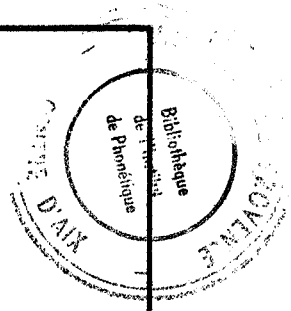
Université de Montréal
Département de linguistique

Bibliothèque de Phonétique

XIIe Journées d'études sur la parole

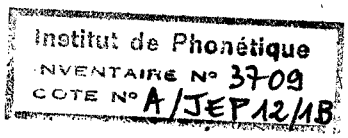
organisées à l'Université de Montréal
avec la collaboration du GALF
les 25, 26 et 27 mai 1981

Laboratoire de phonétique
Département de linguistique
Université de Montréal



XIIèmes Journées d'études sur la parole

Organisées à l'Université de Montréal
avec la collaboration du GALF
les 25, 26 et 27 mai 1981



Comité d'organisation

Danièle Archambault
Raymond Descout
Alain Marchal
André Marti
Philippe Martin
Laurent Santerre



Nous remercions les organismes suivants
qui ont rendu possible la tenue des
XIIèmes Journées d'études sur la parole

Bibliothèque de l'Institut de Phonétique
CENTRE D'AIX

Le ministère des Affaires étrangères de la France

Le Centre national d'études des télécommunications

Le Groupement des acousticiens de langue française

L'Insitut de phonétique d'Aix-en-Provence

Le Consulat de France à Montréal

L'Association canadienne-française pour l'avancement des sciences

Le Conseil de recherches en sciences humaines du Canada

Le ministère de l'Education du Québec (FCAC)

Le ministère des Affaires intergouvernementales du Québec

L'Université de Montréal

Bibliothèque de l'Institut de Phonétique
CENTRE D'ARC

TABLE DES MATIERES

page

THEME 1 : Codage - Synthèse de la parole

P. GRAILLOT -
Commandes pour la synthèse à partir d'une entrée par mots clés 1-11

J.Y. AULOGE - R. DESCOUT - B. GUERIN
Modelisation élastique des cordes vocales, aspects modaux 12-25

J.P. LEFEVRE - B. TOUSIGNANT - M. LECOURS
Utilisation des méthodes acoustiques pour l'évaluation des fonctions d'aires du conduit vocal 26-41

J.P. ADOUL - Ph. MABILLEAU
Représentation de la parole entre 4 et 8 bits/sec. utilisant un dictionnaire de formes d'onde 42-53

L. SANTERRE - J.L. CHANDON
Fonctions morphologiques des paramètres suprasegmentaux en français québécois 54-65

THEME 2 : Reconnaissance

H. MELONI - J. GUIZOL
Un système de reconnaissance automatique de locutions 66-79

J.P. HATON - J.M. PIERREL
Analyse et reconnaissance automatique de la parole au C.R.I.N. : approches, motivations et résultats obtenus 80-94

J. LEBRAS
Utilisation de deux méthodes de reconnaissance automatique de la parole pour la détection des fautes de prononciation 95-112

D. RAMASSE
Indices acoustiques secondaires de lieu et contraintes distributionnelles 113-130

M. MICHEL - J. LEROUX - Y. GRENIER
Expériences en reconnaissance par prédiction linéaire. 131-141

A. LELIEVRE
Classification ascendante hiérarchique automatique pour la création de phonèmes ou de mots de référence en reconnaissance de la parole 142-154

G. MERCIER
Analyse acoustique et phonétique dans le système DEAL 155-171

page

D. GILLET
Syntaxe, sémantique et dialogue dans le système DEAL.
Application au renseignement automatique 172-187

N. TIRANDAZ - C. BERGER-VACHON
Recherche des segments stables dans les comptines parlées en continu 188-201

THEME 3 : Analyse Fo

W. HESS
Algorithmes et méthodes pour la détermination du Fo 202-220

Ph. MARTIN
Détection de Fo par intercorrélation avec une fonction peigne 221-232

R. CARRE
Couplage conduit vocal - source vocale - conséquences et analyse de parole 233-245

G. JOUANELLE - L.J. BOE - G. BELLET
Variations intra et interlocuteur de Fo : Procédures de normalisation. Etude préliminaire 246-262

J.S. LIENARD - F. MANCERON
Analyse impulsionnelle de la parole : expériences préliminaires 263-272

M. CONTINI - G. BELLET - L.S. BOE
Variations intralocuteurs de Fo dans la réalisation d'une phrase énonciative en français : normalisation temporelle 273-288

THEME 4 : Pathologie et réhabilitation

M. DI BENEDETTO - F. DESTOMBES - J.P. TUBACH
Utilisation de la théorie de l'information pour une étude quantitative de l'ambiguité en lecture labiale 289-303

F. EMERARD
Application de la synthèse de la parole à la communication des handicapés 304-314

D. O'SHAUGHNESSY - M. LENNIG - PL MERMELSTEIN
Simulation d'un lecteur automatique du français 315-325

THEME 3B : Analyse - Articulation

J.P. ZERLING
Quelques particularités articulatoires et acoustiques de l'occlusive /g/. 326-338

M. GENTIL
Etude EMG de quatre muscles labiaux dans les syllabes finales de mots en français : l'effet de coarticulation 339-353

page

L.J. BOE
Espace articulatoire mandibulaire, premiers résultats 354-363

Ch. ABRY
Vers un espace de la labialité vocalique : apports de l'analyse des correspondances et des nuées dynamiques 364-376

THEME 5 : Perception

J.P. GOUDAILLER
Exemple de traitement de l'opposition de sonorité par des enfants de cours préparatoire. Utilisation de la méthode électroglottographique 377-391

E. LHOTE - M. STRICKLAND - P.W. LEE
Influence du système linguistique original sur la perception : cas du voisement 392-405

F. LONGCHAMP
Analyse multidimensionnelle de l'espace perceptif des sons vocaliques 406-418

Y. NISHINUMA - S. BARBER - D.J. HIRST
Estimation de la durée intrinsèque des voyelles 419-428

J.F. BONNOT - A. BOTHOREL
Rencontre des consonnes identiques à la frontière de mots en néerlandais 429-442

Ch. SORIN
Fonctions, rôles et traitements de l'intensité dans la parole continue 443-461

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

COMMANDES POUR LA SYNTHESE A PARTIR D'UNE ENTRÉE PAR MOTS CLÉS

P. GRAILLOT CNET – LANNION

.**RESUME**

Dans le cadre de l'aide que peut apporter la synthèse de la parole à la communication des handicapés, nous avons réalisé un système permettant de passer de la suite des mots (désignés sur un tableau de symboles Bliss) d'une phrase à la synthèse vocale de cette dernière. Les fonctions réalisées consistent à générer certains mots (par ex. les articles), à effectuer les conjugaisons et la génération des marqueurs prosodiques ainsi que les liaisons et les élisions.

Ce système sera expérimenté en janvier 1981 dans un centre d'enfants I.M.C.

. ABSTRACT

Together with the development of studies on speech synthesis, the production of vocal communication systems for severely handicapped people - with essentially normal hearing but unable to speak - has considerably progressed. We present the results of a current study which uses a talking computer, a Blissymbol Scanning board, and a grammar especially prepared for that purpose : with the selected symbols, the computer produces syntactically reconstructuted strings with both orthographic and phonetic outputs. Then, the phonetic string is translated into a vocal output with an automatic prosodic processor.

Experiments are already being conducted with disabled children in a few Rehabilitation Centers.

COMMANDES POUR LA SYNTHESE A PARTIR D'UNE ENTRÉE PAR MOTS CLÉS

Un article précédent (EMERARD et al.) rend compte des objectifs de cette étude : synthétiser une phrase grammaticalement correcte à partir de la désignation - sur un tableau de symboles - d'une suite de mots clés.

Le tableau se compose de 486 cases correspondant d'une part aux mots français les plus courants - quelle que soit leur catégorie lexicale - pour les utilisateurs visés et d'autre part à des informations d'ordre syntaxique (marques de l'interrogation, du pluriel, du féminin, du passé, du futur et de la négation).

La terminologie «mots-clés» est préférée à celle de «concepts» : alors que cette dernière (YOUNG, FALLSIDE, 1979 ; MEMMI, LIENARD, 1979) utilise comme entrée une suite de prédicats logiques et s'intègre dans un système complet de dialogue homme-machine (reconnaissance puis synthèse), une entrée par mots clés suppose la désignation de mots isolés dans l'ordre où ils apparaîtront dans la phrase reconstruite et la compression de l'information en utilisant l'option par défaut (génération de l'article défini et du genre le plus fréquent pour les substantifs, du présent comme temps pour les verbes).

Exemples de fonctionnement du système :

Entrée (suite des cases sélectionnées)	Sortie orthographique
chat manger soupe	le chat mange la soupe
pluriel chat manger soupe	les chats mangent la soupe
chat manger futur soupe	le chat mangera la soupe
mon mère aller a bureau	ma mère va au bureau
docteur venir pour je	le docteur vient pour moi

I — LES OPTIONS

Un certain nombre de contraintes, liées au particularisme des utilisateurs envisagés sont à l'origine d'un certain nombre de questions et de décisions :

I - 1 : Outil d'apprentissage ou de communication ?

Une prothèse de parole peut s'envisager de deux façons différentes selon que la personne handicapée :

— est en cours d'apprentissage d'un moyen de communication (stade de la pré-lecture). Pour cette population, l'expérience pédagogique montre qu'il est illusoire de faire débuter l'enseignement d'une langue par l'apprentissage de son écriture. Plus efficace semble être l'utilisation de langages picturaux (comme le BLISS) qui donne accès de façon simple au niveau sémantique. L'expérience visuelle et auditive de ces enfants est propre à leur inculquer les notions d'ordre des images et des mots dans des messages.

En ce sens, le système proposé ici peut être utilisé à des fins d'apprentissage à condition bien sûr que, en même temps que la synthèse vocale, la visualisation de la suite orthographique correcte soit réalisée. L'idéal étant évidemment, quand le message d'entrée est inacceptable, qu'une correction soit effectuée ou qu'un diagnostic sûr permette de préciser l'erreur commise. Mais cette option conduit inévitablement à une limitation de la syntaxe autorisée. Un tel système, conçu dans un but pédagogique, sera utilisé dans une classe, et dans cette optique, il n'est pas indispensable de prévoir un appareil individuel et portable.

— maîtrise l'écriture orthographique. Il ne s'agit plus ici de raisonner en termes de pédagogie mais en termes de vitesse de communication. L'avantage d'un système permettant la composition d'un message à l'aide d'unités linguistiques plus grandes que le caractère orthographique se situe évidemment au niveau de l'économie d'information réalisée. Destiné à faciliter la communication, le système se doit d'être pratique, c'est-à-dire individualisé (vocabulaire adapté à l'utilisateur) et portable ; il se doit également d'autoriser la syntaxe la plus riche possible, mais ceci ne sera possible qu'au détriment des possibilités de récupération des erreurs.

Ce projet se voulant bivalent quant aux objectifs, c'est la seconde option (richesse syntaxique) qui a été retenue. Un seul système similaire (EULENBERG, RAHIMI, 1978) existe à notre connaissance ; son but est également d'être bivalent.

I - 2 : Utilisation d'un programme de transcription orthographique-phonétique ou non ?

Compte tenu du but poursuivi, une double sortie simultanée - vocale et écrite - est souhaitable. Le système de base suppose toujours un programme et un lexique mais la décision de créer deux sorties laisse le choix entre deux options :

— stocker un seul lexique sous forme orthographique et introduire un programme de transcription orthographique phonétique.

— stocker deux lexiques différents dans leur forme (orthographique et phonétique) ce qui permet de se dispenser du module de transcription.

Les deux traitements étant exactement similaires, la seconde option a été adoptée parce qu'elle présente l'avantage de s'abstraire de toutes fautes de transcription.

I - 3 : Analyse syntaxique ou non ?

Il existe des difficultés propres à chacune des deux sorties envisagées (vocale et écrite).

— pour la sortie orthographique, il s'agit surtout des problèmes relatifs aux accords et aux élisions.

— pour la sortie phonétique, les problèmes concernent davantage la génération automatique des marqueurs prosodiques (EMERARD, 1977), les liaisons et les élisions. En fait comme il ne s'agit de travailler que sur un vocabulaire limité, les informations nécessaires aux liaisons, aux élisions et à une partie des marqueurs prosodiques (liés aux mots grammaticaux) ont été introduites dans les lexiques.

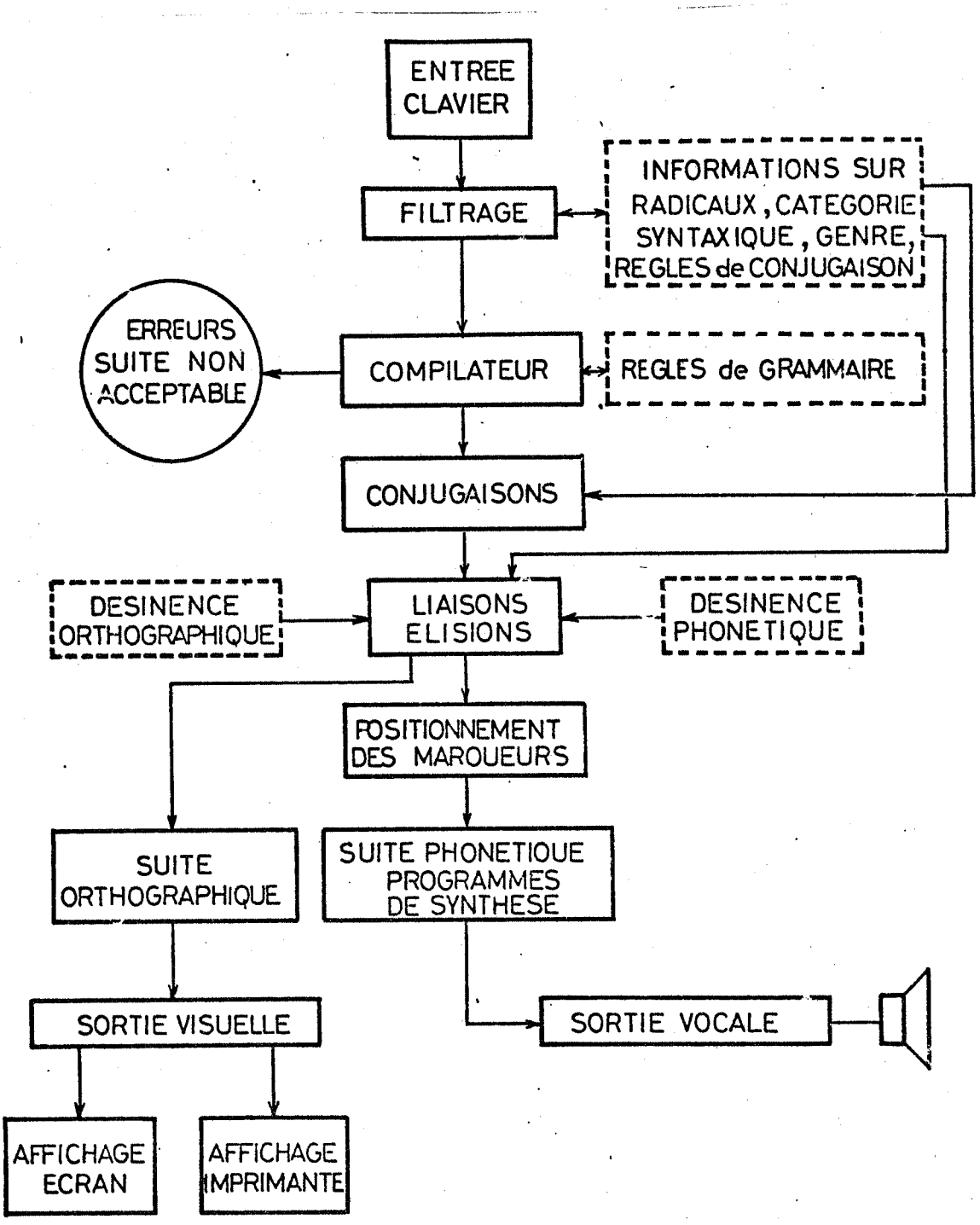
La génération des marqueurs prosodiques liés à la syntaxe s'effectue quant à elle en même temps que les accords (GRAILLOT et EMERARD, 1981) dans la mesure où la suite (déterminant - substantif) ou (déterminant - adjectif - substantif) ne réclame aucun marqueur prosodique.

En ce qui concerne les accords, il a été décidé d'effectuer une «mini» analyse syntaxique qui offre le double avantage d'accepter des structures multiples et de filtrer les suites lexicales impossibles. Cette analyse syntaxique est essentiellement destinée à déterminer la «fin du groupe du nom», la «fin du groupe sujet», la «fin du groupe complément» et la «fin du groupe verbal». Les règles grammaticales autorisées (BERTHIER, 1980) ont été traduites en une grammaire de précédence (AHO et ULLMAN, 1972).

I - 4 : Adaptation à d'autres vocabulaires :

Les langages picturaux de communication susceptibles d'être utilisés avec le système proposé sont nombreux ; aussi le programme a-t-il été fait de façon a être le plus indépendant possible du lexique de référence (tableau Bliss). Dans cette optique, mis à par un sous-programme préliminaire (nommé «PREFIL») de préfiltrage qui regroupe les exceptions, l'ensemble du programme demeure absolument indépendant du vocabulaire et de son étendue.

Les considérations qui viennent d'être données conduisent à l'organigramme général présenté ci-après (figure 1).



II – ORGANISATION DES LEXIQUES

Un fichier regroupe, pour chaque mot, les radicaux orthographiques et phonétiques, la catégorie lexicale, le genre pour les substantifs et les numéros des règles de désinence en orthographique et phonétique (même en omettant les liaisons, la correspondance n'est pas objective, par ex. manger est sur le modèle d'aimer en phonétique, mais pas en orthographique : nous mangeons).

Deux autres fichiers contiennent les désinences orthographiques et phonétiques.

Exemple :
Pour beau et nouveau.

Fichier général

Radical orthog.	Catégorie lexicale	Règle orthog.
Be	Adjectif	8
Nouve	Adjectif	8

Fichier des désinences orthographiques :

Règle 8 des adjectifs :

au	masculin singulier devant consonne (le beau bureau)
l	masculin singulier devant voyelle (le nouvel an)
lle	féminin singulier
aux	masculin pluriel
lles	féminin pluriel

III – LE PRÉ-FILTRAGE

Le but de ce sous programme est à partir de la suite, la plus condensée possible, des mots à l'entrée de générer la véritable suite de mots pour l'analyse syntaxique (compilateur). Les deux principales fonctions de ce sous-programme sont donc de traiter les exceptions et de générer l'article défini suivant le processus explicité en introduction.

Parmi les exceptions, citons le cas des verbes posséssifs (prenons l'exemple de s'asseoir). La première solution est de stocker 5 radicaux différents (m'ass, t'ass. . .). Il est plus astucieux de ne stocker que le radical «ass» et de générer dans le pré-filtrage le pronom personnel complément correspondant au sujet.

Exemple

Entrée	Pluriel	chat	qui	asseoir	sur	lit	être amusant	
sortie du pré-filtrage	Les	chat	qui	se asseoir	sur	le	lit	être amusant.

IV – LE COMPILATEUR

Rappelons que le tableau d'entrée comprend environ 480 mots parmi les plus courants du français. Il nous fallait déterminer quelle complexité syntaxique nous permettions. Les principales difficultés étaient posées par les phrases ne commençant pas par le sujet (cet hiver, je n'irai pas à la piscine), par les relatives (le chien qui a des poux se gratte) et les conjonctions de coordination (principalement le «et») qui permettent indifféremment de relier deux phrases, deux substantifs,

deux adjectifs et deux verbes. Dans les deux derniers cas évoqués on aborde le principal problème d'une analyse syntaxique : la récursivité. Mais les phrases sélectionnées étant forcément courtes (vu le temps d'entrée), nous avons décidé de permettre l'emploi de ces difficultés à condition qu'une seule y figure par phrase. Celà permettait de traduire les règles permises en une grammaire de précédence (pour les cas ambigus, par exemple la coordination de deux phrases, nous générons une virgule fictive dans le module de pré-filtrage), ce qui constitue une solution pratique et rapide au point de vue programmation. La définition des éléments terminaux et non-terminaux est sur la figure 2, les lecteurs avides de connaître l'ensemble des règles permises pourront consulter la publication de F. BERTHIER.

Cette analyse syntaxique fournit, pour chaque phrase, un arbre complet. Nous avons vu que l'information fournie par l'emplacement des quatre non-terminaux fin du groupe du nom (fGN), du groupe sujet (fGS), du groupe complément (fGC) et du groupe verbal (fGV) suffit pour résoudre le problème des accords.

17 éléments terminaux

- Invariant
- Adverbe
- Conjugaison de coordination
- Proposition
- Conjonction de surbordination
- Négation
- Signe de ponctuation
- Adverbe et adjectif interrogatif
- Substantif
- Verbe
- Adjectif qualificatif
- Pronom personnel complément
- Pronom personnel sujet
- Déterminant
- Qui
- Que
- Il y a

12 éléments non-terminaux

- Phrase (axiome)
- Phrase affirmative simple
- Groupe sujet
- Groupe verbal
- Groupe verbal simple
- Groupe complément
- Groupe du nom
- Groupe du nom simple
- Déterminant
- Début d'interrogative
- Coordination de phrases simples
- Début et fin de phrase

Fig. 2 : Liste des éléments de la grammaire

Exemple :

Entrée	Gentil	chat	négation	manger	bon	soupe						
Sortie pré-filtrage	Le	gentil	chat	négation	manger	la	bon	soupe				
Sortie compilateur	Le	gentil	chat	fGN	fGS	Neg	manger	fGV	la	bon	soupe	fGN	fGC

V — LES MODULES DES ACCORDS ET DE GÉNÉRATION DES MARQUEURS PROSODIQUES

Les accords sont assurés par un programme principal appelant quatre sous-programmes : SUJET, NOM, COMPLÉMENT, VERBE. Dans chacun d'eux nous traitons les cas particuliers (ils sont assez nombreux, mais nous les épargnerons aux lecteurs). Dans le programme principal nous examinons les cas où un retour en arrière est nécessaire : d'une part pour les relatives (les chiens que je vois dans la cour, sont gentils) où le deuxième verbe ne s'accorde pas sur le dernier groupe sujet rencontré et d'autre part pour les interrogatives avec inversion du sujet (viendrez-vous ce soir ?).

La sortie de ces programmes d'accords consiste à fournir, pour chaque mot non-invariant, le format (nombre, genre et personne) dans les désinences possibles.

Exemple :

Sortie compilateur	Le	gentil	chat	fGN	fGS	Neg	manger	fGV	la	bon	soupe	fGN	fGC
Sortie conjugaison	Le	gentil	chat	fGN	fGS	Neg	manger	fGV	la	bon	soupe	fGN	fGV
	M-S	M-S	M-S				3ème personne du présent		F-S	F-S	F-S		

M = masculin, S = singulier

Dans le module de génération des marqueurs prosodiques, ces derniers prennent la place dans la suite sortie de l'analyse syntaxique des quatre non-terminaux (fGN, fGS, fGC, fGV).

Pour les affirmatives, la définition des marqueurs prosodiques utilisés dépend du fait que soit :

a) la phrase est syntaxiquement simple.

b) la phrase contient une virgule, que celle-ci soit entrée (cas 1) ou générée (cas 2 et 3) avant le compilateur. On se trouve alors dans l'un des trois cas :

1 - groupe complément en tête : cet été, j'irai à la mer.

2 - relative : le chien qui dort dans la niche, est gros.

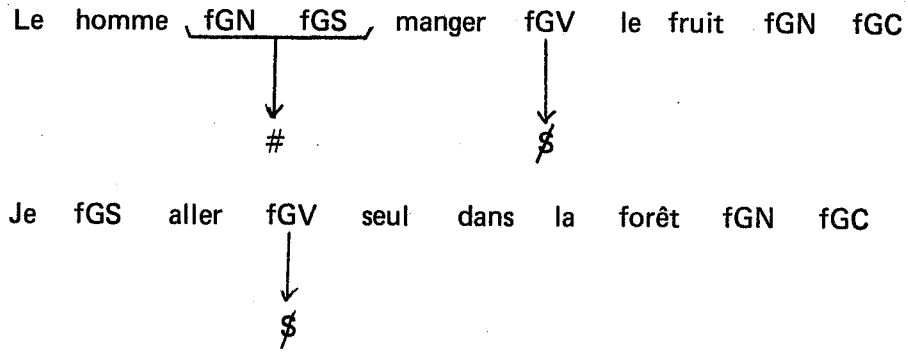
3 - coordination de deux phrases : le vent est fort, mais la mer est calme.

Pour la partie de la phrase située après la virgule, on retrouve le cas a).

Les interrogatives sont traitées suivant le cas b), avec une exception pour le groupe sujet.

En outre pour certains mots (prépositions, qui, que. . .) le marqueur prosodique est inclu dans le radical phonétique, ce qui permet de simplifier le programme.

Exemples (données en orthographiques) :



Pas de marqueur # devant le verbe (il n'est généré que par la suite fGN fGS, ce qui évite les pronoms sujets), par contre un marqueur sera généré devant car son radical phonétique stocké est #DAN".

VI – LA CONCATENATION RADICAL-DÉSINENCE

Le module des accords ayant choisi la désinence appropriée il reste à la concaténer au radical. C'est à ce stade que sera résolu en orthographique le problème des élisions (sous ce terme nous regroupons d'autres problèmes : le bel enfant - le beau meuble, mon histoire - ma sœur) et en phonétique celui des élisions et celui des liaisons.

C'est à cette étape que nous effectuons un traitement, car les informations nécessaires se trouvent dans les lexiques sous forme de marqueurs différents :

– pour les liaisons (en Z, en T et en N)

– pour les élisions (marqueur %). Signalons que ce marqueur n'est pas situé à la fin du mot (je, te, de, . . .) subissant l'élision mais au début du radical suivant la permettant (% aim, %ir, . . .).

– pour les contractions (à le → au, de la → du).

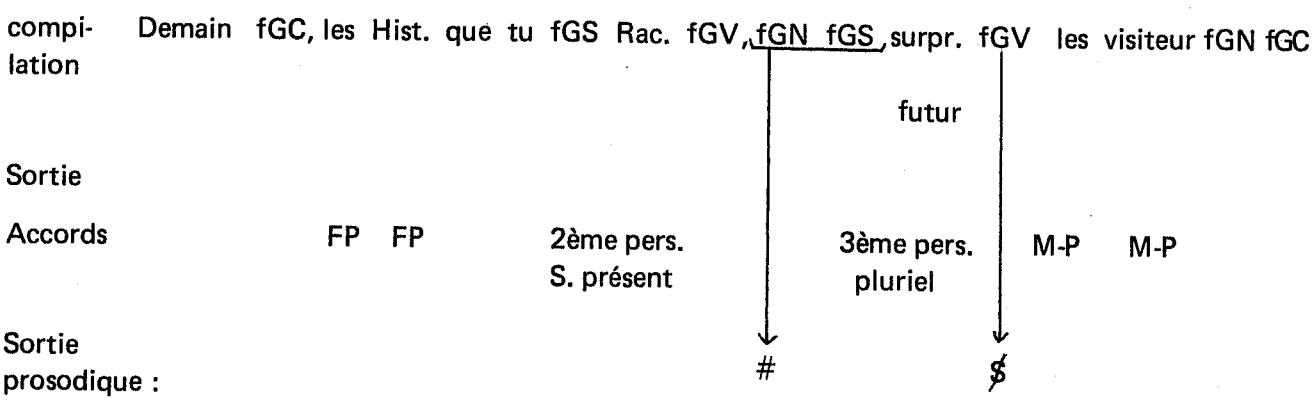
A la fin de ce module, nous disposons de deux suites grammaticalement correctes : l'orthographique est visualisée sur un écran pendant que la phonétique est synthétisée.

La figure 3 donne un exemple de traitement complet.

Entrée : demain, pluriel histoire que tu raconter surprendre futur pluriel visiteur

sortie pré-filtrage : demain, les histoire que tu raconter, surprendre les visiteur

Ce module a précisé les articles définis et généré une virgule, nécessaire à la compilation, à la fin de la relative.



Pas de marqueurs prosodiques avant «raconter» (pronom sujet), ni après (à cause de la virgule).

concaté-nation	Radical ...	L	%ISTWARE	#KEU"	....	SURPR	L	VIZITOE" ...
phoné-tique	Désinence ...	AI"+			....	AN"DRON"(	AI"+	R+ ...

Signification des différents marqueurs :

" : le phonème est caractérisé par les deux caractères précédents
\+ : possibilité de liaison en Z
(: .. .. en T
% : possibilité d'élision
# : marqueur prosodique

Sortie phonétique

DEU "MIN", LAI" ZISTWARE #KEU" TU RACON"TE SURPRAN"DRON" S LAI" VI ZITOE"R.

Fig. 3 : exemple complet de traitement d'une phrase

VII – RÉALISATION ET CONCLUSIONS

Ce système fonctionne sur un mini-calculateur équipé d'une mémoire 64 K octets et d'un disque magnétique, le programme comporte un millier d'instructions FORTRAN. Le classique dilemne «temps-mémoire» bien connu en informatique s'est évidemment posé et la réduction la plus large possible de la taille mémoire a été recherchée pour deux raisons : d'une part la durée nécessaire à la confection du message ne rend pas critique le temps d'exécution et d'autre part l'objectif ultime étant de réaliser un système portable autour d'un micro-processeur, il convenait de minimiser la mémoire utilisée.

Actuellement le programme et l'ensemble des lexiques (dictionnaire des diphones non compris) prennent environ 65 Koctets, le temps nécessaire à la visualisation de la suite orthographique et à la synthèse vocale est de 5 secondes pour une phrase de 10 mots.

L'objectif n'ayant pas été d'écrire une grammaire exhaustive du français parlé mais un logiciel permettant d'avoir un outil de communication souple et utile, ce sera donc l'expérimentation en centre de réadaptation qui dira si ce but a été atteint.

VIII – BIBLIOGRAPHIE

AHO, A.V., ULLMAN, J.D «The theory of parsing, translation, and compiling»
Vol. I, Prentice-Hall, 1972.

BERTHIER, F., «Analyse syntaxique pour un projet de prothèse de parole»
Rapport de stage CNET, mars 1980.

EMERARD, F., «Synthèse par diphones et traitement de la prosodie»
Thèse 3ème cycle, Grenoble 1977.

EMERARD, F. et al. «Outils de communication verbale pour les handicapés» 12ème J.E.P. Montréal, 1981.

EULENBERG, J.B., RAHIMI, M.A. «Toward a semantically accessible communication aid» Proceedings of the national electronics conference. Vol. XXXII. Chicago, octobre 1978.

GRAILLOT, P., EMERARD, F., «Synthèse par mots-clés : génération automatique des marqueurs prosodiques». Symposium Prosodie, Toronto 1981.

MEMMI, D., LIENARD, J.S. «Generation automatique de phrases en phonétique à partir de formules sémantiques» 10ème J.E.P., Grenoble 1979.

YOUNG, S.J., FALLSIDE, F., «Speech synthesis from concept» JASA. Vol. 66, nº 3, septembre 1979.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

MODÉLISATION ÉLASTIQUE DES CORDES VOCALES — ASPECTS MODAUX

Jean-Yves AULOGE Laboratoire Mécanique - École Centrale de Lyon - ECULLY (France)
Bernard GUERIN École Nationale Supérieure Électronique - Radioélectricité GRENOBLE (France)
Raymond DESCOUT Centre National Études Télécommunications - LANNION (France)

.RESUME

Le but de ce travail est de donner un nouveau modèle de vibrations des cordes vocales. On considère qu'elles peuvent être modélisées par deux poutres continues, reliées entre elles par des ressorts et des amortisseurs et excitées par l'air provenant des poumons. L'écoulement est un écoulement de Bernouilli.

Nous écrivons les équations du mouvement de flexion compte-tenu des conditions aux limites. Le système est non linéaire à cause des forces aérodynamiques agissant sur les poutres. Nous résolvons ces équations sur la base des modes propres pour la solution libre et la solution forcée.

La projection sur la base modale redonne le classique modèle à «deux masses» pour le premier mode. Une simulation numérique des oscillations de relaxation est réalisée.

. ABSTRACT

The aim of this paper is to give a continuous modelization of vocal cords by using beams instead of masses. The well - known «two - mass» model gives only informations on the fundamental frequency of the glottal waveform, but not on the harmonics of it.

This problem is solved by using a four continuous beams model. The aerodynamic forces are classical ones. The beams are hinged with tensions applied at the ends.

By projecting the solution on the modal basis we obtain the two - mass model as the first mode that is to say the fundamental mode. The following modes give the Titze model which appear as a higher order model better than the two - mass model.

In this paper, we show that this continuous modelization gives interesting theoretical results ; first results of simulation shall be presented in Montreal.

