développement de nouvelles architectures issues du tableau noir (blackboard) de Hearsay II [Erman, 80]. Les problèmes liés à ces architectures constituent maintenant une matière abondante en Intelligence Artificielle Distribuée [Ferber, 88].

Dans le domaine de la reconnaissance de la parole, la plupart des systèmes orientés-connaissance actuels, sont constitués d'une communauté d'experts qui échangent leurs informations sous le mode vérification/proposition. Ces systèmes se distinguent entre eux surtout par la stratégie mise en oeuvre plus que par la nature des sources de connaissances utilisées: en effet ils manipulent presque tous des informations sur les niveaux acoustique, phonétique, lexical, prosodique, syntaxique et sémantique. Dans ces systèmes, les experts fonctionnent de manière centralisée (C) ou distribuée (D) --sous la direction d'un superviseur-- ou de manière autonome (A). Dans le premier cas (C) la stratégie est déterminée par un seul processus centralisé qui active les experts dans un ordre pré-établi, dans le second cas (D) les stratégies sont distribuées entre les experts (acteurs) qui doivent coopérer entre eux, souvent sous le contrôle d'un superviseur, et dans le troisième cas (A) chaque expert est un agent autonome qui se pose et résoud ses propres problèmes au moment opportun en n'ayant à chaque instant qu'une connaissance partielle sur son environnement (il ne connait pas les décisions prises par les autres experts au même moment). Dans les véritables systèmes "tableau noir" les agents sont guidés par les données, ils obéissent donc aux règles d'autonomie décrites ci-dessus. En théorie c'est l'organisation la plus attrayante: l'autonomie permet en effet une modularité aussi bien au niveau du calcul d'une solution qu'à celui de la décision. En réalité dans sa tâche de planification locale l'expert doit être capable de remettre en question ses propres buts, d'affiner, voire modifier ses prédictions: or il n'a peut être pas au moment opportun tous les éléments ni la compétence pour le faire --par exemple, certaines décisions de nature phonétique dépendent de décisions syntaxiques via la prosodie. Une harmonisation entre les deux approches (D) et (A) semble donc plus réaliste, en ne conservant de chaque méthode que ses avantages intrinsèques, c'est-à-dire:

(a) en gardant la technique du tableau noir pour son rôle de mémoire commune à long terme, d'aiguilleur de données (muni de sémaphores) et de régulateur de flux,

(b) en faisant jouer au superviseur un rôle "d'expert en stratégies" au sein des autres experts. Cela revient à considérer que la stratégie de nature dynamique se trouve décrite dans la base de règles de ce superviseur qui fonctionne alors comme un planificateur ayant à organiser le travail de ses experts (qui trouvent toutes les ressources dont ils ont besoin dans le tableau noir) au moment opportun. Chaque expert garde ainsi une part d'autonomie, mais perd toute responsabilité dans le choix de ses buts ainsi que dans la décision globale.

Dans le domaine de la parole le choix d'une stratégie de reconnaissance est guidé par la taille du vocabulaire, l'étendue de la syntaxe, les contraintes pragmatiques, etc., et dépend donc de l'application: ici, la base de règles du superviseur peut décrire toutes ces situations et permet donc de s'affranchir d'une architecture liée à l'application. Dans le projet DIRA (Dialogue oral Interactif pour le pilotage d'un Robot Avancé) il doit, précisément, y avoir plusieurs

stratégies possibles selon les contraintes induites par le contexte de l'action et les niveaux de langage [Janot-Giorgetti, 87].

Parmi les travaux les plus récents en la matière on peut citer ceux de Y.F Gong et J.P. Haton [Gong, 88] et ceux de R. De Mori [De Mori, 85] [De Mori, 87] et plus généralement ceux des domaines connexes de l'IA [Hautin, 86], [Hayes-Roth, 85], [Hewitt, 81], [Konolidge, 80], etc. Dans [Gong, 88] la "société d'experts" est structurée en plusieurs couches (fig. 1) et les responsabilités des experts sont hiérarchisées. Le blackboard est divisé en mémoires indépendantes: les conférences. Les experts "discutent" au sein d'une même conférence et transmettent leurs résultats au directeur de leur association qui communique avec les autres directeurs à l'aide de messages. Les experts sont vus tantôt comme des concurrents tantôt comme des partenaires dont le directeur arbitre les débats. L'administrateur gère l'ensemble des associations. Le système a plutôt une planification décentralisée.

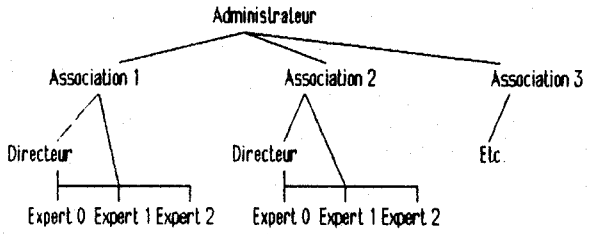


Fig. 1: Structuration de la société d'experts dans [Y.F. Gong, 88]

Pour De Mori [De Mori, 87] les actions des experts sont planifiées à l'aide de "réseaux d'actions" dans lesquels les actions à exécuter sont sélectées par une procédure de décision contrôlée par des règles obtenues à l'issue d'une passe d'apprentissage. Dans cette version les experts sont directement liés à l'administrateur et la planification est plutôt de type centralisé mais la stratégie est figée à l'issue de l'apprentissage.

L'architecture proposée ici se rapproche plutôt de cette dernière mais s'en distingue cependant: dans le projet DIRA le superviseur est un générateur de plans dont la base de connaissances contient un ensemble de règles décrivant la méta-stratégie; cette méta-stratégie est formulée à partir du savoir-faire d'un expert en RAP (Reconnaissance Automatique de la Parole). Les autres experts du système utilisent des connaissances représentées sous forme de réseaux: réseaux phonétiques, réseau lexical, réseaux à noeuds procéduraux syntactico-sémantique et réseaux sémantiques. Le système se présente donc comme un assemblage d'automates gouvernés par un planificateur.

On propose dans cet article de décrire en détail le rôle et l'organisation du planificateur.

1. LE SYSTEME DIRA

1.1. Architecture générale

C'est un système multi-experts distribué organisé autour d'une architecture de blackboard (Fig. 2).

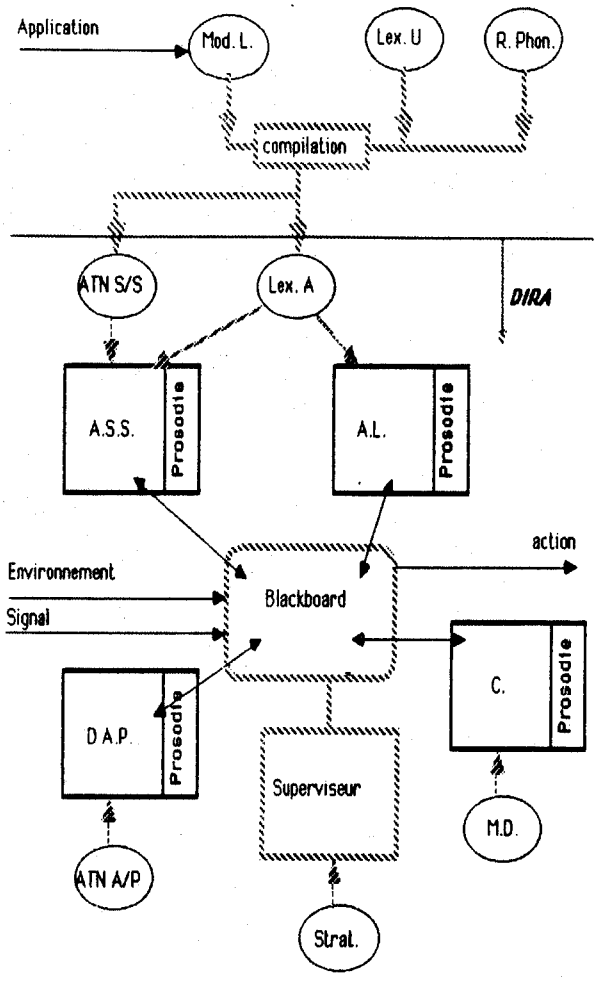


Fig 2: Architecture du système DIRA-RAP

Les experts, au nombre de quatre, communiquent leurs informations à travers le blackboard organisé en couches. Le superviseur est un cinquième expert qui gère le flux des données dans ce blackboard (à l'aide de sémaphores), planifie la stratégie et fixe les points de rendez-vous (synchronisation sur les ilôts de confiance). Les experts restent autonomes dans l'exécution de leur propre tâche dès que celle-ci est définie et activable.

Ces experts 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 [Caelen, 88],
- 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. Les tâches prosodiques sont distribuées à ces 4 experts selon leur spécificité --par exemple la microprosodie est traitée au niveau du DAP tandis que les marqueurs de syllabes et de mots sont traités au niveau de l'analyseur lexical.

Chaque expert dispose de ses propres sources de connaissances qui sont:

- un ATN (Augmented Transition Network) syntactico-sémantique (ATN-S/S) compilé,
- le réseau lexical de l'application (LEX-A), également compilé à partir d'un lexique universel LEX-U (type BDLEX) et de règles phonologiques R-PHON, --ceci est possible dans la mesure où le vocabulaire est limité (moins de 1000 mots),
- les règles acoustico-phonétiques mises sous forme de Réseaux Phonétiques RP, indépendantes de l'application et multilocuteur pour la partie "macro-structure" [Caelen, 88],
- le modèle de langage MD,
- les données sur l'environnement qui transitent par le blackboard (identité du locuteur si elle est connue, conditions d'acquisition, etc.).

1.2. Le modèle de communication

Les experts communiquent entre eux par le blackboard en partageant l'information qui y est contenue. Le dispositif de contrôle du blackboard est pris en charge par le superviseur lui-même. Par contre les experts communiquent avec le superviseur sous forme d'envoi de messages. Ces messages sont de deux sortes: (a) du superviseur vers l'expert X pour lui communiquer les tâches à exécuter, le mode et les contraintes d'action, (b) de l'expert concerné au superviseur pour lui communiquer les variables de contrôle de fin d'exécution. Chaque expert dispose bien sûr d'une mémoire de travail à court terme.

2. FONCTIONNEMENT GENERAL

2.1. Le superviseur gère les données dans le blackboard, fixe les points de rendez-vous, coordonne les échanges d'informations avec l'extérieur et planifie les tâches des experts. Chaque expert peut fonctionner sous un mode donné --proposition, prédiction, vérification-- et sur des tranches de temps distinctes (fenêtres) --le passé, le présent, le futur. Nous noterons:

X(m,f) le message envoyé à l'expert X pour lui spécifier d'exécuter une tâche sur le mode m et sur la fenêtre f avec:

X nom de l'expert
m=[p,v] p=proposition ou prédiction, v=vérification
f=[-,=,+] "-"=passé, "="=présent, "+"=futur

Chaque expert dispose d'une mémoire locale de travail découpée selon ces fenêtres qui sont mises à jour au fur et à mesure de la progression dans le temps (Fig. 3).

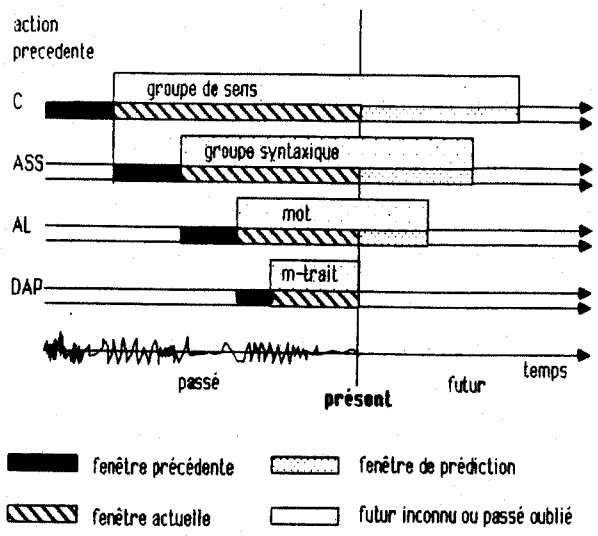


Fig. 3: Diagramme des temps et tailles relatives des fenêtres

Pour l'expert X l'exécution d'une tâche se décompose en:

1. lecture du blackboard: initialisation de ses variables locales en fonction de l'environnement, sélection des problèmes à traiter et de leur degré de profondeur, lecture des hypothèses en cours et des contraintes locales,
2. résolution du(es) problème(s) posé(s),
3. écriture dans le blackboard des résultats (hypothèses et scores), des variables modifiées,
4. désactivation (avec envoi de message), et retour au superviseur.

Pour l'ensemble des experts le fonctionnement apparait comme asynchrone avec progression de gauche à droite et resynchronisation par points de rendez-vous. Cette resynchronisation s'effectue dans les cas suivants:

- lorsqu'il y a une contradiction locale insoluble entre les experts: on tente de rechercher l'information la plus proche dans le temps et la plus sûre pour réactiver le processus de reconnaissance (une pause par exemple),
- lorsqu'il y a une information très sûre provenant d'un expert: on aligne les autres experts sur cette information,
- lorsqu'un problème complet est résolu et vérifié: on désactive la fenêtre de traitement correspondant à ce problème et on réinitialise le processus de reconnaissance.

2.2. La planification des tâches est de nature opportuniste. Elle se fait de manière classique, en respectant les grandes étapes suivantes:

a- analyse de la situation à chaque instant: par raisonnement non monotone après évaluation des réponses aux trois questions suivantes:

Quel est l'expert qui a fait évoluer le plus la situation ?

Où en est-on dans la phrase ?

Quelles sont les informations les plus sûres ?

les réponses à ces questions permettent de développer l'arborescence des actions possibles

b- ordonnancement des actions après pondération des solutions selon les critères contenus dans les trois questions suivantes:

Quelles sont les opportunités ?
Quel est le risque encouru ?
Quel est le but attendu ?

c- exécution du plan en partant des actions les plus sûres

d- organisation des hypothèses reçues, filtrage.

De manière plus précise, l'algorithme de planification est le suivant:

```
Début
   Initialisation du blackboard
   C(p+) ET C/P(p+) (produit les hypothèses de départ)
   Problème <- non_résolu
   Tantque Problème=non_résolu Faire
      PourChaque Hypothèse Faire
         Analyse de la situation (règles AS)
         Développement de l'arborescence des actions
            possibles
         Analyse des choix (règles AO)
         Ordonnancement de l'arborescence des
actions
         Plan <- non_résolu
         Tantque Plan=non_résolu Faire
            Activation expert X, mode m, fenêtre f
            Si résultat conforme au plan
               alors écriture nouvelles hypothèses
                  dans le blackboard
               sinon abandon tâches successeurs
                  recherche chemin suivant
                  si pas de chemin alors effacement
                     hyp. courante
                  Finsi
            Finsi
         Fintantque
      Finpour
      Analyse des hypothèses nouvelles (règles AH)
      Filtrage et hiérarchisation des hypothèses (pose
éventuelle de points de rendez-vous de synchronisation
sur les ilôts de confiance)
      si 1ère Hypothèse='fin_parole'
         alors problème <- résolu; C(v=)
         sinon si 1ère Hypothèse='néant'
               alors situation d'échec: pose d'un
               point de rendez-vous de redémarrage
               Finsi
      Finsi
   Fintantque
Fin
```

2.3. La base de règles du planificateur est découpée en trois parties qui contiennent respectivement:

- les règles AS qui permettent une analyse de la situation,
- les règles AO qui permettent une analyse des opportunités,
- les règles AH qui permettent une analyse des hypothèses.

Il est evident que tout l'intérêt du système de planification réside dans ces règles, dont la forme déclarative autorise facilement une modification ou une adaptation du système au domaine d'application envisagé. L'inconvénient majeur est d'ailleurs comme dans tout système expert, de disposer des connaissances idoines ou d'un expert compétent (ici il s'agit du savoir-faire d'un expert en reconnaissance).

Les exemples de règles décrites ci-après donneront au lecteur une idée des possibilités offertes par une telle méthode.

Exemple: production de plans en fonction de la localisation dans la phrase.

Le tableau I précise les problèmes à traiter selon la nature des unités de décision et la localisation courante du planificateur en reconnaissance:

Localisa unité	avant	début	dans	fin	entre	après
phrase	détection début			vérif Synt/Sém	référent struct. vérif	vérif. textuelle
syntagme			prosodie			pause
groupe			prédict mot suiv.	vérif.synt.	prosodie	
mot		prédict. lex.		vérif. lex.	liaison	prosodie
syllabe			noyau	durée		

Tableau I: nature des problèmes en fonction de la localisation du système de reconnaissance au cours du traitement d'une phrase.

Ce tableau se traduit simplement sous forme de règles par:

AS-R1: Stratégie de début de phrase
si Avant(phrase)
alors but <- détecter(début_parole)

AS-R13: Stratégie entre deux mots m_1 et m_2
si Entre(mots) $\wedge$ non(pause) $\wedge$ dernier_phonème(m_1)='consonne' $\wedge$ (prem_phonème(m_2)='voyelle' V prem_phonème(m_2)='semi-voyelle')
alors but <- traiter(liaison)

etc.

Les buts sont développés en sous-buts à l'aide des connaissances rangées dans la base des règles d'opportunité de la manière suivante:

AO-R1: Détection du début de parole
si Hypothèse='pause' $\wedge$ but=détecter(début_parole)
alors s/but <- DAP(p=)
résultat_attendu='pause' V 'macro-trait'

Remarques:
(a) Cette règle n'autorise pas de réponse "je ne sais pas".
(b) Pour trouver effectivement le début de parole le superviseur doit itérer sur cette règle.

(c) L'expert a indiqué explicitement qu'il est opportun de détecter ce début de parole en appelant le DAP en proposition sur la fenêtre courante.

AO-R13: Traitement de la liaison pluriel
si but= traiter(liaison) ∧ Hypothèse='pluriel'
alors s/but <- vérifier consonne(/z/), score<-4
V s/but <- filtrer phonème, score<-1

sachant que vérifier consonne(/z/) sur le signal peut se faire par DAP(v=) avec un score minimum exigé donné après avoir vérifié dans le lexique qu'une liaison est autorisée pour ce mot. La présence des scores indique ici qu'il vaut mieux vérifier la consonne de liaison (score stratégique=4) mais qu'on ne filtre pas le phonème courant (score stratégique=1).

Une fois l'arborescence des actions construite (par application de toutes les règles) on ordonnance les branches selon le score qu'elles ont obtenu.

3. COMPTE-RENDU D'EXPERTISE SUR UN EXEMPLE

Nous considérons ci-après un exemple pour lequel nous donnons un compte-rendu exprimé par l'expert en fonction des réponses du système, avec les conventions d'écriture suivantes:

Conventions d'écriture et notations:

Sup:	superviseur
Sup:->X(pf)	instruction d'activation de l'expert X en proposition ou prédiction sur la fenêtre f
Sup:->X(vf)	instruction d'activation de l'expert X en vérification sur la fenêtre f avec
f	'-' passé, '=' présent, '+' futur
X:<- H ,s≥n	vérification des hypothèses H par l'expert X (score minimum exigé = n)
X:-> H, s=n	production des hypothèses H par l'expert X (score de prédiction obtenu = n)
C	expert en compréhension
ASS	expert en syntaxe sémantique
AL	analyseur lexical
DAP	expert en décodage acoustico-phonétique
X/P	composante prosodique de l'expert X
GP	générateur de plan du système
s	score ≈ [0,5]
H	hypothèses assorties de contraintes mises sous forme de prédicats
!	commentaires en italique

Ces conventions permettent d'expliciter un langage externe de planification aisément formulable par l'expert. Considérons maintenant un langage de commande d'un robot mobile à consonance naturelle, qui accepte la phrase: "avancer dans le couloir de gauche". Supposons que cette phrase soit à reconnaitre.

Au départ la base de faits contient des données en provenance du robot: il est arrêté au milieu d'une salle, il attend un ordre de commande et (probablement) de mouvement. Chaque expert a pu produire un certain nombre d'hypothèses qui sont:

C: -> sémantique=action
C: -> pragmatique=mouvement.
C: -> syntaxe=G(c) grammaire des ordres de commande
ASS: -> début de phrase
AL: -> début de mot
ASS/P: -> groupe bref et pente globale(Fo)='-'
DAP: -> pause

La trace du raisonnement suivi par l'expert à chaque instant est la suivante:

! pour détecter le début parole on choisit le DAP en mode proposition
Sup:->DAP(p=)
DAP:-> pause, s=5 *! c'est un silence de début, aucune information nouvelle, on réitère l'appel du DAP*
Sup:->DAP(p=)
DAP:-> voyelle+ouvert, s=4 *! (correspond à /a/)*
! cette hypothèse semble avoir un score suffisant, on raisonne sur les opportunités du plan suivant:
! plan: connaitre les mots possibles ou décoder le prochain phonème.
Sup:->ASS(p+) ∧ AL(p+)
ASS:<-1er mot de la phrase *! cette hypothèse est lue dans le blackboard (voir état initial)*
ASS:->verbe de mouvement,s=5
AL:<- verbe de mouvement ∧ premier phonème = voyelle ouverte
AL: -> [avancer, attendre, arrêter, aller, accélérer],s=5
! la liste est retenue car elle est assez réduite
! plan conforme: on détruit le noeud (DAP), on interdit donc tout retour arrière (risque pris par l'expert)
! plan: tous les mots de la liste commencent par /a/: on peut activer le DAP(v-) pour confirmer cette voyelle (retour arrière) ou progresser vers la droite en vérification ou en proposition.
Sup:->DAP(v=),s≥2
DAP<-(/v/V/t/V/r/V/l/V/k/) ∧ (contexte_subséquent=(/ã/V/e/Vgémination)) ∧ (contexte_précédent=/a/)
DAP: -> consonne+vocalique, s=2 *! correspond à /v/*
! hypothèse retenue bien que le score ne soit pas excellent
! plan: prédire une liste de mots ou activer DAP pour plus d'informations.
Sup:->AL(p+)
AL:<- liste ∧ 2ème phonème=consonne+vocalique
AL: -> [avancer, arrêter, aller],s=5
Sup:->DAP(v=) *! suite de la stratégie précédente*
DAP<- ((/ã/ ∧ cont_subs=/s/)V(e)) ∧ (cont_préc=consonne+vocalique)
DAP: -> voyelle, score=5 *! correspond à /ã/*
! la réponse n'est pas concluante pour la discrimination phonétique car le macro-trait obtenu est peu précis mais autorise la poursuite de la recherche. On peut tester l'hypothèse "aller" au niveau prosodique en cherchant l'existence d'un marqueur de fin de mot, d'où:
Sup:->AL/P(v=)
AL/P:<-fin de verbe dans syntagme verbal
AL/P:->? *! la prosodie ne peut confirmer la fin de mot il faut garder toutes les hypothèses qui sont: h1=avan(cer), h2=arrê(ter), h3=alle(r-) c'est-à-dire 3 hps lexicales, 1er mot de la commande=verbe de mouvement, liaison possible 'r- pour h3, score faible pour la consonne vocalique*

! plan: il faut continuer à progresser car un retour arrière buterait au niveau phonétique sur la reconnaissance de la consonne. Il faut d'abord prédire les mots pouvant succéder au verbe "aller" donc:
Sup:->ASS(p+) ∧ AL(p+)
ASS:<-2ème mot après verbe de mouvement
ASS:-> (adv. vitesse V prép. lieu)
AL: <- (adv. vitesse V prép. lieu) ∧ (mot_précédent="aller")
AL: -> lentement, rapidement, vers, dans devant, derrière, jusqu'à, tant que, à
! la liste commence à s'allonger, on peut faire une analyse des catégories phonétiques à tester ou activer le DAP(p=). La combinatoire des hypothèses phonétiques donne: h_1 : /s/, h_2 : /t/, h_3 : /l,r,v,d,t,'r-/, ce qui est trop large pour le DAP(v=). On choisit donc DAP(p=).
Sup:->DAP(p=)
DAP<- (consonne V occlusive V fricative V liaison 'r)
DAP: -> fricative+sourde+stridente,s=5
! le score est excellent on pose un point de rendez-vous, ce qui revient pour les autres niveaux à résorber toutes les hypothèses secondaires
Sup:->PRV(AP) *! la pose de ce point de rendez-vous phonétique sollicite les autres modules pour qu'ils filtrent leurs hypothèses. Si la liste restante est vide on reprend les analyses depuis le dernier PRV en élargissant la recherche ou on se trouve en impasse*
AL: -> [avancer] *! h_2 et h_3 sont effacées*
etc.

On obtient finalement le réseau d'action de la fig. 4 associé à la séquence des hypothèses de la fig. 5.

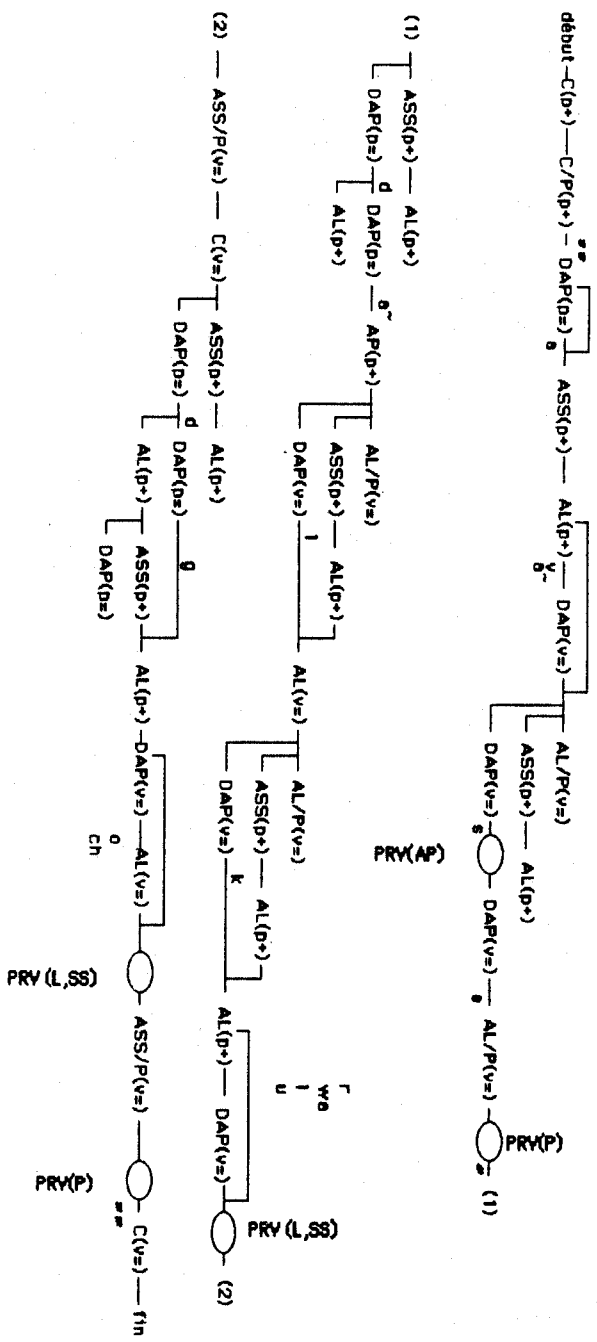


Fig. 4: Réseau d'action pour la reconnaissance de la phrase "avancer dans le couloir de gauche".

CONCLUSION

Ce système présente l'avantage, grâce à la prise en compte de la méta-stratégie, de gérer dynamiquement les tâches des experts: il peut donc être adapté à des univers d'utilisation variés. Le concept multi-expert le rend très modulaire et le concept de superviseur, au sens de planificateur comme décrit ci-dessus, le rend indépendant des experts eux-mêmes: cette architecture est tout à fait applicable à un système de reconnaissance de mots isolés qui aurait un décodeur markovien. C'est aussi un outil de laboratoire qui permet de valider certaines stratégies de reconnaissance et d'effectuer des traces ou des essais divers par simple modification de la base de règles stratégiques. Dans une perspective de dialogue avec un robot il offre enfin l'avantage de présenter une unité de conception proche de celle des plans de navigation utilisés en robotique [Crowley, 87]. Quant aux experts, ils se présentent tous comme des automates dont les bases de connaissances sont des réseaux d'états finis.

DIRA, dans son ensemble, est en cours d'implantation en Prolog et en C.

BIBLIOGRAPHIE

[Allen, 84] J.F. Allen, 1984
Towards a General Theory of Action and Time;. AI, 23, pp. 123-154.

[Brownston, 85] L. Brownston, R. Farrel, E. Kant and N. Martin, 1985
Programming Experts Systems in OPS5. Readding, MA: Addison-Wesley, 1985

[Caelen, 88] J. Caelen et H. Tattegrain
Le décodeur acoustico-phonétique dans le projet DIRA. 17èmes JEP, SFA, Nancy.

[Cerri, 87] S A Cerri, P. Landini and M. Leoncini, 1987
Cooperative agents for knowledge-based information systems. AII, Vol. 1 n°1, pp. 1-24

[Crowley, 87] J. Crowley, 1987
Coordination of Action and Perception in a Surveillance Robot. IJCAI, Milan.

[De Mori, 85] R. De Mori, P. Laface and Y. Mong, 1985
Parallel Algorithms for Syllable Recognition in Continuous Speech. IEEE-PAMI, Vol. 7 n°1, pp. 56-69, 1985

[De Mori, 87] R. De Mori, L. Lam and M. Gilloux, 1987
Learning and Plan Refinement in a Knowledge-Based System for Automatic Speech Recognition. IEEE-PAMI, Vol. 9 n°2, pp. 289-305, march 1987

[Erman, 80] L.D. Erman, F. Hayes-Roth, V.R. Lesser and D.R. Reddy, 1980
The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty. Comput. Surv., Vol. 12, pp. 213-253, 1980

[Ferber, 87] J. Ferber, 1987
Des objets aux agents. Actes du 6èmes congrès RFIA, AFCET.

[Ferber, 88] J. Ferber et M. Ghallab, 1988
Problématiques des univers multi-agents intelligents. PRC/IA actes des journées nationales, Teknea éditeur, pp. 295-320.

[Fikes, 85] R.E. Fikes and T. Kehler, 1985
The role of frame based representation in reasoning. Com. ACM, 28(9), pp. 904-920

[Gong, 88] Y.F Gong and J.P. Haton, 1988
A Specialist Society for Continuous Speech Understanding. ICASSP, New York, April 1988.

[Haton, 84] J.P. Haton, 1984
Present Issues in Continuous Speech: Recognition and Understanding NATO Advanced Study Institute, Bonas, Juillet 1984.

[Hautin, 86] F. Hautin et A. Vailly, 1986
La coopération entre systèmes experts. Journées nationales PRC/IA, Cepadues éditions.

[Hewitt, 81] C. Hewitt et W.A. Kornfeld, 1981
The scientific community metaphor. IEEE Trans. on Man, Systems and Cybernetics, Vol. CMC 11 (1).

[Janot-Giorgetti, 88] M.T. Janot-Giorgetti, J. Caelen, E. Bauer, 1988
DIRA: Integrated dialogue for the command of advanced robot in nuclear surrounding. Actes de la conférence internationale sur l'interface homme/machine dans l'industrie nucléaire. Tokyo, 15-19 février 1988.

[Konolidge, 80] K. Konolidge et N.J. Nilsson, 1980
Multi-agent planning systems. Proc. AAAI-80, Stanford, pp. 138-142.

[Lea, 80] W.A. Lea, 1980
Trends in Speech Recognition, Prentice-Hall, 1980.

[Minsky, 75] M. Minsky, 1975
A framework for representing knowledge. In The Psychology of Computer Vision. P. Winston ed., New York: McGraw Hill, 1975.

[Wilkins, 84] D.E. Wilkins, 1984
Domain-independant Planning: representation and plan generation. Aritificial Intelligence, Vol. 22 n° 3, pp. 269-302, April 1984

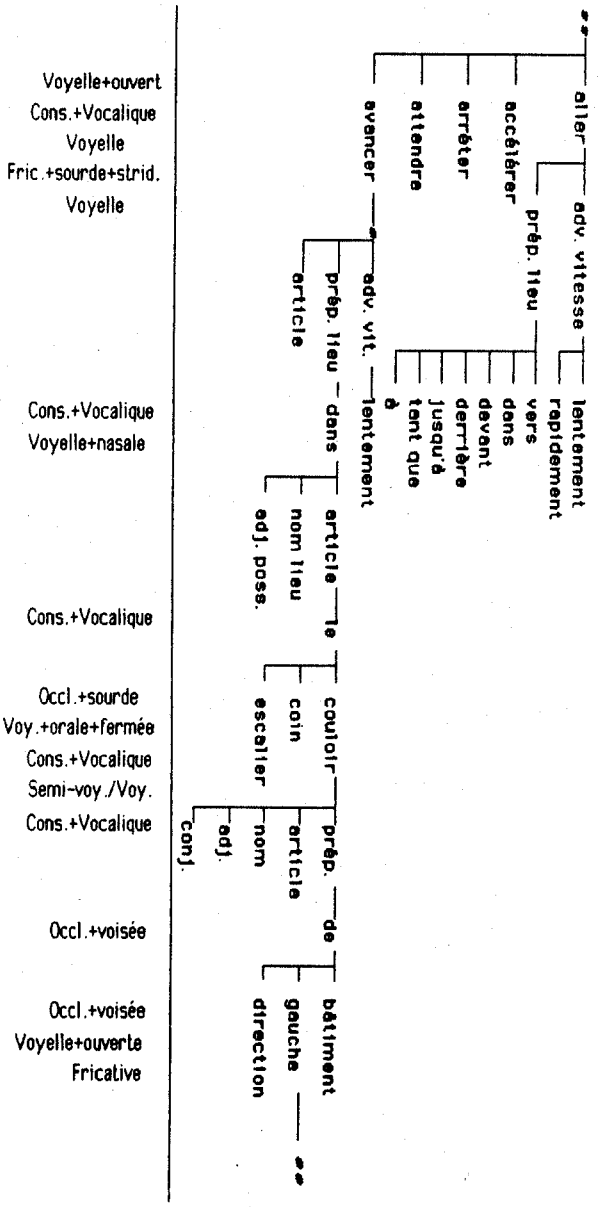


Fig. 5: Contenu du blackboard pour l'exemple considéré

Production articulatoire

LES TROIS DEGRES DE LABIALISATION DES VOYELLES TENUES EN FRANCAIS PREMIERS RESULTATS

Jean-Pierre Zerling

Institut de Phonétique de Strasbourg
22 rue Descartes - 67084 Strasbourg Cedex

ABSTRACT

We study here frontal lip opening parameters for the 14 steady-state French vowels prononced by 16 native speakers: width(A), height(B) and area(S), as well as AxB and A/B. A few well-known common facts are first confirmed.

Interestingly, our data also reveal that although French vocalic labiality is described in terms of a binary phonological feature (+/-round), articulatory processes show 3 degrees of labialization, i.e., non-rounded (-lab), rounded (+lab) and hyper-rounded (++lab). Splitting of the traditional labialized category does not seem to have been considered before.

The articulatory behaviour used for each class is described. Further, the comparison of labial data to x-ray tongue data shows that, for 2 series of vowels a single tongue-shape may be associated with 3 degrees of labiality. This seems to be confirmed by articulatory modelization.

INTRODUCTION

L'importance du rôle joué par les lèvres lors de la phonation a été maintes fois démontrée. Pour le français, des études ont été réalisées dans des domaines différents: phonétique articulatoire, électro-myographie, coarticulation, etc.: ABRY & al(79a,79b), BELL-BERTI & HARRIS(76), BENGUEREL & COWAN(74), DESCOUT & al(78). En 1980, s'est tenu à Lannion un Séminaire International Labialité (Actes,80), et la même année, un ouvrage collectif, Labialité et phonétique, a été publié par l'Institut de Phonétique de Grenoble (ABRY & al,80). Depuis, des études n'ont pas cessé d'être menées, qu'elles soient du domaine descriptif, ZERLING(79b,84a,84b), de l'E.M.G., GENTIL(80), BONNOT & al(83), ou de la modelisation articulatoire, MAEDA(79), ABRY & BOE(86), MAJID & al(87).

Il ressort de ces travaux que derrière l'apparente simplicité du trait binaire ecarté/arrondi couramment adopté en phonétique, semble se cacher en fait une utilisation plus complexe des lèvres, notamment sous l'influence du contexte consonantique. DESCOUT & al(78) distinguaient déjà 3 degrés de labialité phonétique:

- non-labialisé: voyelles non-arrondies /i,e/
- très labialisé: voyelles arrondies /y,ø/
- moyennement labialisé: voyelles non-arrondies /i̯,e̯/ labialisées par les consonnes /ʃ/ et /ʒ/.

Une influence analogue de la consonne bilabiale /b/ sur les voyelles non-arrondies est également observée par ZERLING(80). *

---

* Notons au passage que le problème inverse de l'influence des voyelles labiales sur les consonnes sera abordé dans un travail ultérieur, ainsi que l'étude de la labialité de /ʃ/ et /ʒ/

Ces observations sont à rapprocher de celles que nous avons faites à propos de la délabialisation possible de la voyelle nasale /œ̃/, ZERLING(84b), où nous montrions qu'aux 2 degrés phonologiques habituels: (-lab) pour /ɛ̃/ et (+lab) pour /œ̃/, s'ajoutait quelquefois un degré intermédiaire (0lab) pour /œ̃/ et même parfois pour /ɛ̃/.

Si ces phénomènes sont relativement stables et systématiques, il convient de remarquer qu'ils correspondent à la réalisation de variantes phonétiques non pertinentes phonologiquement.

Nous voudrions maintenant attirer l'attention sur des phénomènes différents qui, contrairement aux précédents, semblent relever de la phonologie et qui, à notre connaissance, n'ont pas fait l'objet d'études antérieures.

Il y a quelques années (ZERLING,84c), nous avons souligné le rôle décisif de la labialisation pour certaines articulations, notamment lorsqu'il s'agit de distinguer des voyelles utilisant une position linguale basse: /a,ɔ,ɑ̃,ɔ̃/. On doit admettre que la position de la langue pour ces 4 voyelles est voisine et que leur différenciation acoustique ne peut provenir que d'autres caractéristiques articulatoires telles que la nasalité et la labialité:

- la nasalité distingue /a,ɔ/ de /ɑ̃,ɔ̃/.
- la labialité distingue /a/ de /ɔ/ et /ɑ̃/ de /ɔ̃/.

Mais, lorsqu'on sait que /ɔ/ et /ɑ̃/ ont des degrés de labialité très proches, force est d'admettre que /ɔ̃/ doit être labialisé encore plus fortement pour se distinguer de /ɑ̃/. Nous avions donc été vivement intéressé de constater que les degrés de labialité s'élevaient à 3 et non à 2 comme il est courant de l'admettre:

(-lab) : /a/,
(+lab) : /ɔ/ et /ɑ̃/,
(++lab): /ɔ̃/ que nous dénommions "surlabialisé".

C'est à partir de cette remarque que nous avons décidé de pousser plus loin l'investigation. En effet, en relisant la littérature concernee, on constate que les études précédentes:

- admettent toutes a-priori le caractère binaire du trait de labialité.
- reconnaissent d'autres degrés de labialité, mais uniquement dans le cas particulier de variantes phonétiques contextuelles.
- ne portent que sur les voyelles /i,e,y,ø/ et plus récemment /a/, en contexte, et pour un maximum de 5 locuteurs.

L'étude que nous avons entreprise a donc pour but de compléter les travaux existant portant sur le français. Citons, parmi nos objectifs, ceux qui diffèrent essentiellement des précédents:

- Décrire toutes les voyelles du français, orales et nasales, à l'exclusion de /ɑ/ et y-compris /œ̃/, dans la mesure du possible.
- Observer de plus près l'opposition traditionnelle des voyelles selon 2 catégories labiales.
- Accorder une plus grande part au nombre de locuteurs des 2 sexes.
- Etudier les caractéristiques des voyelles prononcées individuellement (donc hors contexte), et tenues.

Notons que si les voyelles isolées présentent l'inconvénient d'être parfois sur-articulées, leurs caractéristiques sont vraisemblablement plus proches de celles généralement retenues comme traits distinctifs. Leur étude n'est donc pas à négliger, mais nous avons l'intention de la compléter ultérieurement pour des voyelles en contexte.

METHODE, SUJETS, CORPUS, MESURES.

La méthode utilisée est désormais classique (ZERLING,79a): une photographie du visage du locuteur est réalisée de face, et un miroir placé à 45 degrés permet d'obtenir en même temps une vue de profil. Le miroir est muni d'une réglette de 10 cm servant d'échelle. Pour minimiser l'écart de distance non négligeable entre la vue frontale et la vue latérale (image virtuelle de l'autre côté du miroir), on utilise un télé-objectif de 105 mm permettant de prendre un recul d'environ 4 mètres.

Les sujets sont des étudiants âgés généralement de 20 à 25 ans, et dont l'origine linguistique est prise en compte le cas échéant: les données de ceux dont l'accent étranger était trop marqué ont volontairement été écartées, et elles seront exploitées séparément.

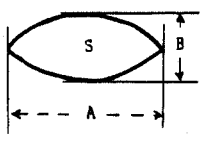
Les locuteurs ont pour consigne de prononcer des sons tenus, en évitant dans la mesure du possible la sur-articulation. Chaque son est suffisamment éloigné temporellement du suivant pour éviter les effets de liste. Dans le cas de /œ̃/ que nous avons ajouté par curiosité, il est demandé de prononcer le plus naturellement possible le chiffre 1.

Une numérisation automatique des données labiales semble être bientôt envisageable (LALLOUACHE & WORLEY,88), néanmoins nos données n'ayant pas été réalisées dans cette optique, notre méthode d'acquisition est manuelle: les négatifs des photos sont projetés agrandis à échelle constante et le contour de l'orifice labial est dessiné sur papier.

Différents logiciels sont ensuite utilisés:
1 - Acquisition des croquis sur tablette graphique. Les paramètres choisis (A,B,S) sont mesurés automatiquement et stockés sur fichier agrémentés d'un code pour le locuteur et d'un autre pour le son prononcé.
2 - Représentations graphiques diverses d'un paramètre en fonction d'un autre.
3 - Analyse statistique: moyenne, écart-type, et écart-type relatif sont calculés pour chacun des paramètres, selon les sons, les catégories de sons, et pour tous les sons.

Dans le cadre de cette étude, seules les vues frontales de l'orifice intéro-labial sont exploitées. Les paramètres retenus sont ceux qui se sont révélés les plus pertinents lors des études précédentes (ABRY & al,80; ABRY & BOE,86), (fig.1):

Fig.1 : paramètres labiaux étudiés:
- A: écartement horizontal
- B: séparation verticale
- S: aire intérolabiale
- AxB: produit des 2 dimensions
- K2=A/B: facteur de forme



RESULTATS ET DISCUSSION

1 Une variation linéaire de l'aire aux lèvres

Tout d'abord, nos résultats confirment qu'il existe une relation linéaire entre l'aire aux lèvres S et le produit AxB des 2 dimensions de l'orifice intéro-labial (fig.2):

$$S = K1 \times AxB$$

Cette équation, déjà proposée par FROMKIN(64,p.224) pour les voyelles américaines prononcées par 5 sujets, a, depuis été vérifiée pour le français: ZERLING(79b,80), ABRY & BOE(86).

Notre étude, qui porte actuellement sur 16 locuteurs et 225 réalisations vocaliques, confirme que cette relation est indépendante:
- du locuteur et de son sexe,
- probablement de la langue parlée,
- du son émis et de son contexte (une étude pour les consonnes paraîtra ultérieurement).

Pour notre corpus, la valeur moyenne de K1 est de 0.70, et on peut admettre que dans la plupart des cas:

$$0.60 \times AB < S < 0.75 \times AB$$

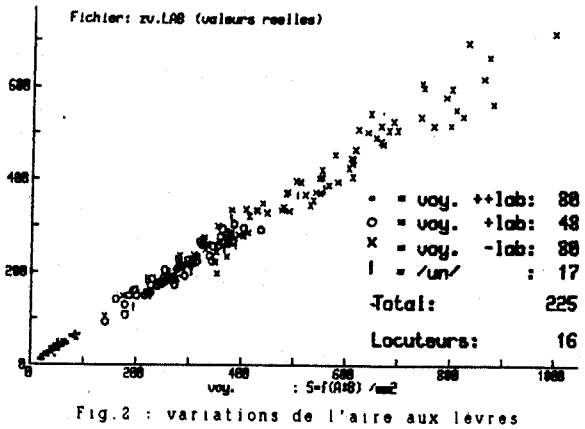


Fig.2 : variations de l'aire aux lèvres en fonction du produit AxB.

2. Trois degrés de labialité vocalique

Pour les réalisations étudiées, le caractère binaire généralement attribué au trait de labialité en français paraît totalement contestable en termes purement articulatoires: ce n'est pas en 2 mais en 3 groupes que se répartissent les voyelles; l'apparition d'un troisième groupe provient de la division des voyelles dites arrondies en 2 catégories indubitablement distinctes: aucun cas de confusion n'apparaît sur la figure 3.