INTRODUCTION

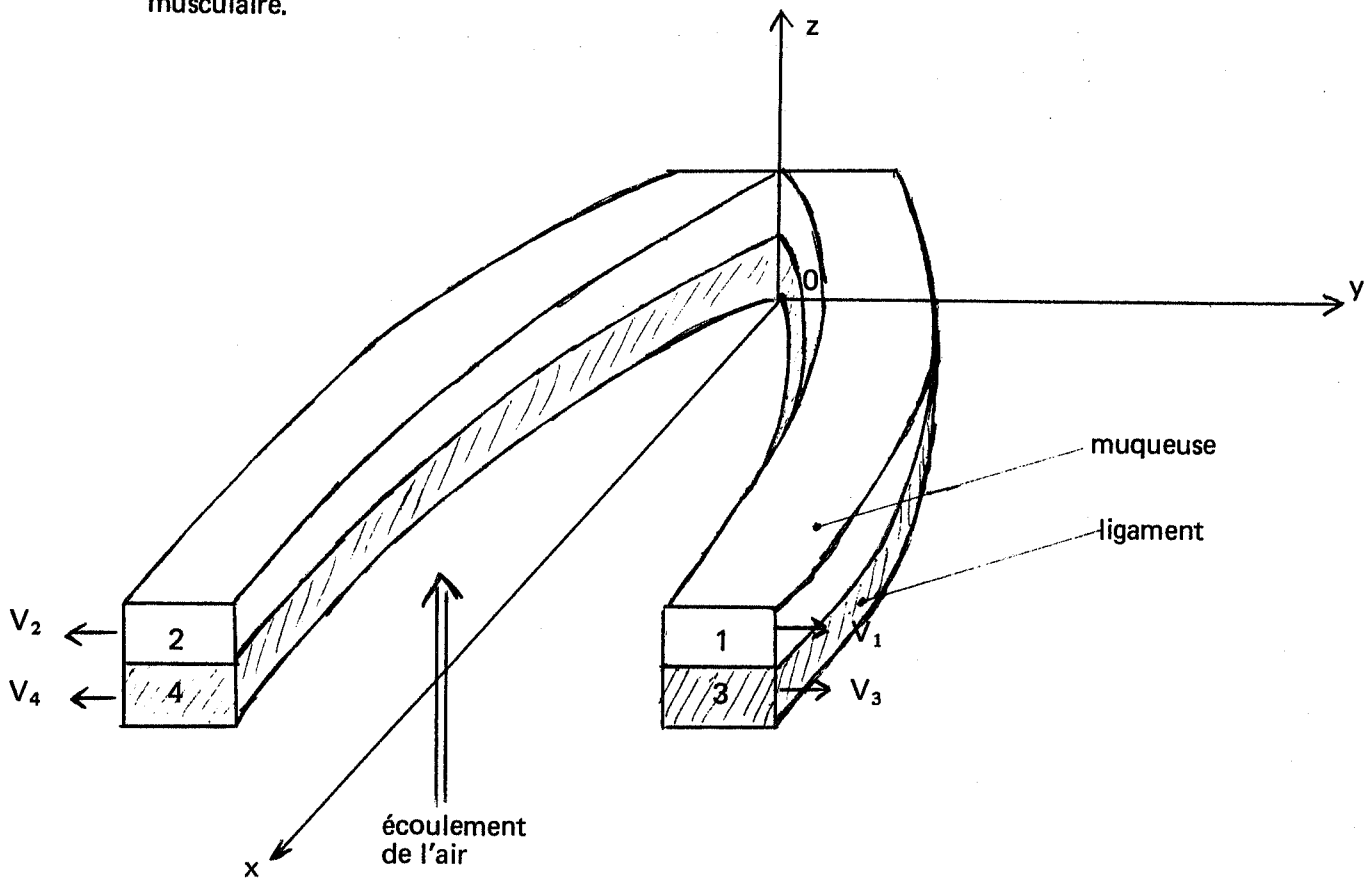
Les modèles classiques de la source vocale sont des modèles fondés sur un nombre fini de degrés de liberté excités par l'écoulement de l'air provenant des poumons. Un modèle «à une masse» a été proposé par Flanagan et Landgraf (1968). Un modèle à «deux masses» a été proposé par Ishizaka et Matsudaïra (1972) et un modèle plus conséquent par Titze (1974).

Dans notre modèle, au lieu de nous servir d'analogies électriques ou de modèles discrétisés, nous considérons que les cordes vocales peuvent être représentées par deux poutres biarticulées soumises à un effort normal de commande couplées par des ressorts et des amortisseurs et soumises aux mêmes forces aérodynamiques que le modèle à deux masses. Cette modélisation respecte les données physiologiques.

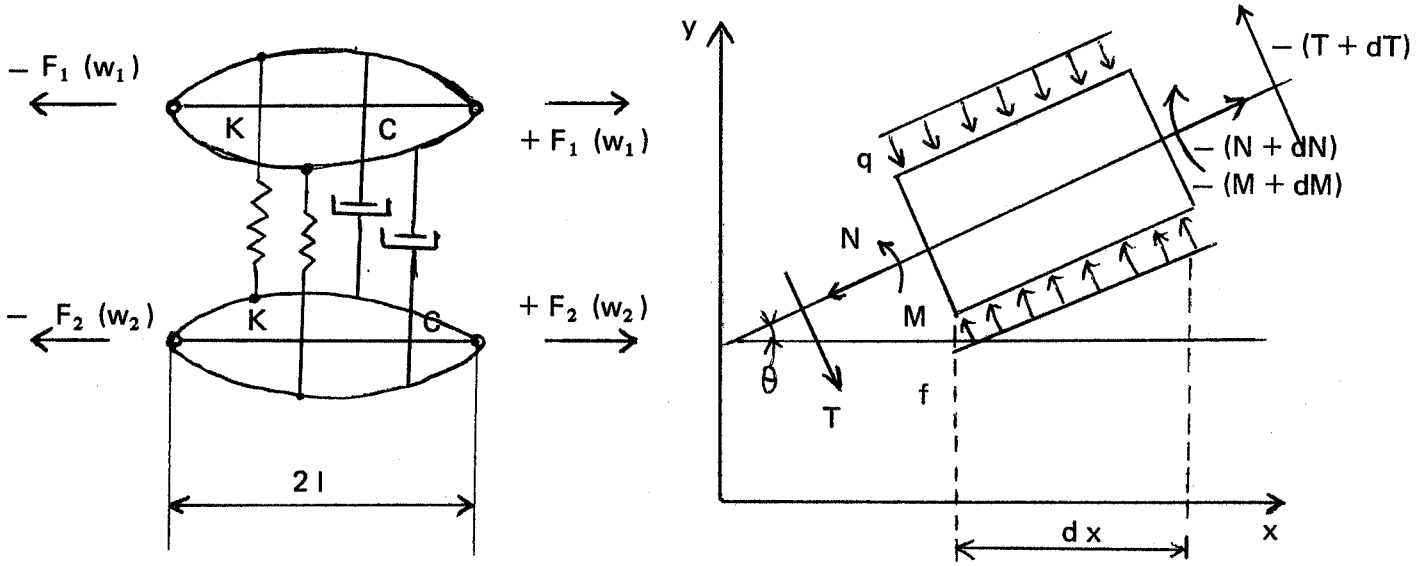
1 – DESCRIPTION DU MODELE

1.1 - Écriture des équations mécaniques du modèle :

Les cordes vocales sont supposées être des poutres biarticulées soumises à des forces de contrôle F_1 (W_1) et F_2 (W_2) qui apparaissent comme étant des efforts normaux d'origine musculaire.



Soient U_1, U_2 et V_3, V_4 les déplacements latéraux de chacune des poutres, T l'effort tranchant, N l'effort normal, M le moment fléchissant, q et f les densités de charge comptées algébriquement.



S la surface de la poutre, ρ la densité par unité de volume, E le module d'Young, I l'inertie géométrique :

En se servant des relations classiques de la résistance des matériaux et des conventions de signe de la R.d.M.

$$T = -\frac{\partial M}{\partial x} \quad ; \quad \theta = \frac{\partial y}{\partial x} \quad ; \quad M = -EI\,\frac{\partial^2 y}{\partial x^2}$$

L'application du principe fondamental de la dynamique permet d'écrire :

$$T-(T+dT) + f(x,t)\,dx - q(x,t)\,dx + N\theta - (N+dN)(\theta+d\theta) = \rho S\,dx\,\frac{\partial^2 y}{\partial t^2}$$

Ce qui conduit à l'équation aux dérivées partielles : en négligeant les effets dus à l'inertie de rotation.

$$EI\,\frac{\partial^4 y}{\partial x^4} + N\,\frac{\partial^2 y}{\partial x^2} + (\rho S)\cdot\frac{\partial^2 y}{\partial t^2} = f(x,t) - q(x,t)$$

Si l'on suppose qu'il existe un amortissement visqueux linéaire réparti le long de la poutre, alors on obtient :

$$EI\,\frac{\partial^4 y}{\partial x^4} + N\,\frac{\partial^2 y}{\partial x^2} + (\rho S)\cdot\frac{\partial^2 y}{\partial t^2} + C\,\frac{\partial y}{\partial t} = f(x,t) - q(x,t)$$

Il faut adjoindre les conditions aux limites

$$y(0,t) = y(2\ell,t) = 0$$: flèche nulle aux extrémités

$$\frac{\partial^2 y}{\partial x^2}(0,t) = \frac{\partial^2 y}{\partial x^2}(2\ell,t) = 0$$: moment nul aux extrémités

ce qui conduit aux équations pour chacune des poutres

$$(EI)_1 \frac{\partial^4 u_1}{\partial x_4} + N_1 \frac{\partial^2 u_1}{\partial x^2} + (\rho S)_1 \frac{\partial^2 u_1}{\partial t^2} + c_1 \frac{\partial u_1}{\partial t} = f_1(x,t) - q_1(x,t)$$

$$u_2 = -u_1 \quad ; \quad u_1 \geq 0$$

avec les conditions aux limites :

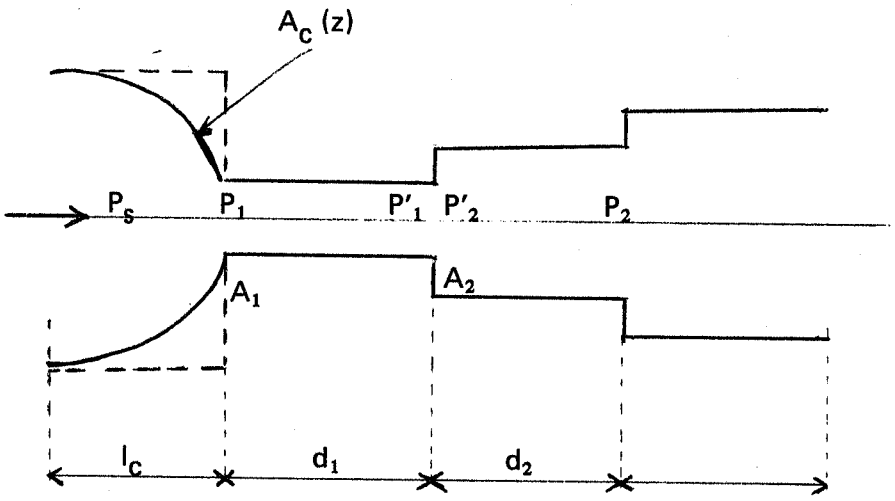
$$u_1(0,t) = u_1(2\ell,t)$$

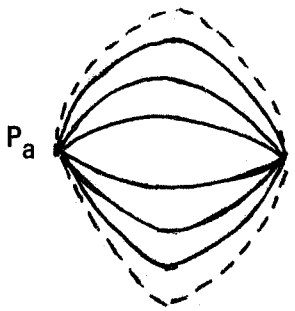
$$\frac{\partial^2 u_1}{\partial x^2}(0,t) = \frac{\partial^2 u_1}{\partial x^2}(2\ell,t)$$

mêmes équations pour V_3, les forces f_3 devenant q_3.

1.2 - Forces aérodynamiques agissant sur les poutres :

Les forces aérodynamiques agissant sur la structure ont des expressions analogues à celles données par Flanagan (1972).





Soient :

- P_i : les pressions aux différents points
- A_i : les aires d'entrée et de sortie
- U : le débit d'air
- d_i : les épaisseurs des deux parties de la glotte
- ρ : la densité de l'air
- μ : le coefficient de viscosité de l'air

Ag_i : les aires des sections droites au niveau de chacune des poutres
$A_c(z)$: la fonction d'aire de l'entrée
P_s : la pression subglottique
P_a : la pression à la sortie du conduit glottique

Les différences de pression sont alors données par les différentes expressions :

$$\bullet\ P_s - P_1 = \frac{1,37}{2}\,\rho\,\frac{u^2}{Ag_1^2} + \rho\left(\int_0^{l_c}\frac{dz}{A_c(z)}\right)\cdot\frac{du}{dt}$$

$$\bullet\ P_2 - P_a = \frac{1}{2}\,\rho\,\frac{u^2}{Ag_2^2}\cdot 2N(N-1) \quad ; \quad N = \frac{Ag_2}{A_2}$$

$$\bullet\ (-1)^i\,(P_i - P'_i) = 0,158\,\rho^{0,75}\,\mu^{0,25}\,l^{1,25}\,d_i\,\frac{u^{1,75}}{Ag_i^3} + \rho\,\frac{d_i}{Ag_i}\,\frac{du}{dt} \qquad i \in [1,2]$$

$$\bullet\ P'_1 - P'_2 = 0,5\,\rho\,u^2\left[\frac{1}{Ag_2^2} - \frac{1}{Ag_1^2}\right]$$

Soient $\overline{p}_1$ et $\overline{p}_2$ les pressions moyennes exercées sur les poutres en cas d'écoulement et soient FA_1 et FA_2 les forces aérodynamiques exercées sur les poutres, alors les forces aérodynamiques par unité de surface ont pour expression :

	$FA_1 / 2 l d_1$	$FA_2 / 2 l d_2$
$Ag_1 > 0 ; Ag_2 > 0$	$\overline{p}_1 = \frac{1}{2}(p_1 + p'_1)$	$\overline{p}_2 = \frac{1}{2}(p_2 + p'_2)$
$Ag_1 \leqslant 0 ; Ag_2 > 0$	P_s	0
$Ag_1 ; Ag_2 < 0$	P_s	P_s

d'où les expressions des forces FA_1 et F_2

1.3 - Forces extérieures appliquées sur les poutres :

Parmi les forces extérieures, il y a les forces de liaison d'origine mécanique et les forces de contact.

1.3.1 - Forces de liaison d'origine mécanique :

Pour la poutre 1, dont le déplacement est U_1 , la force mécanique FM_1 a pour expression :

$$FM_1 = -K(u_1 - v_3)\,\delta(x-\ell) - c\left(\frac{\partial u_1}{\partial t} - \frac{\partial v_3}{\partial t}\right)\delta(x-\ell)$$

$$= -\left[K(u_1 - v_3) + c\left(\frac{\partial u_1}{\partial t} - \frac{\partial v_3}{\partial t}\right)\right]\cdot\delta(x-\ell)$$

Pour la poutre 3 dont le déplacement est v_3

$$FM_3 = -FM_1 = -\left[K(v_3 - u_1) + c\left(\frac{\partial v_3}{\partial t} - \frac{\partial u_1}{\partial t}\right)\right]\cdot\delta(x-\ell)$$

Pour les poutres 2 et 4, il suffit de remplacer U_1 par U_2 et v_3 par v_4.

1.3.2 - Étude des forces de contact :

Lorsqu'il y a oscillation des poutres, il y a une succession de plages temporelles pour lesquelles on obtient des déformés négatives $u_1 \leqslant 0$; $v_3 \leqslant 0$.

Il y a alors contact des poutres (1) et (2) et des poutres (3) et (4). Ces forces de contact apparaissent donc lorsque l'aire est négative et sont nulles lorsque l'aire est positive.

$FC_i = 0$ si $Ag_i > 0$

On suppose que la force de contact est non linéaire avec une non linéarité de type Düffing, de façon que la force de rappel soit d'autant plus grande que la poutre s'enfonce dans sa vis à vis. D'autre part, le contact dissipe de l'énergie : il est intéressant de lui associer une dissipation énergétique obtenue à partir d'un amortissement visqueux de type Rayleigh.

La force de contact est donc l'intégrale des forces de contact élémentaires pour un élément de longueur dx.

$$d(FC_1) = \left[-k_1 u_1(x,t)\, d_1 (1+\beta_1 u_1^2(x,t)) - d_1 D_1 \frac{\partial u_1}{\partial t}(x,t)\right]\cdot dx$$

$$FC_1 = \int_0^{2\ell} -\left[k_1 u_1(x,t)\cdot d_1(1+\beta_1 u_1^2(x,t)) + d_1 D_1 \frac{\partial u_1}{\partial t}\right] dx$$

$$FC_1 = -k_1 d_1 \int_0^{2\ell} u_1(x,t)\,dx - \beta_1 k_1 d_1 \int_0^{2\ell} u_1^3(x,t)\,dx - d_1 D_1 \int_0^{2\ell} \frac{\partial u_1}{\partial t}(x,t)\,dx.$$

$$\text{or} \quad Ag_1 = 2 \int_0^{2\ell} u_1(x,t)\, dx.$$

$$\frac{dAg_1}{dt} = 2 \int_0^{2\ell} \frac{\partial u_1}{\partial t}(x,t)$$

La force de contact peut donc s'écrire sous la forme :

$$Fc_1 = -\frac{1}{2\ell} k'_1 \cdot Ag_1 - \beta'_1 k'_1 \int_0^{2\ell} u_1^3(x,t)\, dx$$

$$- D'_1 \cdot \frac{dAg_1}{dt} \cdot \frac{1}{2\ell}$$

De la même façon, la force de contact FC_3 peut se mettre sous la forme :

$$Fc_3 = -\frac{k'_3}{2\ell} Ag_2 - \beta'_3 k'_3 \int_0^{2\ell} u_3^2(x,t)\, dx - \frac{D'_3}{2\ell} \cdot \frac{dAg_2}{dt}$$

Remarque :

Contrairement à Flanagan, on ne peut pas a priori remplacer $\int_0^{2\ell} u_1^3(x,t)\, dx$. par Ag_1^3 sauf en faisant une hypothèse sur la déformée. Supposons que cette déformée soit celle du premier mode :

$$u_1(x,t) = \sin \frac{\pi x}{2\ell} \quad \alpha(t)$$

Alors

$$Ag_1(t) = \int_0^{2\ell} \sin\left(\frac{\pi x}{2\ell}\right) \alpha(t)\, dx = \frac{4\ell}{\pi} \alpha(t)$$

et

$$\int_0^{2\ell} u_1^3(x,t)\, dx = \alpha^3(t) \int_0^{2\ell} \sin^3\left(\frac{\pi x}{2\ell}\right) dx = \frac{8\ell}{3\pi} \alpha^3(t)$$

D'où :

$$\int_0^{2\ell} u_1^3(x,t)\, dx = \frac{\pi^2}{24\ell^2} \left(Ag_1(t)\right)^3 = \frac{\pi^2 \ell}{3} \left(\frac{Ag_1}{2\ell}\right)^3$$

et de la même façon

$$\int_0^{2\ell} v_3^3(x,t)\,dx = \frac{\pi^2}{24\ell^2}\left(Ag_2\right)^3(t) = \frac{\pi^2\ell}{3}\left(\frac{Ag_2}{2\ell}\right)^3$$

1.3.3 - Équations du modèle :

Les équations du modèle sont obtenues en remplaçant f_1 et q_1 ainsi que f_3 et q_3 par leurs expressions.

$$\bullet (EI)_1 \frac{\partial^4 u_1}{\partial x^4} + N_1 \frac{\partial^2 u_1}{\partial x^2} + (\rho S)_1 \frac{\partial^2 u_1}{\partial t^2} + C_1 \frac{\partial u_1}{\partial t} + \left[K(u_1 - v_3) + C\left(\frac{\partial u_1}{\partial t} - \frac{\partial v_3}{\partial t}\right)\right]\delta(x-\ell) = \left[FA_1 + FC_1\right]\frac{1}{2\ell}$$

$$\bullet (EI)_2 \frac{\partial^4 v_3}{\partial x^4} + N_3 \frac{\partial^2 v_3}{\partial x^2} + (\rho S)_3 \frac{\partial^2 v_3}{\partial t^2} + C_3 \frac{\partial v_3}{\partial t} + \left[K(v_3 - u_1) + C\left(\frac{\partial v_3}{\partial t} - \frac{\partial u_1}{\partial t}\right)\right]\delta(x-\ell) = \left(FA_3 + FC_3\right)\frac{1}{2\ell}$$

$$\bullet\ v_4 = -v_3$$

$$\bullet\ u_2 = -u_1$$

avec les conditions aux limites :

$$u_1(0,t) = u_1(2\ell,t) = 0$$

$$\frac{\partial^2 u_1}{\partial x^2}(0,t) = \frac{\partial^2 u_1}{\partial x^2}(2\ell,t) = 0$$

$$v_3(0,t) = v_3(2\ell,t) = 0$$

$$\frac{\partial^2 v_3}{\partial x^2}(0,t) = \frac{\partial^2 v_3}{\partial x^2}(2\ell,t) = 0$$

2 – DÉCOMPOSITION MODALE – LIEN AVEC LE MODELE A 2 MASSES

2.1 - Étude de la projection modale pour des équations non amorties sans couplage :

Soit l'équation :

$$(EI)_1 \frac{\partial^4 u_1}{\partial x_4} + N_1 \frac{\partial^2 u_1}{\partial x^2} + (\rho S)_1 \frac{\partial^2 u_1}{\partial t^2} = q_1(x,t)$$

avec les conditions aux limites précédentes.

On cherche une solution de la forme :

$$u(x,t) = U(x).\ \alpha(t)$$

pour l'équation sans second membre : ce qui conduit à des fonctions

$$U(x) = A\ ch\ r_1 x + B\ sh\ r_1 x + C\ \cos r_2 x + D \sin r_2 x$$

avec r_1 et r_2 solution de l'équation du 4ème degré

$$(EI)_1 r^4 + N_1 r^2 - \omega^2 (\rho S)_1 = 0$$

telles que

$$r_1^2 = \frac{-N_1 + \sqrt{N_1^2 - 4\omega^2(\rho S)_1(EI)_1}}{2} \quad ; \quad (r_1 \in R)$$

$$r_2^2 = \frac{-N_1 - \sqrt{N_1^2 - 4\omega^2(\rho S)_1(EI)_1}}{2} \quad ; \quad (r_2 \in R)$$

et ω^2 est un paramètre positif qui va être déterminé dans les lignes suivantes en écrivant les conditions aux limites, ce qui conduit à :

$$A = B = C = 0$$

D quelconque dépendant des conditions initiales. Les pulsations propres ω^2 sont telles que :

$$\sin(2\ell r_2) = 0$$

soit $\quad 2\ell r_2 = n\pi \quad$ ou $\quad r_2^2 = \left(\frac{n\pi}{2\ell}\right)^2$

ce qui permet de calculer ω_n.

On constate donc que l'effort normal N_1 a une influence sur les pulsations propres du système.

La fonction $\alpha(t)$ doit donc être indicée n et est sinusoïdale :

$$\alpha_n(t) = \sin(\omega_n t - \varphi_n)$$

La solution générale de l'équation sans second membre est alors :

$$u_1(x,t) = \sum_{n=0}^{\infty} U_n(x) \cdot \alpha_n(t)$$

$$= \sum_{n=0}^{\infty} D_n \cdot \sin\left[(r_2)_n x\right] \sin(\omega_n t - \varphi_n)$$

Notons $\tilde{U}_n(x) = \sin\left[(r_2)_n x\right]$ les formes propres qui satisfont au principe d'orthogonalité des modes :

$$\langle \tilde{U}_n(x), \tilde{U}_m(x) \rangle = 0 \quad \text{si } n \neq m \text{ pour le produit scalaire}$$

$$\int_0^{2\ell} \tilde{u}_n(x) \cdot \tilde{U}_m(x)\, dx = \langle \tilde{U}_n, \tilde{U}_m \rangle$$

Il est alors possible de chercher la solution de l'équation par projection sur la base des modes propres

$$u_1(x,t) = \sum_{n=0}^{\infty} \tilde{U}_n(x) \cdot \alpha_n(t)$$

Alors les fonctions $\alpha_n(t)$ sont solution de l'équation différentielle obtenue en multipliant la solution par $U_m(x)$ et en remarquant que

$$\tilde{U}_m^4(x) = (r_2)_m^4 \cdot \tilde{U}_m(x)$$

$$\tilde{U}_m^2(x) = (r_2)_m^2 \cdot \tilde{U}_m(x)$$

et ensuite en prenant le produit scalaire on obtient :

$$\left[(EI)_1 (r_2)_m^4 \alpha_n - N_1 (r_2)_n^2 \alpha_n + (\rho S_1) \ddot{\alpha}_n\right] \langle U_n, \tilde{U}_n \rangle$$

$$= \langle \tilde{U}_n, q_i \rangle$$

Par conséquent $\alpha_n(t)$ apparaît comme étant la réponse forcée d'un oscillateur masse ressort :

$$\alpha_n(t) = \frac{\langle \tilde{U}_n, q_i \rangle}{\langle \tilde{U}_n, \tilde{U}_n \rangle \left[(EI)_1 (r_2)_n^4 - N_1 (r_2)_n^2 - \omega^2 (\rho S)_1\right]}$$

avec $\langle \tilde{U}_n, q_i \rangle = \psi_n(t) = F_0 \sin \omega t$

α_n (t) présente donc un caractère harmonique et l'amplitude dépend de la pulsation de l'excitation. Cette amplitude dépend aussi de la tension des cordes vocales qui va donc modifier plus ou moins l'amplitude de la vibration.

2.2 - Étude de la projection modale pour des équations couplées

Cas non amorti

Nous supposerons que les équations complètes sont couplées. Nous allons nous servir des résultats de 2.1 pour chercher des solutions de la forme :

$$u_1(x,t) = \sum_{n=0}^{\infty} \alpha_n^1(t)\, \tilde{u}_n(x)$$

$$v_3(x,t) = \sum_{n=0}^{\infty} \alpha_n^3(t)\, \tilde{V}_n(x)$$

où $\tilde{U}_n$ et $\tilde{V}_n$ sont les formes modales pour les poutres (1) et (2).

Les solutions élémentaires $\alpha_n^1(t)\,\tilde{U}_n(x)$ et $\alpha_n^3(t)\,\tilde{V}_n(x)$ doivent vérifier l'équation aux dérivées nulles.

$$(EI)_1 \frac{\partial^4 u_1}{\partial x^4} + N_1 \frac{\partial^2 u_1}{\partial x^2} + (\rho S)_1 \frac{\partial^2 u_1}{\partial t^2} + \left[K(u_1 - v_3) + C\left(\frac{\partial u_1}{\partial t} - \frac{\partial v_3}{\partial t}\right)\right]\delta(x-\ell) = (FA_1 + FC_1)\frac{1}{2\ell}$$

ainsi que celle, identique en V_3.

En projetant sur la base modale et en écrivant que les fonctions $\alpha_n^1(t)$ et $\alpha_n^2(t)$ sont synchrones.

$$\alpha_n^1(t) = \lambda\, \alpha_n^3(t) \qquad : \lambda \text{ constante}$$

On obtient les deux équations couplées que l'on peut écrire sous forme vectorielle dans l'hypothèse des forces extérieures nulles avec $\alpha_n \mid \begin{matrix} \alpha_n^1 \\ \alpha_n^3 \end{matrix}$

$$\underline{M}_n\, \ddot{\alpha}_n + \underline{K}_n\, \alpha_n = 0$$

Les fréquences propres du mouvement de synchronisation sont les racines carrées positives des valeurs propres de $\underline{M}_n^{-1} \cdot \underline{K}_n$

Dès lors pour obtenir les solutions forcées des deux équations couplées, la projection sur la base modale va donner :

$$\underline{M}_n\, \ddot{\psi}_n + \underline{K}_n\, \psi_n = F_n(t)$$

avec $F_n(t)$ définit comme le vecteur : $\langle FA_1 + FC_1 \mid \tilde{U}_m \rangle$

$\langle FA_3 + FC_3 \mid \tilde{V}_m \rangle$

C'est une équation différentielle vectorielles qui donne la projection de la solution forcée sur le nième mode.

Cas amorti

De la même façon, l'existence des termes répartis $C_1 \frac{\partial u_1}{\partial t}$ et $C_3 \frac{\partial v_3}{\partial t}$ entraînerait la résolution d'une équation de la forme :

$$\underline{M_m}\, \ddot{\psi}_m + \underline{C_m}\, \dot{\psi}_m + \underline{K_n}\, \psi_m = G_m(t)$$

avec des matrices $\underline{M}_n$, $\underline{C}_n$, $\underline{K}_n$ et une force $G_n(t)$ obtenue par projection modale.

Pour $n = 1$, on obtient les équations du modèle à deux masses bien connues, dont on peut d'ailleurs identifier les coefficients comme étant respectivement les masses, amortissements et raideurs modales correspondant au premier mode.

Mais on obtient également en sélectionnant les premiers modes de vibration le modèle de Titze, dont le modèle à 16 masses constitue une méthode de résolution numérique des différents modes d'une poutre et qui tient bien mieux compte que le modèle à 2 masses des différentes contributions modales.

L'intérêt d'une telle modélisation est d'expliquer le transfert énergétique entre la vibration de la poutre et le signal acoustique émis mode par mode.

CONCLUSION

La modélisation continue de la source vocale permet de généraliser le modèle à deux masses tout en ayant un aspect plus physique et en intégrant l'aspect d'une commande des cordes vocales sous forme d'un effort normal. Il faudra attendre cependant la fin de la simulation numérique pour pouvoir comparer les deux modèles. Néanmoins le modèle à deux masses apparaît comme un cas particulier du modèle continu à deux poutres.

Des résultats, de simulation numérique, actuellement en cours d'achèvement, seront présentés à Montréal.

BIBLIOGRAPHIE

(1) J.L. Flanagan, L.L. Landgraf,
«Self-oscillating source for Vocal-Tract synthesizer», IEEE Trans audio, AU-16, pp. 57-64 ; 1968.

(2) K. Ishizaka, J.L. Flanagan,
«Synthesis of voiced sounds from a two-mass model of the vocal cords» BSTJ-51, pp. 1233-1268, 1972.

(3) J. Y. Auloge,
«Modélisation continue de la source vocale», Séminaire Larynx et Parole, Groupe de la Communication Parlée, GALF, Ed. Boé, Descout, Guerin, Grenoble, Février 1979.

(4) L.J. Boé,
«Anatomie et Physiologie de la Phonation», Travaux de l'Institut de Phonétique de Grenoble , 1979.

(5) L.J. Boé, B. Guérin,
«Étude du fonctionnement d'un modèle de cordes vocales à deux masses», Recherches sur la Prosodie du Français, Ed. Univ. des Sc. et Lettres de Grenoble, pp. 7-56, 1979.

(6) J.C. Burtin,
«Modélisation de la Source Vocale», Rapport RP/DAS/ETA/64, CNET, 1979.

(7) M. Carcaud,
«Contribution à la synthèse de la parole-Simulation de la source vocale», Thèse de 3ème cycle, Université des Sciences Orsay, 1975.

(8) M.Grojnowski,
«Étude de la source sonore dans un système de synthèse de la parole, modèle des cordes vocales à deux masses», Note technique, NT/CES/101, CNET, juin 1975.

(9) B. Guérin,
«Contribution aux recherches sur la production de la parole. Étude du fonctionnement de la source vocale. Simulation d'un modèle» Thèse d'État, Institut Polytechnique de Grenoble, 1978.

(10) I.R. Titze,
«The human vocal cords : mathematical model», part I : Phonetica, 28, pp. 129-170, 1974,
part II : Phonetica, 29, pp. 1-21, 1975.

(11) I.R. Titze, D. Talkin,
«Estimation of Larungeal Control parameters from glottal waveforms» JASA 66, 1, July 79, pp. 60-74.

(14) R. Descout, J.Y. Auloge, B. Guérin,
«Continuous Model of the Vocal Source», I.C.A.S.S.P. April 80 - Denver.

(15) S. Timoschenko,
«Théorie des vibrations», Beranger / Dunod, 1954.

(16) L. Melrovitch,
«Éléments of vibration analysis», Me Graw Hill, 1975.

(12) I.R. Titze, W. J. Strong,
«Normal modes in Vocal Cord tissues», JASA 57, pp. 736-744, 1975.

(13) I.R. Titze,
«On the mechanisms of vocal fold vibrations», JASA 60, pp. 1366-1380, 1976.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

UTILISATION DES METHODES ACOUSTIQUES POUR L'EVALUATION DES FONCTIONS D'AIRE DU CONDUIT VOCAL

Jean-Paul LEFEVRE, Bernard TOUSIGNANT, Michel LECOURS
Département de Génie Electrique
Université Laval, Québec, Canada, G1K 7P4

.RESUME

Cet article décrit une méthode de détermination de la fonction d'aire du conduit vocal fondée sur des mesures, au niveau des lèvres, de l'onde de pression réfléchie par le conduit vocal en réponse à une impulsion acoustique. Une solution est apportée aux problèmes de quantification et d'interpolation de la fonction d'aire, problèmes issus des contraintes propres à une telle mesure. Les techniques utilisées pour améliorer la convergence et la stabilité des solutions sont présentées. Quelques mesures de fonctions d'aire pour des configurations de conduit vocal tant fixes qu'évolutives, correspondant à des voyelles soutenues et à des transitions entre voyelles font l'objet d'une analyse détaillée.

Vérifier le réalisme des mesures présente certaines difficultés. Un moyen indirect de vérification consiste à utiliser les séquences de fonctions d'aire obtenues comme paramètres de commande dans un synthétiseur. L'accent est mis sur les points originaux de cette réalisation; les possibilités offertes par cet outil étant commentées.

UTILISATION DES METHODES ACOUSTIQUES POUR L'EVALUATION DES FONCTIONS D'AIRE DU CONDUIT VOCAL

Jean-Paul LEFEVRE, Bernard TOUSIGNANT, Michel LECOURS

. ABSTRACT

This paper presents some studies and experiments in vocal-tract area function determination by measurements at the lips of the response to an impulsive acoustic pressure wave. Problems related to the quantization and interpolation of area function are discussed. Required techniques to speed up convergence and stability of solutions are introduced. Some area function measurements for fixed and moving vocal-tract configurations corresponding to sustained vowels and vowel transitions are analysed.

The likeliness of the measured area functions is not easy to verify. An indirect method to check this consists in the utilisation of frames of measured area functions, as control parameters in a speech synthesizer. Only original developments are discussed, and possible uses of this approach are presented.

UTILISATION DES METHODES ACOUSTIQUES POUR L'EVALUATION DES FONCTIONS D'AIRE DU CONDUIT VOCAL

Jean-Paul LEFEVRE, Bernard TOUSIGNANT, Michel LECOURS

I. INTRODUCTION

Cet article décrit les études et expériences que nous avons menées depuis plusieurs années en vue de mesurer les fonctions d'aire du conduit vocal et de réaliser une synthèse de la parole à partir d'un modèle physiologique des cordes et du conduit vocal.