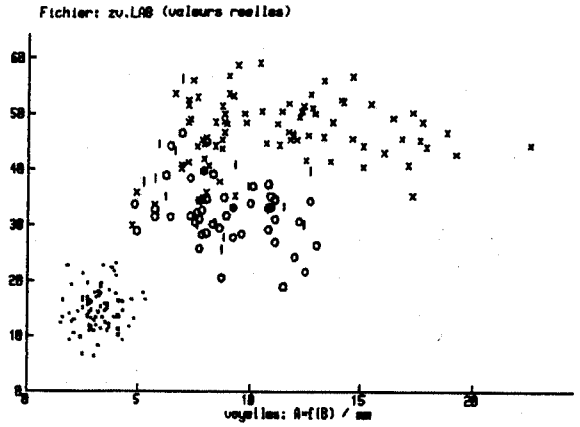


Fig.3 : variations de l'écartement A en fonction de l'espace B

Pour désigner les 3 catégories observées, nous utiliserons indifféremment les termes suivant, dérivés de ceux habituellement admis:

1 - écarté = non-labialisé = (- lab)
2 - arrondi = labialisé = (+ lab)
3 - fortement arrondi = fortement labialisé = (++lab)

Les voyelles contenues dans chaque série sont les suivantes:

1 (-lab) : / ɪ, e, ɛ, a, ɛ̃ /
2 (+lab) : / oe, ɔ, ɑ̃ /
3 (++lab) : / y, ø, u, o, ɔ̃ /

Un symbole particulier est attribué à chacun de ces groupes, ainsi que pour la voyelle /oẽ/ dont l'instabilité était prévisible (fig.2).

Voyons maintenant en quoi, et jusqu'à quel point, s'opposent ces 3 categories, tous sujets confondus.

2.1. Voyelles non-labialisées (-lab)

Il est clair que la série 1 se distingue des 2 autres par une stratégie musculaire et articulatoire bien différente. Une simple observation dans un miroir est parfaitement convaincante: écartement des commissures et écrasement des lèvres contre les incisives pour les écartées, contrairement à l'arrondissement et la protrusion qui caracterisent les arrondies.

Pour la série 1, les variations inter-individuelles de l'écart horizontal A sont relativement faibles (fig.3), et c'est donc surtout le paramètre B qui est responsable des différences de l'aire aux lèvres. Généralement, les valeurs sont telles que:

A > 40mm, 7< B <20mm et 350< S <850mm2 (fig 1)

2.2. Voyelles labialisées (+lab)

Les voyelles labialisées de la série 2 se distinguent des premières tout d'abord par la forme de l'orifice qui est nettement plus arrondi. Cet arrondissement ne touche que légèrement le paramètre vertical B qui adopte souvent des valeurs identiques à celles de la série 1 (fig.3). En revanche, la variable A est systématiquement plus petite, ce qui découle évidemment de l'arrondissement et du rapprochement des commissures. On note en moyenne que:

20< A <40mm, 5< B <13mm et 150< S <400mm2

2.3. Voyelles fortement labialisées (++lab)

Enfin, les fortement labialisées se différencient totalement de celles de la série 2 à la fois pour les valeurs de A et de B, et donc en conséquence de celles de S. La contrainte imposée à A et à B est très forte (fig.3):

S < A < 22mm, 2 < B < 5mm

avec des variations de S: 15< S <75mm2

Les mêmes données ont été représentées après avoir subi une normalisation en fonction de la valeur maximale de chacun des paramètres et pour chaque locuteur (fig.4).

Le résultat est moins convaincant: bien que les 2 séries restent distinctes, la confusion est plus grande. Néanmoins, cette observation présente un intérêt puisqu'elle montre que la contrainte labiale pour l'émission des voyelles ne porte pas sur la forme de l'orifice labial, mais plutôt sur ses dimensions. Cette contrainte, probablement pour des raisons acoustiques, agit sur la valeur absolue de l'aire aux lèvres, indépendemment du sexe du locuteur: que les dimensions de la bouche soient grandes ou petites, le but à atteindre semble être impérativement une aire très precise de l'orifice labial.

Par ailleurs, les valeurs normalisées mettent en relief la faible variation relative du paramètre A dans la serie 1, ce qui confirme que la distinction entre les voyelles non-labiales se fait essentiellement à l'aide de variations de l'espace vertical B.

2.4. Voyelles nasales.

Ces voyelles appellent des remarques particulières (fig.3):

- /ɛ̃/: s'intègre parfaitement à la série des voyelles non-labiales.

- /oẽ/: contrairement à notre attente, la plupart des sujets ont eu tendance à respecter l'arrondissement, mais il convient de rappeler que les voyelles ont été prononcées isolées, donc soumises à un contrôle moteur important, voire à une surarticulation

- /ɑ̃/: sans conteste, cette voyelle est à classer comme labiale, au même titre que /oe,ɔ/.

- /ɔ̃/: cette voyelle a déjà fait l'objet d'une étude de notre part (ZERLING,84b). Les données recueillies ici ne font que confirmer les conclusions antérieures: /ɔ̃/ s'apparente totalement au groupe des voyelles fortement labialisées, comme /u,o/ par exemple. Nous l'avions baptisée "surlabialisée" (ZERLING,84b). Ce terme n'est pas abérrant et pourrait être maintenu dans la mesure ou il reflète que /ɔ̃/ est beaucoup plus labialisée que ne l'évoque son symbole, celui de la voyelle /ɔ/. Dans ce sens le choix du symbole /oẽ/ pour la nasale serait tout-à-fait justifié, comme nous l'avons déjà souligné à plusieurs reprises, et il serait plus évocateur pour toute personne observant les lèvres. En fait, /ɔ̃/ est réalisé avec une position linguale proche de celle de /ɔ/ et de /ɑ̃/, mais necessite un degré de labialisation bien supérieur, notamment pour se distinguer de la dernière.

3. Stratégies labiales: forme ou aire?

ABRY & al (1979a,p.214) avancent que "les valeurs + et - du trait d'arrondissement sont de nature différente, le premier pouvant se caracteriser par un facteur de forme aux lèvres, le deuxième par une constante d'aire qui n'est pas interpretable directement en termes de constante physiologique."

Nous n'avons pas encore pu traiter nos données aprés normalisation , ce qui parait s'imposer pour étudier le paramètre k2=A/B. Les résultats de cette démarche seront donc proposés par la suite.

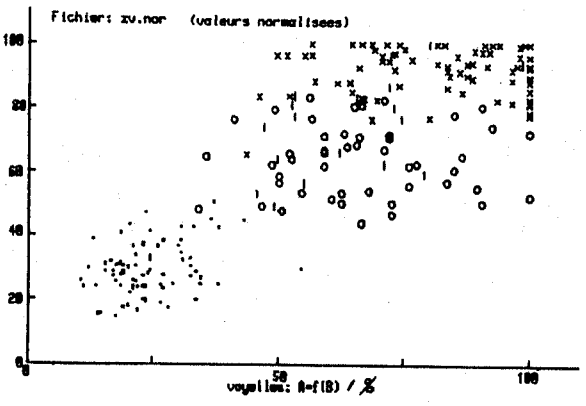


Fig.4 : variations de l'écartement A en fonction de l'espace B (valeurs normalisées).

4 Comparaisons intervocaliques

Il a été effectué pour chaque voyelle, puis pour chaque catégorie labiale un calcul de moyenne, d'écart-type et d'écart-type relatif des 5 paramètres: A, B, S, AxB, A/S (fig 6).

Pour ce qui concerne la distinction entre plusieurs voyelles d'une même catégorie, la fig.5 appelle plusieurs remarques:

4.1. Voyelles (-lab) :

Les valeurs de l'écartement horizontal A sont très voisines. La différenciation labiale se fait donc plutôt par les variations de l'espacement vertical B. Néanmoins, ces dernières sont relativement faibles et variables. Il en résulte des valeurs d'aire qui, bien que regroupées comme il a été dit, sont équivoques pour les voyelles de cette série. Il faut donc admettre que dans ce cas, le rôle des mouvements linguaux est déterminant pour obtenir la distinction acoustique entre les 4 voyelles antérieures.

4 2 Voyelles (+lab) et (++lab):

La subdivision des voyelles labiales en 2 groupes telle qu'elle est proposée ici révéle que 2 degrés de labialité seulement rendent compte de 3 voyelles différentes, tant pour la série /y,ø,oe/ que pour /u,o,ɔ/. Il est donc vraisemble que c'est par la position linguale que cette simplification est compensée, et l'on ne peut que rapprocher cette remarque de celle proposée dans l'ouvrage "Cinéradiographie des voyelles et consonnes du français"(BOTHOREL & al,86, p.287-288): "on observe une position linguale nettement différenciée entre /y/ et /ø/. En revanche, /oe/ est trés proche de /ø/ et même parfois plus fermé". Les auteurs signalent ensuite que /u/ étant "la voyelle postérieure la plus fermée,(...), la position de la langue est parfois identique pour les 2 voyelles /o/ et /ɔ/.

Ainsi, on relève un phénomène intéressant d'économie articulatoire: en combinant 2 stratégies labiales et 2 stratégies linguales, on parvient à créer 3 voyelles différentes, aussi bien en position antérieure qu'en position postérieure.

/y,u/: langue haute / (++lab),
/ø,o/: langue moyenne / (++lab),
/oe,ɔ/: langue moyenne / (+lab).

Il sera évidemment intéressant de vérifier si cette observation est confirmée, et donc applicable, en modélisation articulatoire.

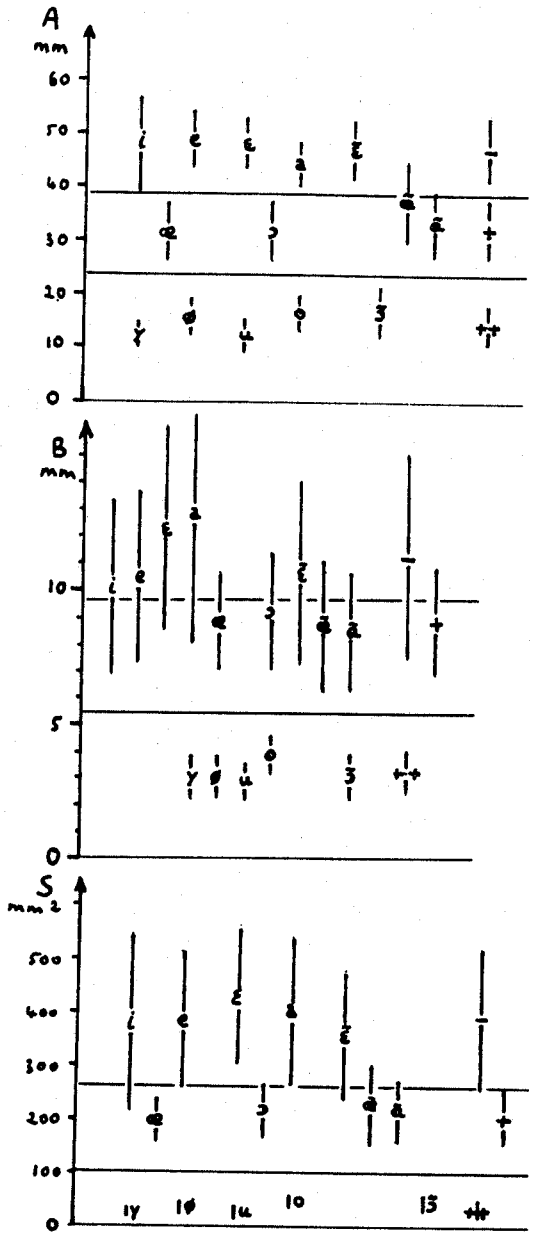


Fig.5 : moyenne et écart-type des parametres A, B, S en fonction des voyelles et des catégories labiales.

31-MAR-88 Fichier: sv LAE
Moyenne, ecart-type et ecart-type relatif des parametres de l'orifice labial

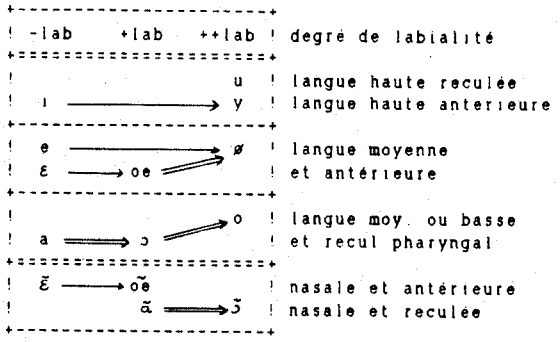
	Nombre	A			B			S			AxB			A/B		
Voyelles:																
[i]	16	(47.5	8.9	0.19)	(10.8	3.2	0.32)	(377.8	164.8	0.44)	(489.1	282.1	0.61)	(5.1	1.2	0.23)
[e]	16	(48.3	5.2	0.11)	(10.5	3.2	0.31)	(382.1	127.2	0.33)	(511.6	166.0	0.32)	(5.0	1.5	0.30)
[e)]	16	(47.7	4.8	0.10)	(12.3	3.8	0.31)	(430.1	133.1	0.31)	(588.3	190.0	0.32)	(4.3	1.5	0.35)
[a]	16	(44.1	4.0	0.09)	(12.8	4.8	0.38)	(402.2	137.0	0.34)	(556.3	198.8	0.36)	(4.0	1.6	0.40)
[y]	16	(12.7	1.8	0.14)	(3.8	0.9	0.29)	(36.4	8.2	0.31)	(37.9	11.4	0.30)	(4.7	1.8	0.38)
[ew]	16	(15.7	3.3	0.21)	(3.1	0.8	0.26)	(36.3	9.2	0.25)	(47.7	12.7	0.27)	(5.5	2.1	0.39)
[oe]	16	(31.7	5.6	0.18)	(6.8	1.9	0.21)	(196.9	48.7	0.21)	(272.6	52.4	0.19)	(3.8	1.3	0.33)
[u]	16	(12.3	3.0	0.24)	(2.8	0.7	0.23)	(26.1	9.5	0.37)	(34.4	11.8	0.34)	(4.6	1.5	0.32)
[o]	16	(16.4	3.2	0.20)	(3.8	0.7	0.17)	(46.5	11.5	0.25)	(62.9	16.5	0.26)	(4.4	1.1	0.26)
[o)]	16	(31.8	5.4	0.17)	(9.3	2.3	0.24)	(214.3	58.7	0.27)	(298.6	78.3	0.26)	(3.7	1.2	0.33)
[ɔ)]	0	(0.0	0.0	0.17)	(0.0	0.0	0.24)	(0.0	0.0	0.27)	(0.0	0.0	0.24)	(0.0	0.0	0.33)
[in]	16	(46.2	5.2	0.11)	(10.6	3.4	0.32)	(351.1	128.7	0.34)	(491.0	171.5	0.35)	(4.7	1.4	0.29)
[un]	17	(37.0	7.8	0.21)	(8.6	2.4	0.28)	(228.8	76.7	0.34)	(316.8	104.5	0.33)	(4.7	1.9	0.40)
[an]	16	(33.3	6.1	0.18)	(8.4	2.2	0.26)	(208.4	55.5	0.27)	(278.9	81.4	0.29)	(4.3	1.5	0.36)
[on]	16	(16.7	4.5	0.27)	(3.1	0.8	0.26)	(38.3	15.6	0.41)	(53.1	21.9	0.41)	(5.6	2.4	0.43)
Sons par categories labiales:																
[++]	80	(14.7	3.7	0.25)	(3.2	0.8	0.26)	(34.7	13.3	0.38)	(47.2	18.2	0.39)	(5.0	1.9	0.38)
[+]	48	(32.3	5.6	0.17)	(8.8	2.1	0.24)	(206.5	51.6	0.25)	(288.7	68.8	0.24)	(3.9	1.3	0.34)
[-]	80	(46.8	5.9	0.13)	(11.2	3.8	0.34)	(388.5	136.3	0.35)	(527.2	185.5	0.35)	(4.6	1.5	0.32)
[un]	17	(37.0	7.8	0.21)	(8.6	2.4	0.28)	(228.8	76.7	0.34)	(316.8	104.5	0.33)	(4.7	1.9	0.40)
Voyelles:																
	225	(31.6	14.6	0.46)	(7.7	4.3	0.57)	(211.8	173.1	0.82)	(288.1	235.8	0.82)	(4.6	1.7	0.36)

Fig.6 : moyenne, écart-type, et écart-type relatif des paramètres de l'orifice labial.

CONCLUSIONS

Pour terminer cette étude, nous proposons sous forme de tableau un rappel des caractéristiques articulatoires des voyelles du français telles qu'elles ont été décrites ici (fig.7).

Fig.7: Caractéristiques articulatoires des voyelles



Notons qu'au plan phonologique, ce tableau ne remet pas fondamentalement en cause le caractère binaire du trait de labialité puisque les voyelles /ɪ,e,ɛ,ɛ̃/ s'opposent respectivement à /y,ø,œ,œ̃/ (flèches simples).

Néanmoins, au plan de la phonétique articulatoire, il faut admettre qu'il existe sans conteste 3 degrés de labialité vocalique en français:

(-lab) : non-labialisé
(+lab) : labialisé
(++lab): fortement labialisé

Nous avons montré la réalité de cette interprétation à partir des paramètres décrivant la forme de l'orifice intéro-labial. Il est remarquable de constater sur le tableau l'apparition d'oppositions non reconnues habituellement et qui reposent pourtant sur le trait de labialité (flèches doubles):
/a-ɔ/ ,et surtout /œ-ø/, /ɔ-o/ et /ɑ̃-ɔ̃/;
à condition que l'on accepte simplement que les positions linguales puissent être très voisines pour chacune de ces paires de sons, ce qui est confirmé par les études citées plus haut.

L'opposition /a-ɔ/ peut paraître étonnante pour le français, mais elle correspond en fait à celle connue en anglais: /ʌ-ɔ/. Les sons /ʌ/ et /a/ sont en effet très voisins (ZERLING,79a, vol.2., fig.IX.5)

L'opposition /ɑ̃-ɔ̃/, qui paraît désormais incontestable, rappelle celle de /ɛ̃-œ̃/, mais avec 2 degrés de labialité respectivement différents.

Enfin, si l'on veut bien admettre qu'il existe une ressemblance articulatoire entre les voyelles d'une même case du tableau (fig.7), alors on est en droit de dire qu'il existe en français 2 séries pour lesquelles des voyelles de même articulation linguale s'opposent essentiellement selon 3 degrés de labialisation différents (fig.8):

Fig.8: Rôle des lèvres pour des positions linguales proches

-lab	+lab	++lab	degré de labialité
ɛ	œ	ø	langue moyenne
a	ɔ	o	recul pharyngal

Les observations avancées ici sont particulières au français. Elle pourraient donc trouver leur application à la fois pour la correction phonétique lors de l'apprentissage de notre langue à des étrangers, ou pour la commande de modèles articulatoires.

Nous espérons pouvoir vérifier, en utilisant des modèles du conduit vocal, que pour une même position linguale, la labialisation puis la surlabialisation permettent bien de passer de /ɛ/ à /ø/, ou de /a/ à /o/. Le diagramme de macro-sensibilité pour le paramètre lèvres proposé par MAJID & al(87,p.351) semble bien confirmer cette hypothèse.

Ce travail avait pour objet de décrire des voyelles tenues hors contexte. Outre une étude en cours de certaines consonnes en contexte, nous pensons poursuivre l'analyse à partir d'un corpus de voyelles réalisées en dynamique pour voir dans quelle mesure nos observations leur seront également applicables.

REFERENCES

ABRY C. & BOE L.J. (1986) " "Laws" for lips", Speech Communication, 5, pp.97-104.

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, Institut de Phonétique de Grenoble. 304 p.

ABRY C.,BOE L.J., GENTIL M., DESCOUT R. & GRAILLOT P. (1979,a) "La géométrie des lèvres en français. Protrusion vocalique et protrusion consonantique", Actes 10e J.E.P., Grenoble, p.99.

ABRY C., BOE L.J., & DESCOUT R. (1979,b) "Voyelles labiales et voyelles labialisées en français", Proceedings of 9th Int. Cong. Phon. Sciences, Copenhagen, p.177.

Actes du Séminaire International Labialité, Lannion, 7-8 février 1980, 250 p.

BELL-BERTI F. & HARRIS K.S. (1976) "An EMG study of coarticulation of lip rounding". Acoustical Society of America.

BENGUEREL A.P. & COWAN N.A. (1974) "Coarticulation of upper lip protrusion in French", Phonetica, 30, pp.41-55.

BONNOT J.F., CHEVRIE-MULLER C., GRENIER G., MATON B. & GUIDET C. (1983) "Etude de l'encodage moteur des traits de nasalité et de labialité à partir de l'activié EMG des muscles orbiculaires (OO) et élévateur du voile (LP)", Actes des IIe I.C.A., Toulouse, 19-27 juillet 83, p.76.

BOTHOREL A., SIMON P., WIOLAND F. & ZERLING J.P. (1986) Cinéradiographie des voyelles et consonnes du français. Recueil de documents synchronisés pour 4 sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques, Travaux de l'Institut de Phonétique de Strasbourg, 298 p.

DESCOUT R., BOE L.J. & ABRY C. (1978) "Labialité vocalique et labialité consonantique en français, premiers résultats", GALF, Actes 9e J.E.P., pp.177-189.

FROMKIN V.A. (1964) "Lips positions in American English vowels", Language and speech, 7, pp.215-225.

GENTIL M. (1980) Labialité en français: étude phonétique et aspects physiologiques des lèvres, Thèse de 3e cycle, Université de Grenoble III, 440p.

LALLOUACHE M.T. & WORLEY C. (1988) "Saisie, edition et traitement d'images et de signaux articulatoires: lèvres et mâchoires", Séminaire Production de la Parole: Modèles et données, 2-3 février 1988, Grenoble.

MAEDA S. (1979) "An articulatory model of the tongue based on a statistical analysis", in Speech Communication Papers, 97th Meeting of Acoustical Society of America, Cambridge, USA, June 1979.

MAJID R., ABRY C., BOE L.J. & PERRIER P. (1987) "Contribution à la classification articulatori-acoustique des voyelles: étude des macro-sensibilitées à l'aide d'un modèle articulatoire", Proceedings XIth ICPS, Tallin, USSR.

ZERLING J.P. (1979,a) Articulation et coarticulation dans des groupes occlusive-voyelle en français. Etude cinéradiographique et acoustique: contribution à la modélisation du conduit vocal, Thèse de 3e cycle en Phonétique, Université de Nancy II, 515p.

ZERLING J.P. (1979,b) "Description de cinq voyelles orales du français en contexte et nouvelle classification articulatoire", Travaux de l'Institut de Phonétique de Nancy, in Verbum, Université de Nancy II, tome 2, fascicule 1, pp.55-87.

ZERLING J.P. (1980) "Coarticulation labiale et aire aux lèvres dans des groupes occlusives-voyelles en français", Actes du Séminaire International Labialité, GALF, Lannion, fév 1980, 12 p.

ZERLING J.P. (1984,a) "Nasalité et oralité vocaliques en français: étude cinéradiographique, premiers résultats", Actes des 13e J E.P., GALF, Bruxelles, 28-30 mai 84, pp.217-218.

ZERLING J.P. (1984,b) "Phénomènes de nasalité et de nasalisation vocaliques: étude cinéradiographique pour deux locuteurs", Travaux de l'Institut de Phonétique de Strasbourg, 16, pp.241-266.

ZERLING J.P. (1984,c) "Aperture, antériorité et labialité sont-elles inévitablement des traits distinctifs binaires pour les voyelles du français ?", 3e Journées de l'ATALA, Paris, 10 mars 1984, non publié.

MESURES DE FONCTIONS DE TRANSFERT DU CONDUIT VOCAL - APPLICATION A LA DETERMINATION DES FONCTIONS DE TRANSFERT DU CONDUIT NASOPHARYNGAL

E. CASTELLI & P. BADIN

I.C.P. (UA CNRS 368)
Laboratoire de la Communication Parlée
ENSERG/INPG, 46 Avenue Félix Viallet 38031 GRENOBLE Cedex

ABSTRACT :

To understand acoustic phenomenons dealing with the nasal vowel production, we have built an experimental setting to obtain vocal tract transfer functions. Our setting is an improved version of FUJIMURA & LINDQVIST's method. The vocal tract is excited, trough the skin of the glottis, by a white noise and the transfer functions are obtained by averaging the F.F.T. spectra. Results are rather complex but a template seems to be drawn for which the main spectral peaks may agree with FENG's predictions : the low nasal poles could be the naso-pharyngeal tract's one.

INTRODUCTION :

Pour la compréhension des phénomènes acoustiques intervenant dans la production de la parole, la connaissance de la fonction de transfert du conduit vocal se révèle extrêmement précieuse. Parmi les différents travaux dans ce domaine (VAN DEN BERG (1955), FANT (1960)), les expériences réalisées par FUJIMURA & LINDQVIST (1971) sont certainement les plus riches : ils ont pu ainsi proposer les valeurs des formants et leurs bandes passantes pour les voyelles orales de l'anglais, et ils observèrent pour les voyelles nasales des "formants et antiformants nasals", et un déplacement des formants associés à la cavité buccale correspondante. Pour cela, ils avaient adopté la technique suivante : ils excitaient le conduit à travers la peau, en appliquant la membrane d'un petit haut-parleur au niveau de la glotte ; le signal d'excitation consistait en un balayage en fréquence d'un son pur ; l'énergie récupérée aux lèvres était retranscrite sur un graphique en fonction de la fréquence. Il était alors facile de mesurer sur les courbes obtenues les résonances et bandes passantes des fonctions de transfert du conduit vocal, sans être gêné par le spectre de la source, qui entâche les fonctions de transfert obtenues par des méthodes plus classiques, (L.P.C., Cepstre...).

C'est pourquoi, à la fois pour vérifier certaines hypothèses émises sur la production des nasales (FENG et al. 1986), et pour étudier les phénomènes à l'origine des fricatives, nous avons mis au point un dispositif expérimental s'inspirant largement de leur méthode.

I PROCEDURES EXPERIMENTALES :

Nous avons gardé la méthodologie générale de ces expériences :

1 - l'enregistrement est réalisé dans une chambre anéchoïque;

2 - le sujet applique la membrane d'un petit haut-parleur directement contre sa peau dans la région du bas pharynx;

3 - le sujet tient une configuration stable du conduit vocal (voyelles) mais en gardant la glotte fermée.

I.1 Les aspects matériels :

Nous avons choisi d'exciter le conduit vocal par un bruit blanc sur la bande de fréquence parole : c'est d'un point de vue technique plus simple à mettre en oeuvre qu'un balayage en fréquence d'un son pur. Le spectre de ce bruit est plat entre 20 et 20000 Hz, (moins de 3db de fluctuations). La figure n° 1 est un schéma fonctionnel de notre réalisation.

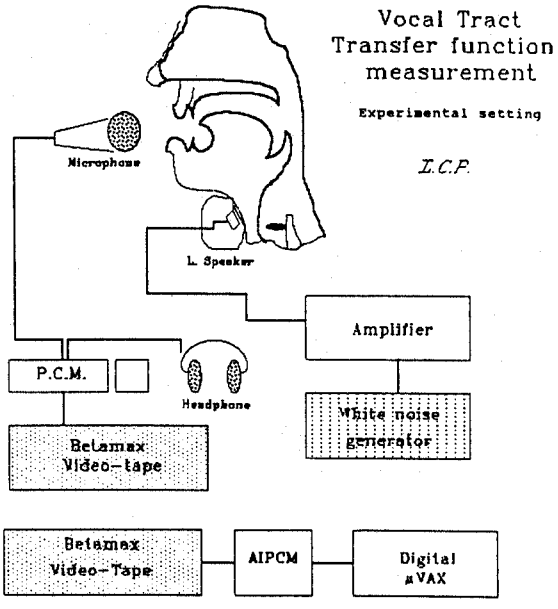


Fig. 1 Schéma fonctionnel

Le bruit blanc, produit par un générateur Brüel & Kjoer de bonne qualité, est amplifié avant d'attaquer le petit haut-parleur (de 3 cm de diamètre). Celui-ci est encastré dans un petit tube de plastique de 2 cm de long, afin d'en faciliter la préhension. Le tout est enrobé d'une boule de pâte à modeler pour diminuer les rayonnements parasites du haut-parleur.

Le sujet applique directement la membrane du haut-parleur contre la peau de son cou au niveau du cartilage thyroïde. Nous n'avons pas retenu l'idée de maintenir la position de l'excitateur par un mécanisme plus sophistiqué car les résultats sont remarquablement constants.

Le signal obtenu, en sortie du conduit vocal, aux lèvres ou aux narines, est enregistré grâce à un microphone "Electret" très directif et très sensible, à travers un P.C.M. CODER, sur un magnétoscope du type BETAMAX. L'enregistrement numérique permet de travailler sur un signal de très bonne qualité et de grande dynamique.

Une sortie son (préampli du P.C.M. CODER) offre l'immense avantage de renvoyer le signal enregistré sur un casque ou sur une enceinte, pour générer un "feed-back" très utile au sujet. Cela lui permet, en effet, grâce au contrôle auditif, de garder une configuration stable durant toute l'opération. Le signal entendu ressemble à des voyelles "chuchotées".

Une fois les enregistrements effectués, un deuxième magnétoscope BETAMAX et un deuxième P.C.M. CODER nous permettent de les stocker sous forme de fichiers sur un MicroVAX, afin de pouvoir les traiter.

I.2 Les traitements des signaux enregistrés

Si on suppose que le conduit vocal se comporte dans ces conditions d'expérimentation comme un filtre acoustique linéaire, l'excitation par un bruit blanc nous permet, en calculant la moyenne sur un nombre suffisant de Transformées de Fourier Rapides, d'obtenir directement la fonction de transfert du conduit vocal. Après plusieurs essais, nous avons conclu qu'une moyenne de 128 F.F.T., calculées sur 1024 points sans recouvrement sur du signal échantillonné à 10 kHz, était suffisante pour obtenir des fonctions de transfert exploitables. Ce calcul impose aux sujets de tenir stable leur conduit pendant 13,2 secondes, ce qui, même si ce n'est pas aisé, reste tout à fait possible.

Notre premier travail a d'abord été la vérification de la cohérence de la chaîne de mesure. Nous avons contrôlé, en le recalculant, le spectre du bruit, donné comme plat sur 20 à 20000 Hz par le constructeur. Ces mesures ont été réalisées en sortie de l'amplificateur, puis, en enregistrement direct du haut parleur (Fig. 2). Celui-ci, en excitation libre, a une réponse en fréquence plate sur une bande passante de 200 à 5000 Hz, mais son comportement, une fois sa membrane plaquée contre la peau, restait à caractériser, les masses de la peau et des tissus du cou interférant sans aucun doute.

Pour cela, nous avons agi comme FUJIMURA & LINDQVIST (1971) en calculant une différence de fonctions de transfert. Ainsi un enregistrement a été réalisé pour une voyelle orale connue [a] et sa moyenne de F.F.T. nous a fourni une fonction de transfert (fig. 3a) ; nous avons alors pu déterminer les fréquences de résonance et les bandes passantes, puis grâce à ces valeurs, nous avons simulé une fonction de transfert théorique avec un modèle du type "filtres du second ordre" (fig.3b) ; la différence entre les deux fonctions nous a renseigné globalement sur le comportement du haut-parleur (fig. 3c), comportement relativement plat de 50 à 5000 Hz. Bien entendu, nous avons validé ce test plusieurs fois.

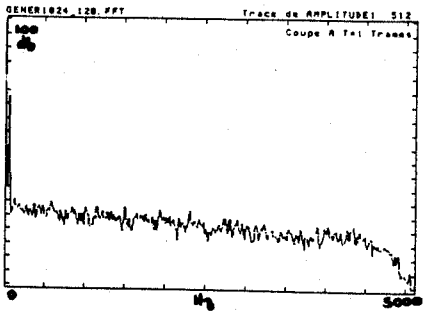


Fig. 2 Spectre du bruit blanc.

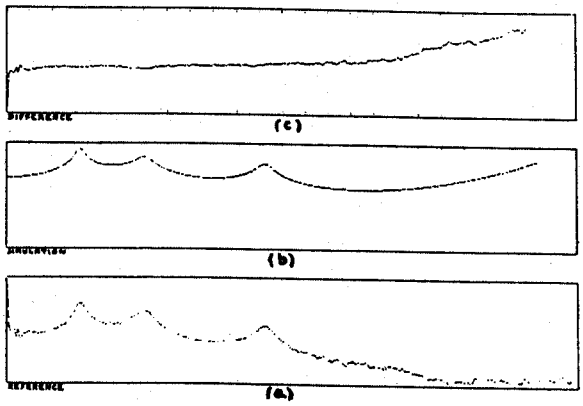


Fig. 3 Evaluation de la fonction de transfert de la source d'éxcitation.

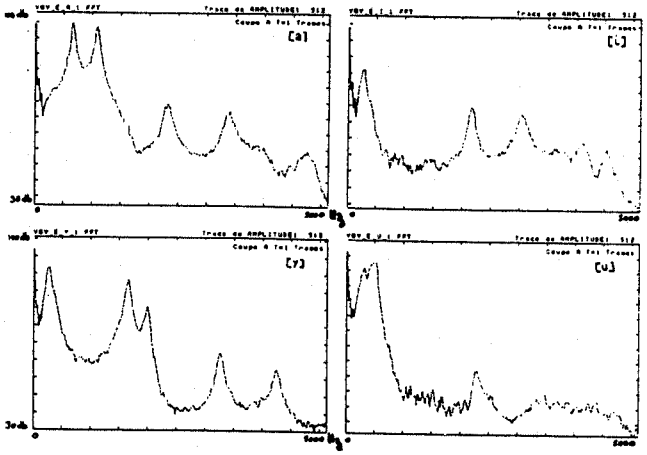


Fig. 4a Exemples de fonctions de transfert de voyelles orales (sujet E.C.)

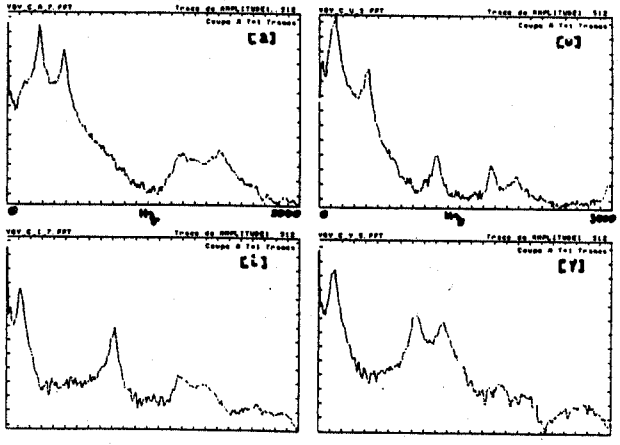


Fig. 4b Autres voyelles orales (sujet T.B.)

Pour compléter ces vérifications, nous demandons à nos sujets d'enregistrer, à chaque séance, le signal pour sept voyelles orales, [a], [i], [u], [o], [e], [Ø], [y], afin de comparer les formants mesurés à des valeurs standard (MRAYATI (1976), DEGRYSE (1981)). Nous pouvons détecter rapidement, de cette façon, un éventuel disfonctionnement de notre petit excitateur (Fig. 4).

II DETERMINATION DES FONCTIONS DE TRANSFERT DU CONDUIT VOCAL :

Pour synthétiser des voyelles nasales avec le modèle temporel de KELLY & LOCHBAUM (1962), nous avons besoin de connaître les caractéristiques du conduit vocal, et plus particulièrement celles du conduit nasal. Or, si l'on connait assez bien l'anatomie de ce conduit (GARDNER, GRAY, O'RAHILLY (1986)), fort compliquée par ailleurs, les influences respectives de ces cavités dans la production de la parole ne sont pas encore déterminées. Nous pensons qu'une analyse systématique de fonctions de transfert obtenues dans des conditions parfaitement connues de production, est une voie intéressante pour avancer sur ces grandes questions.

II.1 Notre but :

C'est ainsi que nous voulons évaluer expérimentalement l'hypothèse émise à l'I.C.P. par FENG et al., qui attribuent au premier formant nasal (250 Hz) une origine du type résonateur de HELMHOLTZ (cavité nasopharyngale + narines) ; en effet, une large cavité se terminant par une constriction voit sa première résonance dépendre directement de l'aire de cette constriction. La formule de la fréquence de résonance nous est donnée par :

$$F0 = (C/2\pi).\sqrt{\frac{A}{l.V}}$$

Où V est le volume du résonateur,
I la longueur de la constriction,
A l'aire de la constriction.

Plus l'aire de la constriction diminue, plus la fréquence de cette résonance s'abaisse pour tendre vers zéro. FENG et al. pensent ainsi que la fréquence du premier formant des voyelles nasales dépendrait essentiellement de l'aire, assez faible, du seuil narinaire.

Nous rappelons que cette idée est contestée par LONCHAMP (1988) qui, rejoignant Maeda, préfère associer le premier formant nasal bas aux cavités sinusales.

II.2 La démarche expérimentale :

Nous avons tenté de réduire artificiellement le seuil narinaire. Pour cela, une série de petits tubes en plexiglas a été réalisée (Fig. 5), leur diamètre extérieur est juste suffisant pour que le tube rentre dans une narine, par contre le diamètre intérieur a été percé à différentes valeurs, 4.5 mm, 3.0 mm et 2.0 mm. Si on considère que le seuil narinaire à un diamètre d'environ 6.0 mm (valeur déduite du diamètre maximum du tube en plexiglas que nous avons pu rentrer dans le seuil narinaire des sujets), nous obtenons donc des séries de 4 fonctions de transfert.

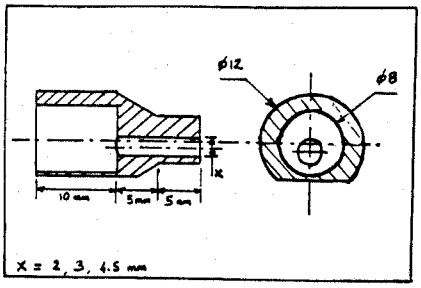


Fig. 5 "Narine" en plexiglas.

Le sujet positionne son conduit vocal, mâchoire, langue et lèvres, pour une configuration de voyelle orale ([a], [i], [u], [o]), puis il doit, SANS BOUGER par ailleurs, abaisser complètement son velum : le conduit mesuré est donc le conduit nasopharyngal, formé du pharynx et du conduit nasal, le conduit oral n'intervenant théoriquement pas, puisqu'on le suppose bouché par le velum complètement abaissé.

Le signal est enregistré pour chaque narine, le micro se trouvant très près de l'ouverture de la narine. On peut considérer, dans ce cas, que le signal enregistré ne provient que d'une narine, le rayonnement de l'autre narine étant négligeable. Il faut, si nous ne voulons rien négliger, ne pas oublier que le conduit mesuré est , en fait, constitué du conduit nasopharyngal mais sur lequel est couplée la deuxième narine. LONCHAMP (1988) montre que la légère dissymétrie des deux narines ne produit qu'une faible paire de pôle zéro aux environs de 1,7 Kz. Nous pouvons donc affirmer que les deux narines fortement semblables se comportent ensemble comme un seul conduit d'aire double.

En fait, le velum n'est pas parfaitement étanche, sa faible épaisseur laisse passer une partie de l'excitation, mais le signal résultant, rayonnant à la bouche, reste toutefois très faible devant celui enregistré. Ce signal, une fois analysé, donne, si le sujet a bien positionné son conduit vocal, des fonctions de transfert analogues à celles des voyelles orales.

Parallèlement, des simulations harmoniques ont été calculées avec les fonctions d'aires du pharynx correspondant aux voyelles orales de départ et la fonction d'aire du conduit nasal proposée par BOE (1972) et FENG (1986).

III. LES RESULTATS :

Une première constatation nous est imposée : l'abaissement de la fréquence du premier formant du conduit nasopharyngal mesuré est nettement moins important que celui que l'on attendait. En effet, alors que les simulations prévoyaient une diminution de plus de 100 Hz, nous mesurons sur les fonctions de transfert obtenues une décroissance de l'ordre de 30 Hz maximum. Le premier formant, d'une valeur moyenne de 230 Hz, ne descend pratiquement jamais en dessous de 200 Hz.

En fait, en étudiant de plus près cette décroissance, nous constatons que tout se passe comme si la valeur de ce premier formant tendait vers une valeur inférieure limite se situant aux alentours de 200 Hz.

Le nombre important de résultats obtenus nous a permis de constater que la plupart des courbes présentent une structure beaucoup plus compliquée que les fonctions simulées. Malgré cette complexité apparente, un grand nombre de pics spectraux et de vallées, pour ne pas parler de zéros, semble garder des valeurs constantes. Un "portrait robot" de la fonction de transfert du conduit nasopharyngal peut être dessiné : jusqu'à neuf pics spectraux, nous les appellerons formants, entre 200 et 4000 Hz, et de un à trois zéros (Fig. 6). Pour un même sujet, les valeurs des fréquences de ces formants ou de ces zéros sont absolument constantes d'une manipulation à l'autre. Une petite étude statistique montre par exemple que pour un même sujet, l'écart type maximum du formant à 2100 Hz est seulement de 60 Hz (pour environ 40 courbes).

Un tableau de moyennes des valeurs mesurées est donné ci-dessous pour le sujet P.B. :

	Fn1	Fn2	Fn3	Fn4	Fn5	Fn6	Fn7	Fn8	Fn9
Hz	225	450	620	925	1245	1460	1680	2100	3080

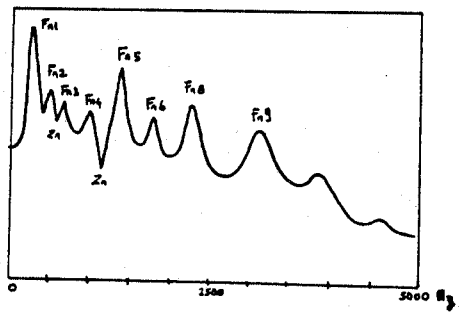


Fig. 6 Fonction de transfert du conduit nasopharyngal.

En fait, neuf formants est un nombre maximum, la plupart des courbes présentent un nombre de formants inférieur. On peut, par exemple pour le sujet P. B., différencier les courbes obtenues en enregistrant séparément le signal à la narine droite, et à la narine gauche. Les formants Fn3 et Fn6 n'apparaissent pas distinctement sur les courbes de la narine droite, en revanche ils sont bien présents sur celles de la narine gauche. Mais le formant Fn7 est affecté à la narine droite, et non à la narine gauche (Fig. 7a et 7b).

IV. DISCUSSIONS :

La limite inférieure du premier formant peut être expliquée par les vibrations des parois (ISHIZAKA & FLANAGAN (1972), FANT (1972)). Considérant que la formule de Helmhotz donne la valeur de la résonance pour une cavité SANS vibration de parois, une correction doit être apportée en tenant compte de la fréquence de ces vibrations :

$$F0 = \sqrt{(Fhelm.^2 + Fvib.^2)}$$

Un rapide calcul montre que si on choisit Fvib = 190 Hz, valeur identique à celle du conduit oral, pour une diminution de l'aire du seuil narinaire de moitié, la décroissance n'est plus que de 28 Hz (au lieu de 63 Hz).

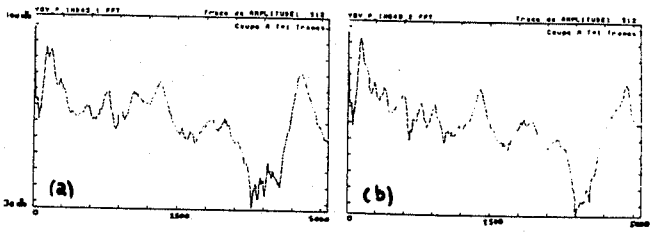


Fig. 7 Fonction de transfert pour la narine droite (a) et pour la narine gauche (b) sujet P.B.

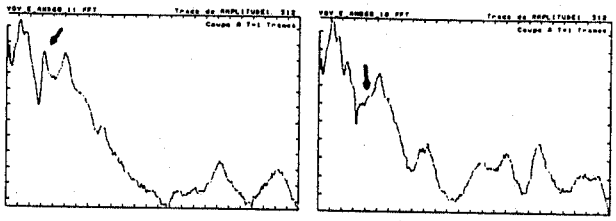


Fig. 7bis Le pic marqué n'apparait pas pour la narine gauche (sujet E.C.)

Cette correction tenant compte des effets des parois peut expliquer la faible décroissance que nous avons mesurée, mais nous restons actuellement incapables de confirmer formellement cette thèse du "Helmoltz". Nous ne pouvons pas écarter les hypothèses suggérées par LINDQVIST & SUNDBERG (1976), MAEDA (1982) et plus récemment LONCHAMP (1988), affirmant l'influence prépondérante des cavités des sinus.

Rapidement, nous pouvons dire que Fn6 et Fn7 sont liés à un même phénomène acoustique (résonance d'une cavité nasale), qui prend des valeurs différentes suivant les narines, du fait de leur dissymétrie incontournable.