Connaître la forme du conduit vocal et être capable de suivre son évolution durant une élocution continue sont deux éléments très importants pour une meilleure compréhension des phénomènes de phonation et présentent un intérêt majeur pour toutes les recherches s'intéressant à la synthèse et à la reconnaissance de la parole. En vue d'arriver à une synthèse par règles, il est nécessaire de développer des modèles acoustiques de production de la parole, réagissant à un nombre limité de paramètres assujettis à des contraintes physiologiques. Pour atteindre ce but, la connaissance de la fonction d'aire réelle apparaît nettement préférable à une simple configuration donnant la même réponse acoustique, telle qu'obtenue par exemple avec les méthodes d'analyse par prédiction linéaire. L'objectif global et à long terme de tous ces travaux sera donc beaucoup plus d'améliorer nos connaissances sur les phénomènes de production de la parole, plutôt que la reproduction proprement dite de sons, et en ce sens pourra avoir des applications très importantes dans plusieurs domaines, tel par exemple la rééducation des handicapés.

La méthode d'évaluation de la fonction d'aire, développée dans cet article, consiste à mesurer la réponse impulsionnelle du conduit vocal aux lèvres dans le domaine temporel (SONDHI et GOPINATH, 1971). Les difficultés de cette approche sont essentiellement liées aux conditions expérimentales. Mentionnons en particulier que seules les positions articulatoires, présentant une ouverture suffisante de la bouche, pourront être analysées, et que l'expérience requiert du sujet une simulation de l'articulation sans phonation. Signalons également que la précision du résultat obtenu va quelque peu décroître à mesure que l'on s'éloigne des lèvres en direction de la glotte. Par contre, parmi les avantages, il est fondamental de noter que ce type de mesure fournit de façon unique la fonction d'aire moyennant certaines hypothèses réalistes, tel que l'a démontré théoriquement SONDHI (1971). Il est de plus intéressant de constater que la méthode ne nécessite la connaissance, ni de la longueur du conduit vocal, ni des conditions aux limites de la glotte. La méthode permet également d'obtenir une aire réelle grâce à l'utilisation d'un tube de mesure servant de référence absolue, ce qui résout par conséquent les problèmes de normalisation rencontrés par exemple en prédiction linéaire. Enfin, on peut signaler que les résultats sont peu affectés par le manque d'information relative aux fréquences élevées.

Dans une première partie, cet article présente et évalue la performance d'une méthode de détermination de la fonction d'aire du conduit vocal modélisée sous forme de coefficients de réflexion. L'identification des coefficients est basée sur la connaissance de l'onde réfléchie par le conduit vocal, mesurée aux lèvres, en réponse à une impulsion acoustique envoyée elle aussi aux lèvres. Nous nous attacherons en particulier à préciser

des stratégies de quantification et d'interpolation de la fonction d'aire, permettant de respecter les conditions d'application de la méthode. Nous présenterons également les techniques utilisées pour améliorer la convergence et la stabilité de la solution.

Dans une seconde partie, nous décrirons la configuration d'un synthétiseur de parole basé sur un modèle physiologique des cordes et du conduit vocal, synthétiseur qui utilise comme paramètres de commande les fonctions d'aire mesurées au préalable. Les différents éléments constituant un ensemble de synthèse de ce type ont déjà été abondamment décrits dans la littérature. Aussi mettrons-nous seulement l'accent sur les points originaux de notre réalisation.

Des mesures de fonction d'aire pour des configurations de conduit vocal tant fixes qu'évolutives, correspondant à des voyelles soutenues et à des transitions entre voyelles, ont été réalisées. Les résultats expérimentaux mettent en évidence la continuité se dégageant des mesures de configurations évolutives du conduit vocal et illustrent la détermination de la position des lèvres et de la glotte. Il est possible de vérifier le réalisme des fonctions d'aire mesurées en les comparant directement aux rares résultats accessibles dans la littérature. Certaines vérifications indirectes peuvent également être entreprises à partir des sons synthétisés à l'aide des fonctions d'aire obtenues.

II. MESURES DES FONCTIONS D'AIRE

Après avoir explicité la modélisation utilisée pour le conduit vocal, nous allons décrire brièvement le processus expérimental, ce qui nous conduira à aborder les problèmes liés à la quantification et à l'interpolation des mesures. Enfin, nous expliciterons les techniques utilisées pour améliorer la convergence et la stabilité de l'algorithme.

II.1 Modèle numérique du conduit vocal

Du point de vue acoustique, le conduit vocal humain, composé des cavités pharyngales et buccales, se comporte comme un tube dont la section est variable dans le temps et dans l'espace. La propagation du son peut y être supposée du type onde plane. Une telle approximation reste valable tant que la longueur d'onde des signaux en cause est supérieure aux dimensions transversales du conduit, ce qui correspond à une fréquence maximum d'environ 5 kHz. Ainsi apparaît la première contrainte sur le modèle qui devra respecter cette hypothèse de largeur de bande limitée.

Si de plus on assume que les variations d'aire du conduit vocal sont lentes par rapport à celles des autres grandeurs, il est alors possible de représenter la transmission de l'énergie acoustique de la glotte vers les lèvres par les équations différentielles de Webster. L'une des méthodes les plus utilisées pour résoudre ces équations est celle faisant appel à la notion de coefficients de réflexion telle que proposée par KELLY et LOCHBAUM (1962), méthode dans laquelle les pertes sont supposées indépendantes de la fréquence. Dans ce modèle à paramètres localisés, les variations de géométrie du conduit vocal se représentent alors par une succession de coefficients de réflexion qu'il est possible de visualiser comme une succession de tubes cylindriques élémentaires de longueur faible devant la longueur d'onde de la plus grande fréquence transmise. Dans chaque tube élémentaire, les ondes de pression incidente p' et réfléchie p se propagent indépendamment l'une de l'autre en onde plane. Cette propagation peut ainsi être re-

présentée comme un retard pur. Pour prendre en compte les pertes, nous avons adopté un coefficient d'atténuation empirique γ variable en fonction de l'aire de la section, tel que suggéré originellement par STEVENS (1953).

A chaque discontinuité de section, une partie de chaque onde est transmise, l'autre étant réfléchie. Si on admet que les variations temporelles des aires ne sont pas trop brutales, il est possible de négliger les termes en $\partial A/\partial t$ et par conséquent d'exprimer les coefficients de réflexion r; r_n représentatif de la jonction n entre les aires A_n et A_{n+1}; sous la forme;

$$r_n = \frac{A_n - A_{n+1}}{A_n + A_{n+1}} \tag{1}$$

L'ensemble de la structure, à chaque jonction, peut alors être exprimé par les relations suivantes.

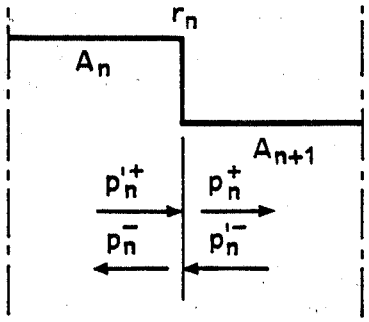


Fig. 1

A l'instant iT :

$$p^{+}_{n,i} = (1+r_n)p'^{+}_{n,i} - r_n p'^{-}_{n,i} \tag{2}$$

$$p^{-}_{n,i} = (1-r_n)p'^{-}_{n,i} + r_n p'^{+}_{n,i} \tag{3}$$

A l'instant (i+1)T :

$$p'^{+}_{n,i+1} = \gamma_n p^{+}_{n-1,i}\ ;\ p'^{-}_{n,i+1} = \gamma_{n+1} p^{-}_{n+1,i+1} \tag{4}$$

L'indice + est attaché aux ondes se propageant de la glotte vers les lèvres.

Le conduit vocal étant modélisé sous la forme d'une succession de tubes cylindriques élémentaires, il est possible de définir un coefficient de réflexion pour chaque jonction entre deux sections adjacentes du tube. L'identification de ces coefficients va être réalisée à partir de la mesure de l'onde réfléchie par le conduit vocal en réponse à une impulsion acoustique. Par ailleurs, il est important de souligner dès maintenant que ce modèle à paramètres localisés ne conserve pas de trace de la contrainte originale de largeur de bande limitée. Par conséquent cette contrainte devra être réintroduite lors du processus de reconstitution du conduit.

II.2 Mesure des fonctions d'aire

On introduit dans la bouche, au niveau des lèvres, une onde de pression acoustique plane et de courte durée e(t) et l'on recueille l'onde s(t) réfléchie par la cavité. Le processus dure environ une milliseconde, temps assez court pour pouvoir considérer le conduit vocal comme étant stationnaire. Différents essais ont montré que les réflexions secondaires s'amortissent suffisamment rapidement pour que l'on puisse répéter l'expérience au maximum cinquante fois par seconde, ce qui permet aisément de suivre l'évolution dynamique du conduit vocal. Le montage expérimental est reproduit à la figure 2.

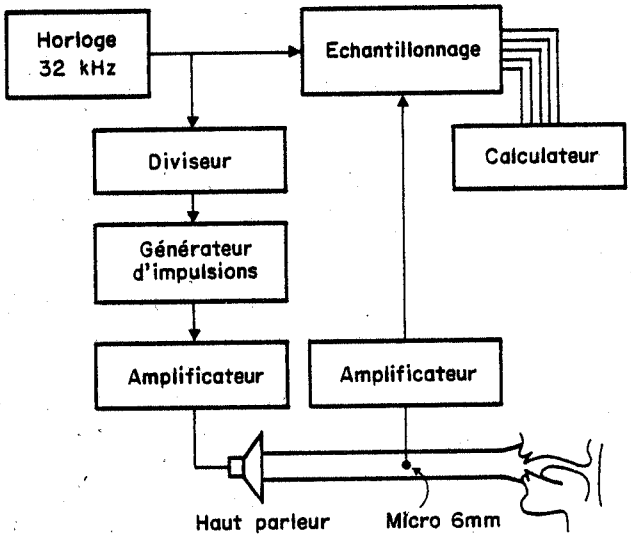


Fig. 2 - Montage expérimental
- *Experimental set-up*

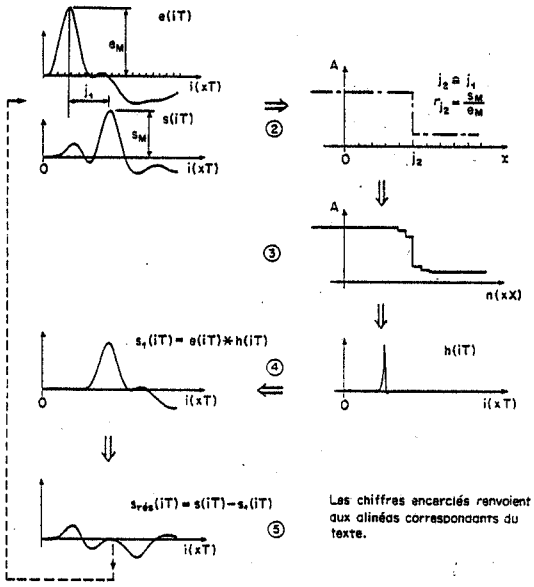


Fig. 3 - Algorithme de reconstitution
- *Reconstruction algorithm*

Pour obtenir la fonction d'aire A(x) du conduit vocal à partir de e(t) et de s(t), la séquence de calcul, schématisée à la figure 3, est la suivante:

1. L'excitation e(t) et la réponse s(t) sont échantillonnées en s'assurant que les deux signaux sont à bande limitée: les échantillons e(iT) et s(iT) sont ainsi obtenus.

2. La variation principale de la fonction d'aire A(x) est localisée. L'amplitude de cette variation est évaluée.

3. Cependant, afin de respecter les hypothèses de départ sur la nature des signaux temporels et sur A(x), il s'avère nécessaire de modéliser la variation détectée par un profil de transition. L'introduction de cette transition conduit à une modification de quelques sections élémentaires de part et d'autre de la variation détectée. Ce point particulier va être développé dans le paragraphe suivant.

4. En utilisant le modèle de transmission précédemment décrit, il est possible d'obtenir la réponse impulsionnelle discrète h(iT) de la fonction d'aire A(nX) calculée. Par convolution de cette réponse impulsionnelle et du signal d'entrée, la réponse de la fonction d'aire A(nX) à l'excitation e(iT) est obtenue.

5. En soustrayant la réponse calculée du signal observé, il subsiste un signal résiduel, qui est utilisé pour trouver une autre variation de section en reprenant la séquence au point 2 ci-dessus. Par passages successifs la fonction d'aire est ainsi modélisée. Un attrait évident de la méthode est qu'elle détermine les constrictions par ordre d'importance.

II.3 Lissage des variations détectées

L'un des points majeurs du processus exposé concerne l'introduction des variations de section détectées dans le modèle numérique du conduit vocal. Pour respecter les hypothèses de départ, ces variations doivent être modélisées par des transitions, elles-mêmes numériquement représentées par un jeu de coefficients de réflexion. En effet, les travaux théoriques de SONDHI (1971) ont montré la nécessité de supposer la fonction d'aire finie, positive et deux fois dérivable. De plus, la propagation doit s'effectuer en ondes planes. Il est à noter que ces hypothèses sont, au départ, posées sur des fonctions continues.

Le problème va donc se ramener à appliquer aux coefficients de réflexion r_n et aux aires A(nX) des contraintes dérivées des conditions imposées à A(x), e(t), s(t) et h(t). Lors de la discrétisation des fonctions temporelles, la simple application du théorème d'échantillonnage permet le passage du système continu au système discret. Par contre, une loi similaire reliant la fonction continue A(x) à la séquence discrète A(nX) n'est pas connue. Une démarche indirecte s'avère donc nécessaire lorsque l'on veut relier A(x) à A(nX). Elle consiste à s'assurer que les lois de reconstitution, utilisées pour A(x), respectent bien les contraintes imposées aux fonctions temporelles.

Satisfaire aux exigences de largeur de bande limitée et aux hypothèses faites sur A(x) se ramène, lorsque l'on passe du système continu au modèle numérique, à effectuer une opération de lissage des A(nX) autour de la variation de section détectée, ou, en d'autres mots, à répartir les coefficients de réflexion. Dans une précédente communication (LEFEVRE et al. 1981) nous avons défini une loi de lissage, relativement complexe, qui respecte les hypothèses de départ tout en modifiant au minimum la variation de section détectée. Nous nous contentons ici d'en rappeler la forme finale. A toute variation de section détectée, est associée une courbe de transition composée de deux branches se raccordant au niveau de la variation détectée. Chaque branche est définie par une fonction de la forme:

$$T_j(x) = A_t \left[k_j^{-2} + \frac{1-k_j^{-2}}{k_j \tanh(k_j bx)+1} \right]^2 \quad (5)$$

où A_t est l'aire correspondant au point de raccordement, elle-même fonction de l'importance de la constriction.

b est une constante faisant intervenir les grandeurs physiques du conduit vocal et la largeur de bande.

k_j est un coefficient fonction du rapport entre A_t et l'aire à atteindre asymptotiquement pour la branche considérée.

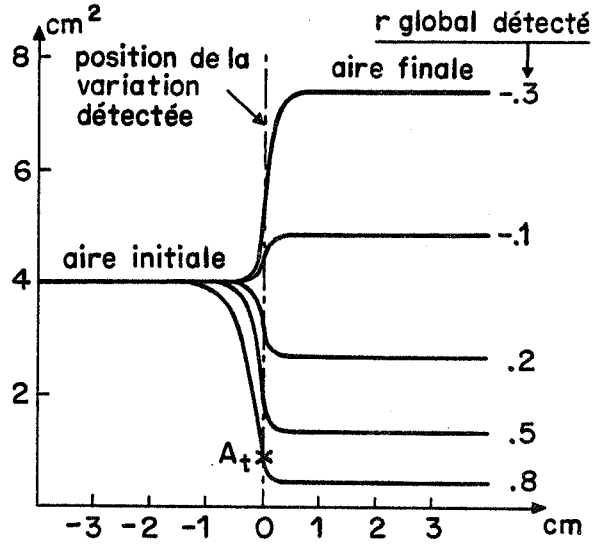


Fig. 4 - Allure de transitions pour différents coefficients de réflexion détectés

- *Shapes of the transition profile for different values of the detected reflection coefficient.*

L'importance d'une constriction détectée est numériquement représentée par un nombre que nous désignerons sous le nom de coefficient global de réflexion détecté. La figure 4 donne l'allure de quelques transitions pour différentes valeurs de ce coefficient global. Notons qu'à toutes fins pratiques, la transition est confinée à un intervalle de 0,5 à 1 cm centré sur le point A_t.

La distribution des coefficients de réflexion, ainsi réalisée, nous a permis de respecter toutes les contraintes imposées au départ. De plus, cette approche réduit les inconvénients causés par le fait qu'une variation de section de la fonction d'aire continue n'est qu'accidentellement localisée à la jonction de deux sections élémentaires du modèle. Dans le domaine temporel, cette remarque correspond au fait que les maxima des fonctions échantillonnées peuvent être légèrement décalés par rapport aux maxima des signaux continus. En ce sens, l'utilisation de la loi de lissage conduit à une convergence plus rapide des calculs.

II.4 Traitement numérique

La période d'échantillonnage des ondes de pression incidente et réfléchie est de 32 µs. Cette valeur correspond à une longueur de section élémentaire égale à 0,55 cm. Ainsi le conduit vocal requiert quelque 30 à 35 sections pour être complètement modélisé. Un tel taux d'échantillonnage peut paraître élevé, spécialement en regard de la bande passante du signal mesuré. Cependant, on en comprend mieux la nécessité lorsque l'on se rappelle que l'objectif poursuivi est l'obtention de la fonction d'aire réelle, ce qui implique une localisation précise des paramètres physiologiques (lèvres, glotte, ...) et des différentes constrictions. De plus, pour que la technique de lissage précédemment exposée remplisse adéquatement son rôle, il est indispensable que la courbe de transition couvre plusieurs sections élémentaires. Utiliser une longueur élémentaire de 0,55 cm revient à dire que le coefficient global de réflexion détecté a une influence significative lors de la répartition sur deux coefficients de part et d'autre, la courbe de transition étant ainsi définie par cinq points.

Au cours du processus itératif, la vraisemblance physique des aires obtenues est testée dans le but d'ajuster automatiquement la longueur totale du conduit et ainsi d'accélérer les calculs.

L'arrêt du processus itératif est l'un des problèmes les plus délicats à solutionner avec ce genre de méthode: une décision prématurée conduit à une définition incomplète du conduit vocal, alors qu'un trop grand nombre d'itérations fait prendre beaucoup d'importance au bruit contenu dans les échantillons et à la précision des calculs numériques. A la limite, ces deux causes d'erreur ont tendance à provoquer des oscillations dans les profils de fonction d'aire. Aussi avons-nous défini quelques critères pour améliorer la décision d'arrêt du processus.

La technique la plus simple consiste à continuer la recherche de nouvelles variations de section jusqu'à ce qu'un minimum de l'erreur quadratique moyenne soit atteint. A chaque pas, cette erreur est obtenue en utilisant les échantillons de la réponse observée et ceux de la dernière réponse calculée. Toutefois ce critère global est insuffisant. On conçoit en effet facilement que la prise en compte d'une nouvelle variation de section, va améliorer la concordance entre les séries de réponses calculées et mesurées dans le voisinage correspondant à cette modification; mais qu'il est par contre possible qu'une détérioration apparaisse au niveau d'échantillons plus lointains. Or, cette détérioration n'est que temporaire puisqu'elle

sera automatiquement reprise par les itérations subséquentes. Aussi l'itération ne peut vraiment être considérée inutile que si elle détériore la convergence même dans le voisinage de la modification qu'elle introduit (critère local). De plus, l'algorithme utilisé évalue la sensibilité de la fonction d'aire à l'itération en cours, et finalement détecte l'éventuelle apparition des oscillations parasites auxquelles nous avons fait précédemment allusion.

Ces quatre critères convenablement pondérés, adjoints aux vérifications relatives aux valeurs minimale et maximale vraisemblables pour la fonction d'aire, permettent d'obtenir une solution après un nombre d'itérations variant en général entre quinze et vingt-cinq.

Le choix d'un coefficient d'atténuation indépendant de la fréquence, bien que non valide en théorie, a donné des résultats très acceptables. Il est toutefois important de choisir adéquatement le paramètre empirique suivant la nature de la mesure effectuée: conduit vocal humain ou tubes de différents matériaux plus ou moins réfléchissants.

L'évaluation des résultats obtenus est rendue difficile de par la nature même de la méthode. Rappelons en effet qu'elle est basée sur la simulation sans phonation d'une articulation. Plusieurs possibilités de vérification nous sont néanmoins offertes. Tout d'abord, l'utilisation d'un vibrateur externe (appareil utilisé par les laryngectomisés) permet au sujet de s'exercer à simuler correctement la transition recherchée. Les résultats peuvent être comparés à d'autres mesures prises en parallèle; cinéradiographie, coupes au scanner, etc... La transposition, en terme de fonction d'aire, des coupes sagittales ainsi obtenues, présente toutefois certaines difficultés. Une autre méthode de vérification indirecte, celle retenue pour nos expériences consiste à comparer un son synthétique issu de la simulation et le son naturel correspondant réalisé par le même sujet dans les mêmes conditions. La synthèse des sons simulés est obtenue à l'aide d'un modèle utilisant les fonctions d'aire calculées.

III. SYSTEME DE SYNTHESE

Le synthétiseur développé, fondé sur un modèle physiologique des cordes et du conduit vocal, utilise les fonctions d'aire préalablement mesurées comme paramètres de commande. Comme nous l'avons mentionné précédemment, nous nous contenterons ici de mettre l'accent sur les points originaux de notre réalisation.

Le système de synthèse, de même que le système de mesure décrit dans la première partie de cette communication, ont été implantés sur un calculateur hybride EAI-690. La calculatrice analogique comprend une trentaine d'intégrateurs, tandis que la calculatrice numérique offre une capacité de mémoire de 32 K mots de 16 bits. L'interface entre les parties analogique et numérique de la machine comprend en particulier des convertisseurs A/D et D/A, dont la liaison avec le calculateur numérique peut être réalisé via un canal d'accès direct à la mémoire. Ces convertisseurs ont une capacité de 13 bits plus signe et un taux de conversion maximum dans le sens analogue vers digital de 33 kHz.

Sur ce système, nous avons implanté un modèle analogique de source vocale et un modèle numérique de conduit. Pour le conduit, une réalisation analogique aurait nécessité un nombre excessif de composants et aurait manqué de souplesse pour d'éventuelles modifications. Cependant, il nous faut

noter ici qu'un tel choix élimine pratiquement la possibilité de synthèse en temps réel. Toutefois, les valeurs numériques calculées et stockées sur une mémoire de masse pourront facilement être "entendues" en temps réel, en utilisant le canal d'accès direct mémoire.

III.1 Modélisation de la source vocale

A la lumière des travaux de plusieurs auteurs, nous avons opté pour le modèle des cordes vocales issu de la structure à deux masses proposés par ISHIZAKA et FLANAGAN (1972). La simulation analogique réalisées'apparente à celle retenue par CARCAUD (1975).

Les paramètres fondamentaux de cette source sont la pression sous-glottique et la tension des cordes vocales, elle-même proportionnelle à la raideur des cordes et inversément proportionnelle à leur masse. Dans la simulation, ces paramètres sont introduits par l'intermédiaire de générateurs de fonction à diode, commandés numériquement. Ils peuvent ainsi être très commodément modifiés. L'implantation sous forme analogique du modèle proposé a nécessité un traitement efficace des phénomènes de saturation et d'indétermination (essentiellement au niveau des divisions par zéro).

III.2 Simulation du conduit vocal

La modélisation du conduit vocal, sous forme d'une succession de tubes élémentaires, telle que décrite dans un paragraphe précédent, est de nouveau utilisée dans notre ensemble de synthèse. Afin d'utiliser au maximum les informations obtenues sur la fonction d'aire, dans le modèle implanté, toutes les sections élémentaires sont de même longueur, longueur correspondant à une demi-période d'échantillonnage. De plus, cette disposition nous permet de mieux positionner le raccordement avec la cavité nasale et nous permet également d'éliminer les problèmes liés aux sections terminales de longueur différente, que l'on rencontre dans les schémas utilisant une longueur élémentaire correspondant à une période complète d'échantillonnage. En contrepartie, la solution retenue demande un volume de calculs plus important et nous oblige à enchevêtrer le calcul des ondes de pression incidentes et réfléchies p' et p, ce qui toutefois ne présente pas de difficulté particulière dans une simulation numérique.

Une interaction entre la source et le conduit s'établit au niveau de la glotte. En effet, la source délivre un débit U_g à la glotte, tandis que la charge créée par la connexion du conduit produit une pression P_g, pression qui va influencer le comportement de la source. Les deux paramètres U_g et P_g sont reliés par les relations

$$U_g = \frac{1}{Z_g}\,(p_0^+ - p_0'^-),\quad P_g = p_0^+ + p_0'^- \qquad (6)$$

avec $Z_g = \frac{\rho c}{A_0}$

L'onde de pression réfléchie $p_0'^-$ étant fournie par le conduit, il est alors facile d'obtenir l'onde de pression incidente p_0^+. L'interface entre la source analogique et le conduit numérique est situé à ce niveau, $p_0'^-$ et p_0^+ étant échangés respectivement par un convertisseur D/A et par un convertisseur A/D.

A l'autre extrémité du conduit vocal, l'effet de rayonnement aux lèvres est incorporé au modèle numérique. Avec les hypothèses préalablement mentionnées, l'aire de rayonnement peut être assimilée à une surface en vibration dans un baffle sphérique (MORSE, 1948). Une étude systématique des différentes approximations utilisables pour modéliser une telle charge, dans le domaine de validité du système, nous a conduit à adopter un modèle simple dans lequel la charge correspond au rayonnement d'une source sphérique, et est indépendante de l'aire aux lèvres. Cette aire intervient toutefois au niveau de la reconstitution du signal de parole. En fait, ce signal qui représente une pression recueillie à une certaine distance des lèvres, est proportionnel à la dérivée du débit aux lèvres.

Lors de la synthèse de transitions, le modèle doit prendre en compte les mouvements tant longitudinaux que transversaux du conduit vocal, permettant le passage d'une configuration à une autre. Afin de minimiser les risques de perturbation brutale au niveau des ondes de pression, les modifications de fonction d'aire sont introduites progressivement dans le modèle lorsque l'énergie dans le conduit est minimale. Cet état est atteint immédiatement avant le début de l'impulsion acoustique, instant facile à détecter sur le modèle analogique de source vocale.

III.3 Contribution de la cavité nasale

La cavité nasale a des dimensions fixes dans le temps, sauf au niveau de la cavité rhino-pharynx, influencée par la position du voile du palais. Pour les fréquences envisagées, les deux canaux parallèles et identiques du nez peuvent être considérés comme un tube acoustique unique. En pratique et d'après les dimensions proposées par House et Stevens, nous utilisons des aires variant progressivement sur une longueur de 3,5 cm, le paramètre de commande représentant le rapport d'énergie devant traverser le conduit nasal et le conduit vocal.

Compte tenu du facteur de forme, il est généralement admis que l'amortissement dans les cavités nasales est plus important que dans le conduit buccal (FANT, 1960). Pour tenir compte de cette réalité, nous avons simplement donné au paramètre empirique, utilisé dans le calcul du coefficient d'atténuation γ, une valeur un peu plus élevée.

Le couplage acoustique entre les deux conduits est supposé s'effectuer uniquement au niveau du voile du palais: la possibilité de couplage à travers les parois ainsi que les interactions éventuelles entre les extrémités nasales et buccales ayant été négligées. Comme dans le cas d'une discontinuité de section simple, l'application des équations de continuité de débit et de pression a permis d'établir les relations liant les six ondes de pression et les trois aires apparaissant au niveau du voile du palais.

La charge au niveau du nez a été étudiée de façon identique à celle aux lèvres, et le même modèle simple a été retenu. En pratique, le rayonnement au niveau du nez n'est pas pris en compte lors de la constitution du signal de parole synthétique, car il n'est pas aisé de combiner simplement l'influence des deux sources de rayonnement non coplanaires (lèvres et nez). Toutefois, dans la plupart des cas, l'influence du conduit nasal semble beaucoup plus significative de par les cavités supplémentaires ainsi introduites dans le modèle, que de par le rayonnement nasal.

Le modèle d'appareil phonatoire qui vient d'être décrit va être utilisé en particulier pour vérifier la qualité des fonctions d'aire préalablement mesurées.

IV. RESULTATS ET PERFORMANCES

A la suite d'un grand nombre de mesures de fonction d'aire, mesures effectuées d'abord sur des moulages en plastique mou et en plâtre (DESCOUT et al., 1976), puis par la suite sur des conduits réels stationnaires et en dernier lieu sur des simulations de transitions, nous avons pu dégager un certain nombre de remarques:

- un choix raisonnable du coefficient d'atténuation, indépendant de la fréquence, n'influence pas de façon critique le résultat;

- la fonction de transfert du tube de mesure peut être assimilée à une constante, à tout le moins pour les fréquences considérées ici;

- la rigidité du couplage entre la source et le tube peut soit générer des vibrations mécaniques, soit amplifier certains modes vibratoires dans le tube de mesure;

- lors des mesures sur un conduit humain, de nombreux facteurs liés à l'expérimentateur peuvent être la cause d'erreurs non négligeables: relâchement de la tension musculaire dans le conduit vocal, simulation incorrecte de l'articulation souhaitée, pressurisation du tube, fuite au niveau des lèvres, mauvais contrôle de la fermeture de la glotte. Toutefois, après apprentissage, le sujet est capable de contrôler d'une manière satisfaisante ces différents facteurs (TOUSIGNANT et al., 1979);

- le système de mesure n'est guère adapté à l'étude des configurations correspondant soit à des sons nasalisés, soit à des sons caractérisés par une occlusion labiale importante (en pratique inférieure à 0,7 cm^2), soit à des articulations nécessitant des mouvements importants au niveau des lèvres (l'étanchéité entre les lèvres et l'embout étant alors difficile à maintenir). De plus, il est évidemment impossible d'obtenir quelqu'information que ce soit sur toute portion de conduit située en arrière d'une occlusion totale.

La figure 5 fournit un exemple de résultats obtenus avec des configurations simulant les voyelles */a/, /e/, /i/, /o/*. Les tracés continus correspondent à nos mesures de la fonction d'aire, alors que les pointillés représentent les fonctions d'aire "normalisées" obtenues pour les mêmes voyelles françaises par MRAYATI (1976) dans une étude s'appuyant sur les fréquences et les bandes passantes des formants et sur certaines contraintes articulatoires. La concordance entre ces deux séries de courbes peut être qualifiée de bonne, ce que confirme la synthèse des sons correspondants.

Tel que nous l'avions anticipé dans la première partie de cet exposé, les expériences ont confirmé que la précision des résultats décroit lorsque l'on se rapproche de la glotte. Toutefois, la glotte est surtout importante de par sa position, puisqu'elle conditionne la longueur du conduit vocal, or la perte de précision n'est pas d'une importance telle qu'elle puisse masquer cette position. La forme de la fonction d'aire au voisinage de la glotte est quant à elle pratiquement constante, et éventuellement une forme standard peut être substituée aux mesures lors de la synthèse.

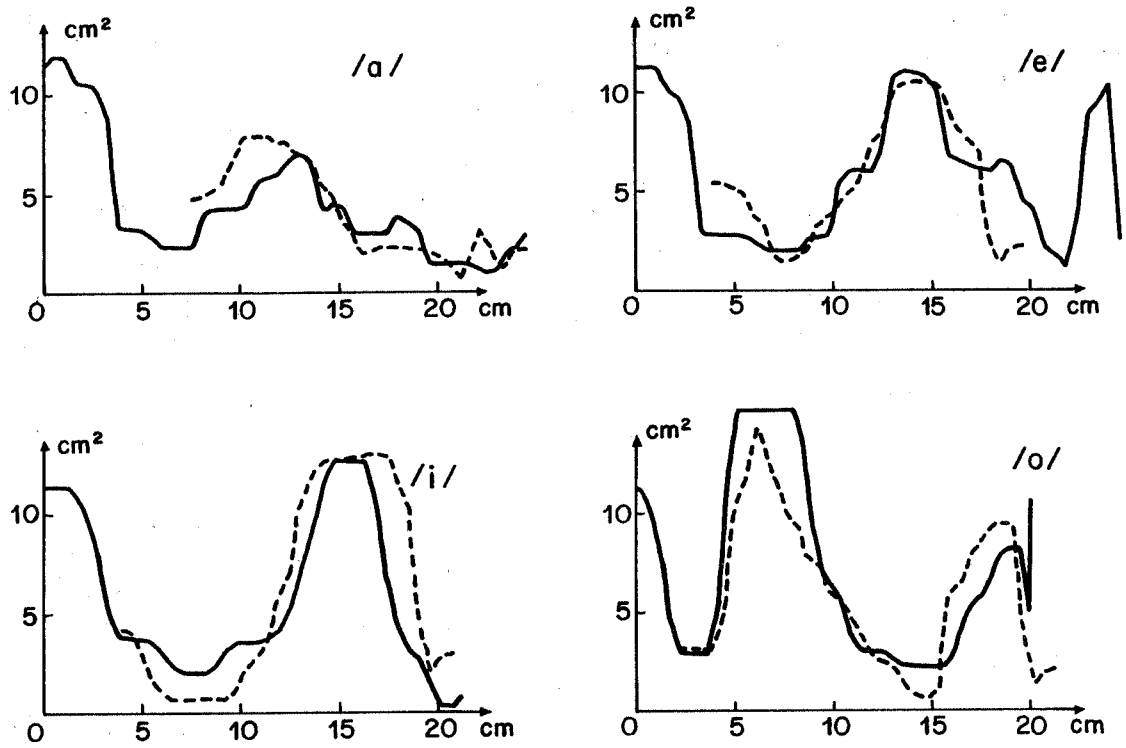


Fig. 5 - Fonctions d'aire de voyelles françaises
———— : embout et conduit vocal
-------- : estimation de l'aire selon Mrayati

- *Static french vowels area functions*
———— : *vocal tract and connexion with measuring system*
-------- : *Mrayati's area estimation*

La série de mesures de fonction d'aire du conduit vocal, représentée sur la figure 6, correspond à une simulation de l'articulation */oej/*. Sur cette figure, les intervalles entre chaque trame sont de quarante millisecondes, les chiffres entre parenthèses représentent le nombre d'itérations nécessaires au processus de reconstitution pour obtenir une solution, et les lettres L et G correspondent aux positions estimées des lèvres et de la glotte.