En revanche Fn3 pourrait s'expliquer par la présence d'un sinus maxillaire. Les petits formants Fn2, Fn3et Fn4 donnent à la courbe l'allure de celles qu'a simulées MAEDA (1982) quand il a étudié le rôle des sinus en faisant varier l'importance du couplage. Le sinus serait-il bouché pour la narine droite ? Il est communément reconnu que les sinus sont souvent plus ou moins bouchés...

Quoi qu'il en soit, l'hypothèse fondamentale de FENG & al. semble être vérifiée. Le conduit nasopharyngal serait la cible des voyelles nasales pour le maximum de couplage. Même en tenant compte des variations du pharynx pour les différentes voyelles, FENG, rejoignant MAEDA (1984), constatait à partir de simulations que les pics spectraux de la fonction de transfert n'évoluent presque pas et tombent toujours dans une région relativement petite située entre les voyelles [w] et [u], ("trou des nasales"). Effectivement, les formants du conduit nasopharyngal, mesurés par notre expérience, sont très peu sensibles aux voyelles orales de départ. Seuls les formants hauts, Fn8 et Fn9, varient d'environ 200 Hz, selon que la voyelle de départ est un [i] ou un [a] ; l'évolution des autres formants dépasse rarement 5 %.

V. CONCLUSIONS :

Toutes ces manipulations ont été réalisées avec deux sujets masculins. Nous sommes actuellement en train de vérifier le bien-fondé de nos observations sur neuf sujets, six hommes et trois femmes. Chaque personne enregistrera au minimum six fois la même configuration. Peut-on dire que la fréquence de résonance du conduit nasopharyngal, d'environ 200 Hz, est un paramètre acoustique pertinent ?

Parallèlement, nous essayons, par des simulations et par l'étude des fonctions de sensibilités du conduit nasopharyngal d'expliquer ce foisonnement de formants. Il serait très intéressant de définir à quoi correspondent certains des pics, (Helmholtz, Sinus ?...).

De plus, comme à chaque séance les sujets enregistrent le signal correspondant à des voyelles orales, une base de données de fonctions de transfert s'établit. Nous envisageons de calculer les dispersions sur les formants et bandes passantes des fonctions de transfert obtenues.

Nous remercions C. ABRY, G. FENG, B. GUERIN et P. PERRIER pour les nombreuses et fructueuses discussions que nous avons eues sur ce travail.

REFERENCES :

BOE L.J. (1972)
Etude acoustique de couplage larynx-conduit vocal.
Revue d'Acoustique, 27, 235-244.

DEGRYSE D. (1981)
Temporal Simulation of Wave Propagation in the Lossy Vocal Tract
4th FASE Symposium, 193-196.

FANT G. (1960)
Acoustic Theory of Speech Production.
Mouton & Co. -'s-Gravenhage.

FANT G. (1972)
Vocal Tract Wall Effects, Losses, and Resonance Bandwidths.
STL-QPSR 2-3/1972, 28-52

FENG G. (1986)
Modélisation acoustique et traitement de la parole. Le cas des voyelles nasales.
Thèse Dr. I.N.P. GRENOBLE

FENG G. ABRYC. & GUERIN B. (1985)
How to Cope with Nasal Vowels ? Some Acoustc 'Boudary Poles'.
French-Swedish Seminar, GRENOBLE.

FENG G. ABRY C. & GUERIN B. (1986)
The Nasopharyngeal Tract : a Target for Nasality. Acoustic Simulations vs. Sweep-tone Measurements.
12th Int. Congr. Acous., Paper A3-8

FUJIMURA O. & LINDQVIST J. (1971)
Sweep-Tone Measurements Of Vocal-Tract Characteristics,
J. Acoust. Soc. Am. 19, 511-558.

GARDNER, GRAY & O'RAHILLY (1986)
Anatomy
W.B. Saunders Company 1986

ISHIZAKA K. & FLANAGAN J.L. (1972)
Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords.
B.S.T.J. 51, 1233-1268.

KELLY J. L. & LOCHBAUM C. C. (1962)
Speech synthesis
4th Int. Congr. Acoust., G42

LINDQVIST J. & SUNDBERG J. (1976)
Acoustic Properties of the Nasal Tract.
Phonetica 33, 161-168
STL-QPSR 1/1972, 13-17.

LONCHAMP F. (1988)
Etudes sur la production et la perception de la parole.
Thèse d'Etat NANCY II

MAEDA S. (1982)
The Role of the Sinus Cavities in the Production of Nasals Vowels.
I.C.A.S.S.P. 82, 911-914.

MAEDA S. (1984)
Une Paire de Pics Spectraux comme Correlat Acoustique de la Nasalisation des Voyelles.
13ème J.E.P. BRUXELLES 1984, 223-224.

MRAYATI M. (1976)
Contributions aux Etudes sur la Production de la Parole.
Modèles Electriques du Conduit Vocal avec Pertes, du Conduit Nasal et de la Source Vocale - Etude de leurs Interactions - Relations entre Disposition Articulatoire et Caractéristiques Acoustiques.
Thèse d'état, I.N.P. GRENOBLE.

VAN DEN BERG J. (1955)
Transmission of the Vocal Cavities
J. Acoust. Soc. Am. 27, 161-168

EVENEMENTS SUR DISCONTINUITES vs. EVENEMENTS SUR CONTINUITE

De la mise en évidence des patrons de phases

en français

SOCK R. DELATTRE C. ZILLIOX C. ZOHAIR L.

Institut de la Communication Parlée, CNRS UA 368
Institut de Phonétique de Grenoble Université III
38400 St. Martin d'Hères

ABSTRACT

With a view to extracting pertinent timing information from acoustic speech signals, we suggest to carry out measurements of acoustic phases - up to now reserved to articulatory-acoustic discontinuities - on more continuous speech signals. Thus provided with equivalent speech events, such signals would yield cycles and phases as discontinous ones. This should allow for an analysis of any type of speech signal, including continous types like [vowel + glide + vowel]. Moreover, this approach provides a principled link with highly continuous temporal functions of articulatory gestures (like lingual or mandibular position).

Finally, we will examine restructuring of overall temporal phasing patterns as strategic temporal cues, evidencing motor programming.

INTRODUCTION

L'étude de l'organisation temporelle dans la parole, sur le plan des régularités motrices (EMG et kinésiologiques), ou sur celui de la production acoustique, ou encore sur le plan des régularités dans le signal perçu (isochronie), nous offre un moyen privilégié d'apprehender les mécanismes qui peuvent mettre en relation les planifications et les programmations de l'action de parole avec les représentations phonologiques - que ce soit par le décèlement de structures coordinatives produisant des actions invariantes ou de programmes moteurs. Le lien entre la structuration temporelle des éléments de production de la parole et la distribution du *timing* dans leurs représentations phonologiques [1] reste cependant controversé. Certains plaident, au niveau articulatoire, pour une invariance relative de coordinations d'articulateurs oraux [2] (pour le français cf. [3]) ou de coordinations orolaryngées [4]. D'autres confirment retrouver cette invariance relative sur le plan acoustique [5]. D'autres, par contre, réfutent les recherches sur l'invariance de l'organisation temporelle relative pour une seule phase dans un cycle [6], parce qu'elles fournissent des résultats peu concluants [7] (y compris dans le plan de phase, cf. [8]) et préfèrent examiner les différences de structuration pluriphasique ([9] et [10]).

Nous maintenons cette dernière position en nous fondant maintenant sur de nouveaux résultats, et nous nous proposons dans cette étude de transposer la mesure de nos phases acoustiques (que nous n'avions rendue opérationelle jusqu'à présent que sur les discontinuités articulatori-acoustiques [11]) sur un signal de type plutôt continu, l'intensité de la parole. Doté d'événements équivalents, ce signal peut, non seulement nous fournir des cycles et des phases, comme le signal discontinu correspondant , mais il nous suggère encore la possibilité de traiter selon la même approche tout signal de parole, y compris les suites continues [voyelle + semi-voyelle + voyelle], et non exclusivement les suites typiquement discontinues [voyelle + plosive]. Enfin ce traitement nous rapproche de celui que l'on doit effectuer sur les fonctions temporelles fortement continues des gestes articulatoires (type position linguale ou mandibulaire).

La validation de ces événements sur signal continu se fera par une comparaison systématique des résultats obtenus sur les deux domaines d'investigations. Leur fiabilité dépendra, pour nous, du degré de proximité de ces résultats avec les deux types de mesures dans l'étude des structurations pluriphasiques.

Les phasages intersegmentaux (phases vocaliques et consonantiques) et intrasegmentaux (VOT, Etablissement de la Cible Vocalique) seront examinés pour deux classes phonétiquement proches : voyelle + consonne simple vs. voyelle + consonne double (dorénavant VC & VCC) dans deux cycles (cf. infra). L'opposition VC / VCC nous renvoie aux études sur le "doubling" et le "singling" [12], et plus récemment à celles sur les agrégations de gestes en dynamique dite sérielle [13] ou encore aux constellations gestuelles proposées par [14].

Notre paradigme d'étude pluriphasique sera emprunté aux études de psychomotricité basées sur le phasage ("phasing"), celles sur la locomotion en particulier [15] ; résultats repris, non pour leur mise en évidence de l'invariance des phases individuelles dans une tâche donnée (invariance, encore une fois non prouvée, malgré le dire des auteurs, cf. [16]), mais pour la méthode qu'ils nous offrent pour étudier les patrons de phases [9].

1. LES PARADIGMES LINGUISTIQUES

1.1. LE CORPUS

Nous avons utilisé le corpus de verbes français suivant : "empatter / empâter, téter / étêter, coter / écôter (forgé sur : enlever les côtes d'un légume), égoutter / goûter", auxquels nous avons ajouté le verbe "virer", pour l'étude des signaux acoustiques continus (avec les réalisations de type [ʁ]. Dans certains contextes, ces verbes donnent de véritables paires minimales, permettant de tester les effets de la gémination consonantique (consonnes doubles sans détente de la première) sur la quantité vocalique (avec des paires du type : "nous l'empattons ? / nous l'empâtons ?" vs. "nous l'empatt't-on ? / nous l'empâtt't-on ?" ou "nous les virons ? / "nous les vir(e)rons ?" , cf. [17] pour une description détaillée de ce corpus).

Chaque item a été enregistré 12 fois pour chaque locuteur sous deux conditions de débit : normale (conversationnelle) et rapide et ceci pour les voyelles [i], [a], [u], [e] et [o]. Nous avons obtenu 864 items par locuteur. Notre analyse se focalisera sur la voyelle [a] et sur une seule locutrice (V.K.), parmi huit autres locuteurs appartenant à une base de données plus importante. Notons, dès maintenant, que les résultats présentés et discutés ici sont représentatifs de l'ensemble des résultats obtenus dans nos études précédentes sur le phasage en timing acoustique.

1.2. LES MESURES

Les signaux numérisés ont été étiquetés manuellement en événements acoustiques, à l'aide d'un éditeur de signal [18]. Nous avons adopté deux procédures de mesure correspondant à deux champs d'investigations :

1.2.1. une détection d'événements sur le signal acoustique présentant des discontinuités ;

1.2.2. une détection d'événements sur la courbe d'intensité de la parole ne présentant pas forcément de discontinuités mais plutôt des inflexions.

Pour l'étiquetage des discontinuités sur le signal acoustique (Figure 1), nous nous sommes munis de trois événements parmi la série proposée par [11] :

- VVO et VVT (Vocalic Voiced Onset & Termination), liés à la fonction de transfert du conduit vocal sont, respectivement le début et la fin du voisement vocalique. Ils marquent, en effet, le début et la fin de l'état supraglottique vocalique associé à une excitation nettement périodique sans obstruction du conduit vocal ;
- CFO (Consonantal Frication Onset) marque le début de plosion-friction du burst. Il est la conséquence acoustique du relâchement (sur la Figure 1, bilabial pour le [p] et apico-dental pour le [t]).

En ce qui concerne l'étiquetage en événements sur le signal d'intensité (Figure 1) nous avons retenu trois événements :

- VTO (Vocalic Target Onset) est l'atteinte, en intensité, de la cible vocalique ; il caractérise le premier signe que l'on parvient à une intensité relativement stable par rapport aux changements rapides de l'établissement. N.B. : Cet événement est, bien entendu, différent de VVO ;
- CCO (Closure Cycle Onset) marque le début du cyle de closion ; il définit le premier signe d'apparition des conséquences sur l'intensité des gestes relativement rapides de fermeture pour la closion consonantique intervocalique. N.B. : Cet événement est différent de VVT ,
- RCO (Release Cycle Onset) marque le début du cycle détente. Il caractérise le premier signe d'apparition des conséquences des gestes d'ouverture (détente). NB. Cet événement est, au déphasage près, dû à la fenêtre de calcul du signal d'intensité (cf. post scriptum), identique à CFO

L'activité de production de la parole - cyclique dans une large mesure en ce qui concerne la modulation syllabique (avec une moyenne de 150 à 200 ms par syllabe, soit environ une fréquence de 5-6 Hz pour la mâchoire) -, nous a permis de suggérer ainsi deux cycles majeurs dans chaque domaine d'investigation :

- le cycle détente, soit la période s'étalant d'un relâchement consonantique (CFO) à un autre sur le signal acoustique discontinu. Sur l'intensité de la parole, ce cycle s'étend du premier signe d'éléments ouverts (RCO) à un autre ;
- le cycle closion, soit la période allant d'un début de disparition de la structure formantique définie pour la voyelle (VVT) à un autre. Sur le signal continu, ce cycle correspond à la période allant du premier geste fermant de la consonne (CCO) au suivant.

Cette étude se limitera aux cycles détentes continu et discontinu comme champs d'analyse privilégiés de l'organisation temporelle des phasages inter et intrasegmentaux (cf. [19] et [9]), en particulier pour étudier le "doubling" et le "singling" (ici l'opposition entre la structure géminée et non géminée).

Parmi les phases possibles, nous en avons retenu trois dans chaque cycle et pour les deux procédures de mesure :

- le VOT, sur le signal présentant des discontinuités (Voice Onset Time ou délai d'établissement du voisement [20], est la coordination temporelle qui s'étale du relâchement consonantique (CFO) à l'établissement d'une structure formantique définie pour la voyelle (VVO)). Voir [21] pour sa modélisation auditive en français et [22] pour sa caractérisation dans le système auditif périphérique. Sur le continu, cette phase "correspond" à l'Etablissement de la Cible Vocalique (dorénavant ETAB. Cib. Voc.) et marque la relation temporelle entre RCO et VTO.
- D.VOC., sur le signal présentant des discontinuités, est la phase dite Durée Vocalique. Elle présente une structure formantique définie et note la relation temporelle entre apparition et disparition de cette structure formantique (de VVO à VVT) sur le signal acoustique. Ce qui "correspond"

à la phase Cib. Voc. (de VTO à CCO) sur le signal d'intensité.

□ La phase Consonantique, habituellement appellée tenue consonantique s'étend de la disparition de la structure formantique (VVT) au relâchement consonantique (CFO) sur le signal à discontinuités. Sur le signal continu, la phase consonantique "correspond" à la coordination temporelle entre CCO et RCO.

2. RESULTATS ET DISCUSSION : DISCONTINUITES vs. CONTINUITE

Les phasages dans le cycle détente sont présentés sur les Figures 2 et 3, avec les dispersions des observations et les tendances des régressions.

2.1. Phasages Intersegmentaux

2.1.1. Phases vocaliques

Les effets de la vitesse d'élocution sur l'organisation temporelle relative de nos deux classes phonétiques (VC et VCC) se présentent comme suit :

□ Pour D.VOC., on constate une différence significative de phase en pourcentage (autour de 11% ; t=12.43) entre VC et VCC, accompagnée d'une différence entre les moyennes des deux cycles (autour de 69 ms) pour les deux classes ;

□ Pour la Cible Vocalique, la différence de phase est du même ordre en moyenne (autour de 11%), elle aussi significative (t=11.82), avec une différence (d'environ 71 ms), sur le cycle, entre les classes VC et VCC.

Notons que, dans les deux conventions de mesures, les phases vocaliques restent relativement invariantes pour les classes des non-géminées mais varient pour leurs homologues géminées. Contrairement aux autres locuteurs de cette même base de données qui maintiennent la phase vocalique relativement stable quelle que soit la vitesse d'élocution, notre locutrice a tendance à réduire en proportion la phase vocalique avec augmentation du cycle. Néanmoins, comme pour les études précédentes ([9] et [10]), la discrimination entre nos deux classes phonétiques reste efficace, puisque les classes VCC en vitesse d'élocution rapide ne se confondent jamais avec les classes VC en vitesse d'élocution normale (séparation des populations à 100%). Ceci contribuera donc à deux patrons temporels nettement différents pour les deux entités linguistiques.

2.1.2. Phases Consonantiques

□ Sur le plan du discontinu, on observe une différence significative de la phase consonantique entre la classe VC et VCC (autour de 12% ; t=13.53), accompagnée, bien entendu, de la même différence de cycle notée plus haut pour la phase vocalique ;

□ Sur le plan du continu, les différences sont semblables (autour de 12% aussi ; t=14.47).

Le comportement des phases prend une même allure dans les deux conventions de mesures : la classe VC reste relativement invariante alors que la classe VCC tend à augmenter avec le cycle ; cette évolution est significative seulement pour le discontinu.

2.2. Phasages Intrasegmentaux

□ Pour le VOT - donc sur le plan du discontinu - la tendance est de réduire cette phase avec l'augmentation du cycle pour la classe VC et de la maintenir stable pour la classe des géminées, VCC, malgré le changement de vitesse d'élocution. La différence de phase entre les deux classes phonétiques (2%) est tout juste significative (t=2.53) ;

□ Pour l'ETAB. Cib. Voc., sur le signal d'intensité, le scénario est le même et la différence des valeurs moyennes entre les phases (2%) est significative (t=4.06).

Il s'agit pour ces deux phases de petites différences de moyennes (2%), entrainant tout de même des effets significatifs, et non, contrairement aux autres phases, de séparations à 100% des classes.

**
*

L'examen des patrons de phases pour les deux méthodes de mesures nous révèle l'interchangeabilité des deux procédés, vu la similitude des résultats en structurations pluriphasiques.

Si l'on compare les deux classes linguistiques, on constate en effet que les patrons de phases présentent, d'une entité à l'autre, globalement deux structures nettement différentes. En détail, le passage d'un pattern à l'autre ne se fait pas de la même façon pour toutes les phases. Pour D. VOC (ou Cib. Voc.) et les phases consonantiques, l'évolution est plutôt discontinue : on observe une séparation à 100% des classes en passant de VC à VCC. Par contre, le passage d'une classe à l'autre pour le VOT (ou Etab. Cib. Voc.) est assuré de façon continue.

CONCLUSION

Avec une telle convergence des résultats pour les deux conventions de mesures, nous estimons pouvoir passer à l'étude du phasage sur tout signal continu. Ce fait révèle, en effet, l'avantage d'une lecture du signal de parole qui permet un passage plus flexible d'un domaine d'observation à un autre : du discontinu au continu acoustiques, du continu acoustique au continu de mouvement articulatoire.

**
*

Nous pouvons aussi réitérer ici, à partir de ces résultats, nos conclusions sur une étude précédente [9], posant qu'il est plus rentable d'examiner les restructurations globales des patrons de phases, en passant d'une catégorie phonétique à une autre, au lieu de rechercher simplement une invariance en timing relatif d'une seule phase. Les contenus des programmes moteurs (invariances et paramétrisations) inférés à partir d'indices stratégiques de timing, se trouveraient plutôt reposer sur des changements de structures pluriphasiques que sur la stabilité supposée de telle ou telle phase.

Il serait intéressant maintenant de pouvoir spécifier par modélisation dynamique, la participation des différentes phases à l'organisation temporelle de leurs cycles, examinant la compétition et la coopération entre articulateurs suite à la distribution des activations/inhibitions dans l'exécution d'une tâche spécifique de parole ([23] et [24]).

POST SCRIPTUM

Note sur le traitement du signal pour obtenir l'intensité : ce dernier est, bien entendu, déphasé par rapport au signal d'origine, cela étant dû à la fenêtre d'analyse (sur 15 ms). On vérifiera que les durées des deux cycles de détente détectés manuellement sur le signal d'intensité sont très voisines (avec une différence moyenne de 18 ms et un écart-type de +/- 16 ms).

REMERCIEMENT : A Christian ABRY, qui a suivi ce travail de près, pour ses nombreux commentaires.

REFERENCES BIBLIOGRAPHIQUES

[1] MacKAY D.G. (1987)
Temporal Organization of Perception and Action. A Theory of Language and other Cognitive Skills.
Cognitive Science Series. SEBRECHTS M.M. FISCHER G. & FISCHER P.M. (Eds.)
SPRINGER-VERLAG, New York, Berlin, Heidelberg.

[2] TULLER B. KELSO J.A.S. & HARRIS K.S. (1982)
Interarticulator Phasing as an Index of Temporal Regularity in Speech.
J. Exp. Psychol. HPP 8, 460-472.

[3] GENTIL M. (1986)
Organisation Temporelle du Système Articulatoire : Contributions Musculaires aux Gestes Labiaux, Linguaux et Mandibulaires.
Thèse d'Etat, Strasbourg.

[4] LOFQVIST A. & YOSHIOKA H. (1981)
Interarticulator Programming in Obstruent Production.
Phonetica 38, 21-34.

[5] WEISMER G. & FENNELL A.M. (1985)
Constancy of (Acoustic) Relative Timing Measures in Phrase-Level Utterances.
J. Acoust. Soc. Am. 78, 49-57.

[6] TULLER B. & KELSO J.A.S. (1984)
The Timing of Articulatory Gestures : Evidence for Relational Invariants.
J. Acoust. Soc. Am. 76, 1030-1036.

[7] BENOIT C. & ABRY C. (1986)
Vowel-Consonant Timing Across Speakers.
12th Int. Congr. Acoust. A6-1.

[8] NITTROUER S. MUNHALL K. KELSO J.A.S. TULLER B. & HARRIS K.S. (1986)
Patterns of Interarticulator Phasing Relations.
112th Meeting of Acoust. Soc. Am. Dec. 8th-12th.

[9] SOCK R. OLLILA L. DELATTRE C. ZILLIOX C. & ZOHAIR L. (1987)
Timings Intersegmental et Intrasegmental en Français.
16èmes JEP du GCP de la SFA, 233-236.

[10] ABRY C. ORLIAGUET J.P. & SOCK R. (1988)
Coordinations Temporelles des Phases Articulatoires dans la Parole.
Colloque de la Société Française de Psychologie (Dijon 29-30 janv.) : Automatisme et Contrôle. Les Processus Cognitifs Elémentaires et leur Intégration dans les Activités Cognitives Complexes. p.13.

[11] ABRY C. BENOIT C. BOE L.J. & SOCK R. (1985)
Un Choix d'Evénements pour l'Organisation Temporelle du Signal de Parole.
14èmes JEP du GCP du GALF, 133-137.

[12] STETSON R.H. (1951)
Motor Phonetics : a Study of Speech Movements in Action.
Amsterdam, North Holland.

[13] MUNHALL K. & LOFQVIST A. (1988)
Gestural Aggregation in Speech.
à paraître in PAW Review.

[14] BROWMAN C.P. & GOLDSTEIN L.M. (1986)
Towards an Articulatory Phonology.
Status Rept. Haskins Lab. SR 85, 219-250.

[15] SHAPIRO D.C. ZERNICKE R.F. GREGOR R.J. & DIESTAL J.D. (1981)
Evidence for Generalized Motor Programs using Gait Pattern Analysis.
J. Motor Behav. 13, 33-47.

[16] GENTNER D.R. (1987)
Timing of Skilled Motor Performance : Tests of the Proportional Duration Model.
Pyschological Review 94, 255-276.

[17] ABRY C. SOCK R. BOE L.J. OLLILA L. DOUBLIER D. DELATTRE C. & ZILLIOX C. (1986)
L'organisation Temporelle des Voyelles et des Consonnes du Français. Durée Phonologique et Vitesse d'Elocution.
Rapport CNET LANNION, 102p.

[18] BENOIT C. (1984)
EDISIG : Encore un Editeur de Signal ?!
13èmes JEP du GCP du GALF, 211-213.

[19] LEHISTE I. (1970)
Suprasegmentals.
M.I.T. Press, Cambridge.

[20] KLATT D.H. (1975)
Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters.
J. Speech Hearing Res. 18, 686-706.

[21] DELGUTTE B. (1984)
Codage de la Parole dans le Nerf Auditif.
Thèse d'Etat, Paris 6.

[22] WU Z.L. ESCUDIER P. SCHWARTZ J.L. & SOCK R. (1988)
Caractérisation d'Evénements Articulatori-Acoustiques dans un Modèle du Système Périphérique Auditif : Rôle de l'Inhibition Latérale et de l'Adaptation Nerveuse.
17èmes JEP du GCP de la SFA

[23] RUMELHART D.E. & NORMAN D.A. (1982)
Simulating a Skilled Typist : A Study of Skilled Cognitive-Motor Performance.
Cognitive Science 6, 1-36.

[24] JORDAN M. (1987)
Comm. pers.

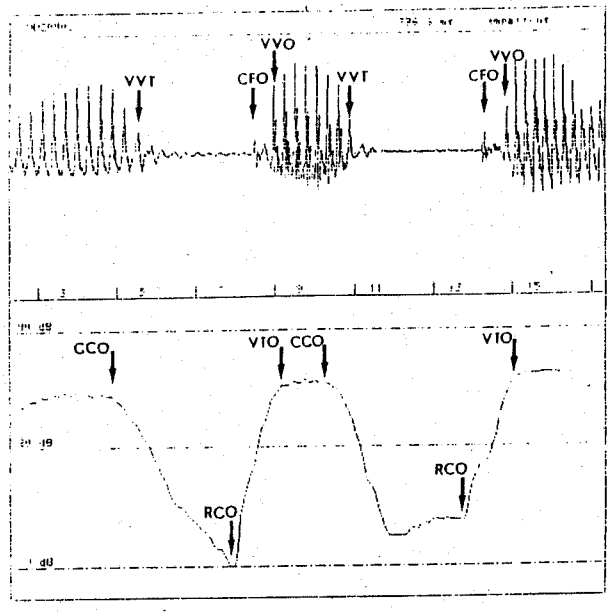


Fig. 1

En haut. Le signal de parole (échantillonné à 16 kHz sur 12 bits) extrait de la réalisation "la grue, nous l'empattons ?". 3 événements, VVT, CFO, et VVO repèrent les discontinuités majeures.

En bas. Le signal de l'intensité dérivé. Sur ce signal relativement continu, 3 événements, CCO, RCO, VTO, ont été repérés.

Pour la définition de ces événements cf. texte. NB.: Seul RCO correspond à CFO au décalage près, dû à la fenêtre d'intégration pour le calcul de l'intensité.

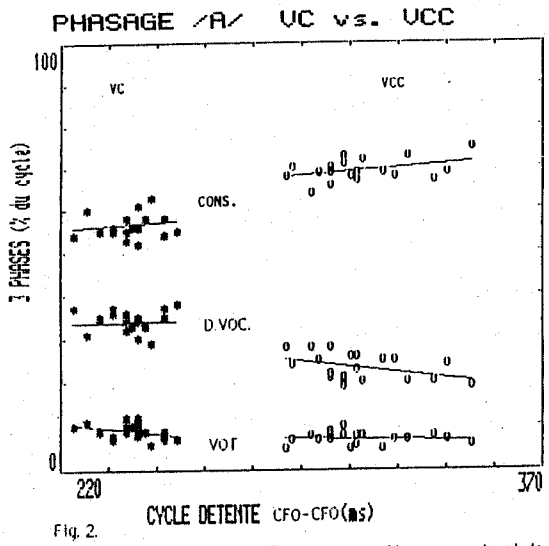


Fig. 2.

Structures de phasages pour 3 phases acoustiques sur signal de parole (discontinu), VOT (CFO-VVO), D.VOC. (VVO-VVT), CONS. (VVT-CFO), données dans le cycle détente (CFO-CFO) pour : "...nous l'empattons ?" (VC=*) contre "...nous l'empatt'l-on ?" (VCC=0), sous deux conditions de débit (normal et rapide). Locutrice V.K.

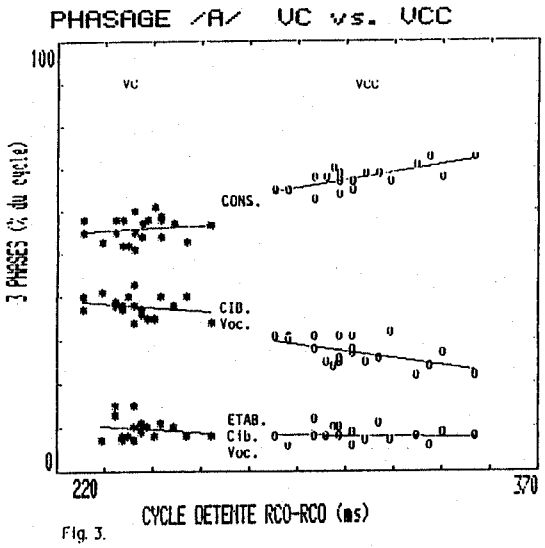


Fig. 3.

Structures de phasages pour 3 phases sur signal de l'intensité (continu), ETAB. Cib. Voc. (RCO-VTO), CIB. Voc. (VTO-CCO), CONS. (CCO-RCO), données dans le cycle détente (RCO-RCO) pour : "...nous l'empattons ?" (VC=*) contre "...nous l'empatt'l-on ?" (VCC=0), sous deux conditions de débit (normal et rapide). Locutrice V.K.

C.F. HELLWAG 200 ans après ou les éléments d'une fibre conductrice.

Louis-Jean BOE[1] & Pascal PERRIER[2]

Institut de la Communication Parlée, UA CNRS n° 368
[1]Institut de Phonétique de Grenoble [2]Laboratoire de la Communication Parlée

ABSTRACT · Theoretical studies and numerical procedures for inverting articulatori-to-acoustic transformations have permitted ATAL & al. (1978) to point out that articulatory regions ("fibers") can map into a single point in the acoustic space. If a given vowel can be produced with different vocal tract shapes, would it be possible to propose a general classification of vocalic systems on the basis of articulatory criteria induced from vocal tract configurations ? Our aim is to show that for cardinal vowels, mainly [i,a,u], different area functions giving the same acoustic output present similar values for three main parameters, namely : the distance from the glottis to the plane of maximum constriction in the vocal tract (Xc), the cross-sectional area (Ac) of the constriction and the lip opening (Al). Such results can be verified on simulation data obtained by ATAL & al. and on our more recent simulations. HELLWAG's propositions 200 years ago have therefore not been discredited.

1. INTRODUCTION

Dans le système de la communication parlée, le processus de l'encodage/décodage passe par une **divergence-convergence** (figure 1) :

➪ divergence due à la variabilité des comportements articulatoires intra- et inter-individuels et aux différentes stratégies contextuelles (les phénomènes de coarticulation) ;
➪ convergence bien obtenue, puisque la variabilité est finalement réduite pour que la communication soit assurée.

Nous nous proposons d'apporter des éléments de discussion, et des précisions permettant de situer le niveau articulatoire (plus précisément celui des fonctions d'aire du conduit vocal) dans ce processus de divergence-convergence, et ceci dans un cas bien déterminé, celui de la production des voyelles cardinales extrêmes [i,a,u].

En effet, entre la première description articulatoire "moderne" de l'espace vocalique et les problèmes soulevés par l'inversion du conduit vocal (la reconstitution de la forme articulatoire à partir du signal acoustique), bon nombre d'éclaircissements ont été apportés, mais d'autres problèmes ont été soulevés, peut-être sans que des relations aient été faites entre les différents champs de connaissance de la parole.

En 1781, il y a donc plus de 200 ans, Christoph Friedrich HELLWAG soutient sa thèse à la faculté de Médecine de TuBINGEN . à partir de considérations sur le lieu

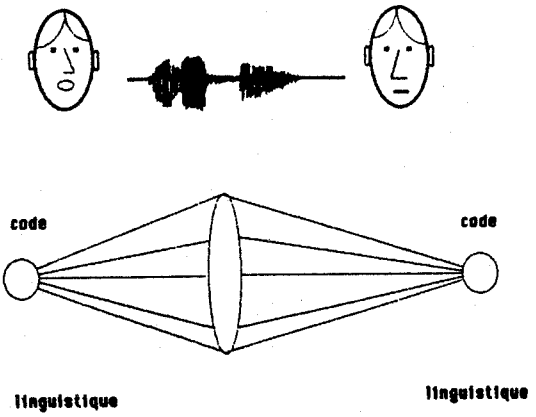


Figure 1

Le divergent - convergent de la Communication Parlée.

d'articulation et son aperture, il présente les voyelles sous forme d'une structure triangulaire qui va être systématiquement adoptée dans les descriptions phonétiques (certains lui préférant cependant un quadrilatère issu de la différence entre le a antérieur et le a postérieur). En 1948, P. DELATTRE dessine le triangle acoustique dans le plan F1/F2 (avec ses échelles inversées), faisant ainsi, à partir de la similitude des représentations, le lien entre l'articulatoire et l'acoustique. Par la suite, et dès les premières analyses acoustiques systématiques (G.E. PETERSON et H.L. BARNEY, 1952), va apparaître la variabilité inter- et intra-individuelle, qu'elle soit contextuelle ou non.

Partant de ces dispersions, et compte tenu de la relation non-linéaire entre l'articulatoire et l'acoustique, on peut se demander, pour une même voyelle :

➪ S'il existe une forme type du conduit vocal de la glotte aux lèvres

➪ S'il n'est pas possible de caractériser une forme du conduit vocal par un nombre limité de paramètres et de reconsidérer ainsi l'apparente diversité des formes qui engendrent un même son acoustique.

2. DES ELEMENTS DE DISCUSSION

Les travaux sur l'inversion (G UNGEHEUER, 1962, M SCHROEDER, 1967, P MERMELSTEIN, 1967, J.M. HEINZ, 1967, B S ATAL, J.J CHANG, M V MATHEWS & J TUCKEY, 1978 ; M.M SONDHI, 1979, F. CHARPENTIER, 1984) vont certes éclairer ces questions, mais, comme nous allons le montrer, ils vont aussi brouiller les pistes

En effet, si ces études vont bien asseoir théoriquement le problème, elles vont, en même temps, le complexifier en insistant sur la multiplicité des fonctions d'aire possibles, sans en faire, dans certains cas, une relecture à la lumière des connaissances de phonétique articulatoire accumulées depuis deux siècles.

La démarche de B S ATAL & al (1978) est, à ce point de vue, caractéristique. Les auteurs explicitent théoriquement la relation de non bi-univocité entre les espaces articulatoire et acoustique décrits respectivement par m et n paramètres. Ils avancent (pour m>n) la notion de **fibre**, ensemble de points ordonnés autour d'une trajectoire articulatoire et associés à une même configuration acoustique. Pour illustrer cette analyse, avec un modèle de production simplifié (m = 4) (K N STEVENS & A.S HOUSE, 1957), ils génèrent 30.720 formes articulatoires tout de même assez réalistes, décrites par les trois premiers formants (n = 3). Partitionnant l'espace acoustique en 1980 cubes (de dimension 50x50x100 Hz), ils extraient les fonctions d'aire de 8 voyelles de l'anglais-américain [i, ɪ, ɛ, æ, ʌ, ɔ, u, u], en utilisant les valeurs formantiques proposées par G.E. PETERSON & H.L. BARNEY (1952). Analysant des exemples de fonctions d'aire des voyelles extrêmes [i, a, u], ils insistent sur leur diversité et concluent à l'extrême variabilité de la forme articulatoire **"Large changes in the shape of the vocal tract can be made without changing the formant frequencies. These changes are consistent with the hypothesis that compensatory articulation is a possibility -that is, different people can produce the same sound with different vocal tract shapes. They are also consistent with the art of ventriloquism. In particular, many sounds can be produced with a wide range of mouth openings".**

C'est aussi une des observations de M.M. SONDHI (1979) qui, reprenant les fonctions d'aire publiées par B.S. ATAL, note **"the area differs considerably from each other and are all 'reasonable' in the sense that they may well be attained by a human vocal tract".**

Aussi, pour trouver un ordonnancement parmi toutes ces possibilités compensatoires, B S ATAL & al suggèrent d'explorer des pistes dans le domaine de la dynamique **"It seems worth investigating whether some minimum motion or minimum energy principle is applied in going from one sound to another".**

Si l'on suit tous ces auteurs dans leurs conclusions, l'espace géométrique des fonctions d'aire se situerait donc nettement dans la partie divergente de la figure n° 1 et la convergence ne s'amorcerait qu'à partir du niveau acoustique. Point de vue dont il faut bien tirer la conséquence : les formes du conduit vocal (fonctions d'aire) ne permettraient pas une classification des voyelles, les variantes de production étant trop importantes.

Faire appel à des principes tels que celui de l'énergie minimale nous semble certes riche en promesses, et en particulier pour l'étude des transitions dynamiques, mais déjà au niveau strictement statique il est possible de clarifier les problèmes posés par cette apparente multiplicité des formes possibles. Avant de passer à une autre lecture des fonctions d'aire de B.S. ATAL & al., et de présenter celles que nous obtenues avec un modèle articulatoire plus affiné, celui de S. MAEDA (1979), il nous semble ainsi important d'en appeler maintenant à deux types de résultats.

1) Ceux qui proviennent de mesures du lieu de constriction obtenues sur des documents radiographiques.

A partir de données portant sur 13 langues et 40 sujets, S WOOD (1979) relève, sans exception, quatre zones de constriction (palais dur, palais mou, haut pharynx et bas pharynx), qui permettent d'établir une typologie des voyelles : **"Each location is appropriate for a definable class of vowels qualities".**

2) Ceux qui sont obtenus dans de conditions de contraintes de production statiques ou dynamiques.

T GAY, B LINDBLOM & J LUBKER (1981) et J. LUBKER & T. GAY (1981) ont imaginé et exploité un paradigme expérimental ingénieux : si on fixe (à l'aide de "bite-blocks") la mâchoire d'un locuteur, ou si l'on perturbe dynamiquement sa position, celui maintient quand même le lieu de retrécissement et son aperture. Ils en concluent : **"Speakers are able to achieve articulatory goals at least spatially, and probably temporally, even in the face of rather considerable constraints imposed upon articulatory movement. That is, in the presence of constraints upon the movements of articulator, speaker will, if at all possible, reorganize both the spatial and temporal movements of other articulators in order to accomplish the original articulatory goal".. " The target of a vowel is coded neurophysiologically in terms of area function related information and is specified with respect to the acoustically most significant area function features, the points of constriction along the length of the tract".** En

d'autres termes, la variabilité articulatoire se manifeste bien au niveau du jeu des articulateurs, mais les possibilités compensatoires sont utilisées pour atteindre des buts spatiaux et temporels bien spécifiés au niveau de la fonction d'aire.

Nous pouvons maintenant faire un premier bilan de ces trois types de données et de résultats.

1) Les possibilités compensatoires nous permettent de réaliser un même produit acoustique avec des commandes articulatoires différentes. B. ATAL & al., font bien la preuve, par simulation, que l'on peut se déplacer le long d'une **fibre articulatoire** de commande, tout en obtenant des réalisations présentant des F1, F2, F3 parfaitement identiques. Et T. GAY, J. LUBKER & B. LINDBLOM montrent bien, de leur côté, que les locuteurs arrivent à produire la même voyelle (avec un lieu de rétrécissement et une aperture quasi identiques), malgré les perturbations apportées sur un articulateur. **Il y a donc bien "convergence" entre le niveau des commandes et le niveau acoustique.**

2) Mais apparemment, les résultats semblent contradictoires, lorsqu'il s'agit de localiser la fonction d'aire par rapport à la divergence-convergence. Si l'on s'en tient aux conclusions de B.S. ATAL & al., les fonctions d'aire se situent nettement dans la divergence. Au contraire, les expériences sur les perturbations de T. GAY, J. LUBKER & B. LINDBLOM et les mesures de S. WOOD tendent à montrer le contraire. Pour des manoeuvres articulatoires différentes, des langues différentes, des locuteurs différents, une même voyelle présente, déjà au niveau de la fonction d'aire, des caractéristiques bien repérables au niveau de la constriction et/ou de l'aire aux lèvres. D'ailleurs, si les nomogrammes de G. FANT (1960), établis à partir d'un modèle rudimentaire (le modèle à 4 tuyaux) ont une telle valeur explicatoire c'est bien parce que ces paramètres sont réellement opératoires. Nous avons vérifié, avec ce modèle, que des variations importantes (± 25%) des cavités d'avant et d'arrière par rapport à la valeur standard (8 cm2) n'entraînent que des modifications du second ordre sur les formants (P. BADIN & L.J. BOE, 1987, P. BADIN, L.J. BOE, P. PERRIER & C. ABRY, 1988). Pour pouvoir faire une interprétation des fonctions d'aire, et a fortiori pour pouvoir apprécier les différences entre ces fonctions d'aire, il faut connaître, par rapport à l'acoustique quels en sont les paramètres pertinents, et quelles sont leurs "sensibilités" dans l'espace acoustique.

Comme nous l'avons vu, les "**zones cruciales**" de la fonction d'aire sont bien la position Xc du rétrécissement, son aperture Ac et l'aire aux lèvres Al. En ce qui concerne les caractéristiques acoustiques de ces paramètres, les travaux que nous avons menés sur les macro-sensibilités des commandes articulatoires du modèle de S. MAEDA, (lèvres, mâchoire et pointe, dos et corps de la langue), fournissent bon nombre d'informations (R. MAJID & al., 1987). Ainsi [i] et [a] sont très sensibles aux variations sur Xc et Ac, et [u] à celles de Al.

C'est donc à la lumière de ces connaissances que doit être faite la lecture d'une fonction d'aire et elle doit porter, en premier lieu, sur ces points cruciaux.

Il est possible de revenir maintenant sur les fonctions d'aire présentées par B.S. ATAL & al. pour [i], [a], [u] et jugées si différentes par leurs auteurs (figure 2). On peut noter que :

⇨ Les [i] (F1 = 292 F2 = 2097 F3 = 2800 Hz) ont bien les mêmes Xc (12.5 cm) et Ac (0.3 cm^2), avec des Al très différents.

⇨ Les [a] (F1 = 653 F2 = 1188 F3 = 2177 Hz) ont pratiquement les mêmes Xc (2.5 cm ± 0.5 cm) et Ac (0.3 cm^2) avec des Al variables mais > 4 cm^2.

⇨ Les [u] (F1 = 376 F2 = 806 F3 = 2328 Hz), qui soit dit en passant sont trop postérieurs pour être vraiment représentatifs (leurs Xc varient entre 5-7 cm au lieu des 8-10 cm généralement relevés. G. FANT, 1960, S. WOOD, 1979), ont bien une tolérance à Xc (environ 2 cm), mais aussi à Al (presque 6 cm^2) contrairement à nos remarques précédentes. Mais il faut noter ici qu'une variation aussi grande qui n'affecterait que la toute dernière section aux lèvres ne semble pas être morphologiquement réaliste.

A cette dernière remarque près, les résultats de B.S. ATAL & al. ne présentent donc aucune contradiction avec les mesures radiographiques ni avec celles qui sont déduites des productions avec contraintes.

On peut donc situer la fonction d'aire, caractérisée par les zones cruciales, dans la partie convergente du processus de communication.

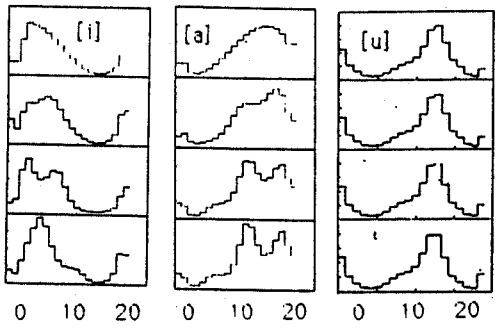


Figure 2 (D'après ATAL et al. 1978)

Exemples de fonctions d'aires pour les voyelles [i], [a], [u].

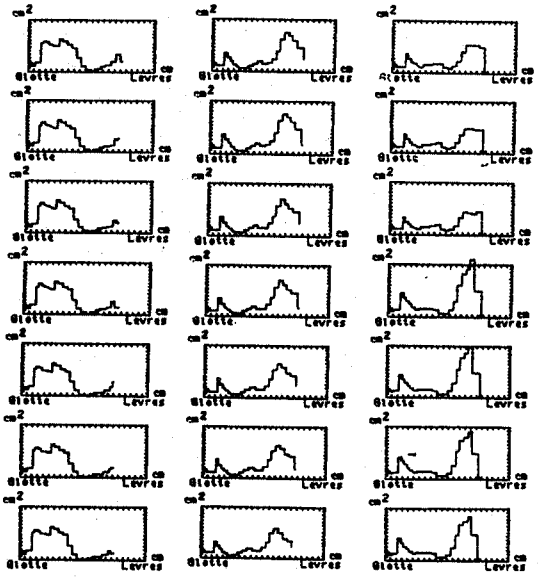
3. DES RESULTATS COMPLEMENTAIRES

Pour leurs simulations, ATAL & al. avaient utilisé un modèle relativement simplifié. Nous avons adopté celui de S. MAEDA, élaboré à partir de données articulatoires et intégrant bien les connaissances phonétiques. Il permet de générer une coupe sagittale à partir de 5 paramètres : trois pour la langue (corps, dos, pointe), un pour la mâchoire et un pour les lèvres.