La position des lèvres est facilement déterminée en raison du changement marqué de pente apparaissant au niveau du raccordement de la bouche au tube de mesure. La variation de l'ouverture des lèvres est clairement mise en évidence.

Le niveau de bruit, la précision finie des calculs et les imperfections du modèle se traduisent par l'addition d'un conduit résiduel en arrière de la glotte. En général, la glotte peut néanmoins être facilement localisée en inspectant la forme obtenue, mais dans certains cas, il peut s'avérer nécessaire d'opérer par continuité d'une trame à l'autre. Les longueurs du conduit vocal obtenues sur la figure 6 varient entre 17 et 18,5 cm.

Sur la figure 6, la fonction d'aire évolue normalement de la trame 1 à la trame 16, mais tous les résultats ne sont pas aussi satisfaisants. Ainsi, dans la dernière colonne de cette figure, la continuité d'une trame à l'au-

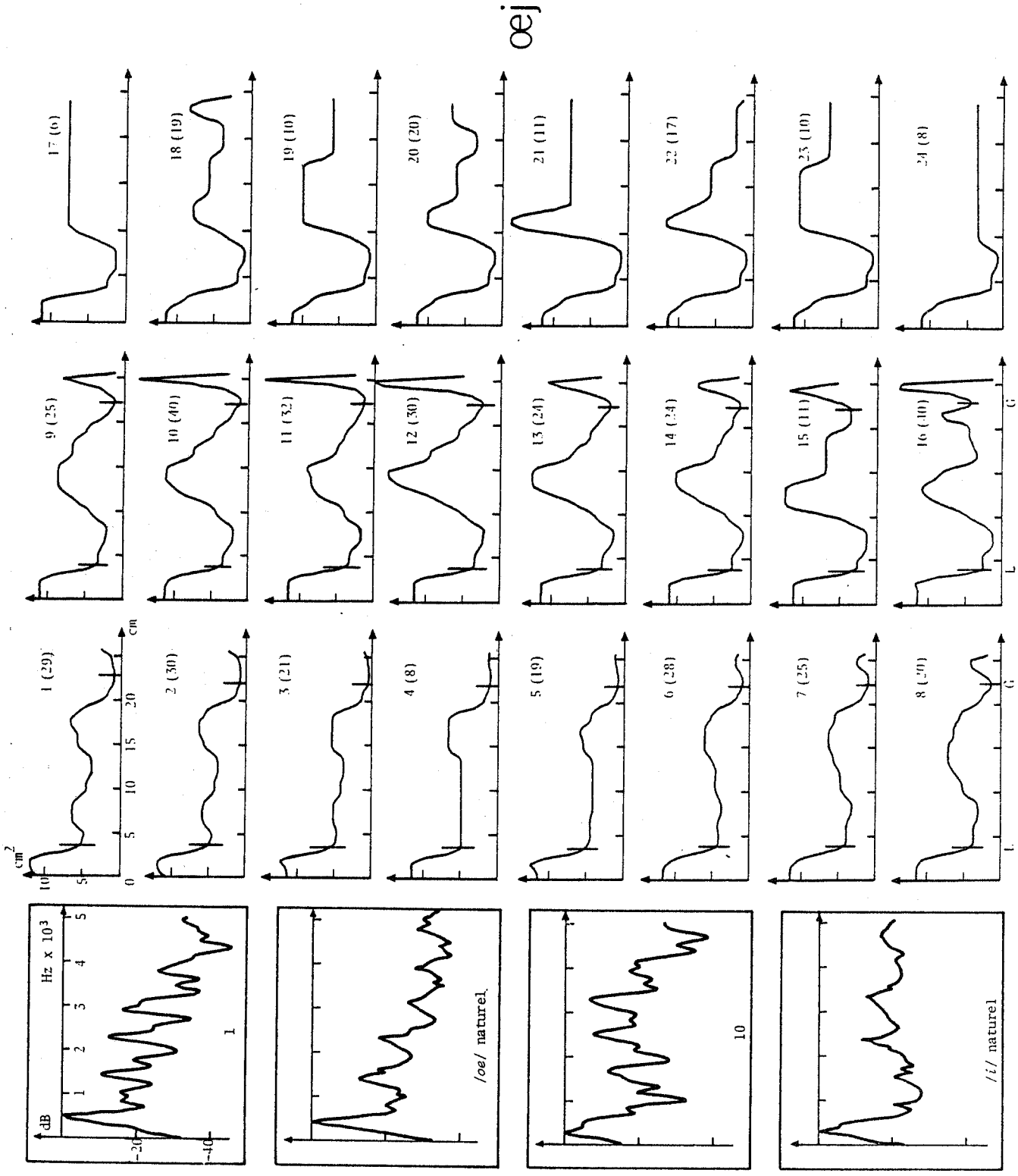


Fig. 6 - Articulation */oej/*: fonctions d'aire et spectres d'amplitude associés

- *Area functions and corresponding amplitude spectra*

tre a complètement disparu. Nous supposons que ces aberrations sont dues à une pressurisation du tube. Il est intéressant de noter que nous avons toujours observé soit une variation continue soit de très importantes discontinuités d'une trame à l'autre, ce qui facilite le rejet éventuel.

La qualité des fonctions d'aire obtenues ne peut être vraiment jugée que par comparaison avec la fonction d'aire réelle du sujet soumis à l'expérience. Comme nous l'avons déjà mentionné, une telle comparaison n'est pas chose aisée, aussi passons-nous par l'intermédiaire du son synthétisé à partir de la fonction mesurée. L'écoute du résultat nous fournit une première vérification essentiellement qualitative et relativement subjective. Il est également possible de comparer les contenus spectraux des sons synthétique et naturel. Ainsi sur la figure 6, les spectres des trames 1 et 10 concordent relativement bien avec ceux des sons */oe/* et */i/* prononcés par le même sujet, du moins en ce qui a trait aux positions et aux amplitudes relatives des formants. Le modèle de production de la parole ne prenant pas en compte les vibrations des parois du conduit vocal, il est normal que les bandes passantes des formants, en particulier celle du premier, aient tendance à être sous-estimées.

V. CONCLUSION

Une méthode d'évaluation de la fonction d'aire du conduit vocal basée sur des mesures acoustiques effectuées au niveau des lèvres a été developpée. Les contraintes propres à cette méthode ont fait l'objet d'un traitement détaillé. Les conditions d'application, de même que les limitations intrinsèques ont été analysées. Les séquences de fonction d'aire obtenues ainsi, directement en valeurs réelles, ont été ultérieurement utilisées comme paramètres de commande d'un synthétiseur, ce qui a permis d'en vérifier le réalisme.

Suite à ces résultats encourageants, d'autres développements font actuellement l'objet d'études plus approfondies. En particulier, l'utilisation de méthodes de prédiction linéaire aux fins de quantification de la qualité des résultats a été envisagée (LEFEVRE et al., 1980). D'autres points, tels la génération automatique de bruits, l'utilisation d'un modèle du conduit vocal à sections elliptiques ou l'établissement de relations entre fonctions d'aire et modèles articulatoires, n'ont encore été qu'explorés. Pour ces études à caractère phonétique, cette méthode d'évaluation de la fonction d'aire peut apporter une contribution intéressante.

REFERENCES

CARCAUD, M., 1975, "Contribution à la synthèse de la parole. Simulation de la source vocale", Thèse de doctorat 3e cycle, Univ. Paris-Sud, Orsay.

DESCOUT, R., TOUSIGNANT, B., LECOURS, M., 1976, "Vocal tract area function measurements: two time-domain methods", *Proc. IEEE ICASSP*, Philadelphie, pp. 75-78.

FANT, G., 1960, "Acoustic theory of speech production", 'S-Gravenhage: Mouton and Co.

ISHIZAKA, K., FLANAGAN, J.L., 1972, "Synthesis of voiced sounds from a two-mass model of the vocal cords", *Bell Syst. Tech. J.*, vol. 51, pp. 1233-1268.

KELLY, J.L. Jr, LOCHBAUM, C., 1962, "Speech synthesis", *Proc. Stockholm Speech Communications Seminar*, R.I.T., Stockholm.

LEFEVRE, J.P., SOUMAGNE, J.C., TOUSIGNANT, B., LECOURS, M., 1980, "Mesures de la fonction d'aire du conduit vocal et estimation indirecte de la qualité des résultats", *Actes de la conférence IEEE sur les communications et l'énergie*, Montréal, pp. 317-320.

LEFEVRE, J.P., TOUSIGNANT, B., LECOURS, M., 1981, "Smoothing vocal-tract area functions obtained by acoustical measurements", à paraître *IEEE Trans. Acoust., Speech, Signal Process.*, (Voir également XIèmes J.E.P., GALF, Strasbourg, mai 1980).

MORSE, P.M., 1948, "Vibration and sound", New-York: McGraw-Hill, p. 325.

MRAYATI, M., 1976, "Contribution aux études sur la production de la parole", Thèse de doctorat d'état, Univ. Grenoble, pp. 25-27.

SONDHI, M.M., GOPINATH, B., 1971, "Determination of vocal-tract shape from impulse response at the lips", *J. Acoust. Soc. Amer.*, vol. 49, pp. 1867-1873.

STEVENS, K.N., KASOWSKI, S., FANT, G., 1953, "An electrical analog of the vocal tract", *J. Acoust. Soc. Amer.*, vol. 25, pp. 734-742.

TOUSIGNANT, B., LEFEVRE, J.P., LECOURS, M., "Speech synthesis from vocal tract area function acoustical measurements", *Proc. IEEE ICASSP*, Washington D.C., pp. 921-924.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

REPRESENTATION DE LA PAROLE ENTRE 4 ET 8 KBITS/SEC UTILISANT UN DICTIONNAIRE DE FORMES D'ONDE

J.P. ADOUL et PH. MABILLEAU

CENTRE DE RECHERCHES SUR LES COMMUNICATIONS
UNIVERSITE DE SHERBROOKE
SHERBROOKE, QUEBEC, (CANADA)

.RESUME

Cet article présente une nouvelle méthode de représentation numérique de la parole entre 4 et 8 Kbits/sec. L'idée de base est l'utilisation d'un dictionnaire de K formes d'onde de durée fixe T (1 à 4 ms). Le signal de parole à coder est divisé en blocs de durée T, chacun de ces blocs est comparé aux formes d'onde du dictionnaire et l'indice de celle qui lui ressemble le plus va être utilisé comme représentation numérique du signal. Un tel procédé demande une définition de la notion de ressemblance entre formes d'onde. C'est à ce niveau qu'interviennent les propriétés du signal de parole.

Notre étude a porté sur la définition de cette distance et sur les moyens utilisables pour construire le dictionnaire de formes d'onde.

1.0 INTRODUCTION

Le codage de la parole à bas et à moyen débit est pertinent pour le multiplexage de celle-ci avec des données ou pour le chiffrage. En effet, il est possible de transmettre sur le réseau téléphonique des signaux de modems jusqu'à un débit de 9600 bits par seconde. Une représentation de la parole avec moins de 8000 bits par seconde va donc permettre un multiplexage temporel de celle-ci avec d'autres signaux numériques (terminaux, tablettes à digitaliser, télécopieurs ext...).

Avec un nombre de bits aussi faible les méthodes de codage par quantification directe des échantillons de parole donnent de mauvais résultats (FLANAGAN,1979). Il est alors nécessaire d'utiliser la corrélation qui existe dans le signal de parole en effectuant un encodage par blocs. C'est ce qui est réalisé par la prédiction linéaire qui consiste à extraire la partie prédictible du signal sous forme d'un filtre linéaire donnant un estimé de l'échantillon courrant en fonction des M échantillons précèdants (MAKHOUL,1975) (ESTEBAN,1978). La différence entre cet éstimé et la vraie valeur du signal constitue le signal résiduel ou l'erreur de prédiction. Cette méthode demande donc un encodage du filtre prédicteur et du signal résiduel.

La méthode de représentation du signal de parole que nous proposons ici permet un codage avec des débits numériques allants de 4 à 8 kbits/sec. Elle s'applique au signal résiduel de la prédiction linéaire mais ayant développé par ailleur un algorithme de codage du filtre prédicteur nous suggèrons aussi un encodage conjoint du résiduel et du filtre. Le principe que nous utilisons est la quantification vectorielle de blocs de signal d'une durée de 1 à 4ms. Ces blocs que nous appelerons formes d'onde sont remplacés par des blocs choisis dans un dictionnaire.

2.0 PRINCIPE DE LA METHODE DE REPRESENTATION

Le principe de la quantification vectorielle par dictionnaire de formes d'onde est exposé ici. Deux exemples de distances utilisables entre les formes d'ondes sont proposées ainsi qu'un algorithme de construction du dictionnaire.

2.1 QUANTIFICATION VECTORIELLE

Cette méthode utilise une représentation du signal par blocs de N (8 à 32) échantillons numériques ce qui correspond à des tranches de parole d'une durée allant de 1 à 4ms. Chaque bloc est codé indépendamment. Le bloc à coder ou forme d'onde est remplacé par celui qui lui ressemble le plus parmi les K formes d'onde d'un dictionnaire. Le signal est alors défini par l'indice du bloc choisi dans le dictionnaire.

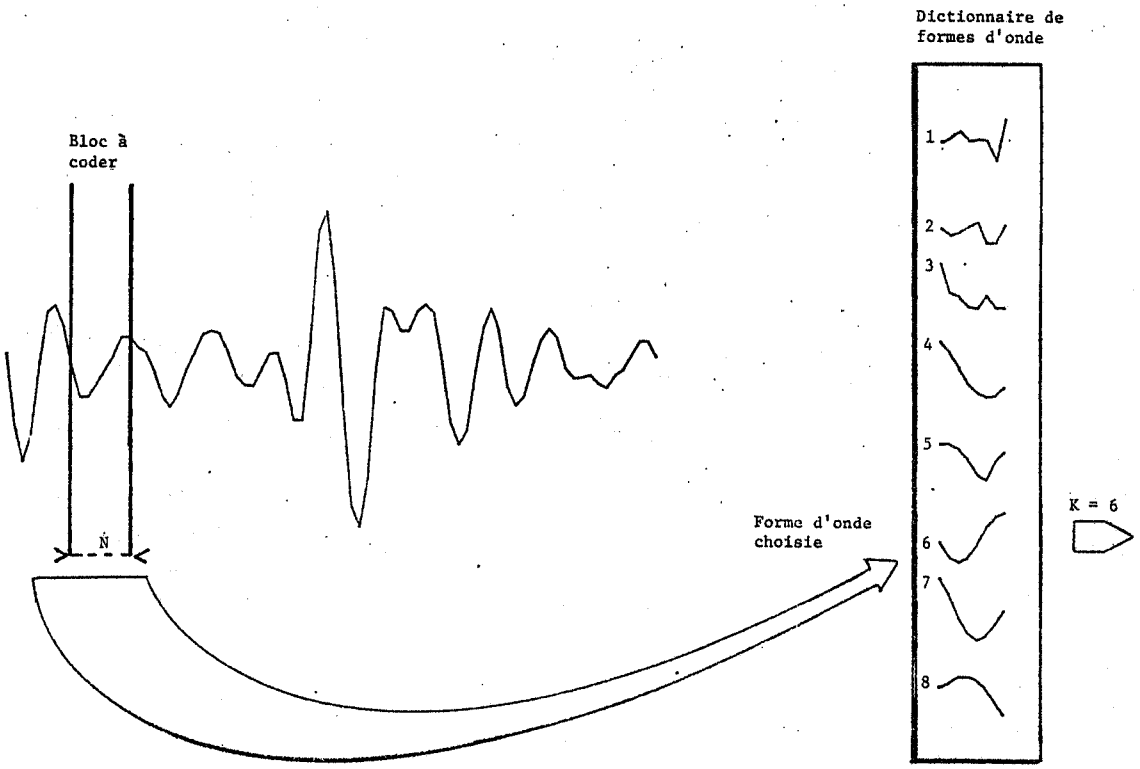


FIGURE 1 Principe de la quantification vectorielle par utilisation d'un dictionnaire de formes d'onde

Cette méthode est qualifiée de quantification vectorielle par opposition à la quantification scalaire qui consiste à remplacer chaque échantillon par un représentant choisi dans une table.

Ceci demande une formulation mathématique de la notion de ressemblance entre deux formes d'onde c'est à dire la définition d'une distance vectorielle dans un espace à N dimensions. C'est cette distance qui va déterminer la qualité de notre codage. Elle peut être basée sur des critères ob-

jectifs de ressemblance comme l'erreur quadratique ou sur des critères résultants d'écoutes subjectives de signaux. Ces derniers peuvent faire intervenir la distribution de l'énergie de la forme d'onde dans le domaine spectral (signal résiduel de sons non voisés) ou la position d'un maximum dans le bloc (signal résiduel de sons voisés).

Il est à noter que le codage nécessite le calcul de K distances alors que le décodage est très simple. Mais si K est une puissance de 2, et en construisant le dictionnaire de façon hiérarchique il est possible de réduire le nombre de distances à calculer pour le codage d'une forme d'onde. Ceci n'est toutefois possible que si l'on admet une augmentation de la mémoire attribuée au dictionnaire.

2.2 EXEMPLES DE DISTANCES ENTRE FORMES D'ONDE

Une transformation du signal permet de définir un nouvel espace de travail où l'on va utiliser une métrique propre aux caractéristiques du signal.

Deux méthodes de traitement ont été étudiées: la division de la bande de fréquence et la prédiction linéaire.

2.2.1 DIVISION DE LA BANDE DE FREQUENCE -

Cette première méthode tient compte des caractéristiques de la perception de la parole, en effet les erreurs qui peuvent être acceptées dans le domaine des basses fréquences ne sont pas les mêmes que celles qui sont acceptées dans le haut de la bande. Les sons dits voisés qui présentent une énergie maximum dans les basses fréquences demandent une bonne transmission du signal dans le domaine temporel alors que les sons non voisés sont relativement bien représentés par la distribution spectrale de leur énergie surtout dans le haut de la bande. En effet ces derniers peuvent être modelisés par une source de bruit blanc attaquant un filtre. Ce sont ces propriétées qui vont être utilisées pour définir la distance entre deux formes d'onde.

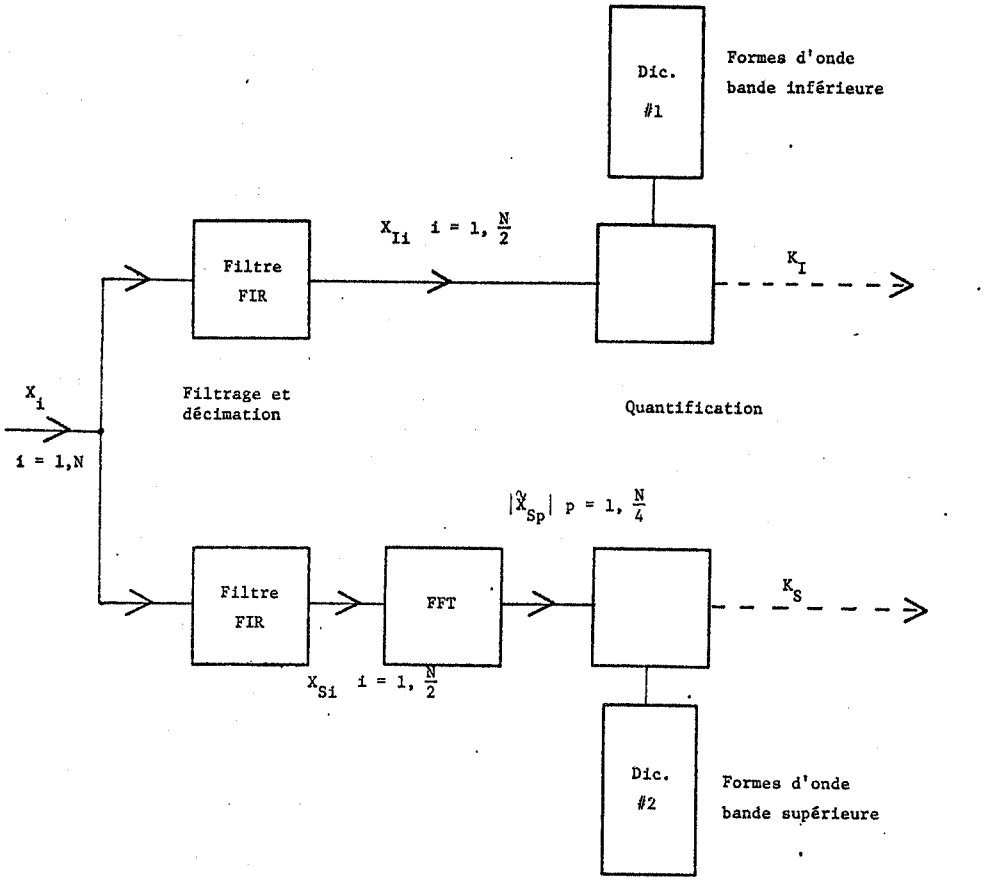


FIGURE 2 Division de la bande de fréquence - codeur

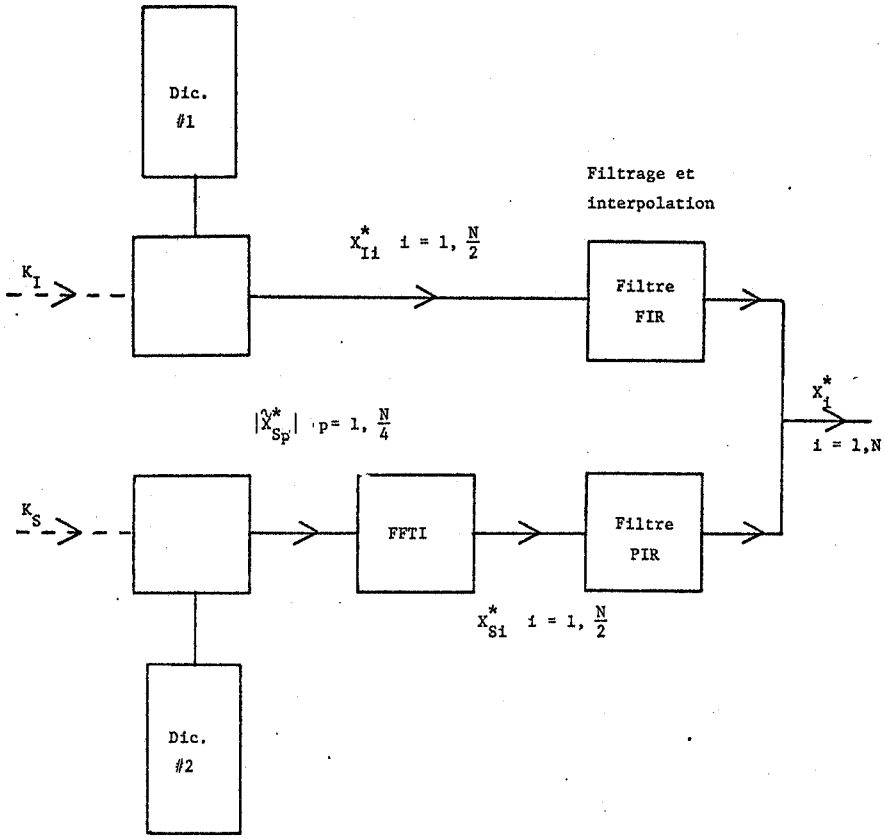


FIGURE 3 Division de la bande de fréquence - décodeur

Partant de ces considérations le signal de parole est divisé en deux bandes à l'aide de deux filtres FIR symétriques d'une fréquence de coupure de 2 KHz. Le signal de la bande inférieure (0-2 KHz) est codé dans le domaine temporel à l'aide d'un premier dictionnaire pendant qu'une transformation de Fourier est appliquée au signal de la bande supérieure (2-4 KHz) de façon à en effectuer le codage avec un dictionnaire de spectres. Ce traitement revient à utiliser une distance spectrale pour la partie haute fréquence de la parole. Etant donné que c'est le module du spectre qui est utilisé l'information relative à la phase est perdue par ce procédé. Une phase aléatoire est utilisée pour reconstituer le signal au récepteur.

2.2.2 UTILISATION DE LA PREDICTION LINEAIRE -

La seconde méthode proposée pour la réalisation de ce type de codage est l'application du codage par dictionnaire de formes d'onde au signal résiduel de la prédiction linéaire.

Le prédicteur utilisé est un filtre en treilli d'ordre 8 à 12. L'encodage des coefficients de prédiction peut de faire d'une façon classique ou en utilisant un algorithme de quantification vectorielle (ADOUL,1980). La détermination du filtre prédicteur ne demande alors que le calcul de l'autocorrélation, les coefficients des filtres étant places dans une table (dictionnaire de filtres) où la recherche est effectuée à l'aide des coefficients d'autocorrélation. Dans ce dernier cas le débit numérique obtenu peut devenir très faible.

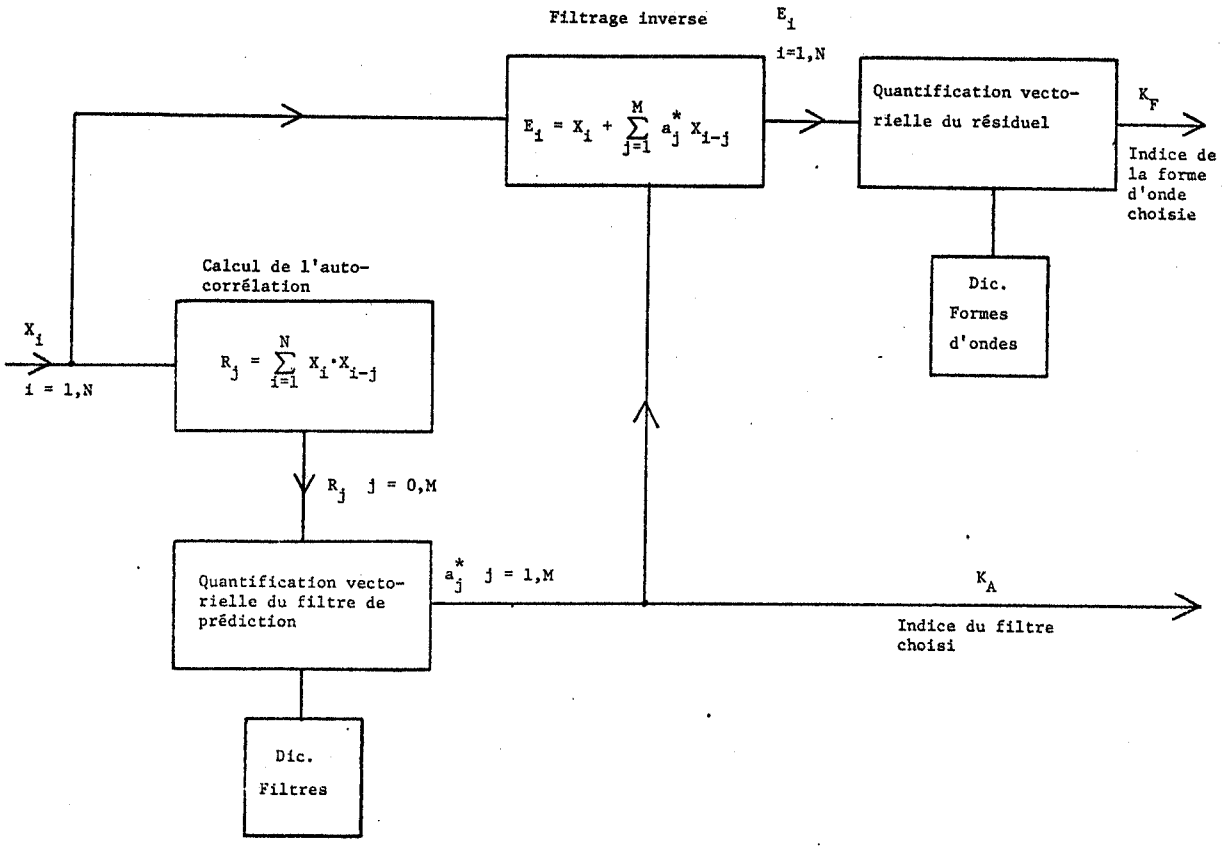


FIGURE 4 Prédiction linéaire - codeur

La distance utilisée dans le domaine du signal résiduel tient compte des propriétées de celui-ci. En effet, la po-

sition d'un maximum dans le bloc à représenter est importante car c'est elle qui va déterminer le fondamental de la parole. Le dictionnaire utilisé va donc comprendre un éventail de formes d'impulsions d'exitation pour les voisés et de bruits pour les non-voisés.

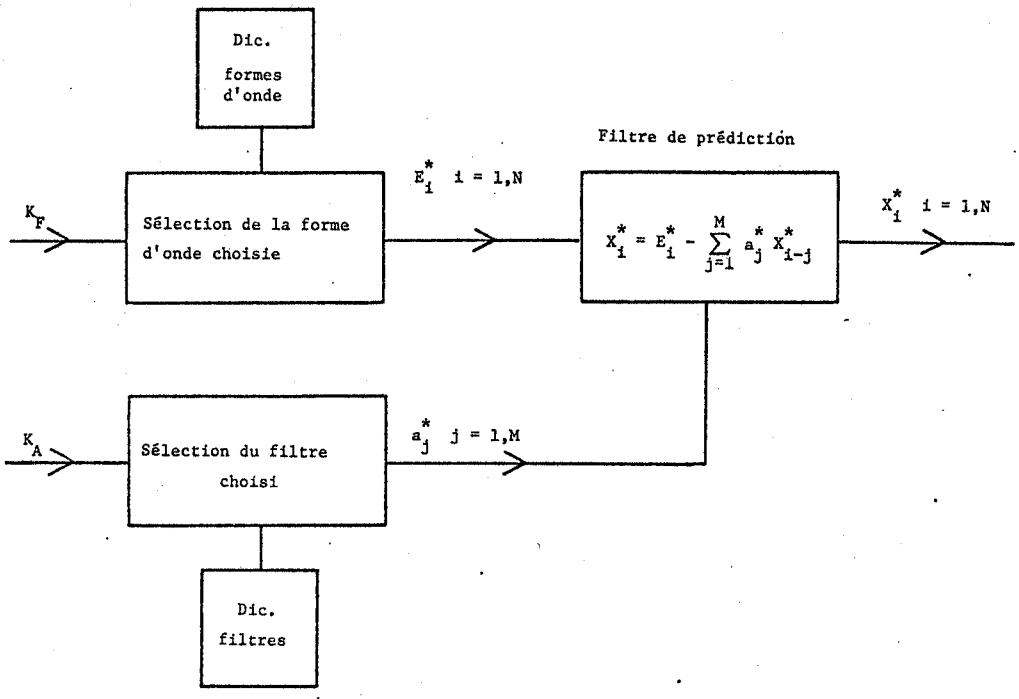


FIGURE 5 Prédiction linèaire - décodeur

2.3 CONSTRUCTION DU DICTIONNAIRE

Le dictionnaire doit être d'une taille suffisante pour permettre une bonne représentation du signal. De plus les formes d'onde qui le constituent doivent offrir un éventail assez large des différents cas possibles.

On voit que pour un débit faible, c'est à dire avec K

petit (inférieur à 256), le choix de ces formes d'onde devient très critique. Nous avons développé des algorithmes permettants de trouver les K meilleures formes d'onde pour représenter un signal donné au sens d'une certaine distance (DIDAY,1976). Il s'agit d'algorithmes itératifs de k-moyenne qui opèrent de la façon suivante:

1. On effectue un tirage aléatoire de K formes d'onde différentes parmis celles dont l'on dispose pour former le dictionnaire. Ces formes d'onde vont constituer les représentants initiaux.

2. L'enssemble de toutes les formes d'onde disponibles est divisé en K classes par association de chaque forme d'onde avec le représentant qui en est le plus proche au sens de la distance utilisée.

3. On calcul K nouveaux représentants en prenant les barycentres des classes.

4. Les opérations 2. et 3. sont recommencées jusqu'à stabilité des représentants.

La distance utilisée intervient au niveau de la détermination des classes et du calcul des barycentres.

3.0 SIMULATION DE CODEURS

Nous avons réalisé des simulations de codages de la parole utilisants la représentation vectorielle par dictionnaire de formes d'onde. Celles-ci ont été effectuées sur un mini-ordinateur PDP11/60 associé avec un processeur de vecteurs MAP200.

3.1 CHAINE DE TRANSFORMATION DU SIGNAL DE PAROLE

La méthode de codage proposée doit être placée dans une chaine de transformation du signal de parole qui peut être la suivante:

1. Filtrage passe bas à 3400Hz et conversion numérique du signal analogique avec une cadence de 8000 échantillons par seconde. Ceux-ci donnent une représentation linéaire de l'onde sonore avec une résolution de 12 bits.

2. Transformation du signal numérique par une méthode de prédiction linéaire donnant pour chaque bloc analysé les coefficients du filtre prédicteur et un signal résiduel auquel nous allons appliquer la quantification vectorielle.