Pour obtenir la fonction d'aire, nous avons élaboré des coefficients de passage (H. SANCHEZ & L.J. BOE, 1984) qui se se sont révélés, aprés évaluation, relativement satisfaisants. En quadrillant linéairement l'espace maximal de commande, et en utilisant une simulation harmonique (D DEGRYSE, 1981), nous avons constitué un dictionnaire de plus de 200.000 configurations vocaliques. Nous avons sélectionné dans ce dictionnaire, toutes les voyelles correspondant à des ellipses de dispersion (distibution gaussienne bi-dimensionnelle), considérées comme représentatives des voyelles du français (tableau I). Nous nous limitons au plan F1/F2, qui ,en français, est suffisamment classificatoire pour qu'il ne soit pas nécessaire de faire appel à F3).

	F1	σ	F2	σ	R (F1,F2)
i	293	20	2247	120	-0.10
e	371	35	2145	105	-0.96
ɛ	538	60	1800	150	0.87
a	650	60	1350	200	-0.67
y	285	20	1777	90	0.40
ø	401	60	1691	150	0.0
œ	531	60	1485	150	-0.71
u	295	25	734	125	0.1
o	419	25	867	175	0.0
ɔ	519	60	1090	175	0.0

TABLEAU I. Les valeurs moyennes des formants, leurs écarts type et coefficients de corrélation, à partir desquelles ont été sélectionnées les fonctions d'aire dans le dictionnaire.



3.a . Voyelle [i] 3.b . Voyelle [a] 3.c . Voyelle [u]

Figure 3

Fonctions d'aire représentatives obtenues pour [i], [a], [u] à partir du modèle articulatoire de S. MAEDA.

A titre d'exemple, nous presentons en figures 3, des fonctions d'aire des voyelles cardinales extrêmes [i, a, u]. Pour l'ensemble de ces 3 voyelles (au total 12 000 environ, avec une grosse majorité de [a] et comparativement peu de [i]) et compte tenu du fait que notre sélection est beaucoup moins sevère que celle de B S. ATAL (qui était, rappelons-le, au Hertz et au Decibel près pour les 3 premiers formants) on vérifie bien l'importance des zones cruciales (figures 4). Pour [i] le lieu d'articulation et l'aperture ont très précisement comme valeurs 11 cm et 0.30 cm^2, pour [a], la zone de constriction peut varier de ± 1 cm autour de 5 cm cette variation peut paraître grande mais nous avons, en fait, ratisse un peu large (des écarts type de 60 et 200 Hz pour F1 et F2) et ceci reste à définir avec plus de précision . pour [u] les lèvres restent toujours fermées (moins de 0.50 cm2), avec un lieu d'articulation situé autour de 10.5 cm (± 0.5 cm).

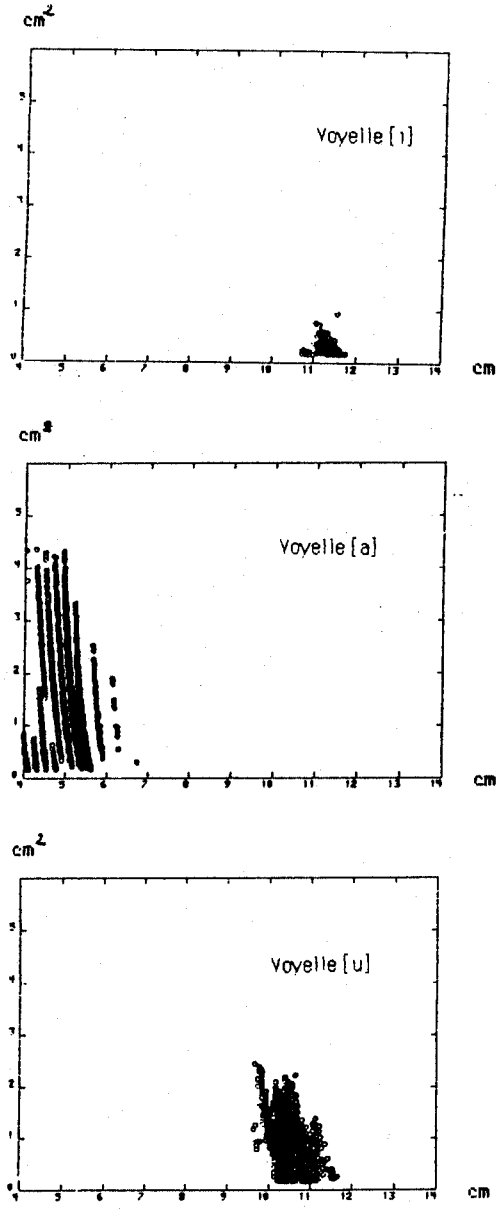


Figure 4.a . Plan (Xc,Ac)

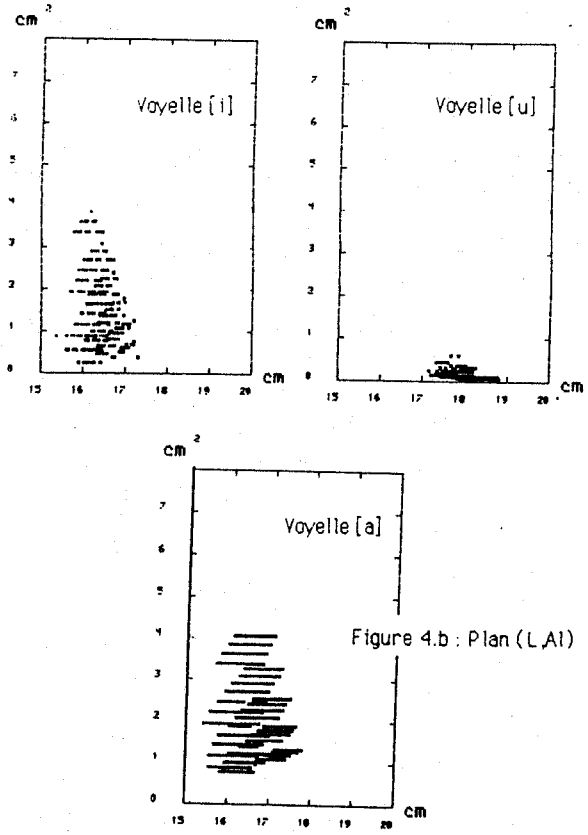


Figure 4.b : Plan (L,Al)

Figure 4

Pour [i], [a], [u] : Ac=f(Xc) et Al=f(L)

Xc, Ac + son environnement

4. CONCLUSION

Du triangle de C.F. HELLWAG, aux fibres articulatoires de B.S. ATAL & al., jusqu'aux simulations obtenues avec le modèle de S. MAEDA, en passant par les résultats de S. WOOD et la démonstration de T. GAY, J. LUBKER & B. LINDBLOM, il n'y a pas de "rupture épistémologique" qui partagerait les descriptions articulatoires vocaliques en descriptions "périmée" et "sanctionnée". La zone de retrécissement, son aperture et l'aire aux lèvres sont des bons paramètres descriptifs et classificatoires pour les voyelles cardinales extrêmes. Il suffit de faire une lecture des fonctions d'aire par rapport aux zones cruciales.

Il nous est donc possible, maintenant, de positionner les fonctions d'aire sur le divergent-convergent de la communication ; pour compléter le schema (figure 5), nous rajouterons, dans la convergence, l'intégration perceptive effectuée sur les masses d'énergies (cf. J.L. SCHWARTZ, 1987, pour plus de détail).

Les possibilités compensatoires des articulateurs (lèvres, mâchoire, langue) se situent dans la divergence, la convergence s'amorce déjà au niveau de la fonction d'aire, à condition d'en faire une lecture par rapport aux zones cruciales, et se poursuit jusqu'au code, avec le niveau acoustique paramétrisé par les formants, masses d'énergie intégrées au niveau perceptif.

Post-Scriptum : En ce qui concerne les voyelles cardinales centrales, il semble cependant sur les toutes premières simulations que pour une même position du lieu de constriction on puisse compenser acousti- quement une augmentation de l'aire aux lèvres par une diminution de l'aire au lieu de constriction.

Remerciements

A Ben LAAUWEN pour nous avoir aidé à traduire des passages de la thèse de C.F. HELLWAG (version hollandaise)

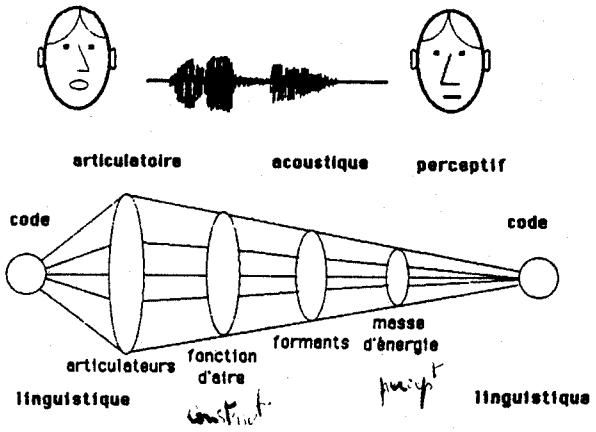


Figure 5

Le divergent - convergent de la Communication Parlée.

REFERENCES

ATAL B.S. CHANG J.J. MATHEWS M.V. TUCKEY J. (1978) Inversion of Articulatory-to-Acoustic Transformation in the VocalTract by a Sorting Technique. J. Acoust. Soc. Am. 63, 1535-1555.

BADIN P. & BOE L.J. (1987) Vocal Tract Vocalic Nomograms Acoustic Considerations. XIth Int. Congr. Phonetic Sci. 2, 352-355.

BADIN P. & BOE L.J. PERRIER P. & ABRY C. (1988) Vocalic Nomograms Acoustic Considerations upon Formant Convergences. Bull. Lab. Comm. Parlée 2, 65-94.

CHARPENTIER F. (1984) Determination of the Vocal Tract Shape from the Formants by Analysis of the Articulatory-to-Acoustic Nonlinearities Speech Comm. 3, 291-308

DEGRYSE D. (1981) Temporal Simulation of Wave Propagation in the Lossy Vocal Tract. 4th F.A.S.E. Symposium I, 193-200

FANT G. (1960) Acoustic Theory of Speech Production. Mouton.

GAY T. LINDBLOM B. & LUBKER J. (1981) Production of Bite-Block Vowels : Acoustic Equivalence by Selective Compensation J. Acoust. Soc. Am. 69, 802-810.

HELLWAG C.F. (1781) Formatione Loqvelae. Dissertatio Inavgvralis Physiologico Medica. Tubingae Literis Fvesianis.

HEINZ J.M. (1967) Perturbation Functions for the Determination of Vocal Tract Area Functions from Vocal Tract Eigenvalues. STL/QPSR, 1-14.

LUBKER J. & GAY T. (1981) Spatio-Temporal Goals : Maturational and Cross-Linguistic Variables In Speech Motor Control, 205-215. Pergamon Press, Oxford.

MAEDA S. (1979) Un modèle articulatoire de la langue avec des composantes linéaires. 10° JEP du GALF, 152-162.

MAJID R. ABRY C. BOE L.J. & PERRIER P. (1987) Contribution à la classification articulatori-acoustique des voyelles. Etude des "macro-sensibilités" à l'aide d'un modèle articulatoire. XIth Int. Congr. Phonetic Sci. 2, 348-351.

MERMELSTEIN P. (1967) Determination of the Vocal Tract Shape from Measured Formant Frequencies. J. Acoust. Soc. Am. 41, 1283-1294.

PETERSON G.E. & BARNEY H.L. (1952) Control Methods used in a Study of the Vowels. J. Acoust. Soc. Am 24, 175-184.

SANCHEZ H. & BOE L.J. (1984) De la coupe sagittale à la fonction d'aire du conduit vocal. Bull. Inst. Phonétique de Grenoble 13, 1-24

SCHROEDER M. (1967) Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements. J. Acoust. Soc. Am. 41, 1002-1010.

SCHWARTZ J.L. (1987) Représentations Auditives de Spectres Vocaliques. Thèse de Doctorat es Sciences, Institut National Polytechnique de Grenoble, p. 393

SONDHI M.M. (1979) Estimation of Vocal-Tract Areas : the Need for Acoustical Measurements IEE Trans ASSP 27, 268-273.

STEVENS K.N. & HOUSE A.S. (1955) Development of a Quantitative Description of Vowel Articulation. J. Acoust. Soc. Am 27, 484-493

UNGEHEUER G. (1962) Elemente Einer Akustischen Theorie der Vokalartikulation. Springer Verlag, Berlin.

WOOD S. (1979) A Radiographic Analysis of Constriction for Vowels. J. Phonetics 7, 25-43.

Perception et production

TONPER, Un test de perception pour langues tonales : Application au bulu (Sud Cameroun)

Jean-Marie Hombert

LAPHOLIA - Université Lumière/Lyon 2 & LACITO - CNRS

ABSTRACT

Data on tone perception are scarce. The purpose of this paper is to describe a perceptual test, TONPER, which will facilitate the collection and comparison of tone perception data. As an exemple of the use of this test we present results from an experiment on mono and bisyllabic tone patterns in Bulu (a Bantu language spoken in Southern Cameroon).

1. Introduction

Bien que plus d'un quart des langues du monde soient des langues à tons [1] [2] [3], les données phonétiques précises sur ces systèmes tonals - en particulier dans le domaine perceptuel [4] [5] - sont rares. Ceci s'explique par le fait que la plupart de ces langues sont parlées par relativement peu de locuteurs et sont, en outre, souvent situées dans des zones difficiles d'accès (Afrique subsaharienne, sud-est asiatique, Nouvelle Guinée...).

Nous proposons ici un test de perception des tons qui, à partir de stimuli synthétiques, permet de mettre en évidence les indices acoustiques qui jouent un rôle déterminant dans la discrimination des tons d'une langue donnée.

Ce test, inspiré d'un protocole que nous avons déjà utilisé pour l'analyse des systèmes vocaliques [6] [7], a été conçu avec une triple préoccupation :

a. Être facile à administrer de manière à pouvoir être utilisé sur le terrain avec des locuteurs n'ayant aucune expérience de ce type de test.
b. Ne pas être lié à un système d'écriture. Une très large majorité des langues à tons sont des langues dites "à tradition orale", il est donc important que le test utilisé ne soit pas dépendant d'une transcription graphique.
c. Être universel, c'est-à-dire que les mêmes stimuli puissent être utilisés quel que soit le système tonal testé.

2. Déroulement du protocole expérimental

2.1 Première étape : analyse tonologique de la langue considérée

Il s'agit ici de déterminer les schèmes tonals contrastifs dans la langue étudiée et de les illustrer par des paires (presque) minimales. Sont choisis de préférence des items qui peuvent être facilement dessinés (cf. 2.3.2).

2.2 Seconde étape : Entraînement des sujets

Le but de cette étape est d'apprendre aux sujets à dissocier l'information segmentale du contour mélodique. Les items retenus au cours de la première étape sont prononcés par le sujet. Il leur est également demandé de murmurer le schème tonal associé à chacun de ces items. Ces productions sont par ailleurs enregistrées et feront l'objet d'une analyse acoustique ultérieure.

2.3 Troisième étape : Test

2.3.1 Stimuli

- Pour les schèmes tonals monosyllabiques, 13 stimuli sont utilisés : 5 tons ponctuels, 4 tons montants et 4 tons descendants (voir Fig. 1). La voyelle porteuse de ces stimuli est un [ə] (F1 = 500 Hz, F2 = 1500 Hz et F3 = 2500 Hz) d'une durée de 250 ms.
- Pour les schèmes tonals dissylabiques, ces 13 stimuli sont précédés par un ton ponctuel bas (110 Hz), moyen (130 Hz) ou haut (150 Hz).

2.3.2 Feuille réponse

Une feuille-réponse est préparée pour chaque langue étudiée. Cette feuille-réponse contient un nombre de cases égal au nombre de tons contrastifs plus un. Dans chacune des cases est dessiné l'un des items retenus lors de l'analyse tonologique ; la case supplémentaire contient une croix.

2.3.3 Tâche des sujets

Après la présentation de chacun des stimuli (chaque stimulus est présenté 50 fois), le sujet doit indiquer sur la feuille-réponse le dessin de l'item qui a le schème tonal correspondant à celui du stimulus. Si aucun schème tonal ne lui convient, il doit indiquer la case supplémentaire marquée d'une croix.

3. Application au système tonal du bulu

3.1 Présentation du système tonal

Le bulu est une langue bantu du sud Cameroun. Elle comporte trois tons contrastifs. Deux d'entre eux, situés dans la partie basse du registre, sont perceptuellement proches. Le premier (ton A, noté par un accent grave au-dessus de la voyelle) semble être légèrement descendant, le second par contre semble avoir une fréquence fondamentale stable (ton B, noté par un accent grave sur la voyelle suivi d'un ˙ sur le segment adjacent). Le troisième ton est un ton haut (ton C, noté par un accent aigu sur la voyelle).

Ces trois tons forment des schèmes contrastifs sur les substantifs mono et dissyllabiques comme l'illustrent les exemples suivants :

Monosyllabiques

Ton A :	kɔ̄ s	*perroquet*
Ton B :	kɔ̀ s̀	*poisson*
Ton C :	kú p	*poule*

Dissyllabiques

Schème A :	è bà ŋ	*banane douce*
Schème B :	è bà ǹ	*pierre*
Schème C :	è bá ŋ	*fagot*

Les valeurs de F0 pour ces exemples sont présentés en Fig. 2 (pour les monosyllabiques) et en Fig. 3 (pour les disyllabiques). Les figures 4 et 5 représentent les feuilles-réponses correspondantes.

3.2 Analyse des résultats

Les réponses du sujet (originaire de Sangmélima) aux stimuli monosyllabiques (tableau 1) indiquent clairement que le ton A se caractérise par une fréquence fondamentale descendante, de faible pente et située vers le bas du registre. Le stimulus le plus représentatif du ton A a une pente de 10 Hz (de 120 à 110 Hz). Si la pente est trop forte (par exemple de 150 à 110 Hz) ou bien si elle se situe dans la partie supérieure du registre (par exemple de 150 à 130 Hz) alors le ton A n'est pas identifié. A noter toutefois que 4% (2 / 50) des stimuli à F0 stable à 110 Hz ont été identifiés comme ton A. Le ton B, quant à lui, nécessite une F0 stable située dans la partie inférieure du registre (le stimulus à 120 Hz étant celui qui donne le pourcentage d'identification le plus élevé : 100%). Enfin le ton C peut être soit stable soit légèrement montant dans la partie supérieure du registre (entre 140 et 150 Hz). Les stimuli qui ont un ΔF0 supérieur où égal à 20 Hz (e.g. 130-150 Hz) ne sont plus perçus comme ton C.

Les résultats de l'identification des schèmes dissyllabiques sont présentés dans les tableaux 2 (lorsque le premier ton du schème est bas : F0 = 110 Hz) et 3 (lorsque le premier ton du schème est moyen : F0 = 130 Hz). Lorsque ce premier ton est haut (F0 = 150 Hz), aucun stimulus - quelque soit le second ton du schème - n'est identifié comme séquence tonale possible du bulu.

Le schème A n'est perçu que dans 2% (1 / 50) des réponses aux stimuli dont le schème tonal comporte un ton moyen sur la première syllabe (cf. tableau 3) et un ton légèrement descendant sur la seconde (120-110 Hz).

La perception du schème B prend clairement en compte la hauteur relative des deux tons qui constituent le schème. Ce schème B sera identifié à 100% lorsque les F0 des deux syllabes sont au même niveau (F0 = 110 Hz pour le tableau 2, F0 = 130 Hz pour le tableau 3). Ces résultats montrent bien l'importance du "contexte" dans la perception des tons. En effet, la même F0 (i.e. 110 Hz sur la seconde syllabe) n'est jamais perçue comme schème B lorsque la F0 de la première syllabe est à 130 Hz (tableau 3) alors qu'elle l'est toujours lorsque la F0 de la première syllabe est de 110 Hz (tableau 2). La même constation peut être faite pour les stimuli dont la F0 de la seconde syllabe a une valeur de 130 Hz.

L'importance du contexte apparaît également dans l'identification du schème C. Comme nous l'avions déjà montré avec les stimuli monosyllabiques, nous obtenons une identification parfaite du schème C lorsque la F0 (ici de la seconde syllabe) est dans la partie supérieure du registre, qu'elle soit stable à 150 Hz ou légèrement montante de 140 à 150 Hz. Toutefois, si l'on compare les colonnes C des tableaux 2 et 3 on constate que lorsque la première syllabe a un ton bas (F0 = 110 Hz, tableau 2) les stimuli à seconde syllabe stable (à 140 ou 130 Hz) ou montante (de 130 à 150 Hz) sont identifiés comme schèmes C alors qu'ils ne le sont pas (ou beaucoup moins) lorsque le ton de la première syllabe est moyen (F0 = 130 Hz, tableau 3).

4. Conclusion

En résumé, les caractéristiques suivantes semblent se dégager : Le ton A a une pente légèrement descendante située dans la partie inférieure du registre. Le ton B, bien que localisé dans la même zone fréquentielle, s'en distingue par une F0 stable. La partie supérieure du registre est occupée par le ton C qui peut être soit stable, soit très légèrement montant.

Lorsque ces tons apparaissent comme second ton dans des schèmes dissyllabiques leur perception est influencée par la valeur de la F0 du ton qui les précède. Le schème B sera d'autant mieux identifié que les deux tons qui le constituent sont à des niveaux très proches (tout en restant dans la partie inférieure du registre). Le second ton du schème C doit être dans la partie supérieure du registre mais débuter avec une F0 supérieure d'au moins 10 Hz à la F0 du ton de la première syllabe. L'examen de la figure 3 indique que le schème A est produit avec une F0 de la seconde syllabe nettement plus faible que celle de la première syllabe. Nos stimuli ne couvraient pas cette zone fréquentielle puisque les valeurs les plus faibles de F0 étaient à 110 Hz. Ceci explique probablement pourquoi le schème A n'a pratiquement pas été identifié. Il est également possible que les identifications de ce schème soient améliorés par une F0 décroissante sur la première syllabe (cf. figure 3). Ceci ne pourra être testé que par la fabrication d'un jeu de stimuli plus étendu.

Afin de permettre la distinction entre les caractéristiques générales et les spécificités individuelles qui conditionnent la perception des tons en bulu, ce test doit évidemment être étendu à d'autres locuteurs de cette langue. Nous espérons toutefois avoir montré par cet exemple que ce test est bien adapté à la collecte de données perceptuelles sur les systèmes tonals.

Remerciements

Je tiens à remercier Monsieur Jean Ndoo Mengo pour sa participation à ces tests perceptuels.

RÉFÉRENCES

[1] RULHEN M., 1975, *A guide to the languages of the world*, Stanford University.

[2] MADDIESON I., 1978, "Universals of tone", in J.H. GREENBERG (ed.), Universals of Human Language, vol. 2, Phonology, Stanford University Press, pp. 335-365.

[3] HOMBERT J.M., 1984, Phonétique expérimentale et diachronie : Application à la tonogénèse, Thèse d'État, Université de Provence.

[4] GANDOUR J.T., 1978, "The perception of tone", in V.A. FROMKIN (ed.), *Tone: a Linguistic Survey*, Academic Press, pp. 41-76.

[5] HOMBERT J.M., 1976, "Perception of tones of bisyllabic nouns in Yoruba", Studies in African Linguistics, suppl. 6, pp.109-121.

[6] HOMBERT J.M., 1979, "Universals of vowel systems: the case of centralized vowels", *Proceedings of the Ninth International Congress of Phonetic Sciences - Copenhagen*, vol. 2, pp. 27-32.

[7] HOMBERT J.M., G. PUECH, 1984, "Espace vocalique et structuration perceptuelle : application au swahili", *Pholia I*, Université Lyon 2, pp. 199-208.

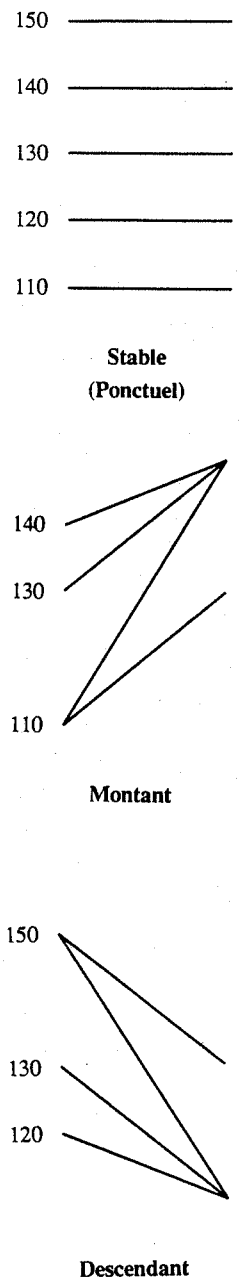
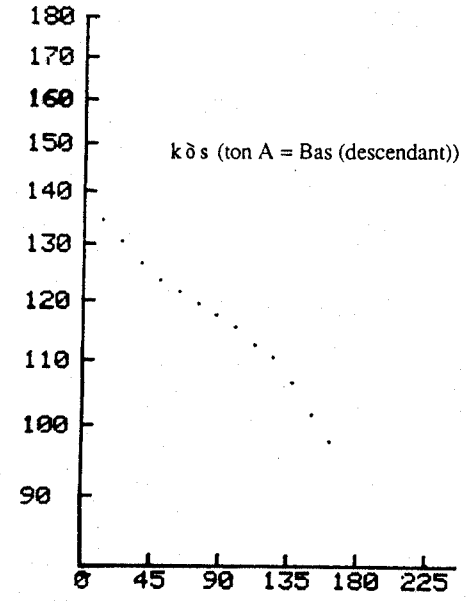
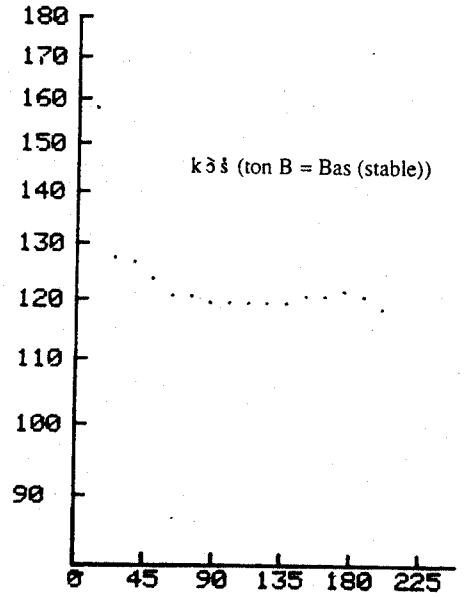
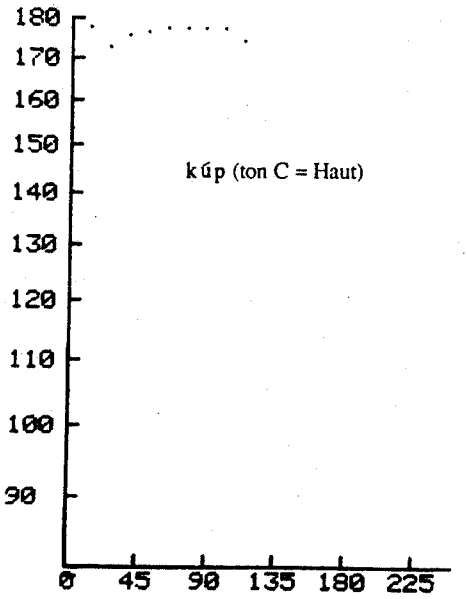
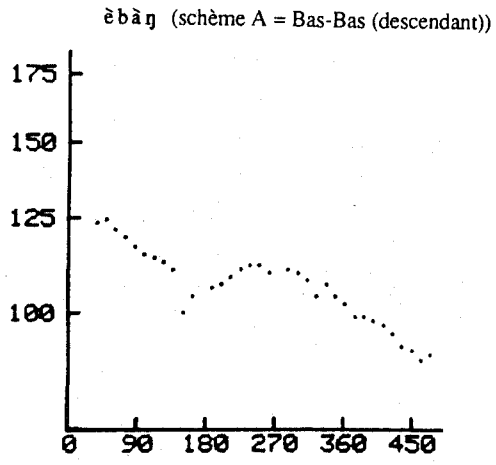


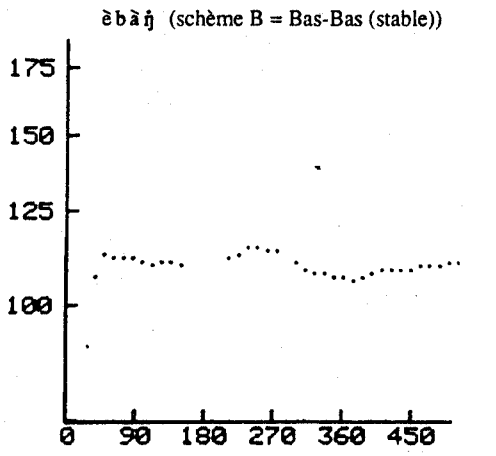
Fig. 1 Fréquence fondamentale (en Hz) des 13 stimuli synthétiques (5 ponctuels, 4 montant, 4 descendants)











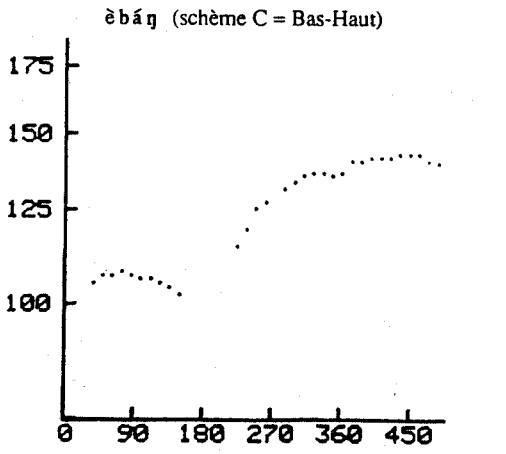


Fig. 2 Fréquence fondamentale (en Hz) en fonction du temps (en ms) des items monosyllabiques du bulu

Fig. 3 Fréquence fondamentale (en Hz) en fonction du temps (en ms) des items dissyllabiques du bulu

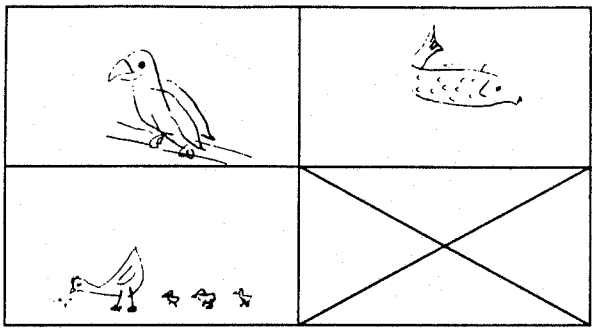


Fig. 4 Feuille-réponse utilisée pour l'identification des trois schèmes monosyllabiques du bulu

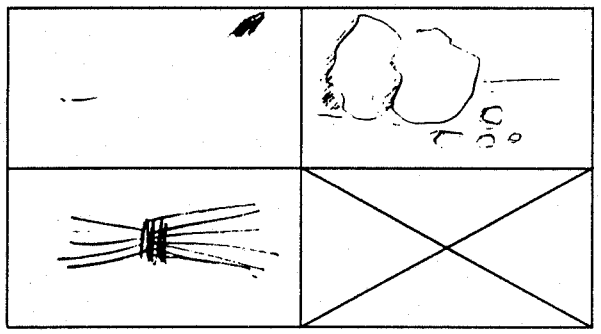


Fig. 5 Feuille-réponse utilisée pour l'identification des trois schèmes dissyllabiques du bulu

Tableau 1 Identification des stimuli monosyllabiques (50 présentations)
(Ton A = Bas (descendant), Ton B = Bas (stable), Ton C = Haut, X=hors système)

Stimuli : Début F0-Fin F0 (en Hz)	*Identifiés comme ton* A	B	C	X
150-150			50	
140-140			46	4
130-130		17	2	31
120-120		50		
110-110	2	38		10
140-150			50	
130-150				50
110-130				50
110-150				50
150-130				50
130-110	31			19
120-110	50			
150-110				50

Tableau 2 Identification des stimuli dissyllabiques avec 1ère syllabe à ton bas (110-110 Hz) (50 présentations)
(Schème A = Bas-Bas (descendant), Schème B = Bas-Bas (stable), Schème C = Bas-Haut, X=Hors système)

Stimuli : F0 de la 2ème syllabe (en Hz)	*Identifiés comme ton* A	B	C	X
150-150			50	
140-140			50	
130-130			49	1
120-120				50
110-110		50		
140-150			50	
130-150			37	13
110-130				50
110-150				50
150-130				50
130-110				50
120-110		15		35
150-110				50

Tableau 3 Identification des stimuli dissyllabiques avec 1ère syllabe à ton moyen (130-130 Hz) (50 présentations)
(Schème A = Bas-Bas (descendant), Schème B = Bas-Bas (stable), Schème C = Bas-Haut, X=Hors système)

Stimuli : F0 de la 2ème syllabe (en Hz)	*Identifiés comme ton* A	B	C	X
150-150			50	
140-140			36	14
130-130		50		
120-120		12		38
110-110				50
140-150			50	
130-150				50
110-130				50
110-150				50
150-130				50
130-110				50
120-110	1			49
150-110				50

IMPLANTS COCHLEAIRES
LA REPONSE IMPULSIONNELLE EN QUESTION

C.BERGER-VACHON, B.DJEDOU

Laboratoire Génie Biologique et Médical
Université Claude Bernard - Lyon I
69622 Villeurbanne - Cédex

ABSTRACT

Preliminary experiments in listening nonsense words through the filters of a cochlear prosthesis had shown that it is possible to recognize words with a high accuracy. Results obtained with implanted patients did not lead to the same efficiency.

As the auditory nerve was excited by the prosthesis by the mean of pulses, a simulation had been set up in order to deliver acoustical pulses to a subject with normal hearing.

The results had shown a strong deterioration of the recognition performances. Furthermore, a rise of the information transmitted to the ear does not guarantee a better recognition.

I - INTRODUCTION

Connu depuis 1957 [6] le principe de l'implantation cochléaire a été longtemps utilisé avant que la technologie ne donne des produits pratiquement utilisables avec les patients. Reprise au début des années 1970, l'implantation cochléaire a longtemps suivi une phase expérimentale avant d'entrer ces dernières années dans un rythme d'interventions beaucoup plus soutenu, avec notamment la multiplication au niveau mondial des sites d'implantation. Pourtant à l'heure actuelle, cette thérapeutique des surdités totales donne encore lieu à beaucoup de discussions et le choix de la prothèse ainsi que le type de patients auquel elle s'adresse est toujours un problème ouvert. Il est vrai que le nombre de principes utilisés pour l'implantation cochléaire est très élevé et que les études multicentriques sont encore trop rares pour qu'on puisse se faire une idée claire des mérites comparatifs de chaque implant.

En schématisant on peut classer les principes de l'implantation cochléaire de la manière suivante [2] :

- excitation par un seul canal,
- excitation par plusieurs canaux fréquentiels,
- stimulation par un signal analogique,
- stimulation par impulsions.

Dans ce dernier cas, le rythme des impulsions peut être fixé par le fondamental de la voix ou par les passages à zéro de l'onde sonore.

Un certain nombre de prothèses semblent à l'heure actuelle donner des résultats de très bonne qualité selon les auteurs ; nous citerons :

- le Nucléus australien qui délivre des impulsions connectées aux paramètres du langage [8],
- le Symbion américain qui excite par un signal analogique sur un petit nombre de canaux fréquentiels,
- le 3M-Vienna qui délivre un signal analogique représentant quatre canaux fréquentiels mais attaquant une seule électrode [10],
- le Multimac qui appartient à la nouvelle génération des produits de la société française Bertin.

Ce rapide survol montre qu'entre une excitation analogique et une excitation par impulsions, le débat est loin d'être tranché.

Nous avons déjà présenté les résultats obtenus par un patient implanté équipé d'une prothèse Chorimac ainsi que les performances réalisées en simulation par un sujet avec une audition normale qui recevait la même information fréquentielle [1]. Néanmoins une différence importante existait entre ces deux situations puisque le nerf auditif du patient implanté recevait des impulsions tandis que l'information spectrale qui parvenait au sujet à audition normale était de type analogique.

Cette simulation a montré que les performances du sujet à audition normale étaient bien supérieures à celles qui avaient été observées avec le patient implanté. On peut se demander si ces performances resteraient aussi bonnes si le sujet à audition normale entendait lui aussi des impulsions.

C'est ce que nous avons voulu étudier dans ce travail qui est en lui-même la première étape vers une étude comportementale du système auditif vis-à-vis de signaux acoustiques plus ou moins complexes se rapprochant de ce qu'on peut imaginer être l'excitation auditive délivrée par ce type d'implant cochléaire.

II - METHODOLOGIE

2.1 Prothèse Chorimac

La prothèse Chorimac [9] effectue une analyse fréquentielle du signal acoustique à l'aide d'un banc de filtres. Toutes les trois millisecondes, l'énergie détectée sur chaque filtre est codée conformément aux notions de physiologie classique de l'oreille [5] pour conduire à une impulsion d'amplitude fixe, mais qui portera dans sa durée l'énergie du filtre auquel elle correspond.

L'émetteur ou partie externe de la prothèse Chorimac comprend 12 filtres dont les bandes passantes s'échelonnent entre 0 et 7500 Hz (tableau I). Au total, toutes les trois millisecondes, un train de 13 impulsions est émis, impulsions que l'on note de 0 à 12.

F1	:	0- 240 Hz
F2	:	240- 350 Hz
F3	:	350- 450 Hz
F4	:	450- 600 Hz
F5	:	600- 800 Hz
F6	:	800-1250 Hz
F7	:	1250-1490 Hz
F8	:	1490-1850 Hz
F9	:	1850-2500 Hz
F10	:	2500-3500 Hz
F11	:	3500-4900 Hz
F12	:	4900-7500 Hz

Tableau I : Bandes passantes des filtres de la prothèse Chorimac.

L'impulsion zéro, de durée fixe, a principalement un rôle énergétique global ; les impulsions 1 à 12 portent l'information détectée par les filtres (figure 1).

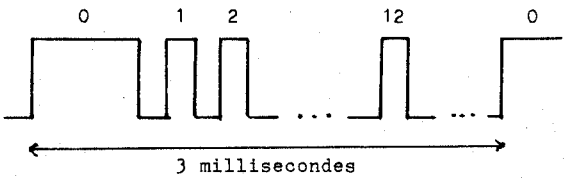


Figure 1 : Représentation d'un train d'impulsions de la prothèse Chorimac.

Ensuite, les impulsions modulent une porteuse de 3,27 MHz qui passe la barrière cutanée et qui est captée par la partie implantée située au contact de l'os mastoïdien. La partie implantée détecte les impulsions, les remet en forme et les distribue sur les électrodes placées dans le labyrinthe osseux au contact des extrémités du nerf acoustique.

On peut, sur l'émetteur de la prothèse Chorimac, fermer les canaux à volonté ; nous avons étudié la reconnaissance par un sujet implanté d'un vocabulaire de logatomes lorsqu'un seul des canaux était ouvert.

2.2 Sélection d'une impulsion

Afin d'étudier les performances d'un sujet normal lorsqu'il perçoit des impulsions de durée variable, nous avons simulé le fonctionnement du récepteur (figure 2).

Le circuit ainsi réalisé permet de délivrer sur un haut parleur des impulsions ayant la même durée que celles qui excitent le nerf auditif du patient.

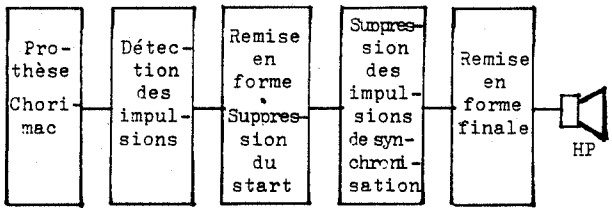


Figure 2 : Synoptique du système de sélection des impulsions porteuses d'information.

2.3 Choix du vocabulaire

Le vocabulaire qui a été utilisé est constitué de douze mots sans signification (logatomes) construits à partir d'oppositions phonétiques simples :

* 3 voyelles (/a/,/o/,/i/),
* 2 consonnes voisées (/d/,/z/) et deux sourdes (/t/,/s/),
* 2 consonnes fricatives (/s/,/z/) et deux plosives (/t/,/d/).

La liste des douze logatomes est indiquée ci-après :

DADA, DODO, DIDI
TATA, TOTO, TITI
SASSA, SOSSO, SISSI
ZAZA, ZOZO, ZIZI.

Des listes de 129 mots construites à partir de ces logatomes ont été établies de façon aléatoire.

Elles permettent d'étudier les reconnaissances suivantes :

- logatomes,
- consonnes,
- voyelles,
- trait voisé-sourd,
- opposition plosive-fricative.

La moyenne pondérée des traits élémentaires a aussi été construite :

$$P = (3*Pv + 2*Pvs + 2*Ppf) / 7$$

avec :

P = moyenne pondérée,
Pv = pourcentage de reconnaissance des voyelles,
Pvs = pourcentage de distinction voisé-sourd,
Ppf = pourcentage de reconnaissance plosif-fricatif.

III - RESULTATS

3.1 Expériences préliminaires

* Reconnaissance avec le patient implanté

En faisant passer successivement, sur une même électrode pour éviter d'introduire des problèmes de sensibilité auditive, un à un (les 11 autres filtres étant fermés) chacun des filtres de la prothèse Chorimac, le patient a obtenu les résultats de reconnaissance indiqués sur le tableau II.

Canal	Voyelle	Consonne	Trait vs	Trait pf	Moyenne
2	28.7	25.6	69.0	45.0	45.2
4	39.5	23.2	52.7	44.2	44.7
6	72.9	29.5	57.4	53.5	62.9
8	60.5	43.4	59.7	74.4	64.2
10	50.4	43.4	55.8	77.5	59.7
12	24.0	49.6	59.7	78.3	49.7

Tableau II : Pourcentages de reconnaissance obtenus avec le sujet implanté.

* Reconnaissance automatique

L'enregistrement de la durée des impulsions dans les trames de 3 millisecondes, pour la prononciation des 129 mots de la liste de logatomes [1], a conduit à des représentations semblables à des sonagrammes, montrant la variation de l'énergie acoustique traitée par la prothèse en fonction du temps.

Une segmentation élémentaire après lissage a permis de représenter chaque prononciation par deux échantillons de 3 millisecondes, un correspondant à la voyelle et un à la consonne. Chaque échantillon contient 12 valeurs numériques (une par filtre), ce qui conduit à 24 valeurs.

En comparant chaque prononciation au barycentre des classes de logatomes, les pourcentages de reconnaissance indiqués sur le tableau III ont été obtenus.

Canal	Voyelle	Trait vs	Trait pf	Moyenne
2	62.7	55.0	51.9	56.5
4	64.3	47.3	60.5	57.4
6	61.3	61.3	57.4	60.0
8	58.9	54.3	55.0	56.1
10	34.1	55.0	55.6	48.3
12	35.6	50.4	49.6	45.2

Tableau III : Pourcentages obtenus en reconnaissance automatique.

La comparaison a été effectuée en utilisant une métrique euclidienne :

$$Dx,j = \sqrt{\sum_{i=1}^{24} (Xi - Mij)^2}$$

où :
- Dx,j est la distance de la prononciation X au logatome j,
- Xi est la i^e valeur de l'image de X,
- Mij est la i^e valeur du barycentre de la classe j.

* Reconnaissance, en simulation, avec l'information spectrale

Le signal traverse successivement chacun des 12 filtres avant d'être envoyé sur un haut parleur.

Les listes de 129 mots ont été écoutées dans ces conditions par le sujet ayant une audition normale. Les résultats sont indiqués sur le tableau IV.

Canal	Voyelle	Trait vs	Trait pf	Moyenne
2	39.5	100	76.7	67.4
4	89.9	100	91.5	93.2
6	100	100	100	100
8	100	100	100	100
10	100	96.1	100	98.5
12	99.2	100	100	99.7

Tableau IV : Pourcentages de reconnaissance obtenus par le sujet ayant une audition normale, avec l'information spectrale.

3.2 Reconnaissance par impulsions

3.2.1 Avec un seul canal

L'oreille est maintenant stimulée par des impulsions acoustiques de durée variable.

Comme la simulation a été effectuée avec un sujet à audition normale, on peut dire que la reconnaissance a été effectuée "sans rééducation" c'est à dire que le sujet est placé dans un monde sonore très inhabituel qu'il aura à analyser.

La phase d'apprentissage consiste donc à écouter des listes témoin et à rechercher les éléments qui permettront de distinguer les logatomes les uns des autres ... quand celà est possible. En général, nous avons observé que les logatomes se regroupaient en quelques classes, la reconnaissance dans chaque classe devenant ensuite aléatoire. L'intérêt des listes traitées de façon statistique apparaît alors clairement ; s'il n'est pas possible d'identifier parfaitement un logatome, la réduction du champ du hasard donnera, sur 129 mots, des résultats qui seront interprétables [3].

Les résultats obtenus, avec six canaux qui balaient le spectre, sont indiqués sur le tableau V.

La comparaison avec les performances du sujet implanté est indiquée sur le tableau VI.

Canal	Voyelle	Consonne	Trait vs	Trait pf	Moyenne
2	34.1	30.2	55.8	54.3	46.1
4	55.8	27.1	46.5	58.9	54.0
6	80.6	28.7	52.7	51.2	64.2
8	40.3	29.5	44.2	60.5	47.2
10	43.4	34.4	49.6	68.2	52.3
12	42.6	31.0	43.4	78.3	53.0

Tableau V : Reconnaissance des logatomes, à partir de la stimulation par impulsions, d'un sujet à audition normale.

Canal	Voyelle	Consonne	Trait vs	Trait pf	Moyenne
2	+ 5.4	+ 4.6	-13.2*	+ 9.3	+ 0.9
4	+16.3*	+ 4.0	- 6.2	+14.7	+ 9.3
6	+ 7.7	- 0.8	- 4.7	- 2.3	+ 1.3
8	-20.2*	-15.9*	-15.5*	-13.9	-17.0
10	- 7.0	- 9.0*	- 6.2*	- 9.3	- 7.4
12	+16.6	-18.4*	-16.3	0	+ 3.3

Tableau VI : Différences entre les performances du sujet à audition normale et comparaison avec le patient implanté, dans le cas d'une stimulation par impulsions (tableau V moins tableau II).
On indique avec "*" les écarts significatifs.

3.2.2 Avec plusieurs canaux

De la même manière que lors du paragraphe précédent, les listes ont été écoutées alors que le nombre de canaux ouverts était variable.

Les résultats sont indiqués dans le tableau VII. Nous n'avons pas utilisé le canal 2 qui ne donne que des pourcentages de reconnaissance situés dans les zones aléatoires.

Nombre de canaux	Canaux	Voyelles	Consonnes	Trait voisé-sourd	Trait plosif-fricatif	Moyenne
2	10 & 12	52.71	31.78	48.83	65.11	55.5
	4 & 6	72.86	26.35	55.03	52.71	60.2
3	8 & 10 & 12	41.86	32.55	48.06	64.34	51.4
	6 & 4 & 12	66.66	29.45	49.61	62.79	59.7
4	4 & 6 & 10 & 12	65.11	29.45	43.41	62.79	57.1
5	4 & 6 & 8 & 10 & 12	65.89	31	51.16	57.36	58.1

Tableau VII : Reconnaissance des éléments phonétiques lorsque le nombre de canaux ouverts varie.

IV - DISCUSSION

4.1 Comparaison des différents tableaux

La comparaison des différents tableaux de reconnaissance montre que :

- la reconnaissance des traits acoustiques est mieux effectuée par les méthodes automatiques que par le patient implanté. L'ordinateur traite donc mieux les variations de durées des impulsions.

- la reconnaissance des traits acoustiques par le sujet à audition normale, lorsqu'il reçoit une information spectrale partielle, est effectuée de façon très satisfaisante. En milieu de gamme (filtres 6,7,8,9) le taux de bonnes réponses est 100 %.

- la reconnaissance des impulsions par le patient implanté est un peu mieux effectuée que par le sujet à audition normale. Un certain nombre de facteurs peuvent expliquer cette différence :

  * les impulsions sont présentes sur le nerf auditif dans un cas et avant l'oreille dans l'autre cas, donc le signal excitateur n'est pas exactement le même,

  * le sujet à audition normale n'a pas eu de rééducation et il n'est pas habitué au nouveau monde sonore auquel il est soumis,

  * les deux personnes ayant participé à l'expérience ne sont pas identiques.

4.2 Facteurs de reconnaissance

L'écoute des impulsions a permis de relever deux critères de distinction entre les logatomes :

- l'intensité sonore : le spectre de puissance, qui admet ici un fondamental à 330 Hz compte tenu de l'échantillonnage effectué par la prothèse, montre des différences entre les logatomes,

- le rythme dans une prononciation et sa durée avec parfois la perception de grésillements ou de continuité très grande dans les sons.

Ces deux critères, en intensité et en temps, sont assez logiques à envisager compte tenu des conditions dans lesquelles nous avons effectué notre expérience.

4.3 Superposition des impulsions

Lorsque plusieurs canaux fonctionnent simultanément, on peut dire que :

. la reconnaissance effectuée par le sujet à audition normale se dégrade régulièrement lorsqu'on part d'un canal efficace et qu'on lui ajoute des canaux efficaces ; la figure 3 précise cette notion.

. par contre, si on part de canaux peu efficaces, l'adjonction de canaux peu efficaces peut augmenter la qualité de la reconnaissance (figure 4).

Ces deux remarques sont faites en ne considérant que des canaux qui conduisent, individuellement, à des pourcentages situés hors de la zone de reconnaissance aléatoire.

Deux phénomènes peuvent donc être évoqués, à partir de cette étude :

- un effet négatif lorsque des canaux efficaces sont considérés,

- une certaine coopération lorsque des canaux peu efficaces sont mis en oeuvre.

L'extrapolation au sujet implanté est délicate, puisque les impulsions sont délivrées à des électrodes différentes ce qui introduit une notion de tonotopie. Néanmoins, l'aspect diffus de la dépolarisation dans l'endolymphe limite la précision de cette excitation tonotopique.

Les résultats décrits ci-dessus se rapprochent de constatations que nous avions déjà faites chez le sujet porteur d'un implant cochléaire [4]. Des travaux sont actuellement en cours pour mieux les préciser.

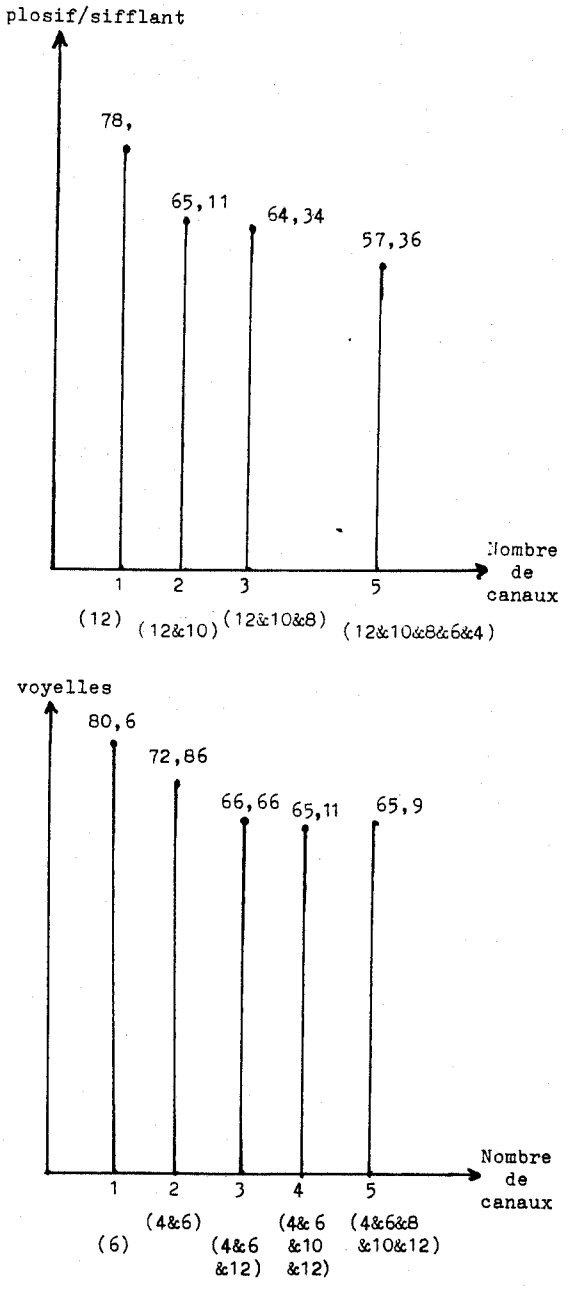


Figure 3 : Evolution des pourcentages de reconnaissance lorsque le nombre d'électrodes augmente, en partant des meilleurs canaux.

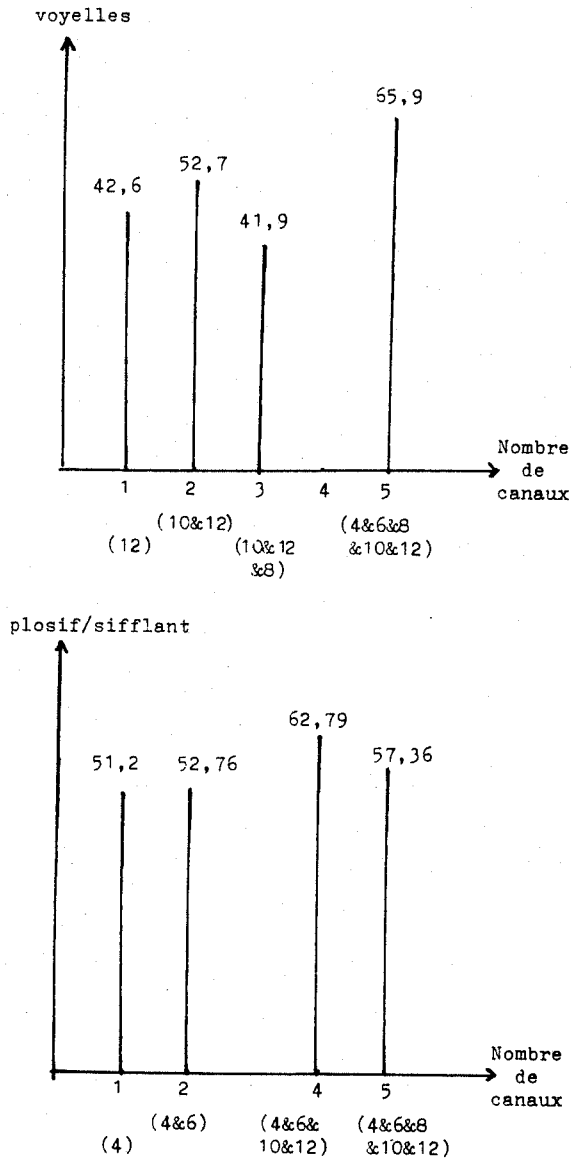


Figure 4 : Evolution des pourcentages de reconnaissance, lorsque le nombre d'électrodes augmente, en partant des canaux les moins efficaces.

V - CONCLUSIONS

L'excitation du système auditif par des impulsions présente un certain nombre d'avantages théoriques intéressants à considérer :

- elle s'adapte bien à l'évolution technologique des implants cochléaires vers les systèmes à microprocesseurs,

- elle permet de bien maîtriser et de bien étudier le signal qui est délivré aux voies auditives.

Par contre, en ce qui concerne les expériences de simulation que nous avons effectuées, le lissage énergétique représenté par les impulsions conduit à une importante perte d'information qui dégrade sérieusement les performances de reconnaissance.

La généralisation au cas des implants cochléaires multicanaux est délicate et actuellement en cours de développement.

VI - REMERCIEMENTS

Les auteurs remercient le Docteur Lionel Collet, responsable du laboratoire d'explorations sensorielles en O.R.L. de l'hôpital Edouard Herriot de Lyon, pour l'aide qu'il leur a apportée dans l'interprétation des résultats.

BIBLIOGRAPHIE

[1] BERGER-VACHON C., DJEDOU B. "Utilisation de l'information acoustique chez le sujet normal et chez le sujet implanté". 16e JEP, 136-139, Hammamet (1987).

[2] BERGER-VACHON C., AMSTUTZ I. "Les implants cochléaires.Principe et répartition". A paraître sur le Bull. Audiophonol. de Franche-Comté.

[3] BERGER-VACHON C., MORGON A. "Evaluation of the Acoustical Performances of Cochlear Implanted Patients using an Artificial Vocabulary". Speech-Com., 7, 87-95 (1988).

[4] BERGER-VACHON C., COLLET L., MORGON A. "What Information is Transmitted by a Chorimac Multichannel Implant ?". Audiol. in Pract., 4, 5 (1987).

[5] DALLOS P. "The Auditory Periphery". Academic Press, New-York (1973).

[6] DJOURNO A., EYRIES C. "Prothèse auditive par excitation à distance du nerf sensoriel à l'aide d'un bobinage inclus à demeure". Presse-Med., 35, 1417-1423 (1957).

[7] MOUHSSINE R. "Prothèse cochléaire. Evaluation objective du codage mis en oeuvre par le système Chorimac". Thèse Spécialité (Electronique) U.C.B. Lyon I (1985).

[8] SELIGMAN P.M., PATRICK J.F. & Al. "A Signal Processor for a Multiple Hearing Prosthesis". Acta Otolaryngol. (Stockh.), Supp 411, 135-139 (1984).

[9] WEBER J.L., CHOUARD C.H., ALCARAS M. "Description of the French 12-Channels Cochlear Implant". Acta Otolaryngol. (Stockh.), Supp 411, 140-145 (1984).

[10] 3M(Sté) "Cochlear Implant System, 3M-Vienna Design, Clinical Trials Manuals". Tech. Report 12/83. Sté 3M-Biomédical. 92245 Malakoff (1983).

CARACTERISATION D'EVENEMENTS ARTICULATORI-ACOUSTIQUES SUR UN MODELE DU SYSTEME AUDITIF PERIPHERIQUE : ROLE DE L'ADAPTATION NERVEUSE ET DE L'INHIBITION LATERALE

Z.L. WU *, P. ESCUDIER *, J.L. SCHWARTZ *, R. SOCK **

Institut de la Communication Parlée - CNRS, UA 368
* LCP, INPG-ENSERG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex
** IPG, ULLG, Domaine Universitaire, BP 25, 38040 Grenoble Cedex

ABSTRACT

In the past few years, several studies have been concerned with speech processing in the peripheral auditory system (PAS), in the light of both physiological data and PAS models. The present work follows this general framework, with a special emphasis on Delgutte's proposals about the effect of neural adaptation on dynamic processing. But we also find Liberman's conception of articulatory gesture as the distal object for speech perception very attractive. This leads us to an original position in so far as we mainly try to reveal a possible adequation between some of the physiological abilities of the auditory system and the acoustical structure of speech induced by articulatory constraints. We limit ouserives here to the auditory detection of the set of articulatori-acoustic events proposed by Abry et al. (1985). We describe first some results obtained on the PAS model developped in our Institute, insisting mostly on the role of neural adaptation and lateral inhibition.

INTRODUCTION

De nombreux travaux ont porté sur le codage de la parole dans le système auditif périphérique (SAP) ou dans des modèles du SAP (en France, voir par exemple (DOLMAZON, 82 a; DELGUTTE, 84 ; CAELEN, 85). Dans ce domaine, notre philosophie est d'abord fondée sur deux idées principales :

1/ Etre aussi près que possible des données physiologiques, sans trop faire de réductionnisme théorique, afin de rester capable de travailler en relation étroite avec les physiologistes.
2/ Rechercher aussi loin que possible les potentialités de traitement logées dans les bas niveaux ("non cognitifs") du système auditif.

D'autre part, sans aller jusqu'à l'affirmation de LIBERMAN et des tenants de la théorie motrice (voir par exemple LIBERMAN, 85) selon laquelle il n'y a pas d'étude de la perception de parole qui ne soit liée à l'étude des mécanismes de production, nous pensons néanmoins que nos études doivent être guidées par la recherche d'événements acoustiques interprétables articulatoirement. Aussi, notre philosophie comporte un troisième point clé :

3/ La perception de parole doit être considérée non comme la transformation de signaux acoustiques en signaux neuronaux mais comme la manipulation de signaux acoustiques par des processus neuronaux afin de "remonter" à la connaissance de gestes articulatoires.

Nous nous sommes donc intéressés à la caractérisation - et, éventuellement, la détection - d'événements articulatori-acoustiques dans un modèle du SAP. Le jeu complet de ces événements a été élaboré en vue de l'étiquetage du signal de parole, et décrit par (ABRY 85).

1. METHODOLOGIE

11. Corpus d'analyse

Il est constitué des 6 logatomes "baki, bouki, kiba, kibou, kapa, koupa", chacun étant réalisé 10 fois par deux locuteurs. Dans un premier temps nous ne nous sommes interessés qu'à un seul locuteur. Ce corpus a été entièrement étiqueté en une suite d'événements, à l'aide du logiciel d'analyse EDISIG (BENOIT, 84). Les principes de cette détection manuelle par des outils classiques sont présentés dans une étude préalable (WU , à paraître).

I.2. Description succinte du modèle utilisé

Notre modèle, largement décrit par ailleurs (voir par exemple DOLMAZON 82 b), comporte deux séries de filtres : l'une simulant en particulier les phénomènes de propagation et ayant une architecture de filtres non résonants en série, l'autre simulant la sélectivité en fréquence et présentant une architecture de filtres résonants en parallèle. Dans notre étude, le nombre de cellules est fixé à 64. Chaque filtre parallèle est suivi d'un module simulant la fibre connectée en ce point. Ce module réalise une compression logarithmique d'amplitude avec seuil et saturation. Dans la suite nous nommerons ce module : modèle de fibre.

Notre problème étant avant tout la détection et la localisation fine de ruptures ou de discontinuités temporelles, il nous a semblé important d'introduire dans notre modèle un module simulant les phénomènes d'adaptation nerveuse, dont DELGUTTE a montré à quel point ils étaient importants pour mettre en évidence de telles discontinuités. Chacun des modèles de fibre est suivi d'un module simulant le phénomène d'adaptation. Ce module, décrit très en détail par WU (87), comporte très peu de paramètres : il est donc facilement réglable.

2. CARACTERISATION SUR LE MODELE DU SAP

Notre objectif est de déterminer ce que deviennent ces évenements dans le système auditif périphérique : comment sont ils transmis, jouent ils le rôle de marqueurs, sont-ils "amplifiés" ?

On peut tenter de dégager dans le SAP un certain nombre de propriétés de base des traitements effectués (SCHWARTZ, 87). Nous nous intéresserons ici à deux mécanismes essentiels : l'adaptation nerveuse et l'inhibition latérale.

2.1. Rôle de l'adaptation nerveuse

Dans chacune des fibres étudiées, le module d'adaptation nerveuse fournit un terme de réponse prépondérante en début d'excitation, ce qui permet de remplacer des changements de pente par des pics d'amplitude - on peut donc apparenter

fonctionnellement le module d'adaptation à une dérivation temporelle - et de préparer ainsi le terrain pour des détections ultérieures à base de seuils. Nous allons constater en effet que la plupart des événements étudiés apparaissent assez clairement sous la forme de pics de réponse.

La figure 1 montre en sortie du modèle l'activité de 6 fibres de fréquence caractéristique (FC) respective 250, 500, 1000, 2000, 3000 et 4000 Hz quand l'excitation est constituée du logatome [baki]. Sur cette figure nous avons laissé en pointillé la matérialisation de l'emplacement des événements détectés avec l'éditeur de signal. Nous avons également reporté, sur chaque fibre, le décalage entre excitation et réponse dû à la propagation de l'information le long de la membrane basilaire (1er étage de filtres en série). Les fibres de basse FC présentent bien sûr un retard assez important, alors que le retard pour les hautes FC est presque nul. En bas de cette figure nous avons reporté les événements détectés avec l'éditeur de signal. La visualisation des événements détectés sur la sortie de chacune des fibres est notée directement sur la courbe. Dans un souci de clarté nous n'avons noté ces événements que sur les sortie des fibres de fréquences caractéristiques extrêmes: 250 et 4000 Hz.

Nous détectons, par ordre d'apparition sur le signal de parole, les événements suivants :

- CVO est le début de vibrations des cordes vocales pour la consonne sonore [b]. Il n'y a pas de problème particulier pour sa détection puisqu'il correspond au début du signal - en tenant compte des délais de propagation des ondes cochléaires. Sur la fibre de fréquence caractéristique 4000 Hz on constate que le signal apparait, en hautes fréquences, légèrement plus tôt que le repère déterminé manuellement sur le signal.

- CFO marque le début de plosion-friction du burst. Il se caractérise par une augmentation de niveau pour les fibres de moyenne et haute FC (entre 1000 et 4000 Hz), ce que l'on constate bien dans la réponse de la fibre FC=4000 Hz par exemple. Cette nette augmentation d'activité ne peut apparaître que si le prévoisement préalable ne fournit que peu de réponse dans ces fibres, ce qui suppose une sélectivité assez grande (voir WU & al., à paraître, pour une discussion détaillée de ce point).

- CFT, fin de plosion-friction de l'occlusive, marque la fin de l'état obstruant du conduit. Nous avons jusqu'à présent concentré notre approche sur la détection de maxima, et nous n'avons pas encore réfléchi à la détection des passages par zéro qui serait nécessaire pour cet événement.

- VVO marque le début de l'état supraglottique vocalique associé à une excitation nettement périodique sans obstruction du conduit vocal. Cet événement, correspondant à l'arrivée d'une énergie ayant une structure formantique, se repère sur toutes les fibres sous forme d'un net accroissement d'activité.

- VVT marque la fin de l'état vocalique : comme le montrent les sorties de toutes les fibres, il est partout détecté dans le cas présent, puisqu'il est suivi de la partie silencieuse de la tenue consonantique de l'occlusive sourde qui suit, c'est-à-dire [k].

- CFO et CFT pour [k] sont très apparents dans les fibres de haute FC et semblent ne pas poser de problèmes quant à leur détection (voir par exemple les fibres FC = 2000, 3000 et 4000 Hz).

- VO représentant l'arrivée du voisement se détecte dans les fibres de basse FC. En effet, l'arrivée du voisement comporte principalement des basses fréquences ; de plus, l'explosion qui précède VO comporte principalement des hautes fréquences et produit peu d'activité dans les fibres de basse FC. Il en résulte que l'adaptation met bien en évidence cet événement, ce que l'on observe dans la fibre FC = 250 Hz, en tenant compte bien entendu du retard inhérent aux fibres de FC faible.

- VVO pour [i] se lit sur les fibres de fréquence caractéristique 3000 et 4000 Hz comme le plus grand pic d'activité après l'explosion.

- VVT, fin du signal vocalique proprement dit, se voit très bien sur les fibres de basse FC ; par contre on remarque sur les fibres de haute FC les bruits fricatifs du [i] qui continuent après VVT.

2.2. Renforcement des pics spectraux par inhibition latérale

Les mécanismes d'adaptation nerveuse ne peuvent fonctionner comme révélateurs d'événements que si l'événement correspondant produit effectivement une modification significative d'excitation dans une zone spectrale donnée, modification qui sera renforcée par l'adaptation nerveuse. Or de telles modifications n'apparaissent pas toujours clairement. Prenons le cas de la voyelle [u] dans le logatome [kibu] : l'événement VVO au début de la voyelle se caractérise par le passage d'une zone de friction consonantique à une structure vocalique sans obstruction, mais la répartition de l'énergie ne change pas beaucoup de la première zone à la seconde. Le changement essentiel est l'apparition d'une nette structure formantique. Nous nous sommes donc intéressés aux mécanismes physiologiques de renforcement de telles structures, et nous donnons ici quelques résultats sur l'inhibition latérale.

Nous avons d'abord simulé un premier module d'inhibition latérale selon le modèle proposé par SHAMMA (85). Sur la Fig.2a, on considère la réponse du seul banc de filtres résonants en parallèle pour une excitation composée de deux sons purs à 0.5 kHz et 2 kHz, avec les basses FC en bas et les hautes FC en haut. On retrouve le pattern spatio-temporel classique, avec notamment un déphasage net et localisé des réponses autour des FC respectivement égales à 0.5 et 2 kHz. SHAMMA propose de tirer partie de ce déphasage par un réseau d'inhibition latérale dans lequel l'activité d'une fibre est diminuée par une fraction de l'activité des fibres voisines. La Fig.2b donne la sortie de notre modèle après une telle inhibition. Les Fig.3a et 3b présentent les mêmes résultats sur le modèle de cochlée complet, avec les phénomènes de propagation produits par l'étage initial de filtres non résonants en série.

On peut alors déterminer, par intégration temporelle, l'activité moyenne des fibres en réponse à cette excitation, respectivement avant et après ce premier réseau d'inhibition (Fig.4a, b) puis effectuer sur la sortie du premier réseau d'inhibition latérale une seconde inhibition par convolution par une fenêtre d'inhibition classique (DANG 87). On obtient alors une très bonne émergence des pics spectraux correspondant respectivement aux deux fréquences d'excitation (Fig.4c). Les Fig.4d, 4e et 4f donnent respectivement les sorties avant inhibition, après le premier réseau et après le second réseau pour une même excitation noyée dans un bruit blanc avec S/N = 0dB. On peut notamment observer dans ce second cas comment le pic à 0.5 kHz ressort finalement (Fig.4f) malgré une émergence initiale très faible, qui rend sa détection a priori (Fig.4d) peu évidente.

Sur la Fig.5, nous présentons une même série de résultats pour une voyelle [i] extraite du logatome /baki/ : la détection des pics correspondant à F1, F2 et au groupe F3-F4 se fait très bien pour le signal pur en sortie de l'ensemble des deux niveaux d'inhibition (voir 5a, b, c) et persiste pour des niveaux de bruit assez importants (ici S/B=10 dB, voir 5d, e, f).

Sous l'action conjointe de ces mécanismes de renforcement des pics spectraux, donc de structures formantiques associées aux zones vocaliques, et des mécanismes d'adaptation nerveuse, la détection de VVO au début du [u] de [kibu] est bien réalisée (Fig.6).

3. CONCLUSIONS

Ces premiers résultats sont positifs si l'on considère que la quasi-totalité des événements semble clairement mise en évidence par le SAP au niveau des fibres du nerf auditif : l'essentiel apparait.

L'adaptation nerveuse joue, comme prévu, un rôle central dans la mesure où le module correspondant transforme le "début de quelque chose" - un événement acoustique - en un pic d'activité. Dans la perspective d'une détection automatique d'événements articulatoires, les maxima d'activité, privilégiés par le phénomène d'adaptation, seront donc des bons candidats pour un futur algorithme. De même, nous avons présenté plusieurs méthodes de renforcement des contrastes spectraux à base d'inhibition latérale, avec des performances qui nous semblent assez prometteuses.

En définitive, notre stratégie consiste à mettre les événements articulatori-acoustiques que nous étudions en correspondance avec des phénomènes localisés à la fois géographiquement (spectralement) et temporellement.

Néanmoins, si l'on veut sortir du cadre uniquement descriptif dans lequel nous nous sommes volontairement placés dans cet article pour entrer dans le domaine du quantitatif, on rencontrera d'autres difficultés liées à l'influence de certains paramètres du modèle sur les résultats des mesures. Le réglage des seuils apparaît dans cette optique tout à fait déterminant. La suite de ce travail consistera donc à préciser davantage la localisation des événements afin d'élaborer des règles. De plus, nous devrons évidemment passer à une phase statistique de nos mesures : alors seulement, nous pourrons espérer démontrer de façon satisfaisante une adéquation entre processus de production et mécanismes auditifs.

BIBLIOGRAPHIE

ABRY C., BENOIT C., BOE L.J. & SOCK R. (1985)
Un choix d'événements pour l'organisation temporelle du signal de parole.
14èmes JEP, SFA, pp.133-138.

BENOIT C. (1984)
EDISIG : encore un éditeur de signal ?!!
13èmes JEP, SFA, pp.211-214.

CAELEN J. (1985)
Space/Time Data-Information in the ARIAL Project Ear Model.
Speech Comm. 4, pp.163-179.

DANG V.C. & CARRE R. (1987)
Etudes sur l'inhibition latérale dans le domaine spectral et dans le domaine temporel.
Bull. LCP, 1B, pp.319-336.

DELGUTTE B. (1984)
Codage de la parole dans le nerf auditif.
Thèse d'Etat, Université Paris 6.

DOLMAZON J.M. (1982)
Representation of Speech-like Sounds in the Peripheral Auditory System in Light of a Model.
The Representation of Speech in the Peripheral Auditory System, edited by R. Carlson & B. Granström, pp.151-164. Amsterdam : Elsevier Biomedical.

DOLMAZON J.M. & BOULOGNE M. (1982)
Interaction Phenomena in a Model of Mechanical to Neural Transduction in the Ear.
Speech Comm. 1, pp.519-554.

LIBERMAN A.M., MATTINGLY I. (1985)
The Motor Theory of Speech Perception Revised.
Cognition 21, pp.1-36.

SCHWARTZ J.L. (1987)
Représentations auditives de spectres vocaliques.
Thèse d'Etat, INP Grenoble.

SHAMMA S.A. (1985)
Speech Processing in the Auditory System. II : Lateral Inhibition and the Central Processing of Speech-Evoked Activity in the Auditory Nerve.
J. Acoust. Soc. Am. 78, pp.1616-1621.

WU Z.L. (1987)
Adaptation in the Response of Auditory Nerve Fibers : a Simulation in Light of a Functional Model.
Bull. LCP, 1B, pp.237-264.

WU Z.L., ESCUDIER P., SCHWARTZ J.L. & SOCK R. (à paraître)
Caractérisation d'événements articulatori-acoustiques sur un modèle du système auditif périphérique : étude préliminaire.
Bull. LCP.

Cette étude a été financée par la D.R.E.T. que nous remercions.
Contrat de recherche n° 86 / 063.
Convention n° 86.34.063.00.470.75.01.

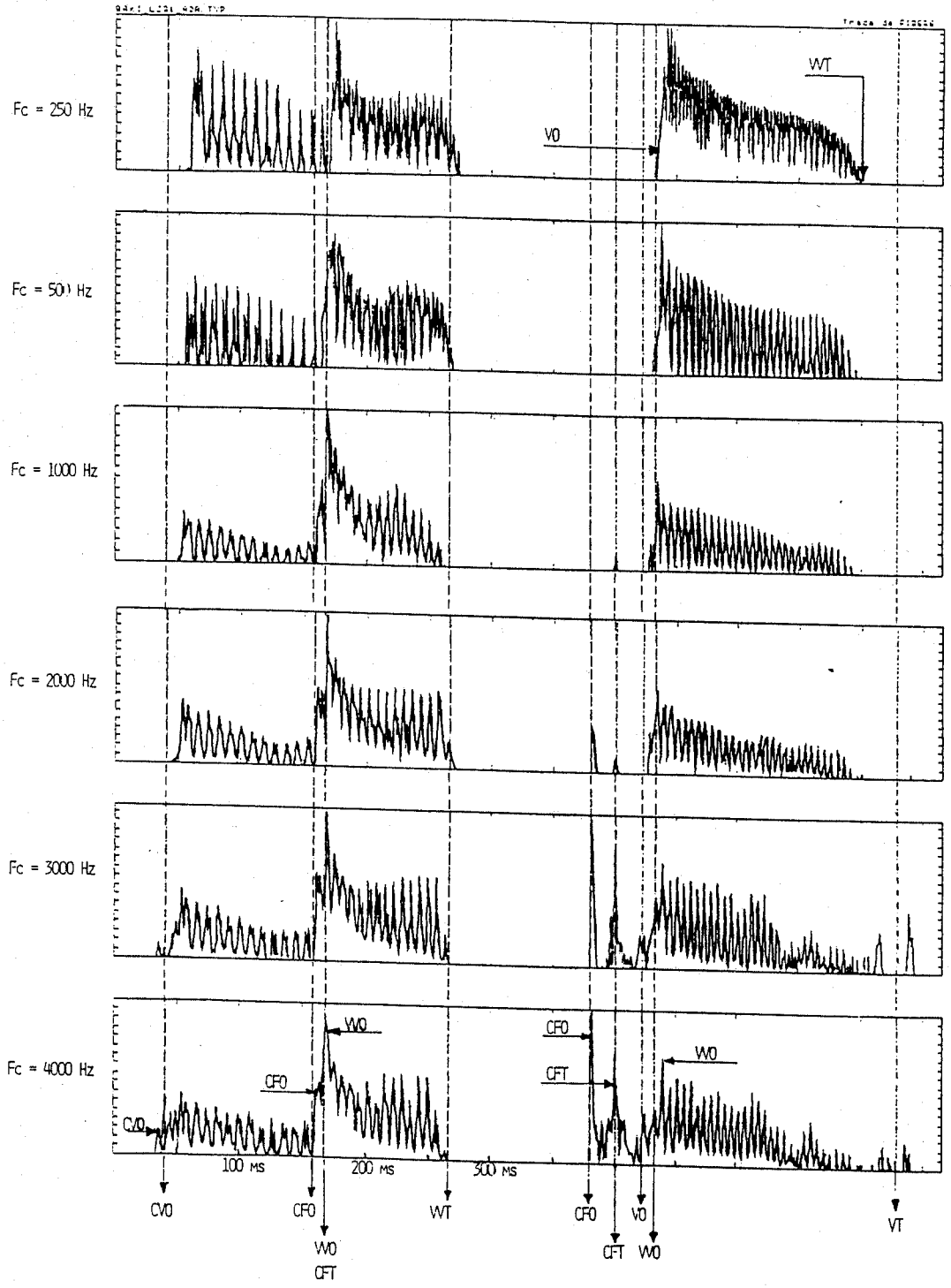


Figure 1 - Réponse du modèle CALSAP (avec adaptation nerveuse) au logatome [baki] (temps en abscisse, activité de sortie de 6 fibres en ordonnée)

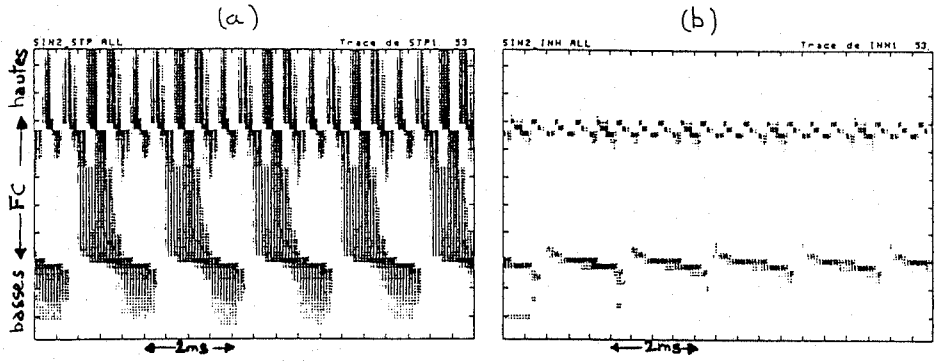


Figure 2 - Réponse du banc de filtres resonants en parallèle à une somme S de deux sons purs (0.5 et 2 kHz) : temps en abscisse, FC en ordonnée
a) sortie directe des filtres b) sortie du module d'inhibition latérale

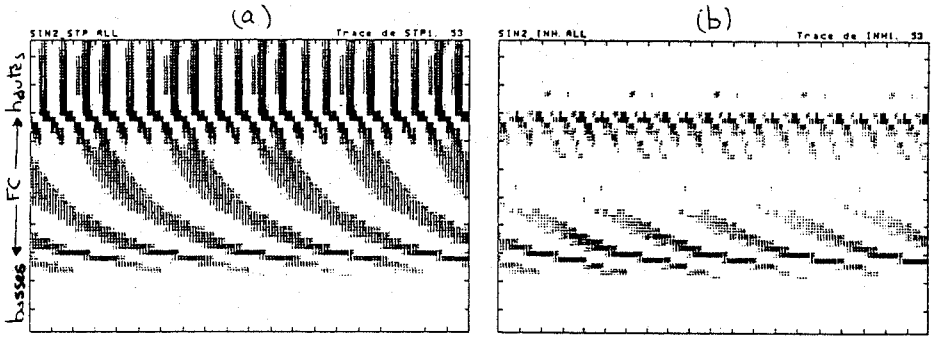


Figure 3 - Réponse du double étage de filtres non résonants en série suivis des filtres résonants en parallèle - même présentation que la Fig.2
a) sortie directe des filtres b) sortie du module d'inhibition latérale

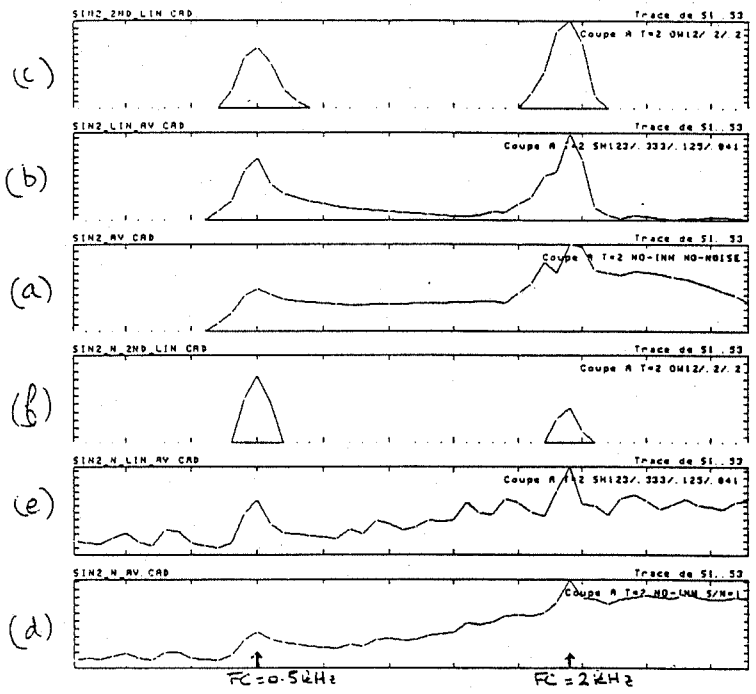


Figure 4 - Reponse de CALSAP à S - FC en abscisse, activité moyenne en ordonnee
sortie directe (a), après le premier (b), puis le second (c) étage d'inhibition
idem pour S + B (bruit blanc) : (d), puis (e) puis (f)

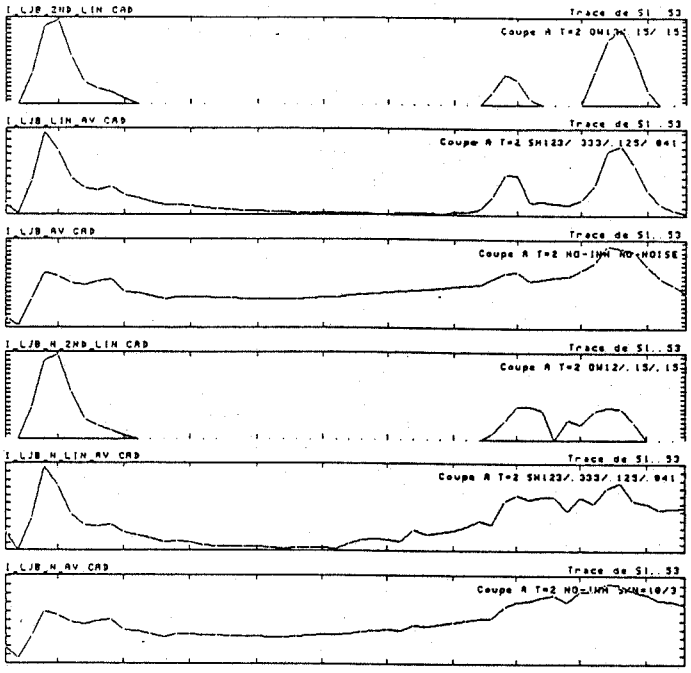


Figure 5 - Réponse de CALSAP à [i] - FC en abscisse, activité moyenne en ordonnée
(a), (b), (c) : sortie directe, premier étage, second étage d'inhibition
(d), (e), (f) : idem pour [i] + B : (d), puis (e) puis (f)

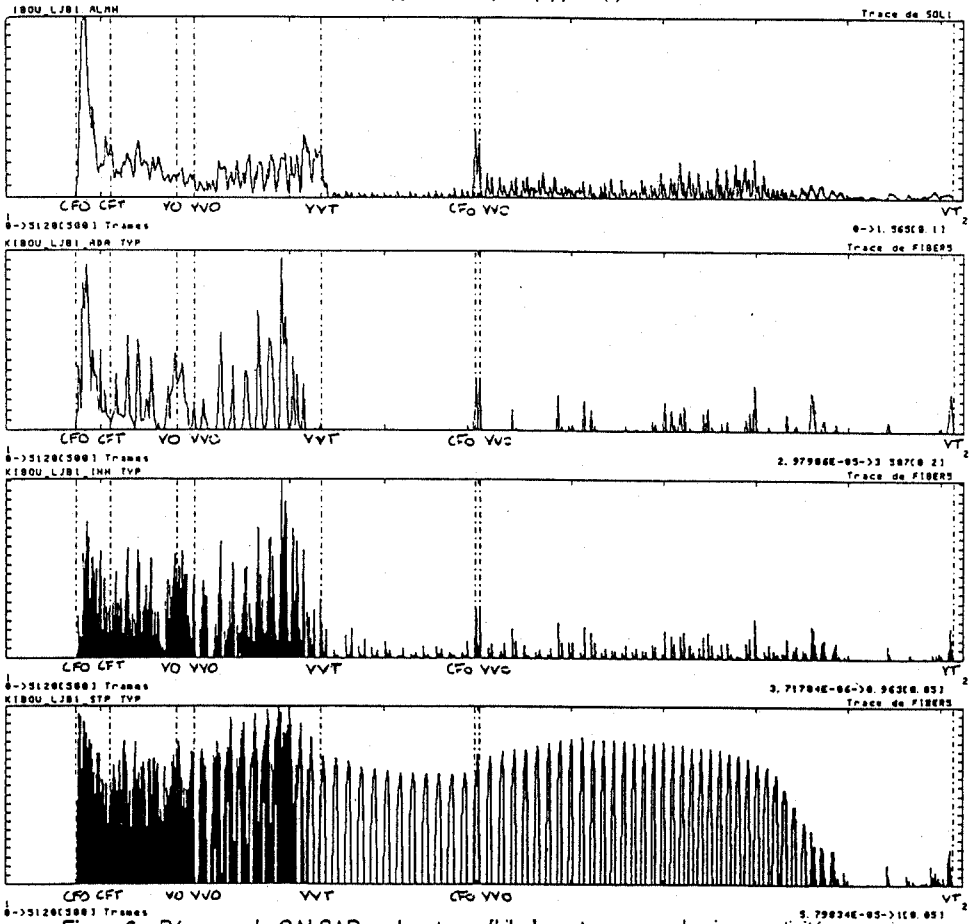


Figure 6 - Réponse de CALSAP au logatome [kibu] - temps en abscisse, activité en ordonnée
(a) sortie après filtrage, pour une fibre de FC = 3 kHz
(b) inhibition latérale appliquée sur la sortie (a)
(c) adaptation nerveuse appliquée sur la sortie (b)
(d) sommation d'activités de type (c) pour toutes les sorties FC ≥ 1 kHz
On remarque la remarquable mise en évidence de la zone de friction du [b] (CFO-VVO) lorsque l'on passe de (a) à (d)

MISE EN RELATION DES INDICES PHYSIOLOGIQUES ET ACOUSTIQUES DE LA NASALITE VOCALIQUE ET CONSONANTIQUE

D.AUTESSERRE [+], N.VIGOUROUX [++], C.BARRERA [++], I.GUAITELLA [+]

[+]Institut de Phonétique, 29,avenue R.Schuman, 13621 Aix
[++]Laboratoire CERFIA, 118,route de Narbonne, 31062 Toulouse

ABSTRACT

A comparative study of the physiologic and acoustic cues of nasality is undertaken: it is based on the simultaneous audio recording of speech signal and video recording of lip and velum movements (fiberoptic). In this purpose, an original method for quantifying endoscopic pictures is developped. Moreover, as data about varying acoustic cues in the time domain are needed, an analysis through auditory modelling is used. After solving the synchronisation problem, a close correspondence between physiologic and acoustic cues is performed. These results show the importance of the dynamic nature of the acoustic cues, relevant to the identification and recognition of nasal sounds.

INTRODUCTION

Une première étude est menée sur l'ouverture labiale et vélo-pharyngée lors de la production des consonnes bilabiales et des différentes voyelles orales et nasales du français. Les résultats obtenus, représentés sous forme de courbes, donnent les trajectoires des mouvements du voile et des lèvres.

En parallèle, les travaux réalisés à partir d'un modèle d'oreille, fournissent les évolutions temporelles d'indices spectraux.

Il nous a paru intéressant de mettre en correspondance ces deux types d'information. La méthode est applicable à la mise en relation de tout paramètre acoustique et physiologique. Nous nous limitons ici à la comparaison des mouvements labiaux et vélo-pharyngés avec les corrélats issus de l'analyse acoustique retenue.