3. Normalisation de l'amplitude du signal résiduel.

4. Quantification vectorielle par dictionnaire de formes d'onde du signal résiduel.

5. Transmission via le cannal de communication ou un organe de stockage des informations suivantes:

   1. coefficients du filtre prédicteur

   2. Amplitude du signal résiduel

   3. Indice de la forme d'onde choisie dans le dictionnaire

6. Reconstitution du signal résiduel à l'aide du dictionnaire.

7. Dénormalisation du signal résiduel.

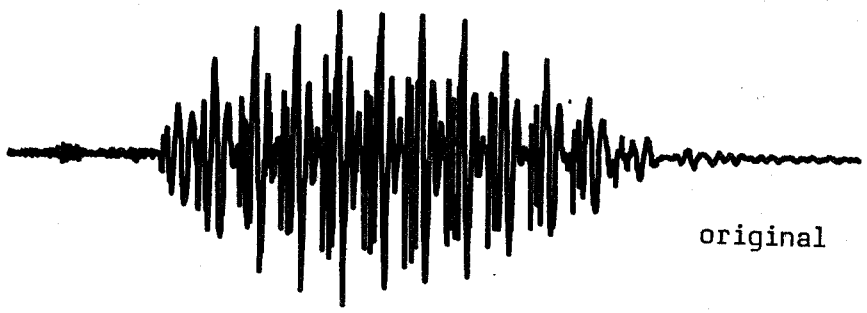
8. Filtrage du signal résiduel à l'aide du filtre prédicteur.

9. Conversion analogique du signal obtenu.

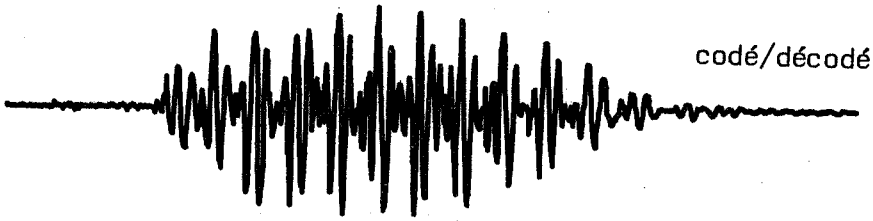
3.2 RESULTATS DES SIMULATIONS

Par le principe de la simulation il est très facile de faire varier les paramètres du codeur pour en étudier les performances. Le débit numérique résultant dépend directement de la taille du dictionnaire utilisé, d'où une étude possible des résultats de cette méthode à différents débits.

Nous proposons ici quelques représentations de parole originale d'une part et codée/décodée d'autre part pour les deux distances proposées. La simulation permet l'écoute de la parole obtenue après codage et décodage.



Division de la bande de fréquence (7 kbps)



Utilisation de la prédiction linéaire (3.5 kbps pour le résiduel)

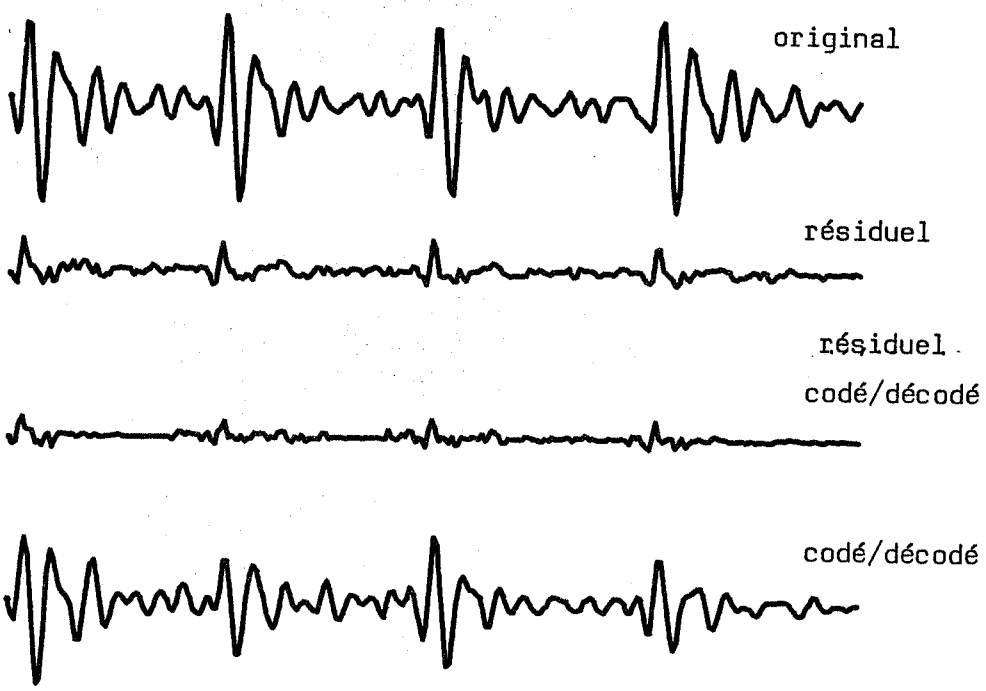


FIGURE 6 Exemples de signaux obtenus

REFERENCES

1. ADOUL (J.P.), DEBRAY (J.L.), DALLE (D.), 1980, SPECTRAL DISTANCE MEASURE APPLIED TO THE OPTIMUM DISIGN OF DPCM CODERS WITH L PREDICTORS, ICASSP, Danver(USA), april 1980, pp. 512-515

2. DIDAY (E.), SIMON (J.C.), 1976, CLUSTERING ANALYSIS IN DIGITAL PATTERN RECOGNITION, K.S. FU editor Springer-Verlag 1976, pp. 47-94

3. ESTEBAN (D.), GALAND (C.), MAUDUIT (D.), MENEZ (J.), 1978, 9.6/7.2 KBPS VOICE EXITED PREDICTIVE CODER (VEPC), ICASSP, Tulsa(USA), april 1978, pp. 307-311

4. FLANAGAN (James 1.), 1979, SPEECH CODING, IEEE transactions on communications vol com 27 no 4 april 1979, pp. 710-736

5. MAKHOUL (John), 1975, LINEAR PREDICTION: A TUTORIAL REVIEW, proceedings of the IEEE vol 63 no 4 april 1975 pp. 561-580

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

Fonctions morphologiques des paramètres suprasegmentaux en français québécois

Laurent Santerre
Département de linguistique
Université de Montréal

Jean-Louis Chandon
Département des sciences administratives
Université du Québec à Montréal

.RESUME

Dans le langage populaire, on remarque que trois formes temporelles du verbe être, employé comme auxiliaire ou au sens de "aller", peuvent être réduites à la même suite segmentale [teta]; ainsi [tetapo] signifie : 1. T'es à Pau - 2. T'étais à Pau - 3. T'as été à Pau. L'analyse acoustique montre que la production des montréalais varie en durée, en intonation et en intensité selon qu'ils se proposent de faire entendre un sens ou un autre. Dans le but de savoir si la variation de l'intonation ou de la durée prise isolément pouvait suffire à distinguer les trois "temps", j'ai synthétisé, au moyen des programmes de Klatt à M.I.T., trois énoncés /tetapo/ où seule varie l'intonation ou la durée. Des tests de compréhension ont été faits auprès de trois groupes d'auditeurs montréalais; les premiers, 24 étudiants placés dans les conditions d'écoute bruyantes, ont montré qu'ils reconnaissaient les trois phrases au moyen de l'intonation, avec une probabilité qui se situe entre 90% et 95%; un 2e groupe de 40 autres étudiants, cette fois dans une petite salle moins bruyante, a réussi le même test sur l'intonation avec un P=.999. Pour ce qui est de la durée, un groupe de 103 étudiants dans un amphi bruyant a reconnu correctement les trois temps avec un résultat très probant : P ≃ 1. Quant à l'intensité, elle ne paraît servir qu'à l'accentuation.

Morphological Functions of Suprasegmentals in Quebec French

. ABSTRACT

In rapid connected speech, we observe that the forms of 3 tenses may be reduced to the same segmental string, for example [teta]; [tetapo] : 1. T'es à Pau (you're in Pau) - Pau is a town of France - 2. T'étais à Pau (you were in Pau) - 3. T'as été à Pau (you've been in Pau or you went to Pau). Sonagraphic analysis of the production of Montrealers asked to say [tetapo] with the intention of meaning the different tenses reveals that both intonation and duration, and, to a certain extent, intensity, as well as the vocalic spectrums, vary with the tense on the three syllables. In order to determine whether the variation of the intonation and of the duration alone is sufficient to distinguish the three tenses, three sentenses /tetapo/ were synthesized by means of the Klatt programs at M.I.T. Tests were done with two groups of Montreal-French listeners (n = 24 and n = 103), and it was found that they were able to distinguish the three tenses on the basis of the variation of intonation ($P \simeq .90$) and of duration on the syllables, $P \simeq 1$. Another test was done for the intonation with 40 listeners in a small and quieter room : $P = .999$. As for the intensity, it only seems to be important for stress.

FONCTIONS MORPHOLOGIQUES DES PARAMETRES SUPRASEGMENTAUX EN FRANCAIS QUEBECOIS*

En français québécois, les fusions vocaliques, qui sont courantes dans le langage populaire, permettent de prononcer au moyen de la même suite homophonique les trois phrases suivantes, à condition de faire varier adéquatement les paramètres suprasegmentaux : [te ta po] = t'es à Pau, t'étais à Pau, t'as été à Pau. On ajoute un t de liaison (comme dans va-t-il, ou Malbrough s'en va-t-en guerre) et on dit t'e-t-à Pau. A l'imparfait, après la fusion de étais à en eta, on obtient la même surface phonétique, avec cette différence qu'ici, contrairement au présent, la deuxième syllabe porte normalement une marque compensatoire de durée et/ou d'intonation. Au passé composé, t'as été à devient teta, les deux syllabes étant susceptibles d'être marquées de traits suprasegmentaux fonctionnels. Quand, au moyen de sonagrammes, on compare les trois syllabes dans ces trois phrases, on voit que l'intonation, la durée et l'intensité varient systématiquement selon les temps, et que c'est la durée qui paraît la plus décisive. Cela n'est pas pour nous étonner, parce que plusieurs voyelles s'opposent encore en français québécois à la fois par le timbre et par la durée (Santerre 1974).

Est-il possible de distinguer les trois temps dans ces phrases en faisant varier au moyen de la synthèse un seul paramètre suprasegmental à la fois? Même si des phrases ainsi conçues ne peuvent jamais être produites humainement, ce serait faire la preuve indirecte du fonctionnement morphologique du trait suprasegmental.

On a donc préparé un premier test avec les trois syllabes [te ta po] où seule l'intonation varie selon les temps, un deuxième, où c'est la durée seule qui varie. Un troisième test a été préparé avec l'intensité variable, mais il s'est dès le début avéré que ce paramètres ne changeait pas les temps; les syllabes sont entendues accentuées comme porteuses d'un accent d'insistance, sans plus, comme dans c'est toi qui est à Pau, c'est à Pau que tu es, ou encore prosodie aberrante. C'est pourquoi, le test sur l'intensité a été abandonné et seuls les tests suivants ont été retenus : test 1, sur l'intonation, test 2 sur la durée, test 3 sur la durée et l'intonation.

Les stimuli

La synthèse des stimuli a été réalisée au M.I.T. en 1980-81 au moyen des programmes de D. Klatt, sous la direction de K.N. Stevens. Pour les spécifications segmentales, on a pris comme point de départ la production d'un seul locuteur québécois, mais pour les variations suprasegmentales, on a adopté la moyenne des mesures prises sur la production de cinq locuteurs qui ont prononcé les trois syllabes avec l'intention de faire entendre le présent, l'imparfait ou le passé composé, quand ils variaient peu le timbre des voyelles.

---

* Subventionné par le Conseil de recherches en sciences humaines du Canada

2.

Formants de voyelles, inchangés dans les trois tests :

	e	a	o
F_3	2500	2500	2375
F_2	1750	1600	1000
F_1	450	700	450

TRANSITIONS

t	e		t	a	p	o
	début	fin	début	fin	début	fin
F_3	2500	2500	2500	2500-2100	2100-2375	2375
F_2	1750	1750	1750-1600	1600-1400	900-1000	1000
F_1	300-450	450-350	400-700	700-450	300-450	450

Affrications : 10 msec, 55 dB à 3.5-5 KHz pour t, et à 2 KHz - 3.5 KHz pour P.

V.O.T. : 10 msec.

Les intensités de voisement (60 dB) sont appliquées et retirées avec une pente de 10 dB sur 10 msec. au début et à la fin, pour éviter les clics et les déclics.

Test 1 (sur l'intonation)

Durée (msec.)

T	e	t	a	p	o
	140	120	140	120	140

3.

Fo(Hz)

T	e	t	a	p	o
stimulus 1.	170 → 195 → 155		145 ——— 130		130 ——— 110
stimulus 2.	115 → 125		210 → 140		130 ——— 110
stimulus 3.	180 → 200 → 150		150 → 200 → 180		130 ——— 110

Test 2 (sur la durée)

Durée (msec.)

t	e	t	a	p	o
1.	190	90	90	120	130
2.	90	90	200	120	130
3.	190	90	190	120	130
Fo(Hz)					
1, 2, 3	150		150		130-110

Test 3 (sur la durée et l'intonation)

Durées : celles du test 2

Fo : ce sont les fréquences du test 1, à ces quelques détails près :

		t		p	
1.	185 → 200–200 → 150		140 ——— 130		125 ——— 100
	130 ——— 145		185 → 200 → 150		125 ——— 100
	185 → 200-200 → 150		185 → 200—200 → 150		125 ——— 100

Les séances de test et les sujets

Les trois tests ont été présentés à la suite l'un de l'autre, dans l'ordre suivant : le test 1 sur l'intonation, le test 2 sur la durée, le test 3 sur l'intonation et la durée. L'ensemble dure cinq minutes. Deux groupes de linguistique ont participé à ces tests : un groupe de 24 étudiants de 2e année

dans une salle de dimensions réduites et comportant 82 dB de bruit en basses fréquences, puis un autre groupe d'étudiants tous différents des premiers, au nombre de 103 et débutants, dans un amphithéâtre capable d'en contenir 250, à la fin de trois heures de cours. Les conditions d'écoute et d'attention étaient tout à fait difficiles.

Les phrases étaient présentées dans les deux cas au moyen d'un magnétophone Sony de médiocre qualité déposé devant la classe. Pour chaque test, les trois phrases étaient jouées autant de fois qu'on le désirait, soit deux, trois et même quatre fois. L'ordre des temps sur la bande était, pour le test 1 : 1. présent, 2. imparfait, 3. passé composé; pour le test 2 : 1. imparfait, 2. présent, 3. passé composé; pour le test 3 : comme pour le test 1. On demandait aux auditeurs d'attribuer un temps différent à chacune des trois phrases, et de ne répondre qu'après avoir jugé par comparaison.

ANALYSE

Si les sujets ne peuvent fournir qu'une seule réponse pour chaque phrase entendue, il y a $3! = 1 \times 2 \times 3 = 6$ possibilités de réponse. Par exemple dans le cas du test 1, on obtient :

- Réponse correcte sans erreur

  a) 1, 2, 3 = Présent, imparfait, passé composé.

- Réponse incorrecte avec 2 erreurs

  c) 2, 1, 3 = seul passé composé est exact
  d) 3, 2, 1 = seul imparfait est exact
  e) 1, 3, 2 = seul présent est exact

- Réponse incorrecte avec 3 erreurs

  f) 3, 1, 2
  g) 2, 3, 1

Si les sujets répondent au hasard, les 6 possibilités de réponse sont équiprobables, soit une probabilité de 1/6.

Pour tester l'hypothèse nulle H_0 : les réponses sont faites au hasard

contre l'hypothèse alternative H_1 : les réponses sont fonctions des durées, des intensités et des intonations.

Il suffit de calculer la probabilité d'obtenir la distribution des réponses observées, sous l'hypothèse nulle. Si cette probabilité est faible, par exemple inférieure à 5%, on rejettera l'hypothèse nulle.

CALCULS

Appellons X_1 l'évènement : réponse correcte sans erreur.

Appellons X_2 l'évènement : réponse incorrecte avec 2 erreurs.

Appellons X_3 l'évènement : réponse incorrecte avec 3 erreurs.

Sous l'hypothèse nulle H_0 on a :

$$P(X_1) = 1/6 = p_1; \quad P(X_2) = 1/2 = p_2; \quad P(X_3) = 1/3 = p_3$$

La probabilité d'obtenir N_1 individus avec réponses correctes, N_2 individus avec 2 erreurs et N_3 individus avec 3 erreurs est donnée par la loi multinomiale :

$$P(X_1 = N_1, X_2 = N_2, X_3 = N_3 / N, P_1, P_2, P_3) = \frac{N!}{N_1!\, N_2!\, N_3!}\ P_1^{N_1}\ P_2^{N_2}\ P_3^{N_3}$$

$$= \frac{N!}{N_1!\, N_2!\, N_3!}\ (1/6)^{N_1}\ (1/2)^{N_2}\ (1/3)^{N_3}$$

La probabilité d'obtenir N_1 réponses correctes et $N_2 - N_3$ réponses incorrectes, sans distinguer les réponses partiellement incorrectes des réponses totalement incorrectes est donnée par la loi binomiale :

$$P(X_1 = N_1 / N, P_1) = \frac{N!}{N_1!\,(N - N_1)!}\ P_1^{N_1}\ (1 - P_1)^{N - N_1}$$

$$= \frac{N!}{N_1!\,(N - N_1)!}\ (1/6)^{N_1}\ (5/6)^{N - N_1}$$

RESULTATS POUR LE TEST 1

On fait varier l'intonation des syllabes.

Pour le premier échantillon on a :

$N = 24$ $\quad N_1 = 6$ réponses sans erreur $\quad N_2 = 11$ réponses avec 2 erreurs

$N_3 = 7$ réponses avec 3 erreurs

donc $P(X_1 = 6, X_2 = 11, X_3 = 7 \;/\; 24, 1/6, 1/2, 1/3) =$

$$\frac{24!}{6!\ 11!\ 7!}\ (1/6)^6\ \ (1/2)^{11}\ \ (1/3)^7$$

Si les individus réponsent au hasard, la probabilité d'obtenir plus de six réponses correctes sur 24 serait de :

$$P(X \quad 6 \;/\; 24, 1/6) = \sum_{k=7}^{24} P(X = k \;/\; 24, 1/6) = 0{,}0912$$

On ne peut donc pas rejeter, au seuil de risque de 5%, l'hypothèse que les réponses ont été données au hasard.

La probabilité de ne pas faire d'erreur peut être estimée à 1 chance sur 4, la probabilité de faire 2 erreurs est estimée par 11 chances sur 24, soit 46%, la probabilité de faire 3 erreurs est de 7 chances sur 24 soit 29%.

Pour le second échantillon :

Le test exact peut être remplacé par l'approximation du Chi Deux car la taille de l'échantillon, 103 sujets, le permet.

On calcule la statistique X^2 avec 2 degrés de liberté.

On a obtenu 17 réponses sans erreur, 62 réponses avec 2 erreurs, 24 réponses avec 3 erreurs.

$$X^2 = \sum_{i=1}^{3} \frac{(n_i - t_i)^2}{t_i}$$ avec t_i = fréquence théorique de réponse sous l'hypothèse nulle.

$$X^2 = \frac{(17 - 103/6)^2}{103/6} + \frac{(62 - 103/2)^2}{103/2} + \frac{(24 - 103/3)^2}{103/3} =$$

$$X^2 = 0.001618 + 2.140777 + 3.110032 = 5.252427$$

On compare le X^2 observé qui vaut 5.25 avec celui fournit par la table du Chi Deux pour un seuil risque choisi et pour 2 degrés de liberté.

$X^2_2\ (.10) = 4.61$ $\qquad$ $X^2_2\ (.05) = 5.99$ $\qquad$ $X^2_2\ (.025) = 7.38$

Le risque d'erreur encouru en rejetant l'hypothèse de réponse au hasard se situe entre 5 et 10%.

Ce test sur l'intonation a été repris dans une petite salle de cours moins bruyante avec 40 étudiants; l'ordre de présentation des phrases a aussi été changé; il est le suivant : imparfait, passé composé, présent. On a eu 14 réponses sans erreur, 23 réponses avec 2 erreurs et 3 réponses avec 3 erreurs.

L'approximation du Chi Deux se lit comme suit :

$$X^2 = \frac{(14 - 40/6)^2}{40/6} + \frac{(23 - 40/2)^2}{40/2} + \frac{(3 - 40/3)^2}{40/3} =$$

$$X^2 = 8.066 + .045 + 8.00 = 16.51$$

Cette fois, le risque d'erreur couru en rejetant l'hypothèse nulle est de 0.1%. Sur la base de cet essai, il semble que l'on puisse conclure que les sujets sont capables, lorsque les conditions d'écoute sont de bonne qualité, de distinguer les trois phrases et de leur attribuer les temps visés par la synthèse, au moyen de l'intonation.

RESULTATS POUR LE TEST 2

On fait maintenant varier la durée des syllabes.

Echantillon 1 : 16 réponses sans erreur, 5 réponses avec 2 erreurs, 3 réponses avec 3 erreurs.

$$P(X_1 = 16, X_2 = 5, X_3 = 3 \;/\; 24, 1/6, 1/2, 1/3) =$$

$$\frac{24!}{16!\,5!\,3!} (1/6)^{16} (1/2)^5 (1/3)^3 = 1.67 \times 10^{-8}$$

Si les individus répondaient au hasard, la probabilité d'obtenir plus de 16 réponses correctes sur 24 serait de :

$$P(X \quad 16 \;/\; 24, 1/6) = \sum_{k=17}^{24} P(X = k \;/\; 24, 1/6) = 6.2 \times 10^{-9}$$

On peut donc conclure, avec moins d'une chance sur un million de se tromper que les sujets n'ont pas répondu au hasard. La probabilité de reconnaître les temps est de 16/24 = 2/3, soit presque 3 fois plus qu'en l'absence de variation de la durée des syllabes.

Echantillon 2 : 68 réponses sans erreur, 27 réponses avec 2 erreurs, 8 réponses avec 3 erreurs.

$$X^2 = \frac{(68 - 103/6)^2}{103/6} - \frac{(27 - 103/2)^2}{103/2} - \frac{(8 - 103/3)^2}{103/3} =$$

$$= 150.53 - 11.66 - 20.20 = 182.38$$

Le X^2 observé est très supérieur au $X^2_{(.005)}$ tabulé qui vaut 10.6 pour 2 degrés de liberté. On peut donc rejeter l'hypothèse de réponse au hasard avec un risque d'erreur largement inférieur à 5 sur 1000. A nouveau, la durée des syllabes permet d'accroître la compréhension de façon spectaculaire puisque la probabilité de ne pas faire d'erreir passe de 0,165 à 68/103 = 0.66 soit une multiplication par 4.

RESULTATS POUR LE TEST 3

On fait varier simultanément la durée et l'intonation des syllabes.

Echantillon 1 : 17 réponses sans erreur, 6 réponses avec 2 erreurs, 1 réponse avec 3 erreurs.

$$P(X_1 = 17, X_2 = 6, X_3 = 1 \;/\; 24, 1/6, 1/2, 1/3) = \frac{24!}{17!\,6!\,1!}(1/6)^{17}(1/2)^6(1/3) =$$

$$= 7.45 \times 10^{-10} \simeq 0$$

Si les individus répondaient au hasard, la probabilité d'obtenir plus de 17 réponses exactes sur 24 n'est que de :

$$P(X \quad 17 \;/\; 24, 1/6) = \sum_{k=18}^{24} P(X = k \;/\; 24, 1/6) = 5 \times 10^{-10}$$

La probabilité d'obtenir au hasard le résultat observé est donc pratiquement nulle.

Echantillon 2 : 70 réponses sans erreur, 30 réponses avec 2 erreurs, 3 réponses avec 3 erreurs.

$$X^2 = \frac{(70 - 103/6)^2}{103/6} + \frac{(30 - 103/2)^2}{103/2} + \frac{(3 - 103/3)^2}{103/3}$$

$$= 162.53 + 8.98 + 28.60 = 200.17$$

A nouveau le X^2 observé est très largement supérieur au X^2 tabulé. On conclue donc au rejet de l'hypothèse de réponse au hasard.

Le surcroît de reconnaissance du temps des phrases apporté par l'intonation paraît assez marginal puisque la probabilité de ne pas faire d'erreur passe de 68/103 à 70/103 soit un accroissement de 2% seulement.

DISCUSSION

Il apparaît clairement que les différences de durée sur les syllabes sont plus efficaces pour distinguer les trois temps que ne le sont les différences d'intonation; les résultats du test 2 sont nettement supérieurs dans les deux échantillons aux résultats du test 1; de plus le test 3 qui comportait une variation de la durée et de l'intonation a des résultats très peu supérieurs au test 2; on voit par là que l'intonation a ajouté peu d'information à la durée dans ces phrases synthétisées. Il n'empêche que l'intonation seule a suffi à distinguer les trois temps et que l'on peut l'affirmer sans trop de risque de se tromper (5% dans le groupe de 24 et moins de 10% dans le groupe de 103, 0.1 dans le groupe de 40).

Le français québécois joue beaucoup sur la durée vocalique et consonantique au niveau morphologique; un autre exemple est celui de la distinction entre Vous êtes malades [vzɛt...] et Vous allez être malade [vuzaleɜtr(e)(1)] qui peut se réduire, après la chute du /l/ intervocalique et la fusion des voyelles brèves avec la longue /ɜ/ : [vzɜt...].

Des tests par la synthèse sont à coup sûr artificiels et ne peuvent donner que des indications sur la parole réelle. Néanmoins, l'intonation semble jouer un rôle morphologique plus fragile que la durée; sa fonction pourrait être plutôt celle de marqueur syntaxique, en particulier aux frontières intersyntagmatiques et intra-syntagmatiques où elle prend des formes différentes (Santerre 1979).

Il est possible que l'ordre de présentation des phrases dans les tests 1 et 3 (présent, imparfait, passé composé) ait introduit un biais dans les réponses; ce biais ne doit pas être grand, puisque le test 2, qui a été très significatif, ne présentait pas le même ordre; cet ordre a d'ailleurs été modifié dans le test 1 avec le groupe de 40.

Quand à l'intensité, il semble qu'elle ait avant tout une fonction d'appoint (mais je ne l'ai pas fait varier) en morphologie, et une fonction directe d'accentuation (accent démarcatif et accent d'insistance) (Warren). Je n'ai pas trouvé de rapport d'intensité entre les syllabes qui avait à lui seul une incidence sur la perception d'un temps plutôt que d'un autre.

---

(1) Il s'agit de la voyelle longue de fête, opposée à celle de faite.

CONCLUSION

Je ne m'arrêterai pas ici à des conclusions proprement linguistiques pourtant importantes du point de vue théorique; dans la description du langage et la définition d'une langue, le moins qu'on puisse dire, c'est que la linguistique ne peut plus s'en tenir à la représentation écrite de la langue parlée, ni même à la transcription phonétique étroite en A.P.I.

Du point de vue pratique, par exemple dans l'enseignement ou l'apprentissage des langues secondes, l'acquisition des traits suprasegmentaux se révèle de première importance. La réalisation des segments pose beaucoup moins de problèmes que celle des faits prosodiques; on ne sait pas enseigner ces faits parce qu'on les connaît mal; la recherche dans ce domaine n'est pas facile mais elle est primordiale.

Du point de vue plus technique de la reconnaissance et de la synthèse de la parole, je n'ai pas à prêcher à des convertis. Les principaux travaux dans ce domaine nous sont connus. Pour la synthèse par règles, on n'envisage pas (du moins pas encore) de faire produire ou reconnaître du langage à surface réduite, mais les fusions vocaliques font partie du discours normal reconnu comme correct. Indépendamment des réductions de surface, un langage synthétisé qui n'aurait pas les justes intonations et les durées appropriées (c'est aussi vrai pour les intensités distribuées dans la phrase) perdrait sans doute beaucoup de son taux de compréhension et son caractère naturel. Ces paramètres peuvent varier d'un dialecte à l'autre en importance et en caractéristiques. Il faut d'abord trouver des mesures qui soient le plus largement utiles.

REFERENCES

Santerre, L. "La fusion des voyelles en frontières inter-syntagmatiques et intra-syntagmatiques", in Current Issues in the Phonetic Sciences, H. and P. Hollien, Amsterdam, John Benjamins, 1979, pp. 1131-1138.

Santerre, L. "Deux E et deux A phonologiques en français québécois", Le français de la région de Montréal, Cahier de linguistique, no 4, P.U.Q., 1974, pp. 117-145.

Warren, R. "Les paramètres acoustiques de l'accent montréalais", Travaux en l'honneur de G. Faure, Studia Phonetica, Didier, sous presse.

Laurent Santerre
Département de linguistique
Université de Montréal

Jean-Louis Chandon
Département des Sciences administratives
Université du Québec à Montréal

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

H. MELONI, J. GUIZOL

Groupe d'Intelligence Artificielle
FACULTE DES SCIENCES DE LUMINY
70, Route Léon Lachamp - CASE 901
13288 MARSEILLE CEDEX 9 - France

"UN SYSTEME DE RECONNAISSANCE AUTOMATIQUE DE LOCUTIONS"

.RESUME

Le système que nous proposons est une application au domaine de la reconnaissance de mots de principes retenus pour la parole continue. La première étape du processus permet une représentation abstraite (proche des phonèmes) des informations contenues dans le signal au moyen d'unités appelées "événements pseudo-phonétiques" ainsi qu'un repérage du noyau accentué. La sollicitation des mots du lexique est limité lors de la reconnaissance, par des informations contextuelles (pragmatiques) ou induites par certains éléments identifiés (noyau vocalique accentué). Un ensemble de règles permet la superposition d'une ou plusieurs des chaînes de phonèmes caractérisant les mots retenus avec la séquence des événements pseudo-phonétiques proposée.

. ABSTRACT

The système proposed here is the application of the principles of continuous speach analysis to words-recognition.

The first step consists in representing the information inclosed in the signal by use of abstract units (quite like phonemes) called "pseudo-phonetic events" and in finding the stressed syllable. The second step consists in the choice of the words of lexicon, which is partly determined by contextual (pragmatic) information and partly induced by some identified elements (stressed vowel).

At last a set of rules matches the sequence of "pseudo-phonetic events" with the phonemic sequence of the words chosen from the lexicon.

1 - INTRODUCTION

Le système de reconnaissance automatique de locutions que nous présentons constitue une étape dans la réalisation d'un projet plus général (schématiquement décrit dans (MELONI, 1980)) concernant la parole continue. Néanmoins, nous pensons que, dans les situations où s'impose une reconnaissance analytique, certains processus utilisés permettent d'apporter une réponse originale au problème de la reconnaissance de mots.

Notre objectif principal se borne à l'élaboration de solutions raisonnables à quelques uns des problèmes posés par la reconnaissance analytique des mots, nous n'avons donc pas envisagé certaines difficultés techniques telles que l'obtention d'un temps de réponse voisin des conditions réelles.

Après avoir présenté le schéma général d'organisation des modules du système, nous détaillerons pour chacun d'eux quelques unes des particularités les plus significatives.

2 - SCHEMA GENERAL DE FONCTIONNEMENT DU SYSTEME

La Figure 1 caractérise les divers éléments du système ainsi que leurs relations. Un premier niveau permet la segmentation de l'unité de parole (mot ou locution pris entre deux silences) en une séquence d'événements assimilables aux diverses étapes de la réalisation d'un ou plusieurs phonèmes. A ce stade de la reconnaissance, le dernier noyau vocalique accentué est également repéré. Ces unités constituent les éléments de base de la reconnaissance; elles caractérisent une description analytique de la parole qui ne sera pas remise en cause dans les étapes ultérieures.

Les informations contextuelles correspondent à des connaissances dont dispose la machine concernant les interventions du locuteur, elles permettent dans certains cas de limiter le nombre des mots du dictionnaire qui peuvent être attendus. Les accès aux mots du lexique sont commandés à la fois de façon prédictive par les contraintes contextuelles ainsi que de manière inductive par les caractéristiques de certains événements (noyaux accentués par exemple).

L'adéquation entre la description phonémique lexicale des mots ou locutions proposés et la séquence des événements reconnue se fait par l'intermédiaire d'un ensemble de règles contextuelles qui rendent compte de toutes les caractérisations possibles d'une suite de phonèmes.

3 - EVENEMENTS PSEUDO-PHONETIQUES

Nous avons donné dans (MELONI, 1979) et (MELONI, 1980) une description de cette étape de notre système; nous rappellerons ici les caractéristiques principales de ces unités.

Notre démarche consiste à identifier dans le signal des segments qui soient aussi proche que possible des unités abstraites que sont les phonèmes. Les nombreux travaux effectués dans ce domaine n'ont pu venir à bout du délicat problème que constitue la segmentation et l'identification de phonèmes dans la chaîne parlée. Cependant, les résultats obtenus attestent que ces unités sont reconnues convenablement dans un pourcentage important

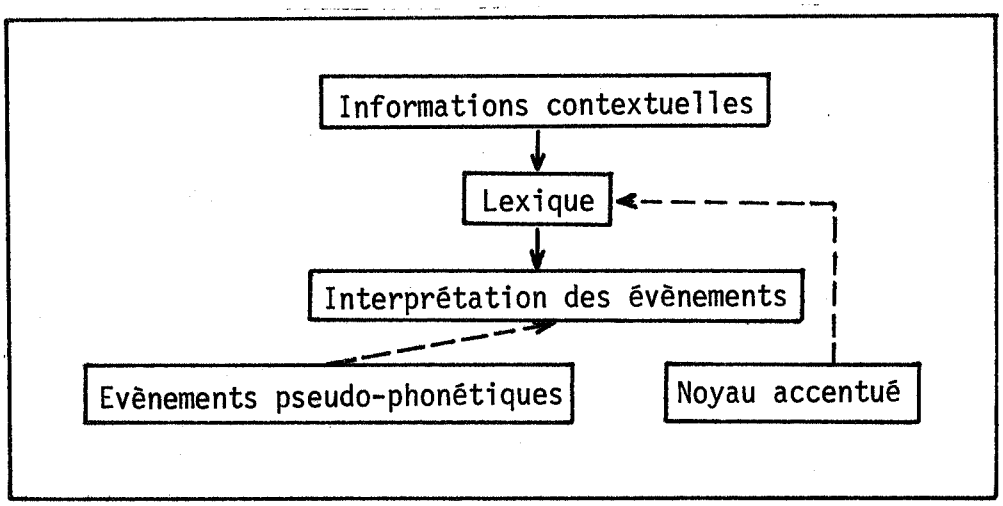


Figure 1

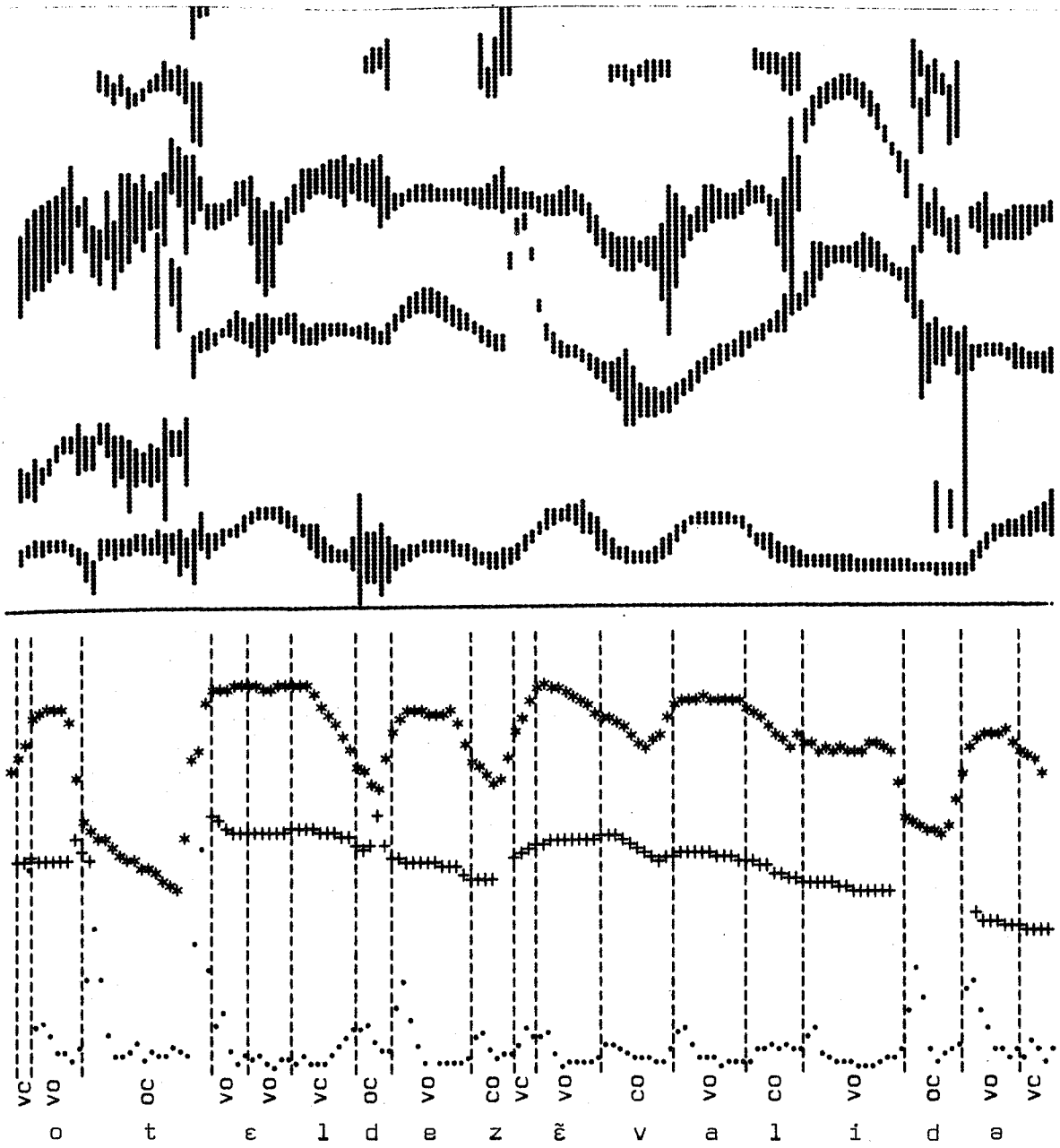


Figure 2

de cas. Il nous a semblé utile de définir des unités plus maléables et susceptibles de recevoir des interprétations contextuelles diverses.

Ainsi que le préconise (LIENARD, 1972), nous avons réalisé la segmentation à l'aide de fonctions d'instabilité spectrale (distance entre deux spectres à 20 ms et 40 ms d'intervalle). Pour chacune des unités ainsi définies, une macro-classe pseudo-phonétique lui est attribuée en utilisant (indépendamment du contexte) un ensemble de paramètres calculés toutes les 10 ms par LPC. Plusieurs segments ont pu être regroupés pour définir certaines macro-classes. Chaque classe peut caractériser un phonème, une partie d'un phonème ou bien la réunion de plusieurs d'entre-eux. La figure 2 représente la segmentation en macro-classes de "Hotel des invalides". Une macro-classe ne représente pas de façon biunivoque une partie de phonème, elle peut avoir des interprétations diverses suivant le contexte (ex : un segment constrictif peut caractériser la transition entre une occlusive stridente et une voyelle aigüe). Cependant, nous pouvons donner la correspondance suivante (la plus générale) entre les macro-classes et les phonèmes : 'VO' → voyelle, 'VC' → voyelle ou consonne vocalique, 'CO' → consonne, 'CC' et 'FR' → constrictive, 'OC' → occlusive, 'TM' et 'TD' → transitions.

A chacune des macro-classes est associé un ensemble de traits pseudo-phonétiques qui contribuent à caractériser le segment. Cette démarche, semblable à celle adoptée par (Caelen, 1979), se singularise par le fait que les traits sont souvent assez éloignés de ceux communément utilisés en phonétique. Les remarques suivantes permettent d'en définir les principaux caractères :

- Leur nombre et leur nature varient suivant la macro-classe qu'ils déterminent; nous en avons retenu 11 (voisé/non voisé, vocalique/consonnantique, interrompu/non interrompu, nasal/oral, ouvert/fermé, grave/aigu, compact/diffus, diésé/non diésé, bémolisé/non bémolisé, strident/mat, intense/non intense).

- Les indices qui les déterminent sont évalués en différents points du segment (partie stable, explosion, début ou fin, etc...).

- Ils ne constituent pas un système minimal et certains peuvent être redondants (les indices associés étant cependant distincts).

- L'affectation d'un trait n'est faite que lorsque la convergence des indices est suffisante, il peut être affecté positivement, négativement ou n'être pas déterminé.

- Bien que le contexte soit pris en compte dans le calcul de certains indices (trait ± compact des occlusives par exemple), la plupart d'entre eux sont affectés indépendamment de l'environnement.

- Leur nature est liée aux conditions de traitement du signal (filtrage par exemple) ainsi qu'aux paramètres choisis.

- Par souci de simplicité, nous n'avons conservé que les indices les plus significatifs pour caractériser un trait, au prix d'une indétermination plus grande.

A titre d'exemple, les traits affectés aux classes définies dans la figure 2 sont les suivants : (la signification des paramètres est donnée au paragraphe 7)

/o/ { VC(MT,V(FE(1).GR.OR.DF.ND.NB),C(CO.VC.IN.*O.GR.CP.ND.NB))
V(FE(1).GR.*N.DF.ND.NB)

/t/ C(OC.NV.*A.*C.ST)

/ɛ/ { V(*F.AI.OR.DF.ND.*B)
V(OU(1).AI.OR.DF.*D.NB)

/l/ C(CD.*V.*I.*O.AI.DF.*Z.NB)

/d/ C(OC.VS.AI.*C.*S)

/e/ V(FE(1).AI.OR.DF.*Z.NB)

/z/ C(CD.CN.IN.OR.AI.DF.*Z.*M)
VC(MT,V(FE(1).AI.OR.DF.ND.NB), C(CO.*V.IN.*O.AI.DF.ND.NB))

/ɛ̃/ V(OU(*O).AI.*N.*C.ND.NB)

/v/ C(CO.VC.IN.*O.*G.DF.*Z.*M)

/a/ V(OU(2).*A.*N.*C.ND.*B)

/l/ C(CO.VC.*I.*O.AI.DF.ND.NB)

/i/ VA(FT,FE(2).AI.OR.DF.DI.NB)

/d/ C(OC.NV.AI.*C.*S)

/ə/ V(*F.AI.*N.*C.ND.NB)
VC(DS,V(OU(1).*A.*N.*C.ND.*B), C(CO.VC.*I.*O.*G.CP.ND.*M))

4 - INFORMATIONS CONTEXTUELLES

Les informations contextuelles concernent toutes les connaissances qui permettent, lors d'une intervention du locuteur, de réduire le nombre des mots du lexique. Ces informations sont indispensables pour la consultation de grands dictionnaires.

Quelques problèmes soulevés par la consultation vocale d'un catalogue de films nous donneront un exemple d'utilisation des contraintes contextuelles pour la reconnaissance de locutions. Les informations codées dans le catalogue sont nombreuses et correspondent à un lexique comportant plusieurs milliers de mots. Pour chaque film, outre son titre, elles concernent des rubriques telles que le personnel (réalisateurs, comédiens, techniciens, etc...), la technique (sonore/muet, durée, couleur, etc...), le thème (comique, social, politique, etc...), etc... Ces caractéristiques peuvent constituer des accès multiples au catalogue. Le dialogue suivant, dirigé par la machine, donne une idée de scénario possible lors d'une demande de renseignements.

- Machine : Vous souhaitez consulter notre catalogue des films 35mm, 16mm ou 8 mm ?

- Locuteur : Films seize millimètres.

- Machine : S'agit-il de films long métrage ou court métrage ?

- Locuteur : Long métrage.

- Machine : Quel type d'accès envisagez-vous ? par nom (du pays, du réalisateur ou d'interprètes), par le thème ou bien par le titre du film ?

- Locuteur : Par le thème.

- Machine : Parmi les centres d'intérêt répertoriés, lequel vous intéresse? comique, politique, fantastique, etc... ?

- Locuteur : Comique.

- Machine : Quels sont les renseignements complémentaires que vous pouvez nous fournir ? Réalisateur, interprètes, pays d'origine, etc...

- Locuteur : Interprètes.

- Machine : Lesquels ?

- Locuteur : Les Marx Brothers.

- Machine : Nous disposons des films suivants :
  "La soupe au canard" USA - 1933 70' (L.Mc.Carey) Vo 1078 F
  "Une nuit à l'opéra" USA - 1935 90' (S. Wook) Vo 1078 F
  .
  .
  .

- Machine : Souhaitez-vous d'autres renseignements ?
- Locuteur : Non.

Cet exemple nous montre les contraintes imposées au locuteur afin d'orienter sa réponse vers un sous ensemble de termes du lexique considérablement réduit. Chaque classe d'éléments concernés regroupe toutes les informations de même nature. Cela suggère donc une partition du lexique en autant de sous ensembles (pas nécessairement disjoints) que de classes.

5 - ORGANISATION DU LEXIQUE

La structuration du lexique permet la cohabitation de deux types d'accès, l'un de nature prédictive à l'aide des informations contextuelles et l'autre de nature inductive à partir d'éléments fiables reconnus dans le message (voyelle accentuée par exemple). Nous retrouvons de telles préoccupations chez (COLE, 1980). Nous illustrerons ces particularités à l'aide de l'exemple précédent de consultation d'un catalogue de films.

5.1 - Accès catégoriel aux mots

Les diverses informations qui concernent un film peuvent être considérées comme autant d'éléments permettant sa catégorisation dans une classe d'autant plus restreinte que les connaissances disponibles sont précises. La catégorie la plus générale regroupe tous les films. Considérons 3 informations simples permettant une sous catégorisation (pays d'origine, thème, réalisateur). L'accès du lexique ne doit être ouvert qu'aux mots (ou locutions) possèdant les caractéristiques définies.

Exemple : En Prolog (MELONI, 1980), les informations concernant un film peuvent être codées de la manière suivante :

```
FILM(FRANCAIS. INFO, TITRE) : FILMFRA ( INFO, TITRE)..
FILM(HONGROIS. INFO, TITRE) : FILMHON ( INFO, TITRE)..
                .
                .
                .
FILMHON(POLITIQUE. INFO, TITRE) : FILMHONPOL( INFO, TITRE)..
FILMHON(COMIQUE. INFO, TITRE) :   FILMHONCOM( INFO, TITRE)..
                .
                .
                .
FILMHONPOL(JANCSO, TITRE) : FILMHONPOLJAN( TITRE)..
FILMHONPOL(TOROK, TITRE) :  FILMHONPOLTOR( TITRE)..
                .
                .
                .
FILMHONPOLJAN(((P.S.AU.M.E.R.).OU.J.E).PSAUME-ROUGE) : ..
FILMHONPOLJAN(((R.OU.J.EI.B.L).AN.§).ROUGE-ET-BLANC) : ..
                .
                .
                .
```

L'accès au lexique est d'autant plus contraint que les informations sont déterminées, mais cela n'interdit nullement d'atteindre toutes les locutions codées.

5.2 - Accès induit par les événements pseudo-phonétiques

Nous pouvons également limiter l'ensemble des mots consultés en utilisant des événements fiables qui constituent des points d'ancrage. Bien que plusieurs types de segments soient susceptibles de constituer de tels éléments (segments occlusifs, constrictifs, etc...), nous n'avons conservé comme événement d'accès que celui constitué par le noyau vocalique accentué (il correspond à la dernière voyelle du mot). En effet, l'événement associé à la dernière voyelle caractérise très fiablement celle-ci et ceci quel que soit le contexte. Si parfois l'indétermination de certains traits conduit à proposer plusieurs phonèmes à cette position, aucun d'entre deux n'est reconnu de façon erronée dans ce contexte.

Dans le cas d'un lexique figé (aucun mot nouveau n'étant ajouté), on pourrait envisager un codage phonétique ou seul les éléments permettant de distinguer un mot des autres seraient représentés. Les informations phonétiques codées seraient dépendantes à la fois du lexique et des performances du système de reconnaissance (dans notre cas, nous devrions privilégier le noyau accentué puis établir une hiérarchie des segments et des traits les plus fiables).

Pour une répartition des locutions du lexique (en fonction de la dernière voyelle) en classes de cardinalités semblables, le nombre des accès autorisés lors de la reconnaissance sera considérablement réduit (divisé par 14 si les traits de la voyelle sont tous déterminés).

Dans l'exemple précédent on peut encore établir une sous-catégorisation à l' aide de la dernière voyelle de la locution, nous aurons alors le codage suivant :

FILMHONPOLJANOU(((P.S.AU.M.E.R).OU.J.E).(PSAUME.ROUGE)) : ..
FILMHONPOLJANAN(((R.OU.J.EI.B.L).AN./Ø).(ROUGE.ET.BLANC)) : ..

Nous n'avons voulu illustrer ici que le principe de tels accès aux mots du dictionnaire, mais on peut envisager, sur ce modèle, des techniques plus sophistiquées permettant une sous catégorisation contextuelle par l'utilisation d'informations "descendantes" ou "remontantes".

6 - INTERPRETATION DES EVENEMENTS

6.1 - Stratégie

Nous disposons pour cette phase de la reconnaissance de la séquence d'événements pseudo-phonétiques à laquelle on doit associer l'un des mots du sous ensemble sélectionné dans le lexique. Une stratégie possible consisterait à inférer, à partir des événements, un treillis de phonèmes ou de diphones. La grande majorité des systèmes de reconnaissance analytique (nous citons uniquement les systèmes opérant sur le français) proposent cette technique (LIENARD, 1972), (CAELEN, 1979), (HATON, 1974), (VIVES, 1978). L'identification d'un mot consiste ensuite à rechercher dans le treillis la séquence la plus probable qui coïncide avec une description phonémique lexicale. On peut envisager une conduite remontante du processus (du treillis vers le lexique) ou bien une recherche descendante (du lexique vers le treillis).

La stratégie que nous avons adoptée consiste à interpréter, au moyen de règles contextuelles, toutes les séquences de phonèmes associées aux mots sélectionnés. Les règles sont appliquées de droite à gauche sur les phonèmes précédant la voyelle accentuée puis de gauche à droite sur ceux qui la suivent. Un mot est abandonné lorsque toutes les règles appliquées sur les phonèmes proposés n'ont pu engendrer la sous séquence d'événements associée. Si un mot est reconnu, le processus est cependant poursuivi pour les éléments restants de façon à proposer tous les mots identifiés à partir de la séquence d'événements (ambiguïtés).

6.2 - Règles d'interprétation des événements

L'ensemble des règles du modèle a été construit manuellement, mais nous envisageons une procédure d'apprentissage automatique dirigé. Le nombre des règles est lié aux performances de l'étape de segmentation en macro-classes et d'identification des traits pseudo-phonétiques. De plus, des variations substantielles interviennent d'un locuteur à l'autre. Afin de limiter les ambiguïtés, les règles doivent être aussi fines que possible. Mais elles doivent cependant permettre d'interpréter toutes les caractérisations d'une séquence de phonèmes en événements. Ces exigences contradictoires rendent délicat le problème de la construction du modèle. Celui-ci ne peut être figé définitivement et l'adjonction de nouvelles règles peut être faite lors d'une reconnaissance qui parait anormalement défectueuse. La version actuelle comporte une quarantaine de règles pour un locuteur.

On peut déterminer également, pour un ensemble de règles et un lexique donnés, les ambiguïtés possibles entre les diverses locutions du dictionnaire en examinant toutes les séquences d'événements produites à partir des représentations phonémiques codées.

La majorité des règles rendent compte de la segmentation d'une séquence de phonèmes en événements. Un phonème est fréquemment représenté par plusieurs événements et quelquefois une séquence phonémique se réduit à un seul événement. D'autres règles permettent un réajustement contextuel des traits. Les règles dépendent parfois du contexte (phonèmes voisins, présence d'un accent, fin de mot, début de mot, etc...). Nous donnons ci-dessous quelques exemples des types principaux de productions du modèle :

Règles non contextuelles

- / ɲ / → CO (+vocalique, +interrompu, +nasal, +aigu) + VC(+vocalique, +aigu, +diésé)
- / k / → OC (-voisé, +compact) + CO(...)
- voyelle nasale → VO (+nasal) + VC (...)
- / b / → CO (+interrompu, +grave) (disparition de l'explosion)
- / ə / + / l / → VC (+vocalique)

Règles contextuelles

- / r / → | voyelle ouverte -
- / t / → OC (-voisé, +aigu, +strident) + CC () | - voyelle aigüe
- / f / → CC (-voisé, +grave, +intense) | - voyelle fermée
- / l / → VC (+vocalique) | occlusive - voyelle
- voyelle → VO (...) + VC (...) | - Fin de groupe

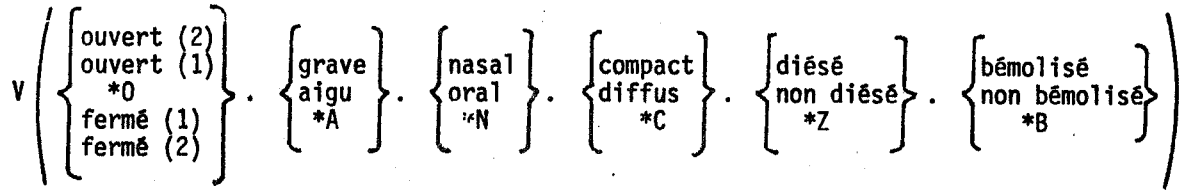
7 - EXEMPLES DE RECONNAISSANCE DE LOCUTIONS

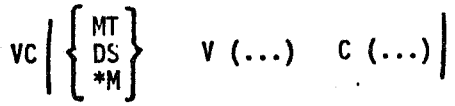
Les locutions soumises à la reconnaissance composent un vocabulaire expérimental d'environ 300 mots. Sans tenir compte des limitations catégorielles, ces mots sont approxivement répartis en 15 classes caractérisées par leur noyau vocalique accentué. Cela signifie que dans la situation idéale (mais peu fréquente) ou tous les traits de la voyelle accentuée ont été déterminés sans ambiguïté, seulement une vingtaine de locutions seront traitées.

Le système ne fonctionnant pas en temps réel, une expérimentation intensive serait fastidieuse. Néanmoins, nous avons enregistré de nombreuses locutions, dans des conditions normales de bruit (pièce calme) et d'élocution, qui ont été soumises pour identification au système. Les mots étant assez différents, nous n'avons pas constaté d'ambiguïtés (reconnaissance de plusieurs locutions pour la séquence d'événements proposée). De plus, l'ensemble des règles d'interprétation des événements constitue un système suffisamment permissif pour tolérer les digressions principales par rapport à la représentation phonétique idéale. Aussi, peu de chaînes identifiées n'ont pu être reconnues comme des locutions permises.

Dans la représentation des événements, aux traits indéterminés sont associées des variables (identificateurs précédés du caractère "*"). Les macro-classes sont du type voyelle (V), consonne (C) ou indéterminé (VC), leurs arguments caractérisent les traits suivants :

$$
C \left\{ \begin{array}{l}
\text{occlusif} \;.\; \left\{\begin{array}{c}\text{voisé}\\ \text{non voisé}\\ *V\end{array}\right\} . \left\{\begin{array}{c}\text{grave}\\ \text{aigu}\\ *A\end{array}\right\} . \left\{\begin{array}{c}\text{compact}\\ \text{diffus}\\ *C\end{array}\right\} . \left\{\begin{array}{c}\text{strident}\\ \text{mat}\\ *S\end{array}\right\} \\
\text{constrictif} \;.\; \left\{\begin{array}{c}\text{voisé}\\ \text{non voisé}\\ *V\end{array}\right\} . \left\{\begin{array}{c}\text{grave}\\ \text{aigu}\\ *A\end{array}\right\} . \left\{\begin{array}{c}\text{compact}\\ \text{diffus}\\ *C\end{array}\right\} . \left\{\begin{array}{c}\text{intense}\\ \text{non intense}\\ *I\end{array}\right\} . \left\{\begin{array}{c}\text{strident}\\ \text{mat}\\ *S\end{array}\right\} \\
\text{consonne} \;.\; \left\{\begin{array}{c}\text{consonnantique}\\ \text{vocalique}\\ *V\end{array}\right\} . \left\{\begin{array}{c}\text{interrompu}\\ \text{continu}\\ *I\end{array}\right\} . \left\{\begin{array}{c}\text{nasal}\\ \text{oral}\\ *O\end{array}\right\} . \left\{\begin{array}{c}\text{grave}\\ \text{aigu}\\ *A\end{array}\right\} . \left\{\begin{array}{c}\text{compact}\\ \text{diffus}\\ *C\end{array}\right\} . \\
\qquad\qquad \left\{\begin{array}{c}\text{diésé}\\ \text{non diésé}\\ *Z\end{array}\right\} . \left\{\begin{array}{c}\text{bémolisé}\\ \text{non bémolisé}\\ *B\end{array}\right\}
\end{array} \right.
$$






```
        ********
****** EXEMPLES ******
        ********..


****** LOCUTION 1 ******..


VC(MT,V(*F.GR.*N.*C.ND.NB),
C(CO.*V.IN.*O.GR.CP.*Z.NB))..
V(*F.GR.*N.*C.ND.NB)..
C(OC.NV.GR.*C.NS)..
V(*F.AI.OR.DF.ND.NB)..
C(CO.VC.NI.OR.*G.CP.*Z.*M)..
V(OU(2).*A.*N.CP.ND.*B)..
C(OC.NV.AI.*C.ST)..
VA(FT,*F.*A.*N.*C.ND.*B)..
FIN..

 **** LOCUTION(S) RECONNUE(S) :  OPERATEUR


****** LOCUTION 2 ******..


VC(MT,V(FE(1).*A.OR.CP.ND.*B),
C(CO.CN.IN.OR.*G.CP.*Z.*M))..
VC(MT,V(*F.AI.*N.DF.ND.NB),
C(CO.VC.*I.*O.AI.DF.ND.NB))..
VC(*M,V(*F.AI.OR.DF.*D.NB),
C(CO.VC.*I.*O.AI.DF.*Z.*M))..
V(OU(*O).AI.*N.*C.ND.NB)..
VC(DS,V(*F.AI.*N.DF.*D.NB),
C(CO.VC.*I.*O.AI.DF.*Z.NB))..
C(CO.VC.*I.*O.AI.DF.*Z.NB)..
V(FE(*O).AI.OR.DF.DI.NB)..
C(CC.VS.AI.DF.*I.ST)..
V(*F.AI.*N.*C.ND.*B)..
C(OC.NV.AI.*C.ST)..
VA(FT,*F.AI.*N.*C.ND.NB)..
FIN..

 **** LOCUTION(S) RECONNUE(S) :  REALISATEUR
```

```
****** LOCUTION 3 ******..

VC(MT,V(OU(*O).AI.OR.DF.ND.NB),
C(CO.*V.IN.*O.*G.DF.*Z.*M))..
V(OU(2).*A.*N.CP.ND.*B)..
C(OC.*V.GR.CP.NS)..
VA(FT,*F.GR.*N.*C.ND.NB)..
VC(DS,V(FE(2).GR.OR.DF.*D.NB),
C(CO.CN.*I.OR.*G.DF.*Z.*M))..
C(OC.*V.GR.*C.NS)..
V(*F.*A.*N.CP.ND.*B)..
FIN..

 **** LOCUTION(S) RECONNUE(S) :  LA BOMBE
```

8 - CONCLUSION

Le système expérimental de reconnaissance de locutions que nous avons présenté nous semble particulièrement adapté pour résoudre le problème d'une consultation vocale (orientée par la machine) de catalogues divers. Les problèmes de segmentation et d'identification des événements ont reçu dans bien des systèmes des solutions satisfaisantes. De plus, le déterminisme d'une telle tâche rend ce processus réalisable en temps réel. Notre expérience nous a montré que l'interprétation contextuelle des événements, malgré son caractère indéterministe, pouvait recevoir des solutions satisfaisantes et relativement efficaces. Cependant, il est pratiquement impossible d'évaluer avec précision les performances de cette étape tant est grande sa dépendance à la fois de la qualité des événements proposés et du contexte général dans lequel s'opère la reconnaissance (contraintes pragmatiques, répartition lexicale des locutions, descriptions phonémiques semblables ou très différentes des mots, etc ...).

Le problème fondamental qui consiste à adapter le système à un locuteur donné doit être séparé en deux opérations de nature très différente. Il s'agit d'une part d'ajuster au mieux les seuils et déterminer le poids à affecter aux différents paramètres servant à la segmentation et à la caractérisation des événements et, d'autre part, d'un apprentissage automatique (dirigé par le locuteur) permettant d'établir l'ensemble des règles d'interprétation phonétique de ces événements. Ces règles peuvent être d'autant plus larges (tolérer de multiples déviations) que les structures phonémiques des mots acceptables dans un même contexte sont plus éloignées. Nous travaillons actuellement à la réalisation d'un tel mécanisme d'apprentissage automatique.

9 - REFERENCES

CAELEN, J. 1979, Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique. Thèse d'Etat. Université Paul Sabatier, Toulouse.

COLE, R.A., JAKIMIK, J., 1980, How are syllabes used to recognize words ? JASA, Vol. 67, n° 5.

HATON, J.P., 1974, Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole. Thèse d'Etat, Nancy I.

LIENARD, J.S., 1972, Analyse, synthèse et reconnaissance automatique de la parole. Thèse d'Etat, Paris VI.

MELONI, H., GUIZOL, J., 1979, Identification d'événements pseudo-phonétiques pour la reconnaissance automatique de la parole continue. G.I.A. Faculté des Sciences de Luminy.

MELONI, H., GUIZOL, J., 1980, Un système de compréhension automatique de la parole continue. Congrès AFCET Informatique, Nancy.

VIVES, R., 1978, Utilisation de l'information phonémique et syllabique pour la reconnaissance de mots prononcés isolément ou dans des phrases. 10èmes J.E.P. pp. 375-384.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ANALYSE ET RECONNAISSANCE AUTOMATIQUE DE LA PAROLE AU C.R.I.N. : APPROCHES, MOTIVATIONS ET RESULTATS OBTENUS

Jean-Paul HATON et Jean-Marie PIERREL

C.R.I.N.
UNIVERSITE DE NANCY I
C.O. 140
54037 NANCY

.**RESUME**

Notre équipe s'intéresse depuis plus de 10 ans aux différents aspects de la reconnaissance et de la compréhension de la parole. Après la reconnaissance de mots isolés, nous avons dès 1973 abordé le problème de l'utilisation d'informations linguistiques dans la reconnaissance de langages de complexité croissante.

Dans cet article, nous tentons de retracer l'évolution de nos recherches en s'attachant, d'une part aux résultats obtenus, et, d'autre part, aux motivations qui ont conduit aux choix effectués en fonction de la complexité des tâches à résoudre.

AUTOMATIC RECOGNITION AND ANALYSIS OF SPEECH AT C.R.I.N. :

APPROACHES, MOTIVATIONS AND RESULTS

. ABSTRACT

Our group has worked for the past 10 years on several aspects of automatic speech recognition and understanding. We have developed several isolated word recognition systems using non linear time warping algorithms by dynamic programming. Since 1973 we have been using linguistic information for the recognition of sentences in languages of various complexity :

- isolated word sentences (numerical control of machine-tools),
- connected speech artificial languages (telephone exchange application),
- connected speech pseudo-natural languages (meteorological date base access application).

In this paper we present the evolution of our research effort for the past 10 years and we try to explain the motivations of the choices which were made in order to solve the problems.

I - INTRODUCTION

Le but de cet article est de présenter un bilan des activités de notre équipe dans le domaine de la reconnaissance et compréhension de la parole, permettant d'exposer et, dans la mesure du possible, expliquer les choix pris et les orientations données aux recherches. Notre équipe travaille depuis douze ans dans le domaine avec, comme objectif général, l'amélioration de la communication homme-machine. Nous avons été ainsi amenés à aborder les divers niveaux de traitement de la parole : acoustique, phonétique, syntaxique, sémantique, etc... Par ailleurs, les diverses réalisations relèvent de deux types différents de motivations :

- la mise en oeuvre de systèmes réels, adaptés :
  . aux mots isolés et aux séquences de mots,
  . puis aux langages artificiels,
  . puis aux langages pseudo-naturels ;

- la réalisation d'outils de recherche permettant de tester aisément divers paramètres, algorithmes et stratégies aux différents niveaux de traitement.

Tout au long de l'article, nous présenterons ces diverses réalisations en les illustrant d'exemples de résultats.

II - ANALYSE ACOUSTIQUE ET TRAITEMENT PHONETIQUE

Le choix à ce niveau est fonction à la fois de la complexité de la tâche à effectuer et d'impératifs de temps réel. Pour la reconnaissance globale de mots isolés, l'utilisation d'un banc de filtres (par exemple analyseur de Vocoder) s'est avérée suffisante, le nombre de filtres pouvant même être réduit (6 ou 8) dans le cas de vocabulaires simples. Dans cette approche globale, cette analyse fréquentielle constitue toute la phase d'analyse.

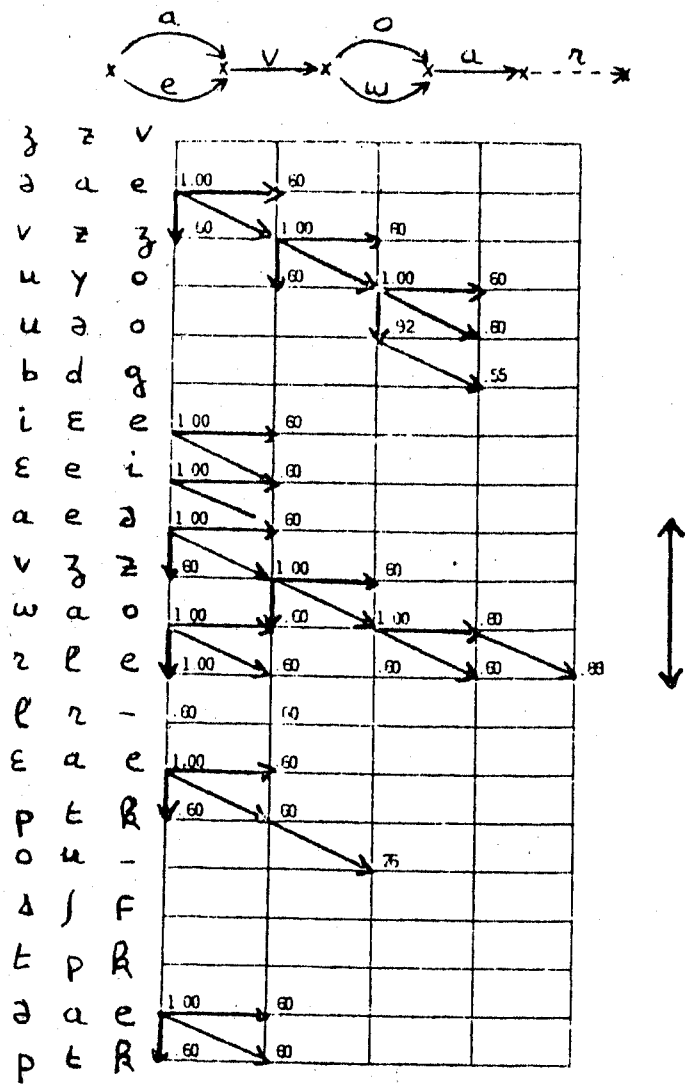
Dans nos approches analytiques, le but est d'obtenir une transcription de la parole sous forme d'un ensemble (chaîne ou treillis) de phonèmes ou d'évènements pseudo-phonétiques. On retrouve dans les choix effectués la distinction entre systèmes adaptés et recherche plus fondamentale signalée en introduction.

Pour les systèmes adaptés à un problème particulier (par exemple MYRTILLE I), pour lesquels les impératifs de temps de réponse sont importants, la solution de l'analyseur spectral a été retenue. Le modèle de transcription phonétique est de type "centiseconde", c'est-à-dire avec segmentation du spectre en tranches de durée fixe (10 ou 20 ms). Ces portions de spectres sont étiquetées par une méthode de reconnaissance de formes et regroupées ensuite en fonction de diverses heuristiques et d'informations prosodiques.