1 ACQUISITION ET QUANTIFICATION DES DONNEES PHYSIOLOGIQUES

1. 1 Protocole d'acquisition des données physiologiques

Les images ont été obtenues à partir d'un enregistrement vidéo des mouvements des lèvres et du voile. Dans ce dernier cas, il a été procédé à un examen fibroscopique par voie nasale (bronchoscope souple). Deux enregistrements du signal de parole ont été réalisés en même temps que les prises de vues: l'un sur bande vidéo par l'intermédiaire d'un magnétoscope, l'autre à l'aide d'un magnétophone bi-piste. Sur la première

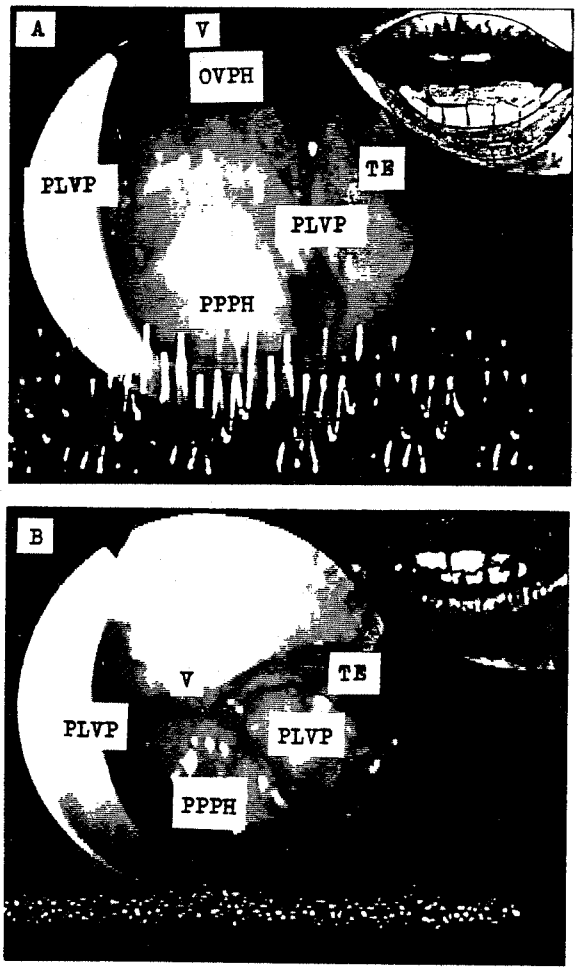


figure 1 : Photographie de l'écran d'un moniteur vidéo en arrêt sur une trame.

En haut à droite, image incrustée des lèvres, au centre, vue fibroscopique du vélo-pharynx, en bas, tracé oscillographique.

A. Position abaissée du voile lors de la prononciation de la voyelle nasale [ɛ̃] dans le mot "main".

B. Relèvement maximum du voile pendant la réalisation de la consonne constrictive non voisée [s] dans le mot "vis".

V, partie centrale du bord du voile, PLVP, parois latérales du vélo-pharynx, OVPH, ouverture vélopharyngée, TE, trompe d'Eustache, PPPH, paroi postérieure du pharynx.

piste, des impulsions de synchronisation ont été enregistrées. Le signal de parole proprement dit et les clics de repérage des séquences phoniques sont présents sur la deuxième piste [Teston 86].

Deux locuteurs, un homme (DA) et une femme (FB) prononcent tour à tour une série de phrases comportant les consonnes bilabiales [p], [b], [m], précédées et suivies de voyelles orales ou nasales Autesserre 86 . Les phrases sélectionnées pour cette étude sont toutes construites sur le même modèle: "il a dit ... six fois", afin que chaque séquence analysée soit toujours du type "-is VCV si-". Ainsi, la consonne bilabiale voisée [b] se trouve placée en contexte vocalique nasal symétrique [ɑ̃] dans la phrase "dis cent bancs six fois", [disɑ̃bɑ̃sifwa]. Les deux réalisations phonétiques [i] et [s] qui encadrent la suite réellement étudiée "VCV" ont pour fonction de fournir un repère de relèvement maximum du voile par rapport auquel sont évalués ses mouvements d'élévation et d'abaissement.

1. 2 La quantification des paramètres physiologiques

Le contour de chaque image du vélo-pharynx et des lèvres (cf. fig.1) est dessiné manuellement sur un calque placé directement sur l'écran du moniteur vidéo. On disposera ainsi d'une image toutes les 20 ms. Les mesures linéaires sont effectuées à partir de chacun de ces calques.

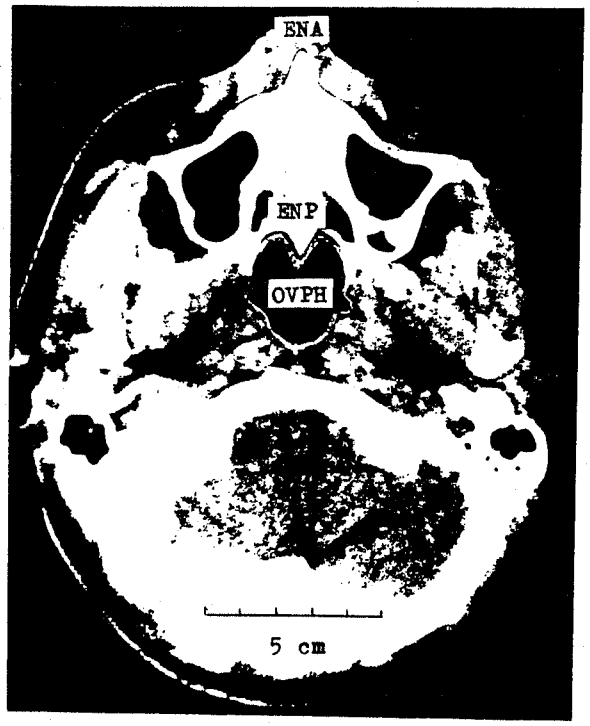


figure 2 : Coupe scanographique, dans le plan palatin, pendant l'émission prolongée de la voyelle [ɛ̃] dans le mot "main".

OVPH, surface d'ouverture vélo-pharyngée, ENA, épine nasale antérieure, ENP, épine nasale postérieure.

Il aurait été plus intéressant de comparer des surfaces qui présentent l'avantage d'intégrer les déplacements non symétriques des articulateurs. Mais il n'est pas possible, encore actuellement, d'obtenir des valeurs exactes des variations de la surface vélo-pharyngée, lors de la phonation. En revanche, on peut arriver à une estimation de cette même surface,pour des réalisations phonétiques tenues, à l'aide de coupes scanographiques (dans des plans proches du plan palatin, fig.2).

1. 2. 1 L'aperture intéro-labiale

Pour les lèvres, nous disposons d'un programme d'analyse automatique de l'image qui permet de mesurer la surface intéro-labiale. Pour des raisons méthodologiques, il nous a semblé plus correct de procéder, dans ce cas aussi, à des mesures linéaires de l'aperture intéro-labiale. Ainsi, la comparaison peut s'établir au niveau physiologique, entre les mouvements d'abaissement et de relèvement du voile, le rapprochement ou l'éloignement des parois latérales du vélo-pharynx et les modifications d'ouverture des lèvres.

Les points de mesure, pour les images labiales, sont déterminés par construction (fig.3): ils se trouvent à l'intersection de deux parallèles au segment intercommissural, tangentes à chaque lèvre de part et d'autre de l'orifice buccal et d'une perpendiculaire abaissée en son milieu. La distance de référence pour le calcul du facteur d'agrandissement de l'image est mesurée directement sur chacun des deux locuteurs: il s'agit de la distance intercommissurale relevée lors d'une phase de repos post-phonatoire. Les images labiales sont agrandies dans un rapport de 1,5 pour le locuteur DA et de 2 pour FB.

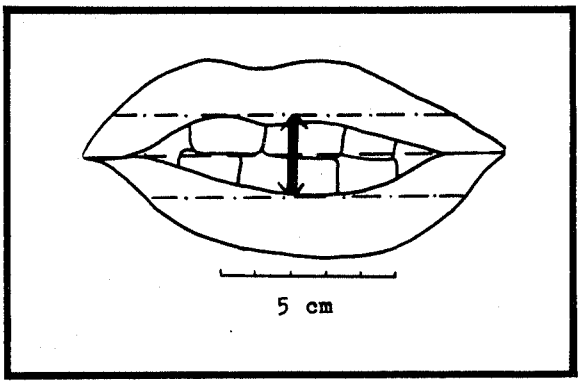


figure 3 : Mesure de l'aperture intéro-labiale sur l'image vidéo agrandie. (La mesure prend en compte l'asymétrie d'ouverture des lèvres.)

1. 2. 2 Le déplacement du voile et des parois latérales

En ce qui concerne le vélo-pharynx, deux séries de mesures sont effectuées pour rendre compte des déplacements du voile et des parois latérales (fig.4).

Le déplacement du voile est mesuré par rapport à la position très abaissée

lors d'une respiration calme pré-phonatoire (prise de souffle). L'axe des mouvements du voile est déterminé par comparaison des positions successives prises par sa partie centrale au moment des phases d'abaissement et de relèvement maxima. Il s'agit alors d'une ligne courbe et l'on a pris la corde de l'arc correspondant.

Les déplacements des parois latérales sont évalués à partir des points de mesure situés d'un seul côté (à droite pour le locuteur FB, à gauche pour le locuteur DA), en tenant compte de l'orientation du fibroscope. La mesure est faite à l'intersection de la tangente à la partie la plus avancée de la paroi latérale concernée et d'une perpendiculaire à l'axe du voile passant par ce point.

L'agrandissement, calculé à partir des coupes scanographiques, est de l'ordre de 4 pour le locuteur FB et de 4,5 pour DA tant pour les déplacements du voile que pour ceux des parois latérales du vélo-pharynx.

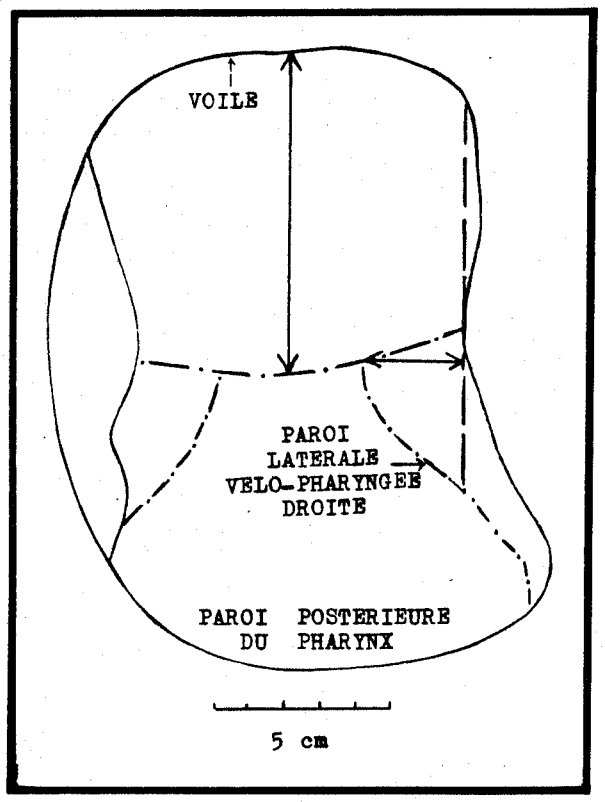


figure 4 : Mesure des déplacements du voile et des parois latérales du vélo-pharynx.

Contour en trait plein: position du voile et des parois latérales lors de la phase de respiration pré-phonatoire.

Tirets et pointillés: position du voile et des parois latérales pendant la prononciation de la consonne [s] précédée de la voyelle [i].

La flèche verticale indique le déplacement maximum du voile, la flèche horizontale rend compte du déplacement maximum des parois latérales vélo-pharyngées. (Pour ce locuteur, FB, la mesure est prise au niveau de la paroi latérale droite.)

L'échelle correspond à l'image vidéo agrandie.

2 LES DONNEES ACOUSTIQUES ET PHONETIQUES

2. 1 Les informations acoustiques

Un magnétophone Radiola de type N 4420 est utilisé pour lire l'enregistrement bi-piste. Une piste est réservée au signal de parole et aux clics. Ces clics sont des repères de synchronisation pour les analyses acoustiques et physiologiques réalisées dans des laboratoires distincts et sur des matériels différents.

Le signal est numérisé à l'aide de la chaîne OROS-AI installée sur un PDP-11/73. Nous réalisons ensuite une analyse spectrale par banc de filtres qui fournit un spectre en décibels sur une échelle de MEL de 24 canaux. Un échantillon spectral correspond à une fenêtre d'analyse de 64 points, soit une durée de 4 ms pour une fréquence d'échantillonnage de 16 kHz. Un vecteur d'indices spectraux est calculé à la sortie de ce modèle: aigu-grave, fermé-ouvert, écarté-compact, bémolisé-diésé [Caelen 81]. Nous disposons aussi d'informations prosodiques: énergie totale du signal, fréquence fondamentale,...

Une fonction de pré-segmentation fondée sur la variation globale [Vigouroux 85] des indices acoustiques et prosodiques permet de découper le spectre en segments infra-phonémiques homogènes (fig.5).

2. 2 L'étiquetage

Cette étape, appelée communément étiquetage [SAM-Report 88] , consiste à mettre en correspondance temporelle des transcriptions, de niveaux variables, avec le signal vocal. Cette correspondance permet d'accéder aux réalisations sonores des unités définies à priori dans la BDAP (Base de Données Acoustico-Phonétiques).

Notre approche est segmentale: l'étiqueteur appose sur les frontières des unités infra-phonémiques, des marques, selon un étiquetage à finesse variable. Pour la mise en correspondance d'informations acoustiques et physiologiques, nous n'utilisons que les niveaux acoustique et phonémique de notre système d'étiquetage [Barrera 87]. En effet, il est nécessaire d'atteindre cette finesse d'étiquetage afin d'accéder à des sous-segments de la production d'éléments phonémiques tels que les parties vocaliques et consonantiques d'une voyelle nasale. Cette segmentation acoustique sera commentée plus bas lors de l'interprétation des résultats (chapitre 4).

2. 3 La synchronisation

L'objectif à atteindre est la mise en correspondance temporelle des informations acoustiques et physiologiques produites par deux chaînes d'acquisition différentes. Nous disposons des informations suivantes:

- des données physiologiques mesurées dans la partie finale de chaque trame image d'une durée d'environ 20 ms,

- des échantillons signal d'une durée constante de 16 ms, ce qui donne,en raison des paramètres de l'analyse spectrale, des

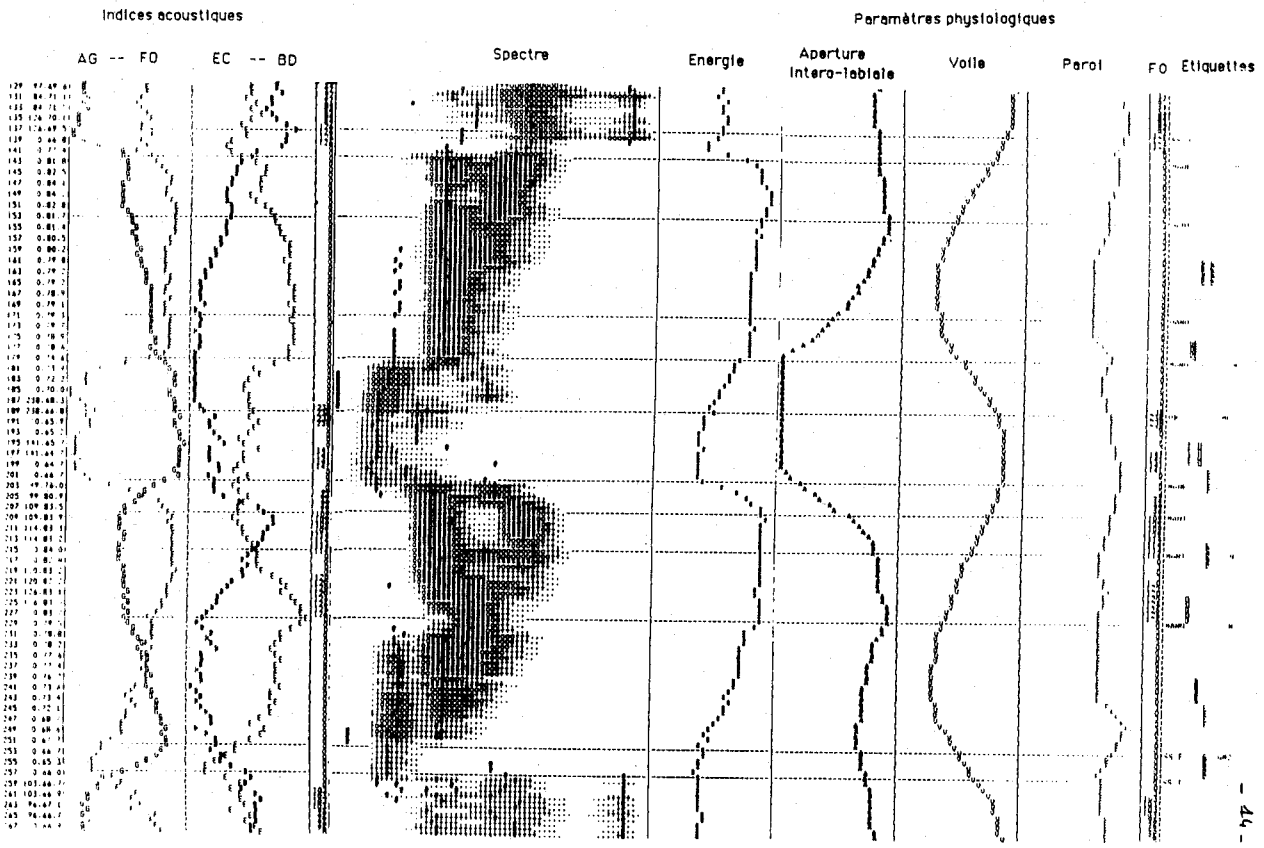


figure 5 : Réalisation de la séquence [sãbã] par le locuteur FB.

Chaque échantillon spectral a une durée de 4 ms (fréquence d'échantillonnage 16 kHz).

Les indices spectraux sont, dans l'ordre: AG, aigu-grave, FO, fermé-ouvert, EC, écarté-compact, BD, bémolisé-diésé.

vecteurs spectraux toutes les 4 ms. Les paramètres de l'analyse spectrale ont été choisis de sorte que la durée d'un échantillon spectral soit le plus petit diviseur de la durée de la trame afin de synchroniser au mieux les deux sources d'informations.

Les clics générés sur la bande son, à chaque début de phrase, nous servent de référence de départ sur l'axe des temps et permettent ainsi la synchronisation d'une trame vidéo avec un ensemble d'échantillons spectraux. Quelques contrôles temporels, comme par exemple la durée entre deux clics nous ont permis de tenir compte de problèmes techniques résultant du transfert de l'information temporelle issue d'analyse analogique dans un système numérique [Teston 86]. Ceci nous amène à une réflexion sur la normalisation des équipements pour faciliter la collaboration inter-laboratoires et, bien sûr, sur la spécification d'un éventuel poste vidéo-parole [Marchal 87].

3 LA BASE DE DONNEES ACOUSTICO-PHONETIQUES

Compte tenu de l'objectif, concernant d'une part l'analyse proprement dite des informations acoustiques et physiologiques et d'autre part la corrélation de ces deux sources d'informations, la BDAP est composée de deux sous-ensembles:

- le sous-ensemble des informations standards,

- le sous-ensemble des outils spécifiques nécessaires à la manipulation, au sens large, des informations.

Le sous-ensemble des informations standards se compose des structures de données suivantes:

<u>au niveau acoustique</u>:

- du signal numérisé,

- des informations spectrales, valeurs fournies par un banc de filtres,

- des tables d'étiquetage phonémiques et acoustiques, informations temporelles placées sur la réalisation de l'onde vocale.

<u>au niveau physiologique</u>:

- des paramètres physiologiques, décrits précédemment,

- des tables d'étiquetage, informations temporelles placées aux positions extrêmes des articulateurs. On attribue les étiquettes suivantes:

. pour les lèvres: O, ouverture labiale, DF, début de fermeture labiale, F, fermeture labiale.

. pour la partie centrale du voile: H, voile haut, B, voile bas.

. pour les parties latérales vélo-pharyngées: E, parois élargies, R, parois rétrécies. Pour plus de détails, se reporter à [Autesserre 87]. Ces informations physiologiques seront considérées comme une vue particulière dans BDAP [Pérennou 88].

A ce noyau d'informations est associé un ensemble de procédures spécifiques. Celles-ci permettent de produire:

- des informations dérivées, par exemple des indices acoustiques [Caelen 81], des paramètres physiologiques dérivés (inertie, convexité,...),

- des informations sur le comportement des paramètres et/ou corrélats acoustiques: taux de corrélation entre des indices acoustiques et paramètres physiologiques...

- des statistiques élémentaires sur le contenu de la base...

4 INTERPRETATION DES RESULTATS

Notre commentaire portera plus spécialement sur la voyelle nasale en position pré-consonantique pour laquelle nous avons remarqué une bonne synchronisation entre la segmentation acoustique et la succession des évènements articulatoires.

Conformément à la méthode de définition du corpus, nous nous intéressons aux séquences correspondant à l'abaissement et au relèvement du voile entre deux phases d'élévations maximales pour [s], dans la séquence [disãbãsi] (cf. tableaux 1 et 2).

L'analyse des courbes obtenues met en évidence 4 phases que nous notons: Vo, voyelle orale, Vn1, voyelle nasale I, Vn2, voyelle nasale II et Vcn, consonne nasale. La séparation des différents segments acoustiques est bien déterminée pour la distinction Vn2/Vcn et Vo/Vn1. Par contre, la marque de pré-segmentation entre Vn1/Vn2 est moins significative.

Paramètres et Indices	Voyelle Orale	Voyelle Nasale I	Voyelle Nasale II	Consonne Nasale
Aperture Labiale	↗	O ↘	↘	→
Voile	↘	↘ B →	→	↗
B1				300
B2	600			
B3		900 ↘ 760	760 ↘ 660	1000
B4	1300		1700 ↘ 1200	
AG	AG	AG	AG	A-
EC	E-	E--	E--	E-
BD	B+	B+	B+	B++
FO	F--	F--	F--	F+
Adresse de début	100	109	117	124
Etiquettes				AN Q N

Présegmentation Acoustique

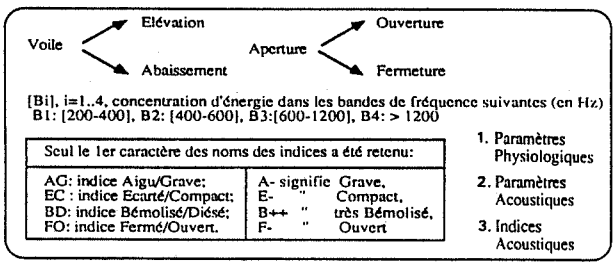
1 2 3

tableau 1 : Réalisation de la lère voyelle [ã] dans la séquence [sãbã] pour le locuteur DA.

Paramètres et Indices	Voyelle Orale	Voyelle Nasale I	Voyelle Nasale II	Consonne Nasale
Aperture Labiale	↗	O ↘	↘	→
Voile	↘	↘ B →	→	↗
B1		330	320	260
B2	580	580	560	
B3				1000
B4	1400 ↘ 1100			
AG	AG	A-	A-	A--
EC	E-	E--	E--	E-
BD	BD	B+	B++	B++
FO	F--	F--	F--	F++
Adresse de début	141	152	170	178
Etiquettes	AN E O	AN T	AN Q	AN Q N

Présegmentation Acoustique

tableau 2 : Réalisation de la lère voyelle [ã] dans la séquence [sãbã] pour le locuteur FB.



Légende des tableaux 1 et 2

- Voyelle orale: Au plan articulatoire, cette phase est caractérisée par un voile non suffisamment abaissé. Pourtant, la commande d'abaissement a déjà été émise dans le dernier tiers de la consonne constrictive précédente. Nous rappelons que notre analyse révèle surtout le mouvement d'abaissement et de relèvement du voile mais que le voile peut s'abaisser dans un premier temps sans se décoller de la paroi postérieure du pharynx (rapport abaissement/tension). Au plan acoustique, nous retrouvons les caractéristiques de la voyelle orale correspondante (B2 = 600 Hz et B4 = 1400 Hz au lieu de 650 et 1500 Hz pour le [a] dans la phrase [disabas(ə)sifwa]).

- Voyelle nasale: La faible séparation entre Vnl et Vn2 au plan acoustique s'explique par le fait que la variation voile/lèvres n'est pas simultanée. Le voile reste en position d'abaissement(même si l'ordre de remonter a été déjà envoyé). Ceci correspond à une bémolisation associée à la réduction de l'ouverture labiale. Nous notons comme Fonsale [Fonsale 84] la disparition de l'émergence des maxima spectraux dans la bande 1600-3400 Hz.

- Consonne nasale: C'est dans cette phase que l'indice de nasalité apparaît le mieux et ce, pour les deux locuteurs: la présence d'une crête d'énergie autour de 300 Hz et d'une masse d'énergie autour de 1000 Hz. Ces résultats confirment ceux de [Marteau 87]. Ceci correspond:

. au plan acoustique, à une variation significative des indices: FO plus fermé (noté F+ ou F++) et BD plus bémolisé (noté B++),

. au plan articulatoire, à une fermeture des lèvres tandis que le voile est encore abaissé.

Acoustiquement, c'est ce dernier segment qui porte l'information de nasalité.

L'évolution sur la partie voyelle nasale (V_{n1} et V_{n2}) est différente pour les deux locuteurs. Par contre, ces deux locuteurs se rejoignent dans la partie consonantique finale où la fermeture labiale favorise au maximum la résonance nasale. L'indice de nasalité est bien individualisable dans cette partie finale consonantifiée. La diversité des locuteurs citée, très souvent, pour la production des voyelles nasales, concernerait surtout le couplage vocalique associé à des coordinations motrices différentes.

CONCLUSION

Par rapport aux grandes études fondamentales menées, tant dans le domaine de la définition des indices acoustiques de la nasalité que dans celui de leur perception*, notre démarche fait ressortir toute l'importance d'une analyse des variations temporelles des indices acoustiques pour l'identification des voyelles nasales.

Le relèvement et l'abaissement du voile entrainent avec un retard, que nous évaluons actuellement, des modifications des indices acoustiques correspondants. Dans le cas de voyelles nasales en position préconsonantique, il convient de ne pas dissocier de la voyelle nasle la partie consonantique subséquente qui apporte une part importante de l'information de nasalité.

Ceci nous conduit à poser le problème de la gestion des étiquettes acoustico-phonétiques au sein d'une base de données des sons, surtout dans la perspective d'une application à la reconnaissance de la parole.

*Il n'est pas possible de citer ici les nombreux travaux consacrés à l'étude des aspects physiologiques et acoustiques de la nasalité et à leurs corrélations, nous renvoyons les lecteurs aux excellentes synthèses de [Feng 87] et [Lonchamp 88].

REFERENCES BIBLIOGRAPHIQUES

[Autesserre 87] Autesserre D., Barrera C., Espesser R., Pérennou G., Rossi M., Teston B., Vigouroux N., "Acoustic-Articulatory Information in Speech Data Base", Edinburgh, pp. 125-132.

[Barrera 87] Barrera C., Caelen J., Caelen-Haumont G., Malet J.F., Vigouroux N., "Towards an Acoustic Labelling System", XIth Int. Cong. of Phon. Sciences, Tallinn.

[Caelen 81] Caelen J., Caelen-Haumont G., "Indices et propriétés dans le projet ARIAL II, processus d'encodage et de décodage phonétiques", Toulouse, pp. 128-143.

[Feng 87] Feng G., "Etude articulatori-acoustique des voyelles nasales du français", Bul. de l'Inst. de Phon. de Grenoble vol. 16, pp. 1-102.

[Fonsale 84] Fonsale P., Simulation informatique d'un système multilocuteurs de reconnaissance de parole (mots isolés) sans apprentissage oral. Analyse par traits phonétiques., Thèse de docteur ingénieur, Grenoble.

[Lonchamp 88] Lonchamp F., Etudes sur la production et la perception de la parole. Les indices acoustiques de la nasalité vocalique. La modification du timbre par la fréquence fondamentale., Thèse de doctorat d'état, Univ. de Nancy II, 345p.

[Marchal 87] Marchal A., "BD-ART", Rapport GRECO-CP 39.

[Marteau 87] Marteau P.F., Caelen J., Janot Giorgetti M.T., "Extraction automatique de caractéristiques dynamiques du signal de parole. Application à l'analyse de voyelles nasales", 16ème JEP, Hammamet, pp. 88-91.

[Pérennou 88] Pérennou G., Vigouroux N., "Préliminaires méthodologiques pour une base de données acoustique phonétique", soumis au comité de lecture des 17ème JEP.

[SAM-Report 88] SAM-Report ESPRIT Project 1541, Final Report, Definition Phase : 2-2-87/31-1-88.

[Teston 86] Teston B., Autesserre D., "Description d'un dispositif d'enregistrement simultané des mouvements des organes articulateurs", 15ème JEP, Aix en Provence, pp. 65-68.

[Vigouroux 85] Vigouroux N., Caelen J., "Segmentation phonétique et organisation d'une base de données acoustiques et phonétiques", 14ème JEP, Paris, pp. 152-155.

LA RESISTIVITE DE LA QUANTITE VOCALIQUE AUX VARIATIONS DE LA VITESSE D'ELOCUTION. LE CAS DE L'ARABE TUNISIEN.

JOMAA M. & ABRY C.

Institut de Phonétique de Grenoble. Institut de la Communication Parlée
(CNRS UA 368), Université Stendhal, BP 25X, 38040 GRENOBLE CEDEX

ABSTRACT

Accelerating speech rate of minimal pairs (embedded in carrier phrases) by a factor of 1.30, for a Tunisian Arabic speaker, we observe an increase in confusion between pairs.

Thus, duration of the short vowel [a] changes from 59% to 78% of the long vowel duration (the latter reducing its mean duration from 113 ms to 63 ms). As concerns relative timing, in the release cycle (voyelle-consonne or VC), the vocalic phase remains constant for short vowels, around 40%, whereas for longs, this phase goes from around 60% to 45%. Consequently, no invariance, even relative is observed (TULLER&KELSO, 1984) for long vowels.

Contrast itself therefore, does not resist, in our dialectal Arabic, to variations in rate - contrary to what has been observed for other languages (for example in Gambian Wolof, SOCK, 1983).

We are able to show here, as regards this issue, that the VC domain (reflecting a well known temporal asymmetry in speech, cf. LEHISTE, 1970) enhances quantity distinctions for all speech rates. This is not true in other domains, like the word for example.

INTRODUCTION

Pour l'arabe tunisien - qui est encore aujourd'hui une langue essentiellement à tradition orale et un outil de communication vernaculaire - l'utilisation de l'opposition de quantité vocalique est partie intégrante du système (GHAZALI, 1979 ; TAMOULDI, 1984; JOMAA, 1987).

Nous savons que l'étude de l'organisation temporelle de la quantité vocalique ne peut se limiter à la mesure de la durée de la phase de manifestation la plus caractéristique de la voyelle. Elle nécessite la prise en compte des manifestations des consonnes adjacentes, non seulement pour la recherche de corrélats éventuels moins "vocaliques" , mais plus fondamentalement pour la définition d'un domaine, permettant de référer l'ensemble des mesures absolues, relativement à ce champ.

Ces mesures absolues sont en effet par nature extrêmement sensibles aux variations de la vitesse d'élocution (débit), et - à moins de compter sur une invariance absolue - elles ne peuvent rendre compte des régularités recherchées, à travers les différentes conditions prosodiques. (par définition de l'invariance en timing relatif , TULLER & KELSO, 1984).

Quels que soient les qualités et les défauts des modèles proposés pour rendre compte de telles régularités dans le comportement moteur, en général, et dans la parole en particulier (MC KAY, 1987), il semble profitable d'en retenir l'expérience accumulée dans le traitement des données de mouvement. Certains concepts en timing sont ainsi profitables, et parmi eux, les notions de cycle (défini par la reproduction d'événements semblables ou par l'achèvement d'un processus) et de phase (relation entre cycles ou simple étape dans le décours d'un processus). En particulier, pour ce qui nous concerne, les cycles permettent de définir le domaine - la base temporelle - par rapport à laquelle seront décrites les phases en timing relatif (cf. pour un examen récent des modèles dits "proportionnels", GENTNER, 1987). Nous rappellerons que rares sont les études acoustiques utilisant ces concepts, d'abord opérationnels sur les signaux EMG ou de mouvement (cf. cependant WEISMER & FENNELL, 1985 et SOCK & al., 1987).

Dans ce contexte, nous examinerons pour l'arabe tunisien, sur le plan acoustique, quelques phasages dans les domaines intersegmentaux (VC et CV) et dans celui du mot(CVCVCV) qui les contient.

CORPUS

Le matériau phonétique comportait 4 mots choisis pour tester l'opposition de quantité en arabe tunisien pour les voyelles basses ([a], [a:]) et hautes ([u], [u:]) et dans des entourages consonantiques d'occlusives.

Nous avons utilisé les deux paires minimales suivantes, dans la phrase porteuse [qult / - - / kamel] " tu as dit : 'pommes de terre' ? , kamel ? " :

□ [ba'tata ~ ba:'tata] (deux variantes lexicales de "pommes de terre", toutes deux disponibles dans le dialecte de notre locuteur)

□ [ba'tuta ~ ba'tu:ta] (deux noms propres)

Les deux positions différentes ainsi obtenues par le hasard du lexique ([a], [a:] sont préaccentuées et [u], [u:] accentuées) ne nous permettront pas de comparer les voyelles entre elles, mais seulement les deux termes de la paire caeteris paribus.

La connaissance des caractéristiques intrinsèques de [a] et de [u] nous permet d'augurer une meilleure résistivité de la première opposition, même mise en contexte peu favorable; l'opposition sur [u] garde, de son côté, toutes les chances d'être mise en évidence dans ce contexte optimal.

La première série d'enregistrement (12 répétitions, en ordre aléatoire) a été faite à la vitesse d'élocution qui a paru la plus normale au locuteur (M. K.E, né dans le sud tunisien, âgé de 28 ans). La deuxième série a été ensuite enregistrée en

exigeant un débit rapide (le résultat obtenu est un facteur d'accélération de 1.30, en moyenne, mesuré, après l'enregistrement, par ajustement auditif d'un métronome sur le rythme syllabique).

Les réponses au test d'identification mené avec notre locuteur ont donné les résultats ci-dessous (exprimés en pourcentage de reconnaissances correctes). Nous constatons que l'identification de [a] et de [u] est généralement supérieure (ou égale pour [u:]) en débit normal par rapport au débit rapide; [u] bref en débit rapide atteint le seuil du hasard.

	A		U	
BREVES	92%	83%	75%	50%
LONGUES	100%	75%	83%	83%
DEBIT	NORMAL	RAPIDE	NORMAL	RAPIDE

MESURES

L'activité pseudo-cyclique de la parole nous a permis de retenir, à partir d' événements répétitifs (ABRY & al., 1985), trois cycles : un cycle détente pour le domaine VC (d'un relâchement à un autre), un cycle closion dans le domaine CV (d'une closion à une autre) et un cycle du mot, à l'intérieur desquels seront examinées deux phases répétitives :

□ La phase vocalique VVO-VVT (de l'événement VVO, ou apparition de la structure formantique vocalique, à la disparition de cette structure , VVT, soit la fin de la réalisation vocalique voisée).

□ La phase consonantique VVT-VVO (pour les consonnes intervocaliques) ou CVO-VVO (pour les consonnes initiales; CVO est le début du voisement consonantique).

Les mesures relatives ne sont données, pour l'opposition sur [a], que dans le cycle détente (nous n'avons pas fait confiance à la durée du prévoisement de la consonne [b] - fortement variable et piètre indicateur de la tenue de cette consonne). Par contre pour l'opposition sur [u] nous avons tenu compte des deux cycles environnants, VC et CV.

Les signaux numérisés (à 16 kHz sur 12 bits) ont été étiquetés manuellement à l'aide d'un éditeur de signal (BENOIT, 1984).

RESULTATS

1. Cycle détente (Fig. 1)

1.1. Opposition [a] ~ [a:]

□ Débit normal :

Les brèves, avec 67 ms de phase vocalique en moyenne, représentent 59% de la durée des longues. Nous constatons une nette différence entre les moyennes (de l'ordre de 15%) pour les phases vocaliques exprimées en pourcentage du cycle détente (Fig. 2).

Cette différence est largement significative ($t=17.10$; $t=1.72$ à $p\leq 0.10$), les deux classes étant séparées à 100%. La consonne suivant la brève est tout juste significativement plus longue que celle suivant [a:]. Cette différence disparaîtra en débit rapide (Fig. 1).

□ Débit rapide :

Les brèves avec 49 ms sont passées à 78% de la durée des longues. La différence reste légèrement significative en pourcentage du cycle détente ($t=3.17$). Mais il n'est plus possible de séparer à 100% les deux membres de la paire (Fig.2).

□ Entre débits :

Il n'y a pas de différence significative de phase vocalique sur les brèves en passant du débit normal au débit rapide : elles se maintiennent autour de 40% du cycle détente. Mais l'effet est très significatif pour les longues ($t=10.70$). Ceci est dû à une compression importante de ces dernières passant de 113 ms à 63 ms, soit de 58% à 46% du cycle.

En d'autres termes, il ne semble pas y avoir d'invariance même relative pour toutes les classe ; seules les voyelles brèves se révèlent relativement stables (si l'on fait abstraction de la dispersion).

Rappelons que dans la situation de débit rapide l'importante intersection des brèves avec les longues aboutit à une large confusion des longues (avec 75% d'identifications correctes), moins bien identifiées que les brèves (avec 83%).

1.2. Opposition [u] ~ [u:]

□ Débit normal :

Les brèves, avec 64 ms, représentent 80% des longues. Dans le cycle détente (Fig. 5), nous observons une légère différence de phase vocalique (autour de 7%) entre les brèves et les longues en débit normal. Cette différence est pourtant significative ($t=6.16$). Cependant nous n'avons jamais ici - contrairement à ce que nous avons observé pour [a] - de séparation à

□ Débit rapide :

L'effet des différences de moyenne entre les phases vocaliques des deux classes reste encore significatif ($t=5.16$). Les brèves affichent toujours 76% de la durée des longues, valeur tout à fait comparable aux 80% observés en débit normal.

□ Entre débits :

La compression pour les deux classes est faible. Elle est de l'ordre de 12 ms. Ceci ne change d'ailleurs rien au rapport de la phase vocalique dans le cycle détente pour les brèves (42% dans les deux débits) et les longues (49%).

Ainsi, dans le cycle détente, nous observons pour [u], et contrairement à ce qui se passe pour [a], une invariance certaine.

Mais une invariance qui, soit dit en passant, ne semble guère réussir qu'aux longues, puisque leur

identification se maintient (à 83 %) en débit rapide, alors que le [u] bref n'est plus reconnu qu'au hasard. Cette invariance serait-elle due à la situation priviligée de [u:] sous l'accent ? Nous ne pourrons répondre qu'en étudiant un autre corpus, comportant d'autres paires.

2. Cycle closion (pour [u] ~ [u:])

Il nous a semblé intéressant de comparer pour la voyelle [u] - avec laquelle cela était possible - le cycle détente au cycle closion (Fig. 8). En débit normal la différence est du même ordre que pour le cycle détente (t=5.30). Mais en débit rapide les choses se gâtent, dans ce cycle, avec une différence tout juste significative (t= 3.36).

3. Cycle du mot

3.1. Opposition [a] ~ [a:]

La phase vocalique présente globalement la même structure dans le mot (Fig. 3) que dans le contexte (VC) du cycle détente (Fig.2), avec une séparation des classes simplement moins nette (t=9.01) en débit normal.

La différence en débit rapide se maintient seulement, là aussi, sous forme d'effet (t= 3.01), les classes étant largement confondues.

La phase consonantique (Fig. 4), de la consonne qui suit la voyelle sur laquelle joue le contraste, permet de distinguer tout aussi bien la voyelle longue (avec une proportion, de la consonne dans le cycle du mot, autour de 18%) de la voyelle brève (avec une proportion de 22%). Cette différence est faible mais nettement significative (t= 7. 42) : c'est la traduction de la tendance "compensatrice" observée dans le cycle détente.

Nous assistons en débit rapide à la même perte de distinctivité pour la phase consonantique que pour la phase vocalique - l'effet est tout juste significatif (t= 2.38).

3.2. Opposition [u] ~ [u:]

Pour [u] les différences de phases vocalique et consonantique, qui étaient nettement significatives dans le cycle détente, en débit normal comme en débit rapide, s'atténuent dans le cycle du mot (Fig. 6 et 7) : elles se maintiennent autour de valeurs allant de 2.66 à 3.83 pour le t de Student. On remarque que les phases vocaliques des voyelles longues (Fig. 6) correspondent à des phases consonantiques relativement brèves (Fig. 7), et vice-versa pour les voyelles brèves.

En conclusion de cette comparaison de trois domaines temporels donnés par les cycles :

1. Le champ VC (cycle de détente) maximise plutôt les distinctions de quantité pour tous les débits; au contraire d'autres domaines, comme celui du mot, auquel on aurait pu d'abord penser, comme un cadre de référence plus naturel pour juger de l'invariance de l'action.

2. Là où nous avons pu comparer les trois cycles - c'est le cas de [u] - c'est bien le cycle de détente (VC) qui est meilleur que celui de closion; et ce domaine VC est encore meilleur que le cycle du mot.

Ceci confirme le choix du champ VC dans de nombreuses études sur l'organisation temporelle (LEHISTE, 1970).

CONCLUSION

Nous terminerons sur une remarque typologique nous permettant de situer brièvement le comportement de la quantité de l'arabe tunisien par rapport aux systèmes pour lesquels nous possédons des résultats quantitatifs.

Notre voyelle basse [a], comme la haute [u], ne peuvent certainement pas se contraster efficacement, en quantité, en maintenant une brève à environ 80% de la longue (78% pour [a] en débit rapide; 80% pour [u] en débit normal et 76% en débit rapide). Seule la voyelle brève [a] en débit normal affiche 60% de la longue .

Nous observons sur ce point des résultats convergeant avec ceux de NORLIN (1987) pour l'arabe égyptien (56%). Mais différents de ceux de PORT & al. (1980) - portant sur l'arabe standard prononcé par des locuteurs égyptiens, koweitiens et irakiens - pour lesquels la voyelle brève [a] ne présente que 39% de la longue (dans un contexte d'occlusives sourdes) en débit normal; alors qu'en débit rapide elle atteint 53% de cette même longue.

Nous assistons, en tous cas pour notre locuteur tunisien, à des oppositions de quantité peu résistives aux variations de la vitesse d'élocution (cas de [a]); ou bien résistives, mais à trop faible contraste (cas du [u]).

Comparé à d'autres langues pour lesquelles nous avons des données sur les variations de débits, cette variété d'arabe apparaît beaucoup plus éloignée des systèmes à voyelles ultra-brèves très résistives, comme le wolof de Gambie (SOCK, 1983), que des systèmes rencontrés dans les langues germaniques, comme le suédois qui affiche des contrastes brèves/longues passant de 60 à 80%, peu résistifs (ELERT, 1964).

Est-ce pour cette raison que l'arabe, à travers ses différentes variétés, semble avoir particulièrement besoin de faire reposer ses distinctions de quantité - face aux variations qu'elles peuvent subir selon le tempo d'élocution - sur des variations concomitantes de qualité vocalique (GHAZALI,1979) ? L'étude typologique du contrôle linguistique de la durée vocalique, en fonction des domaines de programmation, de celle-ci, nous semble, quoi qu'il en soit, de la plus grande importance pour la compréhension des règles de compression et d'expansion de la parole, par rapport aux universaux de son organisation temporelle.

REMERCIEMENTS : A Louis-Jean BOE, Rudolph SOCK, Mahmoud HELAL, Argyro Tseva et Bettina SCHNABEL pour leur aide toujours efficace dans l'utilisation des logiciels de l'IPG.

REFERENCES BIBLIOGRAPHIQUES

ABRY C. BENOIT C. BOE L. J. & SOCK R. (1985)
Un choix d'événements pour l'organisation temporelle du signal de parole.
14 èmes JEP du GCP du GALF, 111-137.

BENOIT C. (1984)
EDISIG : Encore un éditeur de signal ?
13 èmes JEP du GCP du GALF, 211-213.

ELERT C. C. (1964)
Phonologic Studies of Quantity in Swedish.
Almqvist & Wiksells, Uppsala.

GENTNER D.M. (1987)
Timing of Skilled Motor Performance : Tests of The Proportional Duration Model.
Psychological Review 82, 225-460

GHAZALI S. (1979)
Du statut des voyelles en arabe.
In Analyses et Théories, Etudes arabes 2-3, 199-219.

JOMAA M. (1987)
Etude sur l'organisation temporelle de l'opposition de quantité vocalique en arabe tunisien.
Sa résistivité aux variations de la vitesse d'élocution.
Mémoire de D.E.A. Université des Langues et Lettres de Grenoble III

LEHISTE I. (1970)
Suprasegmentals.
The M.T.I. Press Cambridge Mass.