Dans une optique de recherche plus fondamentale, nous avons réalisé un outil conversationnel d'acquisition et d'analyse du signal vocal temporel. Dans ce logiciel est intégré un système de segmentation et de reconnaissance phonétique travaillant directement sur le signal numérisé de façon à ne pas perdre d'informations. La segmentation utilise divers paramètres temporels et fréquentiels pour déterminer des noyaux syllabiques et des zones d'instabilité. Les segments obtenus sont ensuite identifiés par une technique arborescente qui permet la prise en compte d'informations contextuelles.

III - RECONNAISSANCE DE MOTS ISOLES

Ce problème simplifié est le premier que nous avons abordé. La reconnaissance d'un vocabulaire limité (jusqu'à une centaine de mots), monolocuteur peut être résolue dans une approche globale, avec des techniques de reconnaissance de formes. Le principal problème est, dans ce cas, la normalisation en durée des mots. Depuis 1973, nous avons développé plusieurs algorithmes de comparaison dynamique qui fournissent une solution optimale à ce problème. Un algorithme sous-optimal, correspondant à la recherche d'un optimum local, a aussi été implanté sur microprocesseur. Nous n'insisterons pas sur ces techniques, reprises depuis par de nombreuses équipes. Signalons seulement que nous utilisons des techniques analogues pour la parole continue : vérification d'un mot sur une portion d'un treillis phonétique, localisation d'un mot dans une phrase, à partir de son spectre ou de sa transcription phonétique (la figure 1 en donne un exemple, tiré de (MARI-79)).



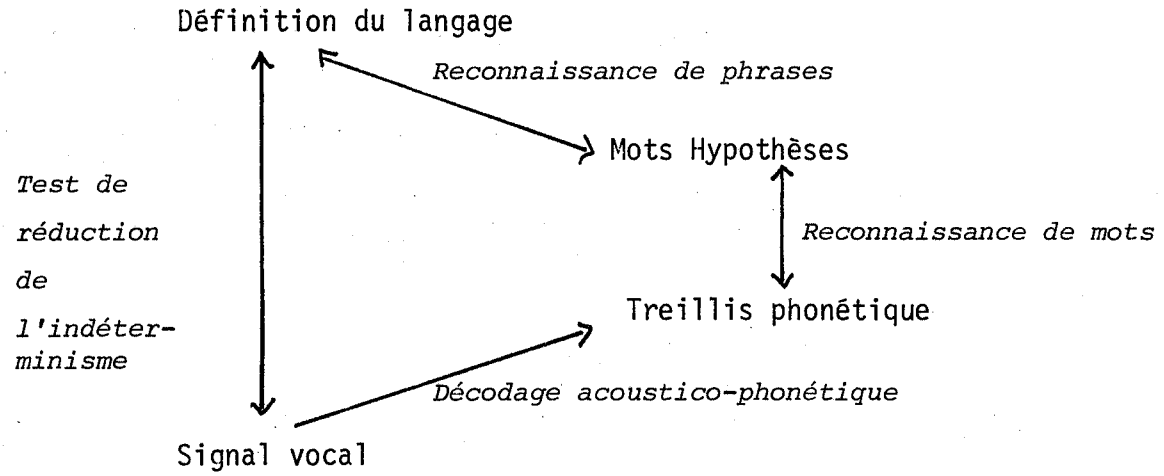
Localisation du mot "avoir /avwar/
dans la phrase "je voudrais avoir le poste 340"
(d'après J.F. MARI). Fig. 1

Les méthodes globales de comparaison dynamique demeurent valables pour des vocabulaires jusqu'à une centaine de mots. Pour de plus grands vocabulaires, nous avons choisi une approche analytique dans laquelle la programmation dynamique intervient au niveau de la comparaison des transcriptions phonétiques données par le système aux transcriptions de référence stockées dans le lexique.

IV - PRINCIPES DE BASE POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE

Pour l'instant, notre approche de la reconnaissance de phrases prononcées continûment est purement analytique (phonétique) ; elle utilise comme élément-pivot une représentation phonétique (chaîne ou treillis) de la phrase. Par ailleurs, les informations de type linguistique sont considérées comme une composante de la communication parlée, elles sont donc prises en compte à ce titre dans le processus de compréhension automatique.

On peut résumer ainsi les intéractions entre les diverses sources d'informations :



Nous avons par ailleurs retenu les principes suivants :

- les systèmes réalisés sont paramétrés par l'application. Ainsi dans le système MYRTILLE I, le langage (grammaire et lexique) est un paramètre qu'il suffit de redéfinir pour changer d'application. Dans MYRTILLE II, le vocabulaire de l'application est un paramètre. Ceci permet d'éviter la réalisation de systèmes trop ad hoc et alloue de plus une grande souplesse d'utilisation.

- Modularité maximale des systèmes de façon à faciliter les corrections et les changements d'algorithmes ou de stratégies.

- Utilisation du schéma général Hypothèse - Test - Validation, bien connu en intelligence artificielle comme base de la stratégie de compréhension. On verra que ce schéma a été très étendu dans le système MYRTILLE II.

- Importance du dialogue homme-machine. Ce dialogue est une composante fondamentale de nos systèmes, précisément pour la phase d'interprétation d'une phrase. De façon plus générale, ceci traduit l'importance accordée aux informations pragmatiques sur l'univers d'application.

V - RECONNAISSANCE DE PHRASES

V.1. EVOLUTION DES TYPES DE LANGAGES TRAITES

La complexité des langages que nous avons traités s'est progressivement accrue au cours du temps. On peut classer ces langages en 3 grandes familles :

(i) Langages d'états finis

Ce furent les premiers langages utilisés en reconnaissance au début des années 1970. Peu de temps après le premier système français utilisant des informations syntaxiques (TUBACH-70), J.P. HATON proposait dans sa thèse (HATON-74) un langage simple de commande numérique de machines outils. Il s'agissait alors de reconnaître des phrases énoncées mot à mot et le rôle des informations syntaxiques consistait à restreindre la reconnaissance à des sous-vocabulaires de taille réduite grâce à des traitements linguistiques très simples.

(ii) Langages artificiels

Forts de cette première expérience, nous abordions ensuite la reconnaissance et la compréhension de phrases du discours continu. Ces phrases à consonance naturelle étaient tirées d'un langage artificiel décrit par une grammaire hors contexte. La reconnaissance était alors dirigée par un analyseur syntaxique, les informations sémantiques étant fortement limitées. Le système MYRTILLE I,présenté en 1975 dans la thèse de J.M. PIERREL (PIERREL-75), permet de traiter de tels langages dont le facteur de branchement statique doit être limité à quelques dizaines.

(iii) Langages pseudo-naturels

Après ces langages artificiels, nous avons abordé la réalisation du système MYRTILLE II pour la compréhension de langages pseudo-naturels à syntaxe proche de celle du français parlé et dont le vocabulaire et la sémantique sont limités par l'univers de l'application envisagée. Pour de tels langages, les informations syntaxiques ne suffisent plus, il est donc nécessaire de mettre en oeuvre des traitements linguistiques plus complexes (HATON-80).

Dans la suite de cet article, nous nous intéresserons essentiellement aux traitements des langages artificiels et pseudo-naturels à travers l'étude des systèmes MYRTILLE I et MYRTILLE II.

V.2. SCHEMA DE PRINCIPE DE DEUX SYSTEMES

a) Le système MYRTILLE I

Nous nous limiterons ici à rappeler l'organisation générale de ce système dont on pourra trouver par ailleurs une description détaillée (PIERREL-75), (PIERREL-78). Utilisant comme paramètres la grammaire du langage et un lexique comportant les chaînes phonétiques de référence pour chaque mot, le système est centré sur un analyseur syntaxique qui émet des hypothèses, validées, dans une étape ultérieure, par un analyseur lexicographique incluant un module de reconnaissance phonétique de mots. Les résultats de

la reconnaissance sont ensuite pris en charge par une procédure de dialogue qui assure l'interface avec l'utilisateur. La figure V.1. fournit le schéma de principe de ce système.

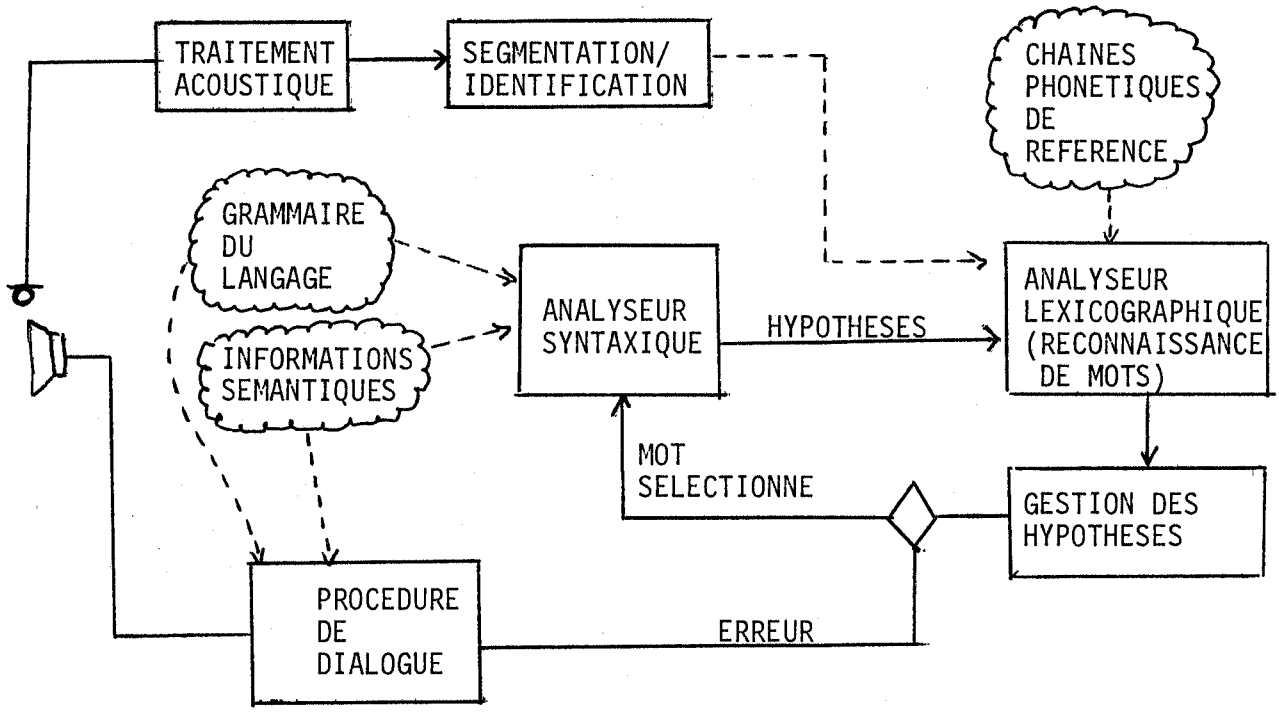


Schéma de principe du Sytème MYRTILLE I (fig. V.1)

b) Le système MYRTILLE II

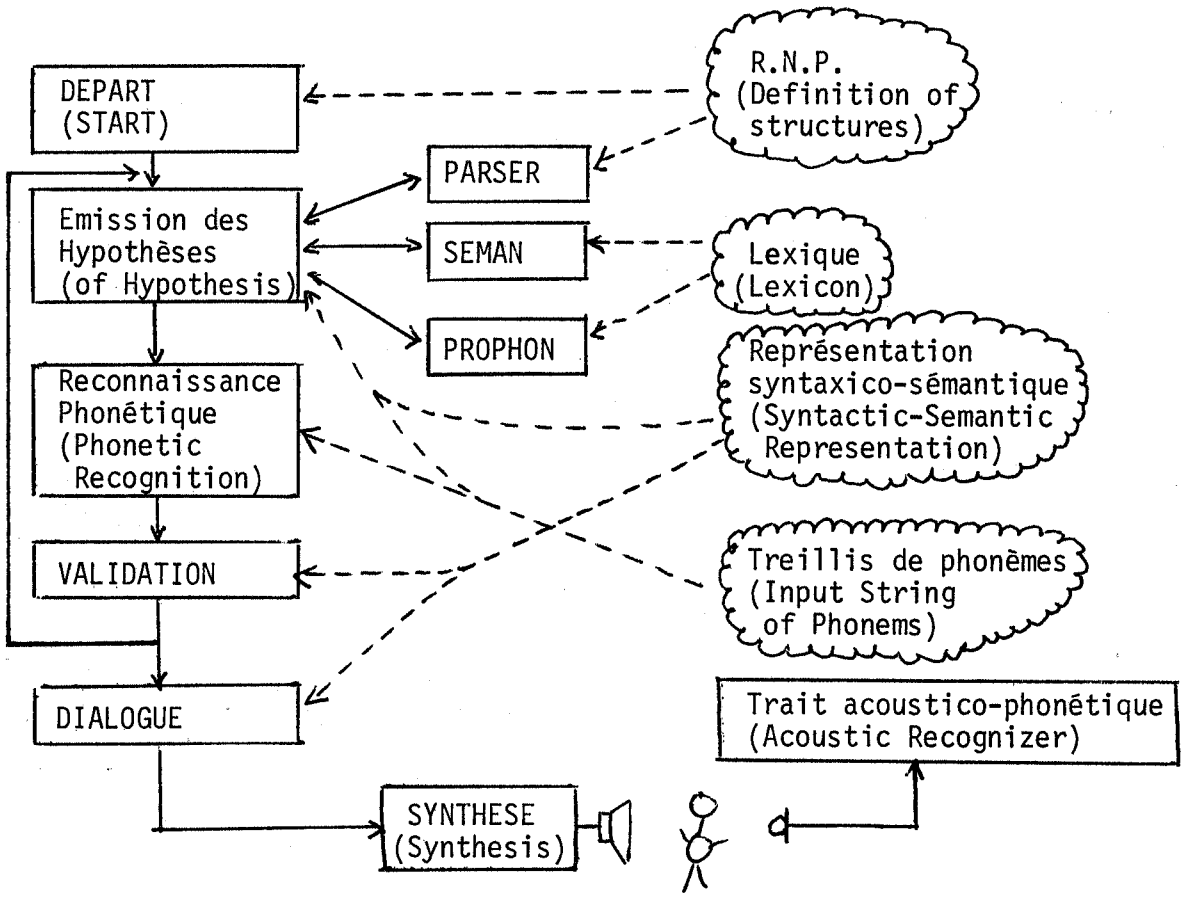
Comme pour le système MYRTILLE I, nous ne faisons ici que rappeler l'organisation générale du système MYRTILLE II dont on trouvera des descriptions détaillées dans (HATON-78), (PIERREL-79 a), (PIERREL-79 b). Le principe de base de la reconnaissance est toujours un principe d'Hypothèse et de Test (cf. par. III), mais les informations linguistiques mises en oeuvre dans ce processus ne se limitent plus seulement aux informations syntaxiques. Interviennent en effet :

(i) des informations liées au langage et à sa structure (PARSER) ;

(ii) des informations liées à l'application mise en oeuvre : le lexique regroupant les définitions syntaxico-sémantiques des mots (SEMAN) ;

(iii) des informations spécifiques à la parole : phonétiques, phonologiques et prosodiques (PROPHON).

La figure V.2. présente la structure générale de ce système qui utilise une représentation originale des informations linguistiques : les réseaux à noeuds procéduraux et le lexique détaillés dans (HATON-78).



Structure générale du Système MYRTILLE II (fig. V.2)

V.3. LES DIFFERENTES APPROCHES TESTEES (PIERREL-79 a)

Lors de la conception d'un système de compréhension automatique de la parole, divers choix sont possibles concernant en particulier : l'organisation générale, le type de l'analyse (ascendante et descendante), son sens (gauche droite ou du milieu vers les côtés) et les stratégies de recherche mises en oeuvre. Dans ce paragraphe, nous présentons les diverses approches que nous avons testées dans les systèmes MYRTILLE I et MYRTILLE II, en précisant les avantages et inconvénients de chacune d'elles.

a) Organisation générale

Les systèmes de reconnaissance peuvent schématiquement être répartis en deux classes :

(i) les systèmes hiérarchisés dans lesquels est déterminée a priori une hiérarchie entre les divers niveaux de traitement ;

(ii) les systèmes non hiérarchisés dans lesquels apparaissent différentes composantes ou sources d'informations mises en oeuvre en cours de la reconnaissance suivant les besoins. Le système HEARSAY II de Carnegie Mellon University (LESSER-75) qui présente diverses sources d'informations indépendantes les unes des autres et dont toutes les intéractions se font grâce à une base de données externe (le blackboard) est de ce type.

Pour notre part, il nous semble que le choix de l'organisation générale d'un système est fortement conditionné par le type de langage que l'on souhaite traiter. Pour les langages artificiels où la seule source d'informations, autre que phonétique, correspond à la grammaire du langage un système

hiérarchisé (la hiérarchie pouvant être différente d'un système à l'autre) nous semble particulièrement bien adapté. C'est en particulier le cas pour le système MYRTILLE I que nous avons développé et pour le système KEAL du CNET (MERCIER-78). Alors que le premier utilise une hiérarchie de type syntaxe-lexique, le second repose sur une hiérarchie inverse lexique-syntaxe. Les résultats obtenus pour des langages artificiels restreints sont comparables, mais une étude rapide du fonctionnement de ces systèmes montrent leur inadéquation à traiter des langages pseudo-naturels à grands vocabulaires (sous-ensemble du français parlé). En effet, lors du traitement de tels langages, on ne peut décider d'une hiérarchie unique et si certaines fois la syntaxe prend le pas sur la sémantique, dans d'autres contextes il faut entièrement inverser le processus, voire occulter complétement l'une des sources d'informations.

C'est pour ces raisons que nous avons adopté dans le système MYRTILLE II une solution intermédiaire où la hiérarchie entre les trois sources d'informations utilisées lors de l'émission des hypothèses (PARSER, SEMAN et PROPHON) est déterminée dynamiquement lors de l'exécution sur la base de critères qui sont eux-mêmes hiérarchisés afin de déterminer une hiérarchie type lorsque le système n'a aucun critère de choix.

b) Traitement ascendant et/ou descendant

Outre le type de hiérarchie mise en oeuvre, on peut aussi distinguer deux philosophies dans les systèmes existants, correspondant à deux approches très différentes :

(i) une approche descendante consistant à s'appuyer sur les niveaux linguistiques (structure du langage,...) ;

(ii) une approche ascendante consistant à partir du signal acoustique ou de sa représentation phonétique.

La première version du système MYRTILLE I fonctionnait de façon purement descendante. Nous avions opté pour cette solution car elle correspondait bien au principe d'Hypothèse et Test que nous avions choisi.

Notons à ce niveau qu'une approche descendante peut utiliser indifféremment un analyseur syntaxique ascendant ou descendant, la seule différence résidant alors dans la manière de construire l'arbre syntaxique.(Les limites de la langue font que, souvent, l'on confonde approche ascendante ou descendante et analyseur ascendant ou descendant, notions pourtant fort différentes : la première fait référence au sens du traitement : du résultat vers les données ou des données vers le résultat ; la seconde aux techniques d'analyses syntaxiques utilisées). Dans sa version descendante, le système MYRTILLE I utilisa d'ailleurs les deux types d'analyse syntaxique.

La seconde version du système MYRTILLE I (MARI-79) alterne les phases de reconnaissance guidées par la syntaxe et celles guidées par les données. En effet, en fonction de la reconnaissance de certains îlots de confiance localisés dans la chaîne phonémique d'entrée et des longueurs des différentes portions de phrases ainsi délimitées, le système ne lance l'analyse que de certains syntagmes. Cet aspect original permet de rejeter d'emblée un certain nombre de solutions et donc d'accélérer le processus de reconnaissance.

Dans le cas où le début de l'énoncé est de bonne qualité, les résultats obtenus par cet analyseur généralisé pour la reconnaissance de langages artificiels sont très comparables à ceux obtenus avec des analyseurs plus classiques; par contre cette méthode apporte beaucoup pour analyser des phrases dont le début est fortement entaché d'erreurs. De plus, nous pensons qu'un

tel analyseur est particulièrement bien adapté au traitement de langages pseudo-naturels où la syntaxe est loin d'être toujours respectée. Aussi, après avoir réalisé une version purement descendante du système MYRTILLE II, nous travaillons actuellement à l'implantation de cet algorithme pour permettre une poursuite de la reconnaissance après un échec syntaxique grâce à la détermination de points d'ancrage dans la partie d'énoncé non encore traité. Ces points d'ancrage ou îlots de confiance correspondent alors à des mots prédits sur critères essentiellement sémantiques et phonétiques.

c) Traitement gauche-droite ou du milieu vers les côtés

Fortement lié au point précédent, on peut aussi parler pour chaque système de reconnaissance de sens de traitement : gauche-droite ou du milieu vers les côtés.

Le premier semble particulièrement bien adapté à la parole car il correspond à un traitement du début vers la fin du message. Néanmoins, comme nous l'avons vu précédemment, un début de phrase peut être entaché d'erreurs et, dans ce cas, il est souvent préférable d'analyser la phrase à partir d'un mot bien reconnu et de poursuivre l'analyse de part et d'autre de l'îlot de confiance ainsi déterminé.

Le système MYRTILLE I peut fonctionner actuellement suivant ces deux principes. A terme, il devrait en être de même pour le système MYRTILLE II bien qu'actuellement seule la version gauche-droite soit opérationnelle.

d) Contrôle et stratégie

On peut à ce niveau distinguer :

. la stratégie générale
. le contexte du dialogue
. les heuristiques particulières.

(i) La stratégie générale

Pour éviter d'avoir à examiner de façon combinatoire toutes les solutions possibles d'interprétation d'une phrase, il est nécessaire d'utiliser une certaine stratégie pour sélectionner les interprétations les plus probables au sens des informations dont on dispose. Nous utilisons actuellement une stratégie de type meilleur d'abord qui, sans être très performante, réduit le plus à chaque instant l'espace de recherche. Son principal inconvénient réside dans le fait que la meilleure solution locale n'est pas toujours la meilleure globalement et que, de plus, cette stratégie comme la plupart des autres ne permet pas en cas d'échec d'analyse d'une partie du signal de fournir des résultats partiels de reconnaissance. C'est pourquoi dans le système MYRTILLE II, nous avons défini (PIERREL-79 b) les notions d'"élément principal" et de "consistance" qui permettent, en cas d'échec de la reconnaissance, de fournir une représentation partielle de la phrase sous la forme d'ensembles d'éléments principaux. Un tel ensemble est certes insuffisant pour comprendre entièrement la phrase mais il peut être utilisé par une procédure de dialogue pour poser des questions pertinentes à l'utilisateur.

Par ailleurs, nous envisageons le test d'autres types de stratégie : la prochaine version du système MYRTILLE II utilisera ainsi une recherche en faisceaux.

(ii) Contexte du dialogue

La procédure de dialogue intéractif,introduite dans nos deux systèmes, supervise aussi les traitements syntaxiques et sémantiques. Ainsi, par exemple, elle détermine le type de reconnaissance à effectuer pendant le dialogue particulièrement pour le traitement d'une réponse à une question en indiquant les mots-clés à rechercher ou en limitant la grammaire à un sous-langage correspondant aux réponses possibles.

(iii) Heuristiques particulières

Outre les deux points précédents, il faut ajouter deux heuristiques particulières que nous avons détaillées dans (PIERREL-75) et qui influent fortement sur les résultats obtenus par nos systèmes :

. le calcul des scores et la notion de seuild'abandon
. la possibilité de prendre certaines libertés par rapport à la grammaire en acceptant des élisions,substitutions ou insertions de mots sémantiquement non caractéristiques.

VI - RECONNAISSANCE ET/OU INTERPRETATION

Les chercheurs américains furent les premiers à faire la distinction entre Speech Recognition (reconnaissance) et Speech Understanding (compréhension).

Pour notre part, nous préférons maintenant parler de reconnaissance et/ou d'interprétation. En première approximation, nous pouvons définir ainsi ces deux termes : nous parlerons de reconnaissance si le résultat obtenu correspond à une liste linéaire des mots reconnus et d'interprétation si ce résultat correspond à une représentation syntaxico-sémantique de l'énoncé de départ. La différence n'est pas toujours aisée à effectuer car si l'on adjoint au système, comme ce fut notre cas, une procédure de dialogue dont le but est de fournir une réponse à l'énoncé de départ,on aura tendance à parler systématiquement d'interprétation. Pourtant, au vu de notre expérience, il nous semble que le niveau interprétatif peut être très divers. De nul ou quasi-nul dans le système MYRTILLE I qui pourtant "simulait" fort bien une standardiste comme le montrent les résultats rappelés dans le paragraphe suivant, le niveau d'interprétation a tendance à prendre de plus en plus de place dans le système MYRTILLE II., bien qu'il ne soit pas envisagé pour l'instant d'y connecter la base de données permettant de répondre aux questions !

Cela tient essentiellement au type des langages traités et aux modèles linguistiques mis en oeuvre. Alors que dans MYRTILLE I, après une phase de reconnaissance, la procédure de dialogue se limitait à une analyse de mots-clés, dans MYRTILLE II on peut véritablement parler d'interprétation car supprimer les incohérences, lever les ambiguïtés ou compléter l'énoncé reconnu nécessite un véritable traitement linguistique. La procédure de dialogue est alors très différente : à partir de la structure de surface de la phrase obtenue comme résultat de la reconnaissance sous forme d'arbre syntaxique, une première étape consiste à déterminer une structure profonde que nous avons choisi de représenter suivant le formalisme des grammaires par cas. Lors de cette phase d'analyse sémantique, nous détectons les ambiguïtés ou les incohérences de la reconnaissance. Dans un second temps, la procédure de dialogue interprète cette structure profonde en utilisant la notion de scénario et est amenée ainsi à engendrer des questions visant à compléter la structure profonde de la phrase. On obtient alors un schéma du type de celui-ci de la figure V.3.et il nous semble

que seul le passage à une représentation plus ou moins profonde de la phrase autorise à parler d'interprétation sans abus de langage.

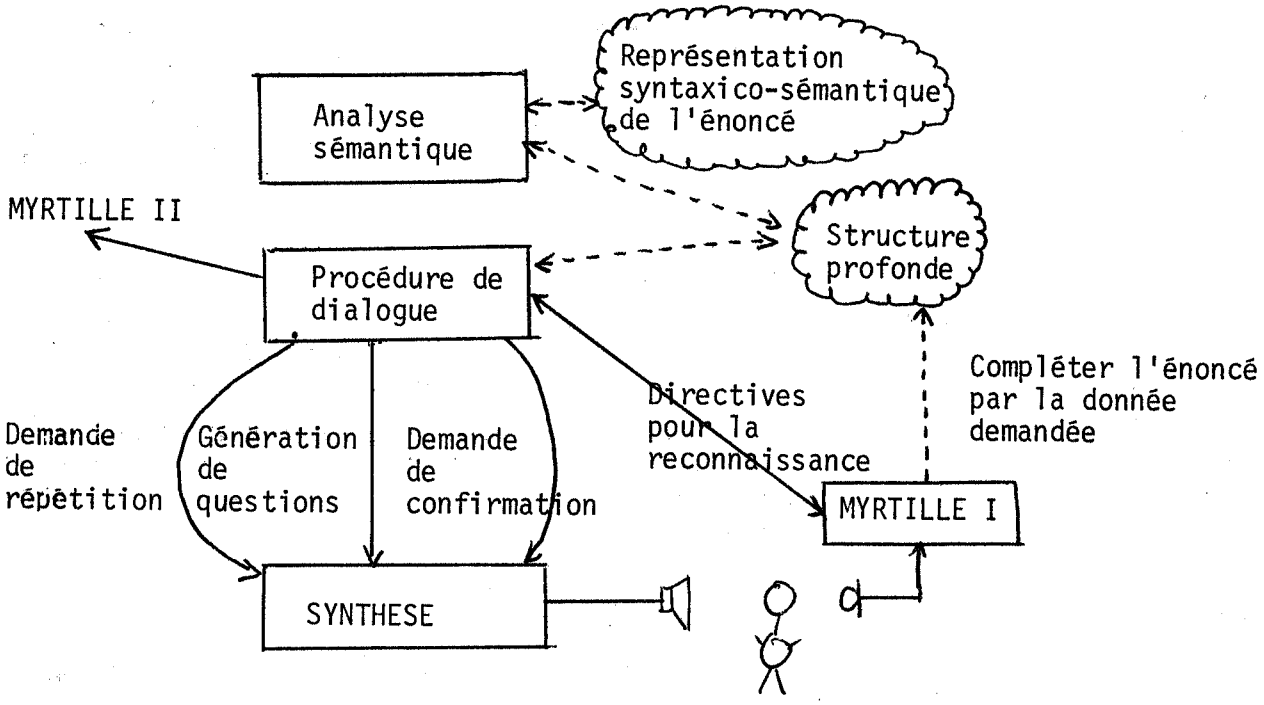


Schéma succint de la procédure de dialogue dans le système MYRTILLE II (fig. V.3.)

VII - RESULTATS OBTENUS

a) Système MYRTILLE I

Le système MYRTILLE I, opérationnel depuis fin 1975 sur IRIS 80, fonctionne maintenant sur le MITRA 125 du laboratoire avec un traitement acoustico-phonétique de type centi-secondes (cf. par.II).Le temps de traitement nécessaire pour la reconnaissance d'une phrase tirée du langage de communication avec un standard téléphonique que nous avons défini dans (PIERREL-75) est de l'ordre de quelques secondes en mode MITRA 15.

Pour les réponses vocales, nous avons opté actuellement pour un système de restitution de parole enregistrée, système largement suffisant pour une application de ce type.

Actuellement, nous sommes en cours d'implantation d'une nouvelle version de ce système pour laquelle nous serons en mesure, lors des journées d'études, de fournir des résultats chiffrés de reconnaissance.

Voici à titre de rappel quelques exemples de dialogue obtenu avec le système MYRTILLE I :

DEBUT DE SESSION

a) DEBUT DE COMMUNICATION
STANDARD AUTOMATIQUE DE .... JE VOUS ECOUTE
... COR : 'JE VOUDRAIS POSTE 339'
DESOLE, JE N'AI PAS COMPRIS. VEUILLEZ REPETER, JE VOUS ECOUTE
... COR : 'JE VOUDRAIS LE POSTE 339'
VOUS AVEZ BIEN DEMANDE LE POSTE 339 ?

...COR : 'OUI'
LE POSTE 339 EST OCCUPE. VOULEZ-VOUS PATIENTER ?
...COR : 'NON MERCI. JE RAPPELLERAI'
FIN DE COMMUNICATION

b) DEBUT DE COMMUNICATION
STANDARD AUTOMATIQUE DE ...., JE VOUS ECOUTE
...COR : 'JE VOUDRAIS PARLER A MADAME BOUCHET'
JE VOUS PASSE MME BOUCHET
VOUS AVEZ EN LIGNE MME BOUCHET
FIN DE COMMUNICATION

c) DEBUT DE COMMUNICATION
STANDARD AUTOMATIQUE DE ...., JE VOUS ECOUTE
...COR : 'JE VOUDRAIS AVOIR LE POSTE 340'
VOUS AVEZ BIEN DEMANDE LE POSTE 339 ?
...COR : 'NON LE POSTE 340'
VOUS AVEZ EN LIGNE LE POSTE 340

b) Système MYRTILLE II

La première version du système MYRTILLE II est actuellement en cours de test et pour elle aussi nous serons en mesure de fournir des résultats chiffrés de reconnaissance pour les journées d'étude.

Les caractéristiques principales de cette version sont :

. application test : centre de renseignements météorologiques

. vocabulaire : 375 mots

. traitement acoustico-phonétique de type centi-seconde

. reconnaissance phonétique par programmation dynamique

. stratégie utilisée : gauche droite et meilleur d'abord.

Les premiers résultats obtenus sont très prometteurs. La phase d'émission des hypothèses sur critères syntaxiques et sémantiques permet pour un tel langage de restreindre le nombre de mots effectivement testés de 1/20 à 1/10 (soit 10 à 20 mots testés pour 1 mot effectivement présent dans la phrase d'entrée, articles exclus) pour des phrases du type :

- Est-ce qu'il pleut sur NANCY ?
- Je voudrais savoir s'il gélera le mois prochain
- Quand est-ce que la température augmentera ?
- Le soleil percera-t-il les nuages demain ?
- Je voudrais connaître la vitesse du vent en Bretagne
- Est-ce qu'il risque de neiger bientôt dans les Vosges ?
- Je voudrais connaître le temps qu'on prévoit pour le mois de Janvier
- Est-ce qu'il faut s'attendre à des risques matinaux de petites plaques de verglas en Lorraine ?
- ...

Les résultats sont fournis sous la forme de listes chainées représentant l'arbre syntaxico-sémantique de la phrase traitée comme nous l'avons montré dans (HATON-80) et (PIERREL-80).

VIII - CONCLUSION

Nous avons rapidement présenté les motivations de nos recherches, les approches testées et les résultats obtenus pour l'analyse et la reconnaissance de la parole au sein de l'équipe "Reconnaissance des formes et Intelligence artificielle" du C.R.I.N. Il faudrait aussi y ajouter des recherches qui débutent au sein de l'équipe tant sur l'étude de la prosodie que sur les systèmes multilocuteurs.

Si l'on veut obtenir assez rapidement des systèmes "intelligents" de communication Homme-Machine, il nous semble qu'un effort soutenu doit se poursuivre sur deux axes complémentaires :

. décodage acoustico-phonétique de la parole, formalisation et utilisation des informations prosodiques ;

. traitements linguistiques en liaison indispensable avec des disciplines connexes fondamentales ou appliquées : psychologie cognitive, intelligence artificielle, etc...

Car si les systèmes commercialisés aujourd'hui sont des systèmes à mots isolés, on peut prédire, sans trop se tromper,que les systèmes de communication orale avec la machine devront accepter, dans la décennie qui vient, des langages de plus en plus naturels pour répondre aux espoirs actuellement formulés.