MAC KAY D. G. (1987)
The Organisation of Perception and Action. A Theory for Language and other Cognitive Skills.
Cognitive Sciences Series. Eds.

SEBRECHT-FISCHER M. M. &
FISCHER P. M. Springer Verlag N. Y, Berlin.

NORLIN K. (1987)
A Phonetic Study of Emphasis and Vowels in Egyptian Arabic.
Working Papers 30, Lund University, 1987, 73-87

PORT R. F. AL-ANI S. & MAEDA S. (1980)
Temporal Compensation and Universal Phonetics.
Phonetica 37, 235-252.

SOCK R. (1983)
L'organisation temporelle de l'opposition de quantité vocalique en wolof de Gambie.
Thèse de 3ème cycle de phonétique. Université des Langues et Lettres de Grenoble III.

SOCK R. OLLILA L. DELATTRE C. ZILLIOX C. & ZOHAIR L. (1987)
Timing intersegmental et intrasegmental en français.
16 èmes JEP DU GCP du GALF, 233-236.

TAMOULDI F. (1984)
The Diglossic Situation in North Africa : A Study of Classical Arabic/Dialectal. Arabic Diglossia with Sample Texts in Mixed Arabic.
Orientalia Gothoburgensia 8.

TULLER B. & KELSO J. A. S. (1984)
The Timing of Articulatory Gestures : Evidence for Relational Invariants.
J. Acoust. Soc. Am. 76, 1534-1543.

WEISMER G. & FENNELL A. M. (1987)
Constancy of Acoustic Relative Timing Measures in Phrase -Level Utterances.
J. Acoust. Soc. Am. 78, 49-57.

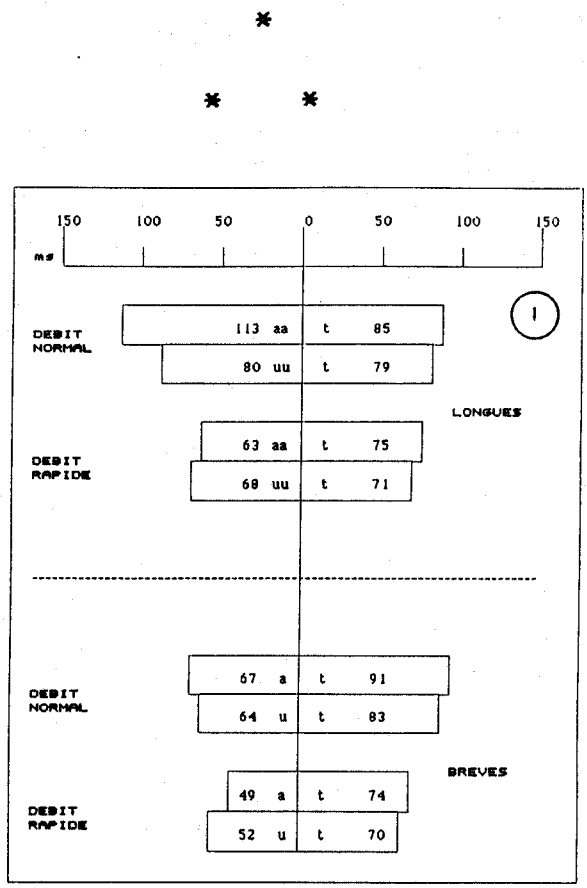
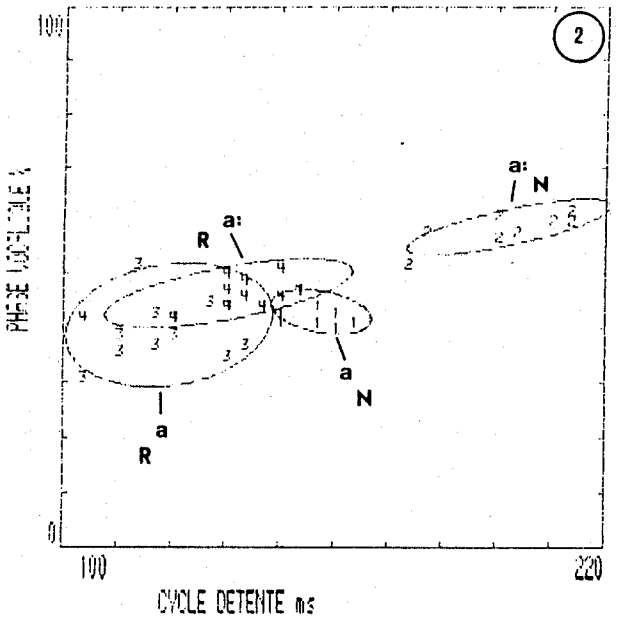


FIG. 1

Comportement moyen des brèves et des longues dans le cycle détente (VC) du débit normal au débit rapide pour les phases vocaliques et consonantiques (valeurs moyennes absolues en ms.). On remarquera la perte, en débit rapide, des différences de durée intrinsèque (significatives en débit normal sur les longues) entre [a] et [u], (t=6.04).



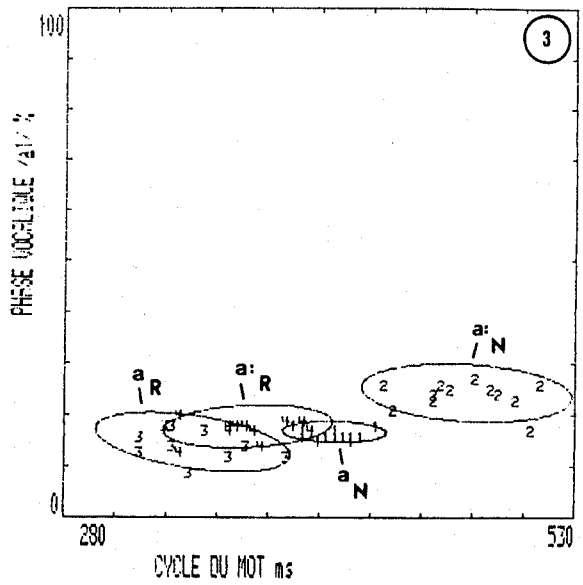
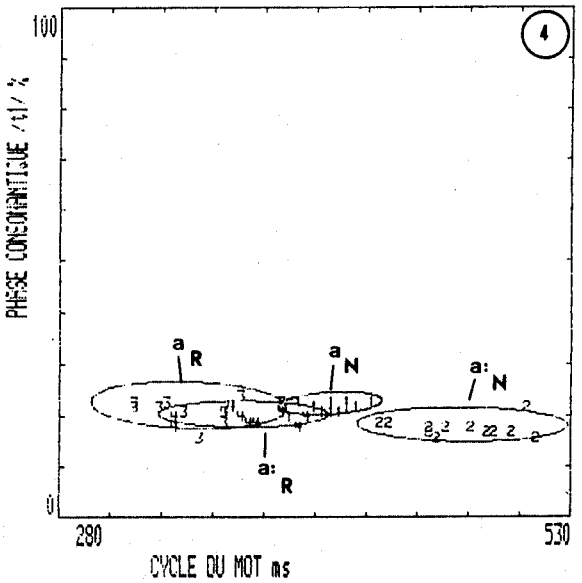
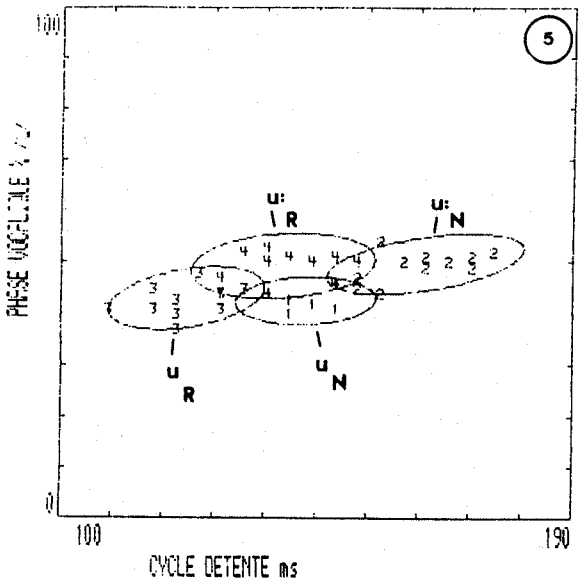


Fig. 2.3.4.
Evolutions pour les voyelles [a] des phases vocaliques en fonction des cycles de détente (fig.2) et du mot (fig. 3) et de la phase consonantique (fig.4) en fonction du cycle du mot suivant les variations des débits normal (N) et rapide (R).





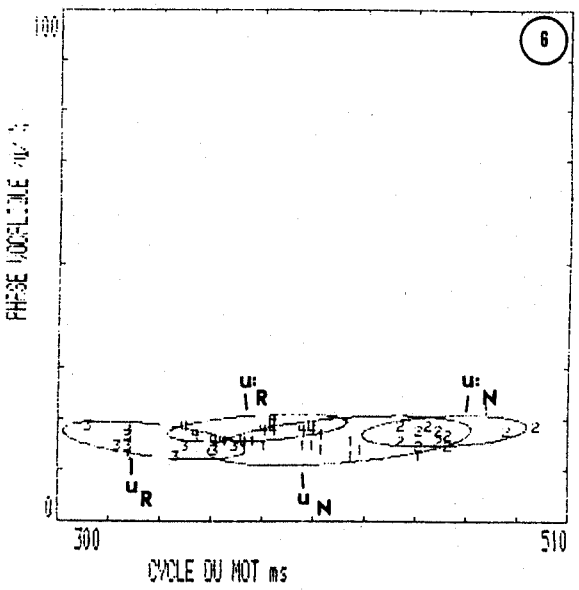
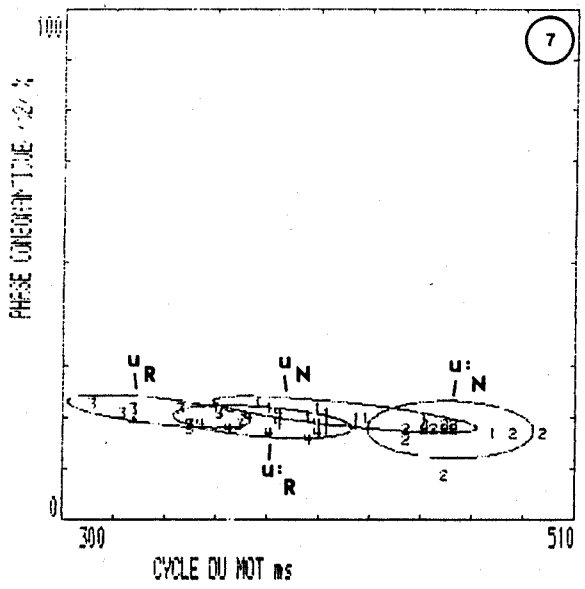
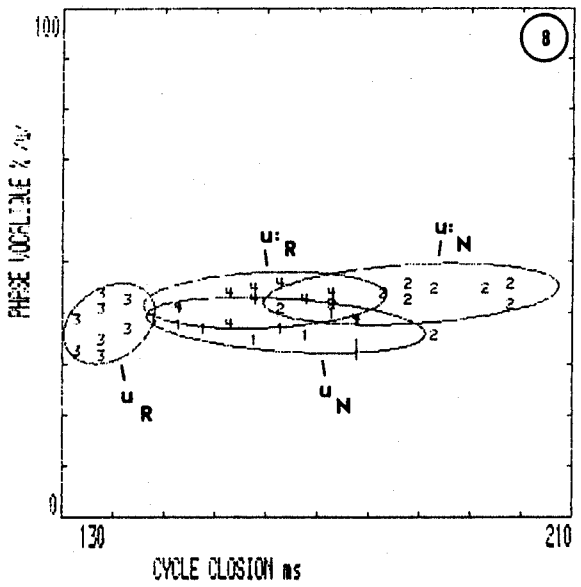


FIG. 5. 6. 7. 8.

Evolutions pour les voyelles [u] des phases vocaliques en fonction des cycles de détente (fig.5), du mot (fig. 6) et celui de closion (fig. 8); et de la phase consonantique (fig; 7) en fonction du cycle du mot, suivant les variations des débits normal (N) et rapide (R).





*

* *

Analyse-synthèse

SYNTHESE DE LA PAROLE PAR CONCATENATION DE FORMES D'ONDES

C. HAMON

Centre National d'Etude des Telecommunications
B.P. 40
22301 LANNION Cedex

ABSTRACT

In order to perform text to speech synthesis by concatenation of diphones we propose an algorithm with very low computational cost, an important but not excessive memory requirement, and good speech synthesis sound quality. Our algorithm, waveform-based, work by overlap-adding pseudo-impulse responses (PIR) created by a special windowing of two consequently pitch periods in voiced PIR case and simply cutting and splicing unvoiced segment in the other case.

1. INTRODUCTION

Cette communication concerne la synthèse de la parole à partir du texte et porte plus particulièrement sur une configuration de traitement de signal de parole qui transforme une chaîne phonétique accompagnée des paramètres prosodiques (durée et hauteur de la voix) pour chaque phonème en un signal audible.

La synthèse de la parole à partir du texte nécessite généralement la transformation d'une chaîne orthographique en une chaîne phonétique, le calcul des paramètres prosodiques pour chaque phonème et la transformation de ces données en un signal numérique de parole puis, après conversion numérique-analogique, en un signal analogique. Ce procédé de génération du signal de parole permet l'interfaçage entre des machines et des auditeurs, en écoute directe ou à travers le réseau téléphonique.

Les configurations de traitement de signal pour la synthèse de la parole basées sur un modèle mathématique du conduit vocal, la synthèse par prédiction linéaire L.P.C [1], la synthèse à formant [2] et la synthèse FFT [3] font intervenir une déconvolution de la source et de la fonction de transfert du conduit vocal. Elles nécessitent en général une cinquantaine d'opérations par échantillon. Lors de la synthèse par diphone, cette déconvolution source-conduit vocal permet d'une part la modification de la valeur de la fréquence fondamentale de l'excitation, et d'autre part la compression des données représentant chaque diphone : le conduit vocal peut être représenté par un modèle considéré comme stationnaire durant plusieurs millisecondes, l'excitation étant modélisée soit par un train d'impulsions à la fréquence fondamentale ou un bruit blanc, soit par un modèle plus élaboré, du type excitation multi-impulsionnelle.

La synthèse micro-phonémique[4] et la synthèse par concaténation de formes d'ondes que nous proposons n'extraient pas du signal de parole la fonction de transfert du conduit vocal mais, implicitement, utilisent l'hypothèse suivante: les sons voisés sont la somme de réponses impulsionnelles d'un filtre (conduit vocal) excité par une suite de Dirac de façon synchrone de la fréquence fondamentale (source).En isolant chaque réponse impulsionnelle et en la replaçant sur l'échelle des temps ,nous modifions ainsi la fréquence fondamentale des sons voisés. La synthèse micro-phonémique n'offre une qualité de son jugée convenable que pour un abaissement du pitch [4], le problème de la compression des données étant abordé et résolu par un lissage des formes d'ondes entre points-clef du signal de parole.L'algorithme que nous proposons permet l'augmentation et l'abaissement de la fréquence fondamentale et la modification dans des limites raisonnables, du rythme de la parole. La qualité de la synthèse est jugée bonne et le nombre de calculs par échantillon est très faible.

2. L'ALGORITHME DE SYNTHESE

La configuration de traitement de parole que nous présentons modifie les unités phonétiques élémentaires (appelées diphones) sur la dérivée première du signal numérique, puis effectue l'opération inverse d'intégration sur le signal transformé, de façon à concentrer la puissance du bruit apporté par ces transformations dans la partie basse du spectre de fréquence, où le spectre de parole de forte énergie masquera l'effet de ce bruit. Une autre spécificité de cette configuration de traitement de signal est l'extraction, pour les sons voisés, de pseudo-réponses impulsionnelles à partir de deux périodes consécutives du signal et d'une fenêtre spécialement conçue pour limiter la perte de naturel du signal de parole due aux transformations opérées sur le signal numérique. Ces pseudo-réponses impulsionnelles sont replacées sur l'échelle des temps, de façon synchrone de la fréquence fondamentale (commandée par le module de génération de la prosodie de synthèse), puis additionnées pour produire le signal de parole dérivé dans le cas des sons voisés.

La méthode de dérivation-intégration permet de réduire les bruits apportés par les transformations du signal original des diphones[5]. Cette opération d'intégration ne coûte qu'une multiplication et une addition par échantillon du signal de parole numérique. Nous l'avons donc retenue pour la synthèse par concaténation de formes d'onde. Cette caractéristique essentielle de la méthode de synthèse lui donne son nom , **K.D.G** , abréviation de "kemmadenn deveradenn ar gomz" qui signifie en breton modification de la dérivée de la parole.

3. DESCRIPTION DE L'ALGORITHME

Le schéma synoptique de la figure 1.a permet de mieux comprendre le déroulement de la méthode de synthèse. Une liste de phonèmes accompagnée des paramètres prosodiques constitue l'entrée de la synthèse [6] [7]. Ces paramètres prosodiques sont:

-la période fondamentale en début et fin de phonème.

-la durée du phonème .

Le dictionnaire de diphone est prétraité: un filtre à réponse impulsionnelle finie, d'ordre 1, dérive le signal numérique de parole:

$$s(k) = e(k) - e(k-1)$$

Sur ce signal dérivé, est marqué semi-automatiquement le début de l'ouverture de la glotte pour les sons voisés (figure 2.1), considéré comme le début de la réponse du conduit vocal à l'excitation des cordes vocales. Ces marques de voisement, ainsi que le début , le milieu et la fin de chaque diphone, sont stockés dans un descripteur du dictionnaire de diphones.Ces traitements sont effectués une seule fois, hors ligne, pour chaque voix de synthèse.

Le nom du phonème courant, du phonème précédent et du phonème suivant permet le décodage du nom des deux diphones qui constituent le phonème courant.

Nous extrayons à partir, des sons voisés, des pseudo-réponses impulsionnelles qui permettent de recréer un nouveau signal de parole avec une structure harmonique et avec les mêmes formants (figure 4.2) que le signal de parole original (figure 4.1). Le module de rythme pour les sons voisés tient compte à la fois de la durée du phonème et du changement de la fréquence fondamentale du signal de parole (fixés par le module de prédiction de la prosodie), et de la durée intrinsèque du phonème original pour déterminer le nombre de périodes de voisement à éliminer ou à ajouter à ce phonème.

Le traitement des sons non voisés est plus simple car, le spectre n'ayant pas de structure harmonique, seuls la concaténation des deux diphones et le changement de rythme sont à réaliser. Une rampe viendra pondérer le début du premier phonème d'une séquence non voisée, la fin du premier diphone et le début du deuxieme diphone pour tous les phonèmes non voisés. Le module de rythme détermine le nombre d'échantillons à éliminer à la frontière des deux diphones mais seul le phonème silence peut être rallongé en intercalant un silence (échantillons de valeur nulle) entre les deux diphones.

Les signaux non voisés et voisés forment le signal de parole numérique dérivé. Celui-ci , intégré puis converti en analogique, est envoyé vers une chaîne d'amplification et un transducteur électro-acoustique.

Nous allons maintenant détailler séparément le traitement des sons voisés et non voisés.

3.1 TRAITEMENT DES SONS VOISES

Le schéma de la figure 1.b illustre le traitement des sons voisés. Une pseudo-réponse impulsionnelle (P.R.I) (*figure 2.e*) est créée à partir de deux périodes consécutives (*figure 2.b*) et d'un fenêtrage.La fenêtre utilisée est représentée par la fonction:

$$h(i) = 0.5 - 0.5.\cos\left(2.\pi.\left(\frac{i}{T}\right)^2\right) \quad 0 < i < T$$

T représente en moyenne deux périodes du signal voisé.

Tout en conservant une partie du signal de la période précédente, nécessaire au lissage entre les P.R.I, cette fenêtre (*figure 2.c*), en accentuant le centre de la deuxième période (période courante), diminue fortement la raucité de la voix, due à une énergie trop importante sur le fondamental et sur la période précédente, déphasée par rapport à la période courante (résidu du voisement original). Le spectre de la fenêtre utilisée est proche de celui de la fenêtre de Hanning (*figure 2.d*) ; La convolution de ce spectre et du spectre de chaque P.R.I. dans le dommaine fréquenciel n'entraîne donc pas de modifications majeures sur ce dernier.

3.1.1 MODIFICATION DE LA FREQUENCE FONDAMENTALE

Deux cas se présentent:

a) diminution de la période courante:

La période T de la fenêtre est égale à la somme de la période précédente du signal naturel et de la période courante diminuée de deux fois le nombre d'échantillons nécessaires à la réalisation de la période commandée par la prosodie, de façon à conserver une répartition de l'énergie prédominante sur la période courante:

$$T(k) = prd\_naturelle(k-1) + prd\_naturelle(k) - 2.(prd\_naturelle - prd\_prosodie)$$

k:numéro de la période courante.

Le signal de la période courante ne déborde pas sur la période suivante, ce qui améliore le timbre de la voix. De plus l'énergie de chaque P.R.I est proportionnelle à la fréquence fondamentale: le signal de parole conserve la même énergie en fonction du temps.

b) augmentation de la période courante:

La période T de la fenêtre est égale à la somme de la période précédente et de la période courante du signal naturel:

$$T(k) = prd\_naturelle(k-1) + prd\_naturelle(k)$$

k=numéro de la période courante.

Les P.R.I sont ajoutées de façon synchrone de la fréquence fondamentale (*figure 3.a*). La continuité est assurée entre les phonèmes voisés de façon à toujours disposer de deux périodes consécutives pour créer les P.R.I .

3.1.2 MODIFICATION DU RYTHME DES SONS VOISES

L'augmentation de la durée peut être considérée comme la mise en correspondance, par déformation de l'axe des temps du signal de synthèse, des n marques de voisement du signal d'analyse et des p marques du signal de synthèse [3]. A chaque marque du signal de synthèse est associée la marque la plus proche du signal d'analyse. Ainsi la duplication de P.R.I également réparties sur tout le phonème modifiera la durée de celui-ci. Toutefois, multiplier la durée du phonème par un facteur important (supérieur à 2) détériore la qualité du signal de parole, plusieurs périodes consécutives se retrouvant après duplication en phase [8]. La suppression de P.R.I à partir du milieu du phonème diminue la durée de celui-ci. Nous imposons de garder au minimum quatres P.R.I pour ne pas introduire de trop brusques transitions du signal de parole.

3.2 TRAITEMENT DES SONS NON VOISES

La concaténation se fait en appliquant une rampe de 3ms (soit 50 échantillons à 16 kHz de fréquence d'échantillonnage):

- au début du phonème, si le phonème précédent est voisé, afin de diminuer la transition avec la dernière P.R.I .
- à la fin du premier diphone et au début du second diphone constituant le phonème afin d'éviter de trop grandes transitions, notamment pour les phonèmes sourds de forte énergie tel que le 'ch'.

La suppression d'un certain nombre d'échantillons au milieu du phonème diminue la durée de celui-ci. Compte-tenu des modalités d'extraction des diphones, il ne s'avère pas nécessaire lors de la synthèse, d'augmenter la durée des phonèmes non voisés, sauf pour le phonème silence qui sera rallongé en intercalant des échantillons de valeur nulle au milieu du phonème.

3.3 INTEGRATION DU SIGNAL

Le signal obtenu après traitement des phonèmes voisés et non voisés est intégré par un filtre RII d'ordre 1 et de coefficient égal à 0.98 (*figure 3.b*).Il est ensuite envoyé à un convertisseur numérique-analogique.

4. MISE EN OEUVRE NUMERIQUE DE L'ALGORITHME DE SYNTHESE

Coût de calcul de la synthèse K.D.G:

Pour un signal voisé, le nombre d'opérations est en moyenne:

- de 2 fois 2 multiplications car chaque échantillon fait partie de deux P.R.I consécutives,

- 2 additions pour la sommation des P.R.I,

-1 multiplication pour l'intégration du signal,

-1 addition pour l'intégration du signal;

soit **5 multiplications et 3 additions par échantillon pour un signal voisé.**

Pour un signal non voisé le nombre d'opérations est en moyenne:

-1 multiplication pour les frontières (10% des échantillons),

-1 multiplication pour l'intégration du signal,

-1 addition pour l'intégration du signal;

soit **moins de 2 multiplications et une addition par échantillon pour un signal non voisé.**

5. MISE EN OEUVRE MATERIELLE DE L'ALGORITHME

Le programme de synthèse fonctionne sur le processeur central d'un ordinateur personnel (IBM PC). Une carte de conversion numérique-analogique au format IBM PC, de fréquence d'horloge égale à la fréquence d'échantillonage (16 kHz), produit le signal de parole analogique à partir du signal numérique calculé par le programme de synthèse. Une chaîne analogique, comprenant un amplificateur et un transducteur électro-acoustique restitue le signal de parole de synthèse. Ce procédé de synthèse fait l'objet d'une demande de brevet.

6. CONCLUSION

Nous avons décrit une configuration de traitement du signal de parole qui ne nécessite qu'un faible nombre d'opérations par échantillon et fournit un timbre de voix de synthèse proche de celui de la parole naturelle; un test effectué par 16 personnes sur un corpus de 12 phrases de synthèse de différentes prosodies, voix homme, montre que la synthèse KDG est préférée par 88% des auditeurs contre 8% pour le synthèse LPC simulée [1] (3% équivalent). Une voix feminine a été synthétisée avec une qualité comparable à celle de la voix masculine. La taille du dictionnaire de diphones peut être diminuée de moitié (soit 2.5 Méga-octets) en échantillonnant le signal de parole à la fréquence de 8 kHz. Un codage classique du signal de parole, à 32Kb/s par exemple, peut être envisagé pour réduire la taille du dictionnaire de diphones. Cependant le décodage 32kb/s est plus "gourmand" en calculs que la synthèse LPC, ce qui limitera l'emploi de cette technique à des cas particuliers.

REFERENCES BIBLIOGRAPHIQUES

[1] STELLA M. , "Speech Synthesis",in COMPUTER SPEECH PROCESSING, F.Fallside and W.A. woods ed., Prentice Hall, 421-460, 1983

[2] FANT G. , "Acoustic theory of speech production (Moutons, gravenhage, the Netherlands), 1961

[3] CHARPENTIER F.J. and STELLA M.G. , "Diphone synthesis using an overlap-add technique for speech waveform concatenation", ICASSP 86 , 2015-2018

[4] LUKASZEWICZ K., KARJALAINEN M. , "Micro-phonemic method of speech synthesis", ICASSP 87

[5] HAMON C.,"Synthèse par concaténation de formes d'ondes" , Note technique n° NT/LAA/TSS/353 CNET, 1988

[6] COURBON J.L., EMERARD F.,"SPARTE: a text-to-speech machine using synthesis by diphones", ICASSP 82, 1597-1600

[7] SORIN C., DESCOUT R. and all ,"text-to-speech synthesis in the French electronic mail environment", Speech Tech, Edimburg, sept 87

[8] KANG G. S. , "Improvement of the excitation source in the narrow-band linear prediction vocoder", IEEE Trans ASSP, vol.33(2), APRIL 1985

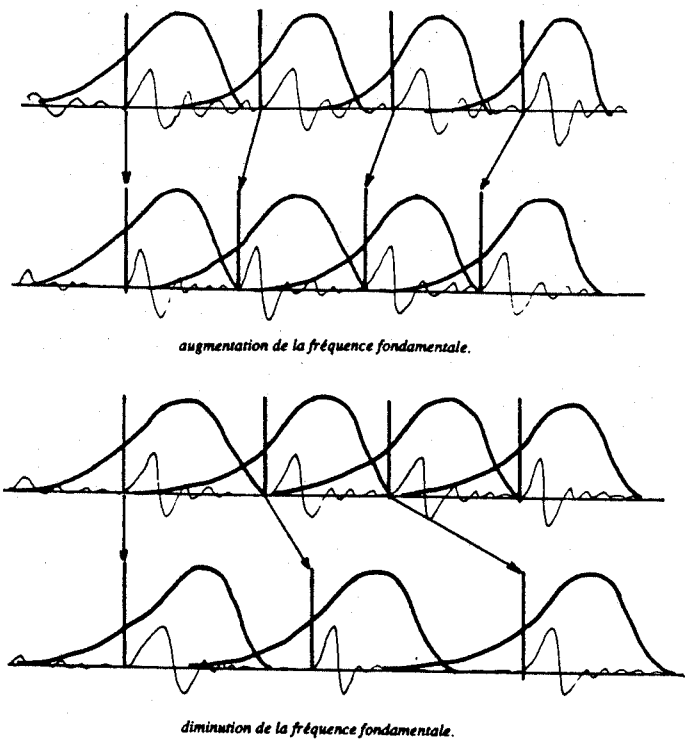


figure 1.b :traitement des sons voisés

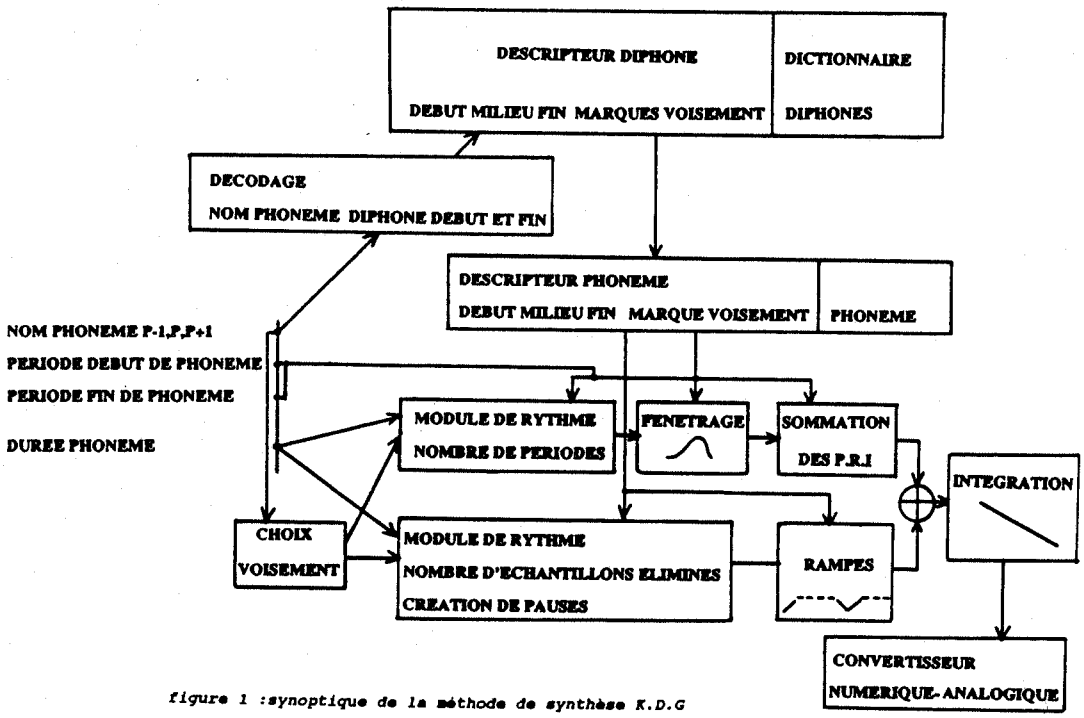


figure 1 :synoptique de la méthode de synthèse K.D.G

convolution sur trop de femmes par fenêtre + petite (= réponse fréquentielle + grande) élargit les pulsations d'où la dégradation de parole

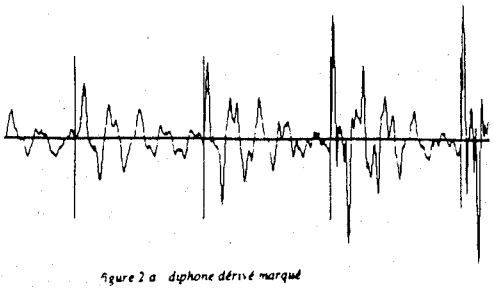


figure 2.a diphone dérivé marqué

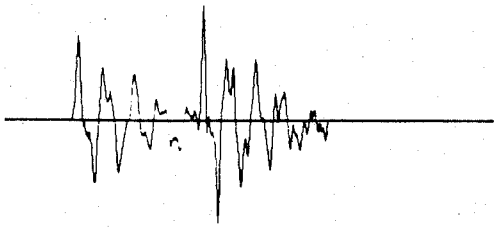


figure 2.b deux périodes consécutives

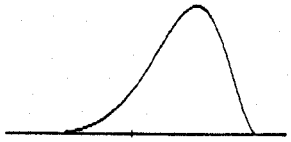


figure 2 : fenêtre

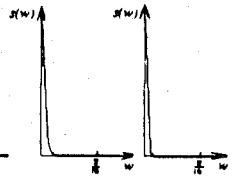


figure 2.d spectre de la fenêtre, fenêtre de Hanning

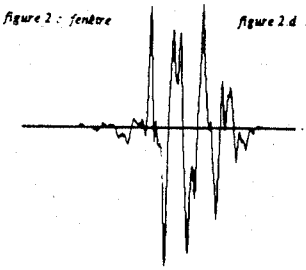


figure 2.e Pseudo-Réponse Impulsionnelle

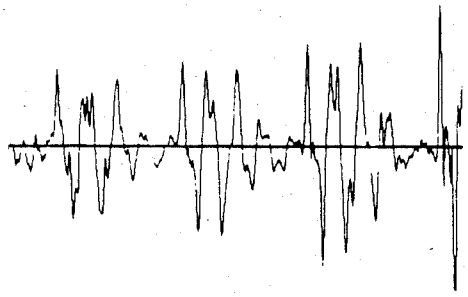


figure 3.a signal de synthèse dérivé

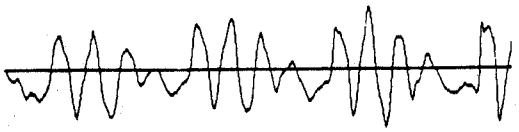


figure 3.b signal de synthèse intégré

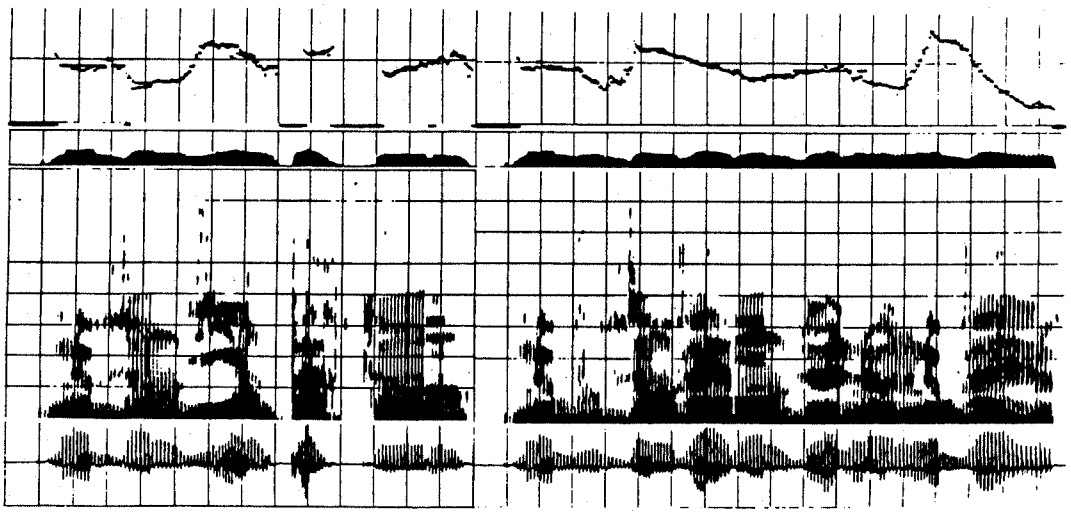


figure 3.c :sonagramme d'une phrase de synthèse KDG

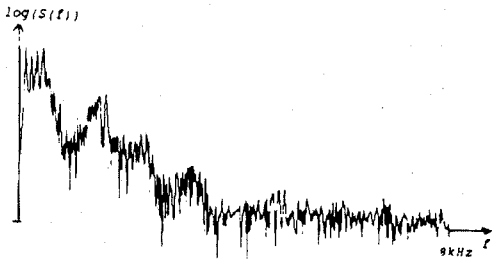


figure 4.a spectre du signal original

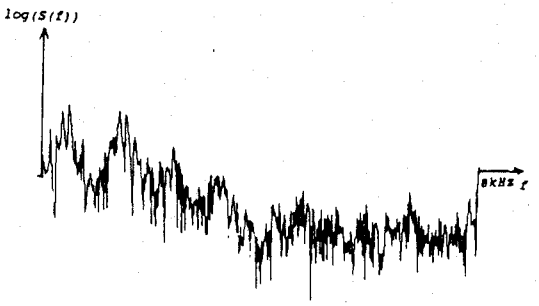


figure 4.b spectre du signal de synthèse dérivé

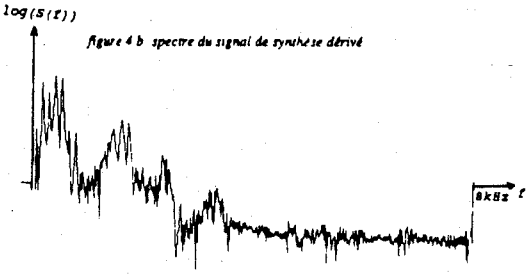


figure 4.c spectre du signal de synthèse intégré

ANALYSE-SYNTHESE DE LA BANDE DE BASE PAR FORMES D'ONDES ELEMENTAIRES

C.d'Alessandro

LIMSI-CNRS: BP 30 F-91406 ORSAY Cedex

ABSTRACT

This paper is a continuation of our work on the representation of speech signal by a set of well-localized time-frequency energy concentrations (elementary waveforms). We present here a method and a system for analysis-synthesis of the speech "baseband" (which will be defined below). The quality problems which were encountered in the lower part of the spectrum during modelling and synthesis are thus circumvented. We use the same kind of method for the processing of both formantic areas and "baseband". After an introduction in section 1, we describe the synthesis formulae in section 2 and the system developed according to these principles in section 3. Some conclusions are presented in section 4.

1 INTRODUCTION

Le traitement automatique de la parole reste tributaire de la représentation préalable du signal qui en est le support acoustique. Parmi les nombreuses méthodes disponibles, l'analyse en formes d'ondes élémentaires se présente comme un moyen neuf, et prometteur dans la mesure où il vise à une représentation permettant une reconstruction parfaite du signal et manipulant des objets pertinents tant du point de vue de la perception que de celui de la production [Liénard 87].

Dans un papier précédent, une méthode de représentation du signal de parole en fonctions d'ondes élémentaires a été développée, en se basant sur une décomposition en parallèle de la fonction de transfert du conduit vocal [d'Alessandro 87]. Les fonctions d'ondes élémentaires apparaissent comme des contributions bien localisées dans le plan spectro-temporel, et permettent ainsi de rendre compte des phénomènes de production (formants, excitations du conduit vocal, explosions ...) de façon explicite, par un ensemble discret d'éléments. L'exploitation de la structure particulière du signal de parole guide la recherche des formes d'ondes élémentaires dans les régions de maximum d'énergie spectrale ("formants", au sens large) et temporelle ("impulsions", au sens large) et permet l'obtention de paramètres perceptivement pertinents [d'Alessandro 88].

Notre système d'analyse-synthèse en fonctions d'ondes élémentaires (s'appuyant sur les fonctions d'ondes formantiques) permettait une bonne représentation du signal de parole, sans perte de qualité du point de vue perceptif, sauf dans la "bande de base" (région spectrale jusqu'au premier formant inclus) où des problèmes de modélisation apparaissaient. Par une démarche semblable à celle adoptée dans les vocodeurs à bande de base, nous présentons ici une nouvelle méthode pour décomposer la bande de base du signal en formes d'ondes élémentaires utilisant un processus d'analyse-synthèse analogue à celui employé précédement et s'appuyant sur une représentation sinusoïdale.

2 REPRESENTATIONS

2.1 représentation formantique

Le modèle linéaire classique de production du signal vocal suppose le filtrage d'une certaine fonction d'excitation par un filtre linéaire évoluant dans le temps.

$$s(t) = e(t) * R(t)$$

- e(t): signal d'excitation.
- R(t): réponse impulsionnelle du filtre.
- s(t): signal résultant.

Dans ce qui suit on suppose le signal stationnaire (sur une tranche de temps assez courte) et donc le filtre de réponse impulsionnelle R invariant. Ce filtre, associé au conduit vocal, peut être décomposé en n sections parallèles, chacune d'elle représentant une résonance (ou formant). Dans le domaine temporel il est ainsi possible d'identifier le signal de parole avec la somme des réponses de chaque section au signal d'excitation. En première aproximation, si celui-ci est constitué d'une série d'impulsions idéales, on peut écrire:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * R_i(t)$$

où R_i représente la réponse impulsionnelle de la $i^{ième}$ section, et $\delta_0(t - t_j)$ une impulsion d'excitation à l'instant t_j.

Si l'on assimile de plus les sections parallèles à des résonateurs du second ordre [Klatt 80], alors:

$$R_i(t) = G_i e^{-a_i t} sin(\omega_i t + \phi_i)$$

où α_i règle la largeur de bande (à -6 dB du sommet), G_i le gain à la résonance, ω_i la fréquence centrale du $i^{ième}$ résonateur et ϕ_i sa phase.

soit:

$$s(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} \delta_0(t - t_j) * (G_i e^{-\alpha_i t} sin(\omega_i t + \phi_i))$$

Pour une représentation en formes d'onde, on peut de plus rendre indépendantes les excitations des différentes sections, ce qui affine le compromis entre précision fréquencielle et précision temporelle en le localisant, et estimer les paramètres pour chaque réponse impulsionnelle; le pavé spectro-temporel où l'on suppose le signal stationnaire est ainsi délimité par la forme d'onde:

$$s(t) = \sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (G_{j_i} e^{-\alpha_{j_i} t} sin(\omega_{j_i} t + \phi_{j_i}))$$

Les fonctions d'ondes élémentaires choisies sont donc ici identiques aux fonctions d'ondes formantiques [Rodet 80].

2.2 représentation sinusoïdale.

Pour la partie grave du spectre (en deçà du premier formant), l'utilisation d'un signal d'excitation trop simple pose de sérieux problèmes de qualité. Pour pallier à ce défaut de nombreux modèles d'excitation ont été proposés, en particulier la représentation sinusoïdale [McAulay 86]:

$$s(t) = \sum_{i=1}^{k} A_i sin(\omega_i t + \phi_i)$$

Le nombre k de sinusoïdes ainsi que l'amplitude A_i, la fréquence ω_i et la phase ϕ_i évoluent dans le temps et doivent donc être estimés sur une tranche de temps pendant laquelle le signal est quasi-stationnaire. Une alternative aux différentes méthodes proposées pour cette estimation est l'utilisation de formes d'ondes élémentaires pour représenter chaque segment de sinusoïde, pendant la durée desquelles on suppose le signal stationnaire:

$$s(t) = \sum_{i=1}^{k} \sum_{j_i=1}^{l_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i}))$$

l'enveloppe temporelle choisie $env_{j_i}(t)$ doit permettre la reconstitution de la sinusoïde initiale; il s'agira par exemple de segments sinusoïdaux.

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_1}} t))$$

pour $0 \leq t < \pi/2\beta_{j_{i_1}}$

$$env_{j_i}(t) = 1/2(1 + cos(\beta_{j_{i_2}}(t - \pi/2\beta_{j_{i_1}}) + \pi/2))$$

pour $\pi/2\beta_{j_{i_1}} \leq t < \pi/2\beta_{j_{i_2}} + \pi/2\beta_{j_{i_1}}$

Les β_i sont calculés de façon à conserver un nombre de cycles constant dans chaque forme d'onde (qui est alors une ondelette au sens de [Goupillaud 85]).

2.3 représentation par formes d'ondes

La représentation par forme d'onde complète s'appuie sur une segmentation spectrale préalable, qui permet de dégager les régions de maximum d'énergie, auxquelles est appliquée une représentation en formes d'ondes élémentaires de type "formantiques" (au delà du premier maximum) ou "sinusoïdales" (en deçà du premier maximum) (fig. 1).

$$s(t) = (\sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \delta_0(t - t_{j_i}) * (A_{j_i} env_{j_i}(t) sin(\omega_{j_i} t + \phi_{j_i})))+$$

$$(\sum_{a=1}^{k} \sum_{b_a=1}^{l_a} \delta_0(t - t_{b_a}) * (G_{b_a} e^{-\alpha_{b_a} t} sin(\omega_{b_a} t + \phi_{b_a})))$$

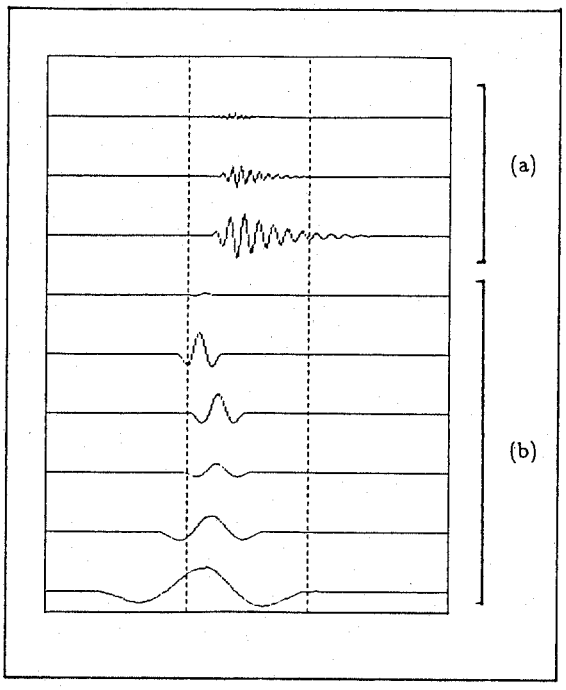


Figure 1: Modèles (a) "formantiques", (b) "sinusoïdaux" de formes d'ondes élémentaires

3 PROCESSUS D'ANALYSE-SYNTHESE

Le processus d'analyse-synthèse en fonctions d'ondes formantiques a déjà été décrit. Rappelons qu'à la suite d'une modélisation de l'enveloppe spectrale, un filtrage à phase nulle permettait d'obtenir des signaux à bande large centrés sur les maxima spectraux. Les formes d'ondes élémentaires etaient ensuite détectées grâce à l'enveloppe temporelle de ces signaux, puis modélisées comme précédemment pour la synthèse.

Pour le traitement de la bande de base, un procédé analogue est mis en œuvre, trame par trame (les trames sont de durée assez courte, 6 ms, pour que l'on puisse supposer le signal quasi-stationnaire):

- Modélisation de l'enveloppe spectrale par prédiction linéaire.
- Détection des maxima spectraux, associés aux formants, et définition de la "bande de base", comme la région spectrale jusqu'au premier maximum inclus.
- Calcul du module de la transformée de Fourier d'une tranche de signal centrée sur la trame.

- Recherche des maxima spectraux, associés aux harmoniques pour de la parole voisée, et segmentation spectrale autour de ces maxima. Tout ce qui suit ne concerne évidemment plus que la bande de base (fig. 2).
- Filtrage à phase nulle dans chacune des bandes ainsi définies, pour obtenir des signaux à bande étroite (fig. 3).
- Dans chaque bande, détection des formes d'onde (qui appartiennent à la trame si leur maximum y appartient) par recherche des maxima du signal.
- Synthèse, par les formules vues précédemment, après estimation des paramètres d'amplitude, de fréquence, de phase et d'enveloppe de chaque forme d'onde.

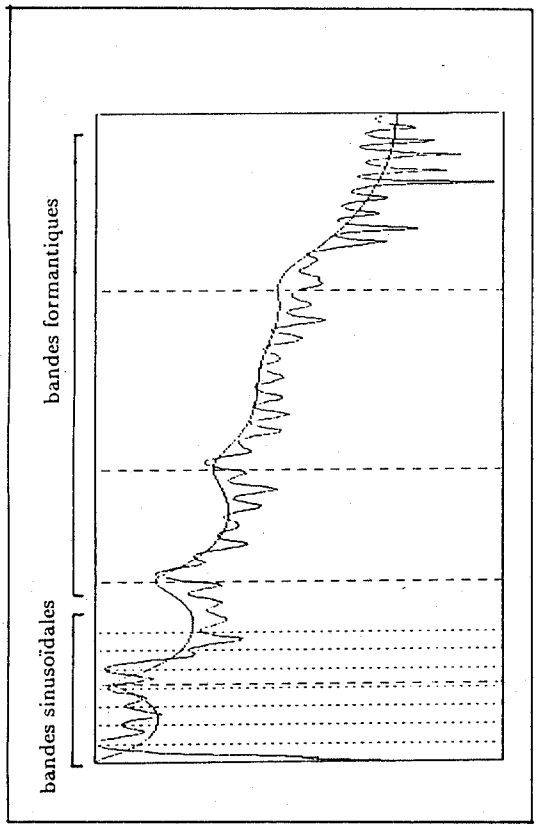


Figure 2: Modélisation et segmentation spectrale.

De par la largeur des bandes spectrales, les signaux obtenus dans chaque bande d'analyse sont des segments de sinusoïdes. Pour la parole voisée, il s'agit bien sur des premiers harmoniques, et l'enveloppe de ces signaux évolue beaucoup plus lentement que celle des signaux issus des bandes formantiques. Par contre, il est de toute première importance d'estimer précisément leur phase et leur fréquence (la vitesse d'évolution de la phase dépend évidemment de la fréquence).

Le choix du modèle de forme d'onde élémentaire adopté (nombre de cycles de la sinusoïde constant quelle que soit la fréquence) est motivé par ce souci de résolution spectro-temporelle, et non par un examen de l'enveloppe temporelle qui perd ici de son importance.

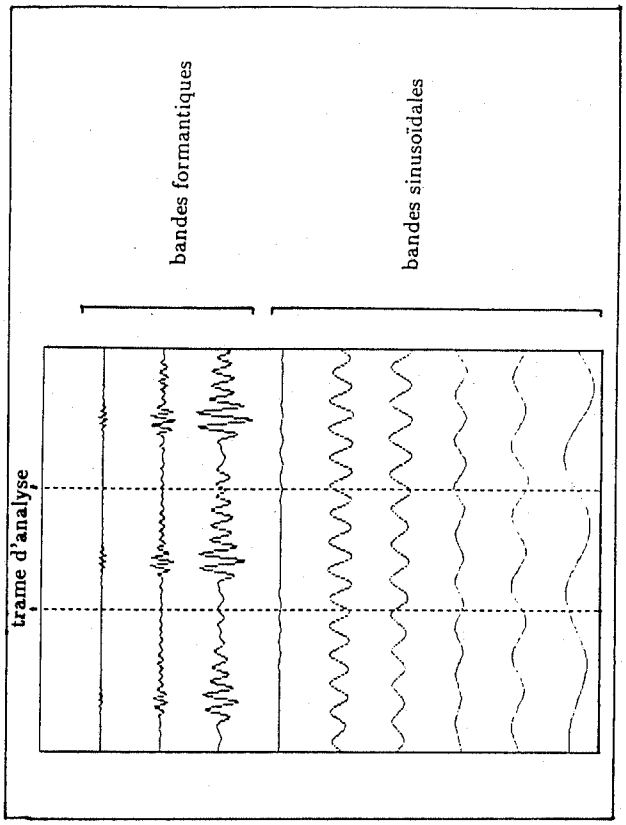


Figure 3: Filtrage dans les bandes précédemment définies.

Ainsi, le critère de segmentation d'une forme d'onde n'est plus basé sur une considération d'amplitude, comme c'est le cas pour les signaux formantiques en bande large, mais plutôt sur une considération de périodicité locale.

De même que dans les bandes d'analyse formantique, Ce processus donne des résultats satisfaisant tant pour de la parole voisée (on détecte alors des segments d'harmoniques) que pour la parole non voisée: le caractère local de la détection des formes d'onde permet en effet de reproduire des signaux transitoires trés brefs ou des signaux aléatoire.

4 RESULTATS

Le système réalisé permet le traitement automatique d'un segment de parole, et a été testé pour diverses voix tant féminines que masculines. La qualité de synthèse est excellente (pas ou très peu de différence avec l'original), mais doit maintenant faire l'objet de tests systématiques.

Le procédé s'appuie sur le modèle linéaire classique de production de la parole (de par la segmentation sur l'enveloppe spectrale) mais il autorise une analyse d'une grande finesse spectro-temporelle tout en ne délivrant qu'un jeu discret d'objets porteurs de toute l'information contenue dans le signal.

L'affichage des formes d'ondes dans le plan temps/fréquence permet de visualiser le résultat obtenu, qui paraît particulièrement intéressant, en ce sens que la plupart des caractères perceptivement pertinents (formants, pitch, voisement, explosions ...) sont représentés par un ensemble réduit de formes d'ondes (fig. 4, fig. 5, fig. 6).

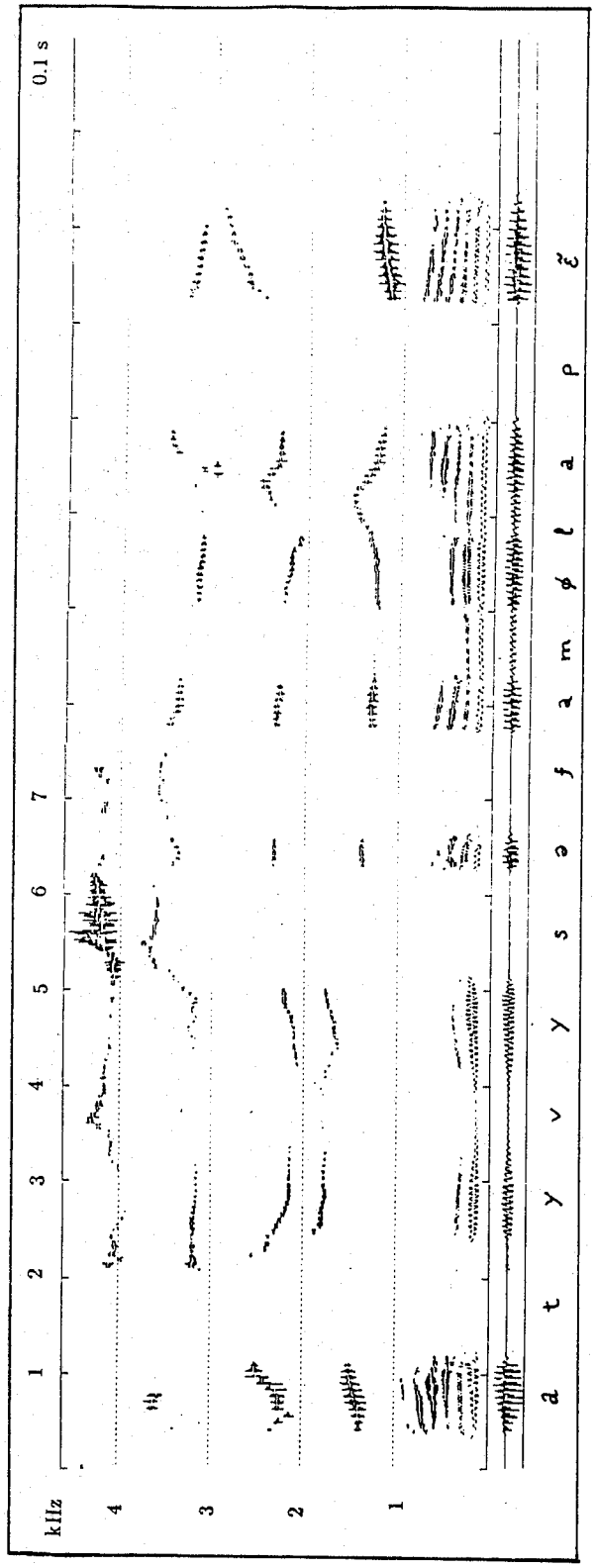


Figure 4: Affichage des formes d'ondes dans le plan temps/fréquence.

5 CONCLUSIONS

Nous avons présenté les fondements et la réalisation d'un système de représentation du signal de parole en formes d'ondes élémentaires. De part le choix opéré sur les formes d'ondes à rechercher, qui dérive du modèle classique de production du signal vocal, un traitement différent doit être appliqué aux différentes régions spectrales: une méthode spécifique semble en effet nécessaire pour rendre compte de la partie grave du spectre, et se trouve ici développée.

Il s'agit désormais, en raprochant cette méthode de modèles de perception, de l'appliquer à l'analyse de la parole. Les paramètres qu'elle fournit semblent également utiles pour une variante de la synthèse à formants en parallèle.Le débit d'information obtenu, qui reste à estimer de façon systématique, paraît offrir de bonnes potentialités en vue du codage: un gain semble en effet possible par rapport au codage sinusoïdal de part l'agglomération de plusieurs harmoniques dans une seule forme d'onde.

Ce travail représente le fruit de nombreuses discussions avec M^{rs} J.S.Liénard & X.Rodet, que l'auteur tient à remercier ici.

REFERENCES

[Liénard 87] Liénard, J.S. "Speech Analysis and Reconstruction Using Short-Time, Elementary Waveforms". IEEE-ICASSP 87, Dallas.

[d'Alessandro 87] d'Alessandro, C. & Rodet, X. "Fonctions d'ondes formantiques: extraction des paramètres et synthèse vocale". $16^{ième}$ JEP, 1987 Hammamet.

[d'Alessandro 88] d'Alessandro, C. & Liénard, J.S. "Decomposition of the Speech Signal into Short-Time Waveforms Using Spectral Segmentation". IEEE-ICASSP 88, New-York.

[Klatt 80] Klatt, D. "Software for o cascade/parallel formant synthesizer". JASA vol. 67(3), Mar. 1980.

[Rodet 80] Rodet, X. "Time Domain Formant-Wave-Function Synthesis". in "Spoken Language Generation and Understanding", J.C. Simon ed., D.Reidel publishing compagny, Dordrecht.

[McAulay 86] McAulay, R. & Quatieri, T. "Speech Analysis/Synthesis Based on a Sinusoidal Representation". IEEE trans. on ASSP, vol. ASSP 34 no. 4 1986.

[Goupillaud 85] Goupillaud, P., Grossmann, A. & Morlet, J. "Cycle-Octave and Related Transforms in Seïsmic Signal Analysis". Geoexploration 1985.

synthèse à formants parallèles
chef de timbre par modif de formants
cf X. Rodet

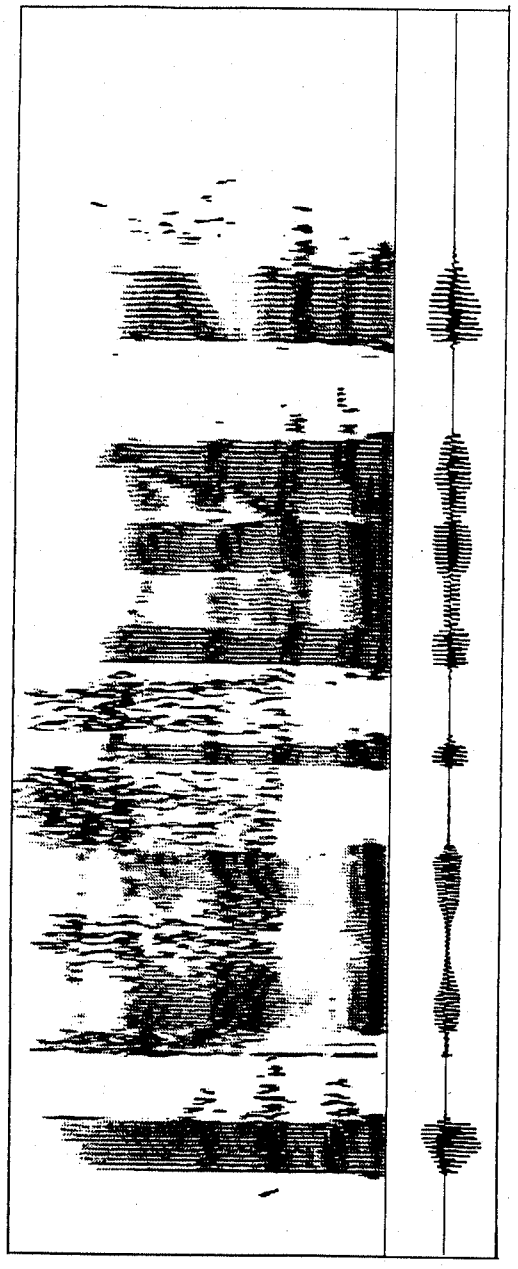


Figure 5: Spectrogramme du signal naturel correspondant à la figure 4.

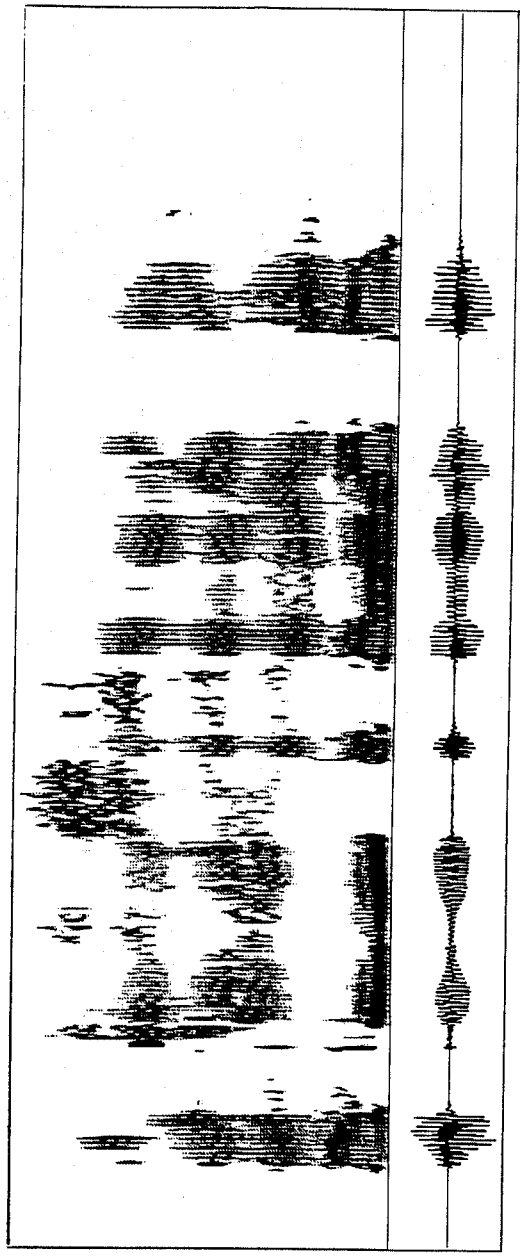


Figure 6: Spectrogramme du signal synthétique correspondant à la figure 4.

UN NOUVEL ALGORITHME DE CODAGE DE PAROLE A 4800 BITS/S

Gang FENG, Jean-Paul LEFEVRE

OROS
Chemin des Prés
38240 Meylan

ABSTRACT

In this paper, a new algorithm, able to encode effciently speech signal at 4800 bits/s, is described. From the well-known multi-pulse linear predictive coding approach, we derive a new method for the determination of the pulses. In fact, it consists of transmitting a drastically limited set of pulses whose locations are restricted within one pitch period. In order to find the optimal sequence of pulses, we propose an efficient two stage procedure. First, locations of the pulses are sequentially computed, then the jointly optimal set of amplitudes is estimated. The relatively low complexity of this algorithm has allowed its entire implementation on a single chip digital signal processor. The evaluation of this realization in terms of both system complexity and speech quality is also given.

1. INTRODUCTION

Les études sur la transmission de parole à bas débit (< 5000 bits/s) - mais de haute qualité - ont beaucoup progressé ces dernières années, pour satisfaire aux nombreuses applications. Certains algorithmes parus récemment semblent fournir un signal codé de très bonne qualité. Par rapport au codage prédictif classique, de même débit, ces codeurs peuvent fournir des voix synthétiques beaucoup plus naturelles. Citons ici comme exemple l'approche CELP /1/. Cependant, presque tous ces algorithmes possèdent une complexité relativement grande en raison de l'optimisation des codes d'excitation et de l'utilisation de la quantification vectorielle. Ceci entraîne des difficultés, notamment pour leur implémentation en temps réel. Souvent alors, des compromis doivent etre trouvés, entraînant de très rapides dégradations dans les performances. Il est donc important d'étudier des algorithmes moins complexes susceptibles de donner également une bonne qualité.

Le codage multi-impulsionnel est maintenant considéré comme l'une des meilleures techniques pour coder le signal de parole à moyen débit (entre 8 et 16 Kb/s). Pour des débits inférieurs, la dégradation commence à devenir importante puisque le nombre d'impulsions utilisable est trop petit. A titre d'exemple, seulement cinq impulsions sont disponibles pour encoder 22,5 ms du signal avec un débit de 4800 bits/s. Ceci est évidemment insuffisant. L'utilisation d'un prédicteur à long terme améliore considérablement les performances des codeurs multi-impulsionnels notamment pour des débits relativement faibles /2/. Cependant la qualité obtenue semble encore insuffisante pour justifier la complexité du codeur.

Une nouvelle approche plus adaptée aux conditions du codage à bas débit, consiste à mieux exploiter la nature pseudo-périodique du signal de parole, tout en conservant la structure essentielle de l'excitation multi-impulsionnelle. Dans un article récent /3/, les auteurs ont proposé une méthode dans laquelle la recherche des impulsions est limitée à une sous-trame dont la longueur est égale à la période fondamentale. Le signal d'excitation de toute la trame est reconstruit ensuite par interpolation à partir de ces impulsions.

Le nombre d'impulsions étant faible, leur concentration sur une période fondamentale permet de mieux représenter le signal d'excitation. On doit toutefois résoudre les problèmes de la détermination de l'intervalle sur lequel s'effectue la recherche, et l'interpolation de ces impulsions.

Nous présentons dans cet article, un nouvel algorithme qui permet de déterminer efficacement un petit nombre d'impulsions localisées sur une période fondamentale, et de reconstruire le signal d'excitation complet à partir de ces impulsions. La nouveauté de cet algorithme réside dans le fait qu'il n'y a pas de pré-détermination de l'intervalle pour la recherche des impulsions. Les impulsions repérées se situent systématiquement au début de la trame. Avec cette méthode, la détermination des impulsions est mieux optimisée. D'autre part, cet algorithme étant relativement simple, nous avons pu réaliser son implémentation en temps réel sur un seul processeur TMS-32020 (sur carte traitement de signal OROS-AU20). La parole resynthétisée est de bonne qualité.

Dans les sections suivantes nous présentons en détail l'algorithme. Nous exposons également une nouvelle méthode de détection du fondamental mieux adaptée à ce codeur, et enfin une technique spécifique aux sons non voisés.

Ce travail a été partiellement réalisé sous financement de l'Agence Européenne de l'Espace dans le cadre du contrat 7107/87/NL/JG.

2. DETERMINATION DES IMPULSIONS

Rappelons brièvement le principe du codage multi-impulsionnel. Sachant qu'un petit nombre d'impulsions (typiquement une impulsion par ms) suffit pour synthétiser une voix naturelle, tout le problème consiste à déterminer efficacement ces impulsions. ATAL a proposé d'utiliser une fonction de localisation qui est en fait la convolution entre le signal résiduel et l'autocorrélation du filtre perceptuel /4/. La détermination des impulsions s'effectue ensuite par une recherche exhaustive sur la fonction de localisation.

On observe souvent une structure pseudo-périodique dans les impulsions trouvées pour les sons voisés. Ceci se manifeste notamment sur les impulsions de grande amplitude. De toute évidence, cette périodicité provient du fait que la fonction de localisation est générée à partir du signal résiduel. Bien entendu, les impulsions correspondant aux différentes périodes ne sont pas exactement identiques malgré la pseudo-périodicité. C'est d'ailleurs pour cette raison qu'elles sont toutes nécessaires dans un codeur multi-impulsionnel classique. Si les impulsions sont réparties périodiquement, celles qui sont comprises dans une période fondamentale suffisent pour représenter les autres. L'idée essentielle de notre algorithme consiste donc à "périodiser" les impulsions d'un codeur multi-impulsionnel classique.

Comme l'amplitude du signal d'excitation varie le long de la trame, il ne serait pas réaliste de "périodiser" les impulsions sans tenir compte de cette variation : des paramètres supplémentaires sont nécessaires. Le débit étant faible, un seul paramètre par trame est autorisé pour représenter la variation d'énergie. Ce paramètre, noté Béta, est défini par le rapport d'amplitude entre deux périodes fondamentales successives.

Nous pouvons résumer notre démarche de la manière suivante : on repère un petit nombre d'impulsions par période fondamentale afin que l'extrapolation dans toute la trame de ces impulsions puisse représenter correctement le signal d'excitation. Nous entendons par "extrapolation" la "périodisation" des impulsions repérées par application du rapport d'amplitude Béta.

Cette approche implique que la détermination de chaque impulsion prenne en compte l'effet de toutes les extrapolations qui lui sont associées. Ceci pose un problème d'optimisation : il n'est pas évident de déterminer en même temps les positions et les amplitudes optimales des impulsions. Pour contourner cette difficulté, nous proposons ici une procédure en deux étapes : détermination des meilleures positions des impulsions puis optimisation de leurs amplitudes.

2.1. Localisation des impulsions

Comme dans un codeur multi-impulsionnel classique, le critère retenu pour déterminer les positions des impulsions est toujours les maxima de la fonction de localisation. Mais une position p est maintenant considérée comme optimale seulement si toutes les positions (p+nM), qui diffèrent de celle-ci par un multiple du fondamental, correspondent aux maxima de la fonction de localisation. Dans la mesure où cette dernière est pseudo-périodique, cette condition d'optimisation n'est pas toujours possible. Dans ce cas, la meilleure position p sera celle qui maximise la somme suivante :

$$1/Np \sum_{n=0}^{Np-1} T\,(p + nM) \qquad (1)$$

Ici, T(.) représente la fonction de localisation, M la valeur du fondamental, Np le nombre de positions situées dans la trame. Pondérée par le nombre de positions Np, cette somme représente la valeur moyenne de toutes les amplitudes correspondant aux positions (p+nM) de la fonction de localisation (figure 1).

On remarque, que selon ce critère, il suffit d'effectuer la recherche d'impulsions sur un intervalle /0,M-1/, puisque les positions en dehors de cet intervalle sont déjà prises en compte dans la sommation précédente. Il est commode de constituer, à partir de la fonction de localisation, une fonction auxiliaire :

$$T'(i) = 1/Ni \sum_{n=0}^{Ni-1} T\,(i + nM), \qquad i=0,1,\ldots,M-1. \qquad (2)$$

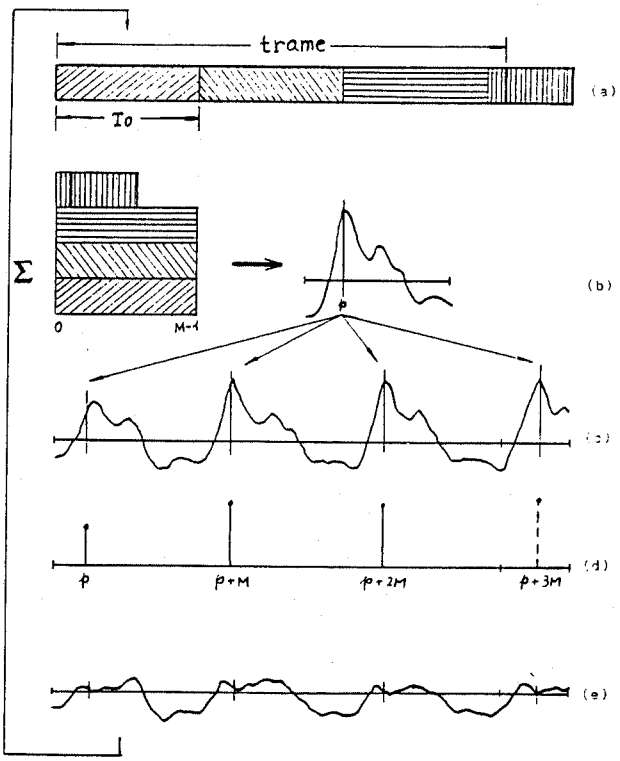


Fig. 1. Processus de détermination d'une position des impulsions.
(a),(c) : fonction de localisation T(i) ;
(b) : fonction auxiliaire T'(i) ;
(d) : impulsions stockées temporairement ;
(e) : fonction de localisation après la mise à jour.

La position p de l'impulsion optimale vérifie alors :

$$|T'(p)| = \text{Max}\ (|T'(i)|),$$

$$i=0,1,\ldots,M-1. \quad (3)$$

Avec cette méthode, toutes les positions optimales déterminées se situent systématiquement dans l'intervalle /0,M-1/. Ceci évite la pré-détermination de l'intervalle de recherche des impulsions qui risque d'introduire des contraintes et de rendre les impulsions sous-optimales.

Après détermination d'une position p, la fonction de localisation est mise à jour, tout comme dans un codeur multi-impulsionnel classique. La seule différence est que cette mise à jour s'effectue sur toutes les positions (p+nM) en fonction de leurs amplitudes T(p+nM). Par ailleurs, ces amplitudes sont stockées afin de déterminer l'amplitude optimale de chaque impulsion, à l'aide de la procédure décrite ci-après.

2.2. Détermination des amplitudes.

Rappelons que dans notre algorithme, les impulsions à transmettre se situent dans un intervalle du fondamental au début de la trame. Le signal d'excitation est reconstruit ensuite par extrapolation à partir de ces impulsions. Lorsque l'on détermine l'amplitude d'une impulsion, il est donc nécessaire de prendre en compte les extrapolations qui lui sont associées. La position des impulsions étant fixée, cette procédure est désormais plus aisée.

En fait, pour chaque position p, toutes les amplitudes correspondant aux positions (p+nM) de la fonction de localisation, c'est-à-dire T(p+nM), sont les meilleures pour ces impulsions. Or, la transmission de toutes ces amplitudes est impossible : un seul paramètre par impulsion est permis !

Nous sommes donc amenés à utiliser un paramètre pour représenter une série d'amplitudes. Ceci est possible grâce au facteur Béta défini précédemment comme rapport d'amplitude entre deux périodes fondamentales. En effet, dans chaque série, les impulsions sont séparées par le fondamental. Il est raisonnable de les approximer par une série d'impulsions dont le rapport d'amplitude est égal à Béta. Pour un Béta donné, cette série d'approximation sera uniquement déterminée par sa première amplitude. Ainsi la transmission d'une seule impulsion suffit pour reconstruire la série dans le décodeur.

Il reste à déterminer la première amplitude pour chaque série d'impulsions (le nombre de séries est égale au nombre des impulsions à transmettre). On minimise l'erreur quadratique entre la série originale et celle d'approximation. Il faut après détermination de chaque position, stocker toutes les amplitudes T(p+nM) de la fonction de localisation.

Notons Vn les amplitudes de la série originale et N le nombre d'impulsions dans la série. L'amplitude de l'impulsion à transmettre (c'est la première de la série) est calculée avec la formule suivante :

$$A_i = \frac{\sum_{n=0}^{N_i-1} V_n \cdot \beta^n}{\sum_{n=0}^{N_i-1} (\beta^n)^2}, \quad i=1,2,\ldots \quad (4)$$

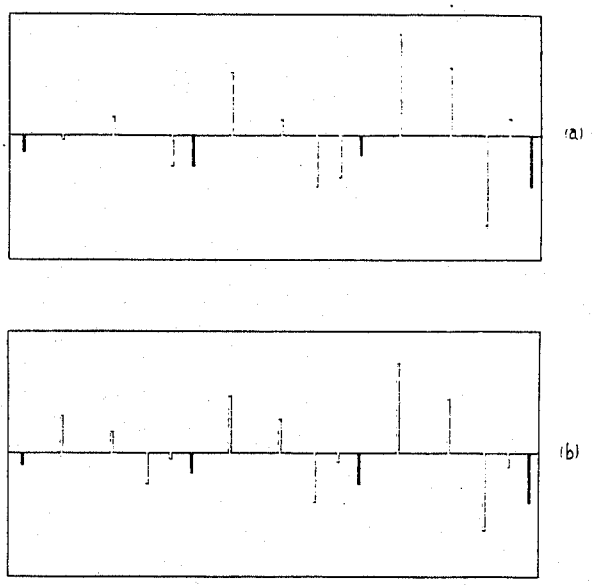


Fig. 2. Exemple illustrant l'optimisation des amplitudes des impulsions.
(a) : impulsions après la détermination des positions ;
(b) : impulsions reconstruites.

Le même calcul s'effectue pour toutes les impulsions à transmettre.

La figure 2 montre un exemple de calcul. La figure 2.a présente les impulsions originales juste après la détermination des positions. Il y a cinq séries d'impulsions dans cette figure et la série no 1 est "noircie". La figure 2.b présente les impulsions reconstituées à partir des cinq impulsions transmises (les cinq premières de la trame).

3. DETECTION DU FONDAMENTAL ET CALCUL DE BETA

Notre algorithme exploitant la pseudo-périodicité du signal de parole, la détection du fondamental prend donc une importance particulière. Il est toutefois difficile de trouver un algorithme qui fournit des résultats toujours satisfaisants. Par ailleurs, il n'est pas question d'adopter un détecteur complexe à cause des contraintes de temps réel. Dans la littérature, certains auteurs commencent à proposer de mesurer le fondamental à partir de la séquence multi-impulsionnelle /5/.

Nous avons développé un algorithme de détection du fondamental adapté aux codeurs de type multi-impulsionnel. L'idée essentielle consiste à utiliser la fonction de localisation comme la base de la détection. En effet, cette fonction présente une allure relativement lissée, puisqu'elle résulte de la convolution du signal résiduel avec l'autocorrélation du filtre perceptuel, cette dernière se comportant comme un filtre passe-bas. Pour la même raison, il y a peu de composantes formantiques dans la fonction de localisation, ce qui est favorable à la détection du fondamental.

La détection s'effectue en deux étapes: on calcule d'abord l'autocorrélation de la fonction de localisation et on cherche ensuite le maximum de celle-ci. Dans la seconde étape, il est nécessaire de fixer certains seuils de détection afin d'éliminer les trames non voisées.

Cet algorithme nous a donné des résultats très satisfaisants. Notons en particulier que la détermination des impulsions exploite la corrélation à long terme de la fonction de localisation. Pour mesurer le fondamental, la meilleure méthode est d'utiliser le maximum de l'autocorrélation de cette fonction.

Une fois le fondamental déterminé, on peut calculer le facteur Béta :

$$\beta = \left(\sqrt{\frac{E_2}{E_1}}\right)^{\frac{M}{L}} \qquad (5)$$

Ici, M représente le fondamental, E1 et E2 les énergies de la fonction de localisation dans deux zones de longueur M, séparées l'une de l'autre par la distance centrale L.

4. A PROPOS DES SONS NON VOISES

Pour les sons non voisés, l'algorithme exposé précédemment semble perdre son efficacité puisqu'il n'y a pas de périodicité. Le faible nombre d'impulsions utilisables entraîne des sensations "granuleuses" dans le signal resynthétisé. Bien que ceci soit limité uniquement aux sons non voisés, l'impression globale est souvent désagréable.

Pour améliorer ce défaut, nous avons essayé une autre approche qui consiste à remplacer dans le signal d'excitation, les impulsions trop peu nombreuses par un bruit blanc. A la place des impulsions, nous transmettons quelques paramètres pour contrôler l'amplitude du bruit blanc. Ces paramètres sont déterminés directement à partir du signal résiduel. Cette méthode se revèle en pratique très efficace pour éliminer la sensation granuleuse du signal resynthétisé, sans toutefois modifier son timbre.

5. RESULTATS ET DISCUSSIONS

Nous avons implémenté notre algorithme sur la carte traitement de signal OROS-AU20 basée sur un processeur TMS-32020. Avec le débit de 4800 bits/s, le nombre d'impulsions est égal à cinq pour une longueur de trame de 22,5 ms. Avec une telle durée de trame, et pour obtenir un débit effectif de 4800 bits/s, 108 bits sont disponibles par trame. 38 sont utilisés pour le codage du modèle LPC (ordre 10), 7 pour le fondamental et 4 pour le facteur Béta, tandis qu'un bit est réservé pour la synchronisation. Les 58 bits restants servent à coder les cinq impulsions.

La figure 3 montre un exemple de fonctionnement du codeur sur une trame de signal. La figure 4 montre le même exemple pour un codeur multi-impulsionnel classique qui utilise 15 impulsions par trame (débit 9600 bits/s). Notons en particulier que le signal d'excitation reconstruit à partir de cinq impulsions (Fig. 3) est très proche de celui du codeur à 9600 bits/s (Fig. 4).

Des tests d'évaluation ont été effectués sur de nombreux corpus avec différents locuteurs. Nous avons notamment comparé cet algorithme, pour un débit de 4800 bits/s, avec un vocodeur LPC et avec un codeur multi-impulsionnel utilisant un prédicteur à long terme.

Les performances objectives des codeurs sont mesurées à l'aide du rapport signal sur bruit segmental. Pour notre codeur à 4800 bits/s, la valeur moyenne de RSBseg se situe entre 7 et 8 dB. Ce résultat est comparable avec celui obtenu par un codeur multi-impulsionnel avec un prédicteur à long terme fonctionnant au même débit. Le RSBseg perdant sa signification pour les codeurs à bas débit, nous l'utilisons uniquement pour indiquer le bon fonctionnement des codeurs.

La méthode de comparaison par paires a été utilisée pour l'évaluation subjective. Par rapport au codeur multi-impulsionnel à 4800 bits, et bien que les RSBseg des deux codeurs soient très proches, la parole resynthétisée de notre codeur est plus claire et moins granuleuse. On peut expliquer ce résultat de la façon suivante : le nombre d'impulsions étant trop faible, le fonctionnement d'un codeur multi-impulsionnel dépend davantage du prédicteur à long terme. Ceci a pour conséquence d'introduire des erreurs dans les transitions pour lesquelles le signal est moins prédictible.

Comparé avec le vocodeur LPC à 4800 bits/s, notre codeur fournit une parole synthétisée qui reproduit mieux le timbre de la voix originale et qui est moins métallique. L'utilisation de cinq impulsions par période fondamentale est sans doute la principale raison de cette amélioration. Un autre avantage de cet algorithme est qu'il est moins sensible aux erreurs du fondamental (saut d'octave par exemple). En effet, notre algorithme étant de type multi-impulsionnel, ces erreurs peuvent être compensées grâce au mécanisme de localisation des impulsions.

CONCLUSION

Nous avons présenté dans cette étude un nouvel algorithme pour coder le signal de parole avec un débit de 4800 bits/s. Ce codeur appartient à la famille des codeurs multi-impulsionnels et une nouvelle conception a permis de l'adapter aux conditions du codage à bas débit. Elle consiste à transmettre les impulsions limitées dans une période fondamentale. Les impulsions localisées se situent systématiquement au début de trame, et le signal d'excitation est reconstruit par extrapolation à partir de celles-ci.

La relative simplicité de notre algorithme nous a permis de l'implémenter en temps réel sur un processeur TMS-32020. La parole resynthétisée du codeur est de bonne qualité (claire et non granuleuse). Le timbre de la voix originale est bien conservé. Certains défauts, donnant l'impression de parole légèrement synthétique, peuvent apparaître dans la voix resynthétisée. Ils sont dus à la monotonie de l'excitation tout au long de la trame. L'utilisation d'une interpolation appropriée pourrait réduire ces défauts.

En résumé, cet algorithme nous a permis de trouver un excellent compromis entre qualité, débit et complexité.

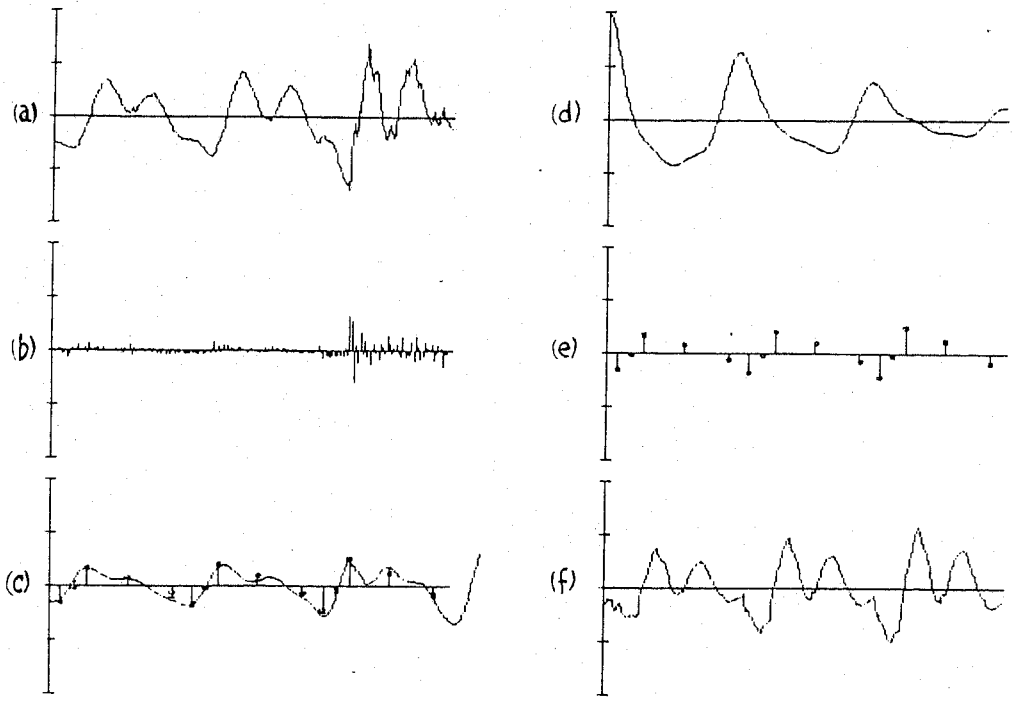


Fig. 3. Fonctionnement de notre codeur à 4800 bits/s sur une trame de signal.
(a) : signal de parole ;
(b) : signal résiduel ;
(c) : fonction de localisation ;
(d) : autocorrélation de (c) ;
(e) : signal d'excitation reconstruit ;
(f) : signal resynthétisé.

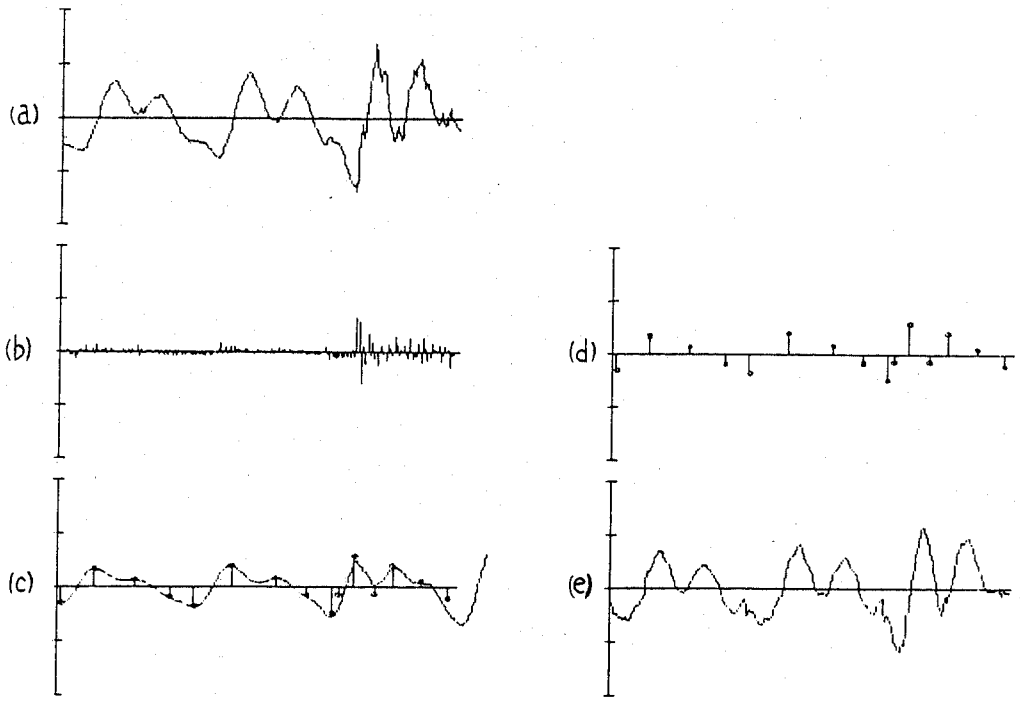


Fig. 4. Fonctionnement d'un codeur multi-impulsionnel utilisant 15 impulsions par trame (9600 bits/s).
(a) : signal de parole ;
(b) : signal résiduel ;
(c) : fonction de localisation ;
(d) : signal d'excitation (les impulsions) ;
(e) : signal resynthétisé.

REFERENCES

/1/. SCHROEDER M.R. & ATAL B.S.,Code-Excited Linear Prediction (CELP) : High Quality Speech at Very Low Bit Rates. - IEEE ICASSP 1985, pp. 937-940.

/2/. LEFEVRE J.P. & PASSIEN O., Efficient Algorithms for Obtaining Multipulse Excitation for LPC Coders. - IEEE ICASSP 1985, pp. 957-960.

/3/. OZAWA K. & ARASEKI T., Low Bit Rate Multi-pulse Speech Coder with Natural Speech Quality. - IEEE ICASSP 1986, pp. 457-460.

/4/. ATAL B.S. & REMDE J.R., Digital Speech Coder. - U.S. Patent no. 4,472,832, sep. 1984 (appl. no. 326,371, dec. 1981).

/5/. BAKAMIDIS S.G., CARAYANNIS G. & SKIADAS N., A New Pitch Detector Based on Preselected Information from the LPC Error Signal. - Communication personnelle, soumis pour publication à IEEE ASSP.

gds sons for studies of transitions
phonetics

INDEX DES AUTEURS