BIBLIOGRAPHIE

(HATON-74) J.P. HATON "Contribution à l'analyse, la paramétrisation et la re reconnaissance automatique de la parole". Thèse d'Etat, Université de NANCY I, 1974.

(HATON-78) J.P. HATON et J.M. PIERREL "Data Structure and Organization of MYRTILLE II System" 4ch IJCPR, KYOTO, Japan, Nov. 1978.

(HATON-80) J.P. HATON, J.F. MARI, J.M. PIERREL, S. SABBAGH "Représentation et mise en oeuvre de contraintes syntaxiques et sémantiques en reconnaissance du discours continu". Séminaire syntaxe et sémantique, GALF-AFCET, RENNES, Sept. 1980.

(LESSER-75) V.R. LESSER and al. "Organization of the HEARSAY II Speech Understanding System" IEEE, TASSP, vol. ASSP-23, pp. 11-23, 1975.

(MARI-79) J.F. MARI "Contribution à l'analyse syntaxique et à la recherche lexicale en reconnaissance du discours continu". Thèse de 3ème Cycle Université de NANCY I, Juin 1979.

(MERCIER-78) G. MERCIER, P. QUINTON, R. VIVES "KEAL : Un système pour un dialogue oral avec une machine" Congrès AFCET, III, PARIS, Nov. 1978, pp. 304-314.

(PIERREL-75) J.M. PIERREL "Contribution à la reconnaissance automatique du discours continu". Thèse de Spécialité, NANCY, Nov. 1975.

(PIERREL-78) J.M. PIERREL "Un système de compréhension du discours continu utilisant des contraintes morphologiques, syntaxiques et sémantiques". Revue RAIRO Informatique, vol. 12, n° 2, 1978.

(PIERREL-79 a) J.M. PIERREL, J.P. HATON "Organisation et stratégie d'un système de compréhension automatique de la parole continue". Congrès AFCET, RF et IA, TOULOUSE, Sept. 1979.

(PIERREL-79 b) J.M. PIERREL, J.P. HATON "The MYRTILLE II Speech Understanding System" in Spocken Language Generation and Understanding edited by J.C. SIMON, D. REIDEL PUBLISHING COMPAGNY, 1980 (acte de l'école OTAN de BONAS 1979.

(PIERREL-80) J.M. PIERREL, J.P. HATON "Syntactic-semantic Interpretation of Sentence in the MYRTILLE II Speech Understanding System". IEEE, ICAPSSP-80, DENVER, Colorado, April 1980.

(TUBACH-70) J.P. TUBACH "Reconnaissance automatique de la parole". Thèse d'Etat, GRENOBLE, Juillet 1970.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

UTILISATION D'UN SYSTEME DE RECONNAISSANCE ANALYTIQUE DE LA PAROLE POUR LA DETECTION DE FAUTES DE PRONONCIATION

JANE LE BRAS

CNET, LAA/TSS/DAT
UER DE LINGUISTIQUE

.**RESUME**

Le travail présenté ici se situe dans le cadre des études menées sur la reconnaissance automatique de la parole ; il s'agit d'une tentative d'évaluation automatique des erreurs de prononciation commises par des sujets anglophones apprenant le français. Après avoir brièvement rappelé les travaux antérieurs dans le domaine de la rééducation vocale assistée par ordinateur, nous exposerons les problèmes liés au choix de la méthode de reconnaissance et évoquerons les contraintes spécifiques à l'utilisation d'une méthode de reconnaissance analytique. Nous décrirons ensuite un algorithme mis au point pour rendre compte de la présence ou de l'absence de l'aspiration sur les consonnes occlusives sourdes. Enfin, nous donnerons les résultats de tests effectués avec des locuteurs anglais et français sur un programme contenant cet algorithme et fonctionnant à partir des résultats du module phonétique de Kéal mis en oeuvre au CNET (MERCIER, 1976).

ON THE USE OF A SPEECH RECOGNITION SYSTEM FOR THE DETECTION OF MISPRONUNCIATIONS

. ABSTRACT

The experiment presented here is a first attempt at automatic evaluation of errors of prononciation made by English subjects learning French. After briefly describing previous work using computers in the teaching of pronunciation, problems connected with the choice of a speech recognition system will be discussed. Then, we will describe the method chosen and the constraints specifics to the using of an analytic speech recognition system. An algorithm for detecting the presence or the absence of aspiration in stop consonants will be presented. In the last part of this work, we will give the results of experiments carried out with five English subjects and seven French subjects using a program employing this algorithm. This program is simply intended to facilitate the testing of algorithms and to allow research with several different speakers. It has not been written in real time, but in batch (IRIS 80). Its principal interest comes from the fact it uses output of the phonetic analyser of Keal, the speech recognition system used in CNET.

UTILISATION D'UN SYSTEME DE RECONNAISSANCE ANALYTIQUE DE LA PAROLE POUR LA DETECTION DE FAUTES DE PRONONCIATION.

Jane LE BRAS

C.N.E.T., LAA/TSS/DAT
U.E.R. de Linguistique de Rennes

1. INTRODUCTION

Les recherches entreprises en reconnaissance automatique de la parole ont surtout été menées dans le but de faciliter le dialogue homme-machine en langage naturel. Ce serait sans doute restreindre les progrès que peut apporter la reconnaissance que de limiter son rôle au simple remplacement des claviers des systèmes informatiques par des micros ou des téléphones. L'introduction de la reconnaissance de la parole devrait accroître l'essor d'un bon nombre de techniques et leur ouvrir de nouveaux débouchés.

Dans la cas de l'enseignement assisté par ordinateur, l'emploi de la reconnaissance doit permettre d'étendre et de perfectionner l'utilisation de l'enseignement assisté à tous les domaines ayant trait à l'expression orale : rééducation des mal-entendants et des handicapés de la parole, cours de diction, apprentissage des langues, etc... L'évolution simultanée de la technologie des ordinateurs et des systèmes de reconnaissance permet de prévoir, à moyen terme, des systèmes d'enseignement assisté de la prononciation, entièrement autonomes, c'est-à-dire fonctionnant sans intervention d'un professeur et pouvant servir à un très grand nombre d'utilisateurs.

Le travail présenté ici se situe dans le cadre des études menées sur la reconnaissance automatique de la parole ; il s'agit d'une tentative d'évaluation automatique des erreurs de prononciation commises par des sujets anglophones apprenant le français. Après avoir brièvement rappelé les travaux antérieurs dans le domaine de la rééducation vocale assistée par ordinateur, nous exposerons les problèmes liés au choix de la méthode de reconnaissance et évoquerons les contraintes spécifiques à l'utilisation d'une méthode de reconnaissance analytique. Nous décrirons ensuite un algorithme mis au point pour rendre compte de la présence ou de l'absence de l'aspiration sur les consonnes occlusives sourdes. Enfin, nous donnerons les résultats de tests effectués avec des locuteurs anglais et français sur un programme contenant cet algorithme et fonctionnant à partir des résultats du module phonétique de Kéal mis en oeuvre au C.N.E.T. (MERCIER, 1976).

2. TRAVAUX ANTERIEURS

L'idée d'utiliser les techniques de traitement du signal de parole en matière de correction phonétique est récente et a surtout été exploitée dans la réalisation de systèmes de rééducation de sujets mal-entendants. Plusieurs équipes travaillent au développement de tels systèmes :

- STEVENS et NICKERSON (1972) en sont les précurseurs avec un système conçu pour les sourds. L'analyse de parole est réalisée à partir d'un spectre sur 19 canaux, d'un détecteur de nasalité et d'un extracteur de pitch. Un mini-ordinateur analyse les défauts de prononciation et les visualise sur l'écran, sous forme de jeu attractif.
- en France, l'équipe de Nancy (HATON, 1973) travaille sur le projet SIRENE depuis 1973. Leur système, analogue à celui de Stevens (analyseur de 32 filtres, détecteur de mélodie), utilise cependant l'apport de la reconnaissance

de la parole pour la prononciation des sons isolés. Les erreurs de prononciation sont évaluées au moyen d'un mini-ordinateur et visualisées sur écran ; des exercices de correction sont présentés sous forme de jeux.

En ce qui concerne l'apprentissage des langues, signalons les travaux de KALIKOW et SWETS (1972), qui ont développé un système de correction des erreurs de prononciation pour des Espagnols apprenant l'anglais (la prononciation de l'élève est comparée à une forme de référence mise en mémoire, le signal de parole est analysé spectralement). L'équipe de Fujimura a également travaillé sur un projet similaire : un professeur intervient pour juger de la prononciation, l'ordinateur sert à mémoriser les performances de l'élève et à choisir la difficulté suivante. Le travail le plus récent a été réalisé par GIORGETTI et LAMOTTE (1980) sur l'apprentissage de l'espéranto. Le système fonctionne à partir d'une méthode de reconnaissance globale utilisant la programmation dynamique, les exercices sont réalisés sur des mots isolés. Plusieurs expériences ont été menées sur l'apprentissage des paramètres supra-segmentaux (FOURCIN et ABBERTON, 1975 ; LEON et MARTIN, 1972).

3. CHOIX ET DESCRIPTION DE LA METHODE

3.1. Méthode de reconnaissance globale ou analytique

On distingue deux approches dans le domaine de la reconnaissance vocale :

- la méthode globale consiste à effectuer une comparaison globale (par exemple de mots) à l'aide de paramètres acoustiques, entre des formes de référence stockées en mémoire et ce qui a été prononcé. Cette méthode est particulièrement efficace lorsque le locuteur est connu de la machine (phase d'apprentissage), mais ne semble réellement adaptée qu'à un vocabulaire limité.
- la méthode de reconnaissance analytique consiste à effectuer une segmentation du signal en unités pseudo-phonémiques (syllabes, diphones, etc...) et à les identifier. Cette méthode se prête mieux que la précédente à la reconnaissance multi-locuteurs et au traitement de la parole continue.

Actuellement, ces deux méthodes se développent parallèlement selon le type de l'application choisie. Les résultats de la méthode globale sont satisfaisants tant du point de vue du taux de reconnaissance que du temps de réponse. Les problèmes posés par l'utilisation d'une méthode de reconnaissance analytique sont liés essentiellement à la segmentation du signal et à l'identification des segments. Ceux-ci impliquent, du fait de la complexité des algorithmes employés, un temps de réponse encore peu compatible avec le temps réel.

Dans le cadre de la mise au point de programmes d'évaluation automatique de la prononciation, la reconnaissance globale présente deux avantages :

- elle ne nécessite au préalable aucune étude acoustique contrastive, car il suffit de mémoriser deux formes référentielles : l'une correspondant à la prononciation du professeur et l'autre étant la prononciation erronée, que l'on compare à la prononciation effective.
- elle est capable defournir à l'élève une réponse quasi-immédiate.

Par contre, elle ne permet ni de localiser, ni d'identifier et de dégager les principaux corrélats de la faute. En partant du principe que tous les évènements perçus comme des erreurs de prononciation sont associés à des modifications d'ordre acoustique, et que ces modifications peuvent être traduites en termes d'algorithmes simples, une méthode de reconnaissance analytique doit être en mesure de diagnostiquer toute erreur. D'autre part, en créant pour chaque faute un seul dictionnaire de référence et un ensemble de règles acoustiques (contextuel-

les ou non) issues d'une analyse comparative, cette méthode autorise une reconnaissance multi-locuteurs sans apprentissage.

On a cherché à ce que les programmes répondent aux contraintes suivantes :

- temps de réponse compatible avec le temps réel ;
- identification du phonème sur lequel la faute est prévue ;
- reconnaissance multi-locuteurs ;
- possibilité de traiter de la parole continue.

Compte tenu de ces objectifs, la méthode analytique nous a semblé la mieux adaptée à notre étude.

3.2. Description de la méthode

3.2.a. Principe

Entre le moment où le locuteur parle et celui où il se voit attribuer une réponse, le signal subit un certain nombre de traitements qui sont schématisés sur la figure 1.

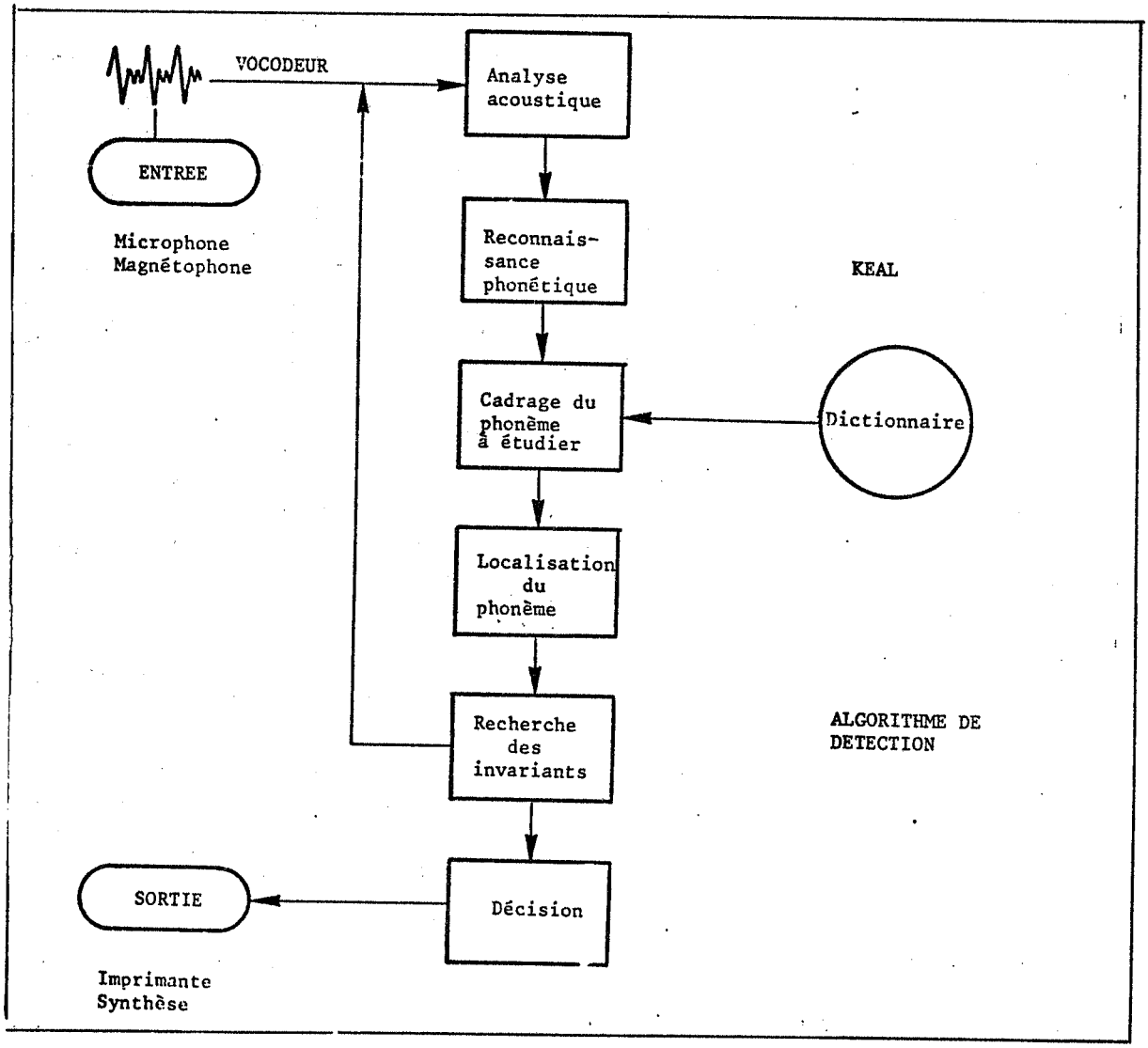


Figure 1 : Principe d'un système de correction phonétique assisté par ordinateur et utilisant le module acoustico-phonétique de Kéal.

Un programme de correction phonétique assisté par ordinateur et utilisant une méthode de reconnaissance analytique (Kéal) ne nécessite pas la reconnaissance de tout ce qui a été prononcé par le locuteur. La méthode consiste à n'extraire du signal que le phonème sur lequel l'hypothèse d'erreur a été émise. Il s'agit donc de programme de reconnaissance d'erreurs, où la machine connaît ce qui a été prononcé, le phonème susceptible d'être erroné, et enfin les paramètres acoustiques devant être inspectés pour juger de la prononciation de ce phonème. En entrée, un locuteur appartenant à une des deux nationalités fixées au préalable prononce des mots ou des phrases appartenant à un corpus donné. Ce corpus est mémorisé dans un fichier appelé dictionnaire. En sortie, une décision prise par la machine est communiquée à l'utilisateur.

A partir du signal, l'enchaînement des traitements se fait de la manière suivante :

- analyse acoustique du signal ;
- reconnaissance phonétique (hypothèses sur les phonèmes prononcés) ;
- cadrage du phonème ;
- algorithme de détection.

3.2.b. Description des différents traitements

a. Analyse acoustique du signal

Le signal est d'abord analysé au moyen d'un vocoder à 14 canaux dont la période d'échantillonage est fixée à 13,3 ms (FERRIEU et PERSON, 1968 ; ZURCHER et FRICHOU, 1972). A l'issue de cette phase, on obtient une représentation spectrale et temporelle du signal, qui couvre un peu plus que la bande passante téléphonique, par l'intermédiaire de 14 canaux répartis de façon non-linéaire.

b. Reconnaissance phonétique

Le spectre est ensuite analysé par les programmes de reconnaissance phonétique de Kéal. Le module phonétique segmente le signal en unités minimales ou phonèmes, et assigne à chaque unité une ou plusieurs étiquettes auxquelles on accorde un taux de probabilité. En fin de reconnaissance, le signal de parole est transformé en une suite pondérée de phonèmes probables sous forme de treillis. Dans Kéal, ce treillis phonémique est ensuite utilisé par l'analyseur lexical, qui a pour tâche d'y retrouver le codage idéal du mot prononcé. Dans notre étude, les résultats de la reconnaissance ne sont pas communiqués au module lexical, mais à un programme de cadrage (QUINTON et LETY, 1977 ; VIVES, 1979).

c. Cadrage du phonème

Le programme de cadrage, destiné à faire de l'adaptation au locuteur, établit la correspondance entre le treillis phonémique et le codage standard (il permet également de faire des statistiques sur le pourcentage de phonèmes correctement reconnus dans le treillis). En fin de cadrage, chaque mot est caractérisé par une suite de phonèmes possédant chacun une adresse de début et une adresse de fin dans la représentation acoustique. L'utilisation d'un tel programme autorise donc la localisation du phonème sur lequel on prévoit une faute et la vérification de la présence des indices attendus.

d. Algorithme de détection

L'algorithme de détection est spécifique de la faute commise. Les étapes intervenant dans la mise au point de tels algorithmes sont les suivantes :

- localisation du phonème sur lequel l'hypothèse d'erreur a été prédite ;
- recherche des indices corrélés à la faute par vérification sur le spectre ;
- décision transmise à l'utilisateur (cette décision est communiquée à l'élève sous forme binaire : bonne ou mauvaise prononciation).

3.2.c. Problèmes soulevés par l'utilisation de cette méthode

Les problèmes posés par l'utilisation d'une telle méthode ont trait à la localisation du phonème sur lequel la faute est prévue et au temps de réponse exigé par la machine.

a. Localisation

Compte tenu des performances actuelles du système de reconnaissance Kéal et du type d'application envisagé, l'utilisation d'un programme de cadrage en amont des programmes du module acoustico-phonétique s'est révélée indispensable. De cette manière, on espère neutraliser les erreurs de segmentation du module phonétique et obtenir une localisation correcte du phonème à étudier. Une fois cette première localisation opérée, il est également possible d'aller vérifier sur le spectre l'information transmise par les modules inférieurs.

b. Temps de réponse

Un programme de détection automatique de fautes de prononciation doit être en mesure de fournir à l'élève une réponse quasi-instantanée. L'utilisation du seul module phonétique et d'un programme de cadrage permet de prévoir, pour l'implantation de tels programmes en temps réel, une diminution importante du temps de réponse actuellement exigé lors de la mise en oeuvre de l'ensemble des modules de Kéal. De même, il est probable que le temps de réponse, étroitement lié aux contraintes informatiques (type d'ordinateur, programmes de traitement de parole très longs), va diminuer de façon considérable dans le futur avec l'essor de nouveaux moyens technologiques.

4. INDICES RETENUS POUR LA DETECTION DE L'ASPIRATION

4.1. Définition de l'aspiration

L'analyse acoustique du trait de voisement a révélé la présence d'une multiplicité d'indices pouvant être hiérarchisés différement selon la langue étudiée. En Français, l'opposition phonologique de voisement repose essentiellement sur la présence ou l'absence de vibrations des cordes vocales dont la principale composante acoustique se traduit par la barre de voisement. En Anglais, l'action des cordes vocales ne contribue pas obligatoirement à la distinction de sonorité, en position initiale et accentuée, l'indice le plus important est celui de l'aspiration, dans les autres positions, la hiérarchie des indices est la même qu'en Français.

D'un point de vue articulatoire, l'aspiration est définie de manière traditionnelle comme un souffle s'intercalant entre la consonne et le début de la voyelle subséquente. Ceci a donné lieu à plusieurs interprétations successives (JONES, 1956 ; Heffner, 1969 ; KIM, 1970, DIXIT, 1979).

D'un point de vue acoustique, l'aspiration se caractérise par une structure formantielle très proche de la consonne H,variable selon le contexte (FANT, 1969 STEVENS et KLATT, 1974). En général, la structure formantielle est réduite aux formants supérieurs car, du fait de sa faible intensité, le F1 n'apparaît pas.

4.2. Indices retenus

La méthode généralement adoptée en correction phonétique, qu'il s'agisse de l'apprentissage d'une langue ou de la correction d'un défaut de prononciation, consiste à analyser et décrire le système d'origine par rapport au système visé, pour les comparer ensuite par le biais d'une analyse comparative. L'étude précédant la mise au point des algorithmes consiste à rechercher, en utilisant le même matériel d'analyse que celui du système de reconnaissance, les indices permettant de rendre compte de l'erreur et les valeurs de seuils correspondants. Bien que le vocodeur à 14 canaux ne soit pas spécialement approprié à une étude

phonétique des consonnes et, à fortiori, à celle des occlusives, ses possibilités nous ont suffi amplement dans le cadre d'une analyse contrastive. A titre d'exemple, on décrit ici les différentes phases d'un algorithme permettant de distinguer l'occlusive bilabiale /P/, de son homologue aspirée /Ph/. L'algorithme se présente ainsi :

- vérification de l'information communiquée par Keal
  On vérifie sur le spectre que la localisation de l'occlusive à étudier est juste, c'est à dire que l'adresse communiquée par le module phonétique de Keal correspond effectivement à la portion de spectre que l'on veut examiner. Cette vérification permet de "recadrer" le phonème et d'éliminer quelques petites erreurs de segmentation (elle ne permet pas d'éliminer des erreurs plus importantes telles qu'une mauvaise segmentation en syllabes). Pour ce réajustement, on est contraint d'utiliser sensiblement les mêmes paramètres que ceux du module phonétique de Keal. On cherche d'abord à délimiter une zone de silence correspondant à l'occlusion de la consonne (plusieurs échantillons vocodeur consecutifs où l'énergie est inférieure à un seuil). On élimine également le bruit qui peut se superposer à la période de silence (par le calcul du centre de gravité du spectre à l'instant t et la valeur de l'énergie dans les hautes fréquences). Dès que l'énergie globale augmente de façon significative, la période d'occlusion est terminée et il est possible de rechercher les indices témoignant de la présence de la faute.

- test sur l'évolution de l'énergie globale (somme de l'énergie codée sur les 14 canaux).
  Une fois l'adresse vérifiée, on teste l'évolution de l'énergie globale durant les 80 msec consécutives à l'occlusion de la consonne. Ce seuil temporel correspond à la durée maximale de l'aspiration observée lors d'une étude préliminaire effectuée sur un corpus de 30 mots où l'on avait remarqué :

  $26 < d < 40$ msec — dans 70% des cas.

  $40 < d < 60$ msec — dans 17% des cas.

  $60 < d < 80$ msec — dans 3% des cas.

  (d = durée de l'aspiration).

  Dans ces 80 msec, on peut identifier deux situations caractéristiques :

  a. montée et chute rapide de l'énergie, avec une valeur de cette dernière faible, un V.O.T. (temps d'établissement du voisement) long, et un F1 peu intense. Dans ce cas, nous sommes en présence d'une aspiration.

  b. montée importante et continue de l'énergie, V.O.T. bref et F1 élevé. Dans ce cas, il n'y a pas d'aspiration.

  Ces deux profils caractéristiques couvrent la majorité des occurences et permet de prendre une décision avec certitude. Dans les autres cas, nous pouvons dire qu'une croissance rapide et continue de l'énergie associée à une valeur relative de l'énergie importante, un V.O.T faible, un F1 élevé, la présence de formants supérieurs, sont des indices d'absence d'aspiration. A l'inverse, une croissance lente de l'énergie ou éventuellement suivie d'une baisse, une valeur relative de l'énergie faible, une répartition de l'énergie dans les hautes fréquences, un V.O.T. élevé, un F1 faible, sont des indices de présence d'aspiration. Si tous ces indices ne convergent pas, une hiérarchie entre eux et la prise en compte du contexte permettent de conclure avec toutefois moins de certitude que précédemment.

5. RESULTATS DU PROGRAMME

5.1 Objectifs du programme

Le programme dont nous présentons les résultats est un programme de simulation, destiné à mettre au point des algorithmes et à les tester sur plusieurs locuteurs. Pour ces raisons, il n'a pas été conçu pour une utilisation en temps réel, comme indiqué sur le schéma de principe, mais a été ecrit en "batch", sur Iris 80, avec en entrée des fichiers constitués d'enregistrement de parole numérisés. Afin de pouvoir être analysés, les résultats ont été édités sur "listing" ainsi que les échantillons de parole correspondants (voir exemple en annexe).

5.2 Déroulement des tests

Pour qu'un programme de correction phonétique soit véritablement efficace et que ses performances soient proches de celles de l'oreille humaine, il faut qu'il soit en mesure de prendre une décision face à deux situations : cas de bonne prononciation et cas de mauvaise prononciation, il faut également qu'il soit capable d'identifier la nature de la faute commise afin de pouvoir s'integrer dans un véritable système d'enseignement assisté par ordinateur et non pas être un simple outil de contrôle.

Nous avons donc soumis le programme à deux types d'essais :

- Locuteurs français, reflet de la prononciation modèle ;
- Locuteurs anglophones, dont la prononciation a été estimée, après un examen auditifs, comme erronée.

Ces essais ont été menés sur un corpus constitué de 20 mots isolés et de 5 phrases (dictionnaire en annexe). Dans le cas de parole continue, le programme ne teste provisoirement que la première occlusive de chaque phrase.

Les erreurs de diagnostic ont pu être classés en deux catégories :

- erreurs dues au module de Keal (programme de reconnaissance phonétique ou programme de cadrage).
- erreurs imputables à l'algorithme.

5.3 Résultats sur 7 locuteurs français

a. Mots isolés

Sujet	Nombre d'occurences	Bonnes réponses Nombre	Taux	Erreurs Keal Nombre	Taux	Erreur Algorithme Nombre	Taux
Locuteur 1	20	16	80%	2	-	2	-
Locuteur 2	20	16	80%	2	-	2	-
Locuteur 3	20	18	90%	0	-	2	-
Locuteur 4	20	16	80%	3	-	1	-
Locuteur 5	20	18	90%	2	-	0	-

b. Phrases

Sujet	Nombre d'occurences	Bonnes Réponses		Erreurs Keal		Erreur Algorithme	
		Nombre	Taux	Nombre	Taux	Nombre	Taux
Locuteur 1	5	5	100 %	0	-	0	-
Locuteur 2	5	4	80 %	1	-	0	-
Locuteur 3	5	2	40%	2	-	1	-
Locuteur 4	5	4	80%	0	-	1	-
Locuteur 5	5	5	100%	0	-	0	-
Locuteur 6	5	5	100%	0	-	0	-
Locuteur 7	5	5	100%	0	-	0	
Total	35	30	85,7%	3	8,6%	2	5,7%

5.4 Résultats sur 5 locuteurs anglophones

a. Mots isolés

Sujet	Nombre d'occurences	Bonnes Reponses		Erreurs Keal		Erreur Algorithme	
		Nombre	Taux	Nombre	Taux	Nombre	Taux
Locuteur 1	18	17	94,4%	0	-	1	-
Locuteur 2	20	20	100%	0	-	0	-
Locuteur 3	19	18	94.7%	1	-	0	-
Locuteur 4	20	15	75%	3	-	2	-
Locuteur 5	20	18	90%	2	-	0	-
Total	97	88	90,7%	6	6,2%	3	3,1%

b. Phrases

Sujet	Nombre d'occurences	Bonnes réponses Nombre	Taux	Erreurs Keal Nombre	Taux	Erreurs Algorithme Nombre	Taux
Locuteur 1	3	3	100 %	0	-	0	-
Locuteur 2	5	4	80 %	0	-	1	-
Locuteur 3	5	4	80 %	1	-	0	-
Locuteur 4	5	4	80 %	0	-	1	-
Locuteur 5	5	4	80 %	1	-	0	-
Total	23	19	82,6 %	2	8,7 %	2	8,7 %

5.5 Synthèse des résultats

Tous essais confondus, le programme a été testé sur 12 locuteurs et 295 occurences (mots isolés et phrases).

Les résultats peuvent être présentés de la façon suivante :

Bon Diagnostic		Erreur Keal	Erreur Algorithme
Moyenne	Ecart-type	Moyenne	Moyenne
86,45 %	2,63 %	7,79 %	5,76 %

La valeur de 86,45% du taux moyen de bon diagnostic nous permet de dire que le programme fonctionne de façon satisfaisante, avec identification du phonème sur lequel la faute a été commise, sur des mots isolés et de la parole continue. La valeur de l'écart-type associé à cette moyenne prouve que ses performances sont indépendantes du locuteur.

6. CONCLUSIONS

Nous avons donc contrôlé que le programme répondait de façon satisfaisante à trois des quatre objectifs que nous nous étions fixés :

- identification du phonème sur lequel la faute est attendue ;
- reconnaissance muti-locuteurs ;

- possibilité de traiter de la parole continue ;

Ces performances sont quantifiées par le taux moyen de bon diagnostic et l'écart-type associé. Elles pourraient être améliorées en analysant l'origine des erreurs :

- Erreurs en provenance de Keal.

Il s'agit soit d'erreur de segmentation, soit d'erreur d'identification, la deuxième source d'erreur étant souvent une conséquence de la première. La connaissance des programmes de Keal a permis de corriger une partie de ces erreurs (c'est le cas dans le programme). D'autre part, il semble raisonnable de penser que pour la correction d'autres fautes de prononciation (voyelles diphtonguées, confusion etc...), ce taux d'erreur sera plus faible - la segmentation et la reconnaissance des occlusives pouvant être considérées comme un des problèmes les plus difficiles à résoudre pour Keal.

- Erreurs de l'algorithme.

Ces erreurs de l'algorithme sont liées au contexte du phonème sur lequel la faute est commise, dans notre cas, lorsque l'occlusive est suivie de consonnes liquides (L,R) ou de semi-voyelles (J, Y, W). Deux solutions sont envisageables : amélioration de l'algorithme en y intégrant les paramètres de ce contexte si celui-çi est identifiable par Keal, ou suppression du corpus des mots sur lesquels l'algorithme ne fonctionne pas.

Pour finir, les améliorations à apporter ne seront pas indépendantes du quatrième objectif que nous nous sommes fixés, à savoir le temps de réponse. En effet, si la version actuelle du programme ne nous permet pas d'avoir des éléments d'appréciation de ce point, on peut quand même affirmer que toute augmentation de la complexité des contrôles des résultats de Keal se fera au détriment du temps de réponse. Un compromis sera alors à trouver. En conséquence, les développements que nous comptons faire à court terme seront orientés vers des tests de validité d'autres algorithmes (effectuées sur d'autres fautes à partir de programmes "batch"), et la réecriture en temps réel des programmes, une fois les algorithmes mis au point.

REFERENCES

DIXIT, R., 1979 : "Aspiration : What is it and how is it produced ?", 97th Meeting of the A.S.A., Cambridge, pp. 83-87.
FANT, G., 1969 : " Stops in CV-Syllabes", STL-QPSR 4, pp. 1-25.
FERRIEU, G., PERSON, J.M, 1968, "Etude d'un Vocodeur : Le projet A.S.P.I.C.", Rapport interne C.N.E.T., Etude CRL 2013/ETA.
FOURCIN, A.J, ABBERTON, E., 1975 : "American English Intonation, Audio-visual Practice Materials", Linguaphone Institute, Text Book, pp. 1-109.
GIORGETTI, M.T, et al, 1980 : "Reconnaissance de mots isolés avec détection de fautes de prononciation", XI J.E.P., Strasbourg, pp.202-203.
HATON, M.C, HATON, J.P, 1975 : "SIRENE : Un projet de système interactif pour la rééducation vocales de non-entendants", Rev. Gale. Ens. Def. Auditifs, Vol. 67, N° 4, pp. 203-209.
HEFFNER, R. M. S, 1969 : "General Phonetics", The University of Wisconsin Press Madison, Milwankee and London.
JONES, D., 1956 : "The prononciation of English", Cambridge University Press, pp. 68-80.
KALIKOW, D.N, SWETS, J.A, 1972 : "Experiments with Computer controlled Displays in second-language learning", IEEE Trans Audio Electro-Acoust, Vol AU-20.
KIM, CHIN-W, 1970 : "A Theory of Aspiration", Phonetica, Vol. 21, pp. 107-116.

LOW, J., 1979 : "L'intonation Anglaise et l'enseignement par Visualisation", Actes du 9ème congrès International des sciences phonétiques, Copenhague, Vol. 1, pp 124-136.

LEON, M., MARTIN, PH., 1972 : "Applied Linguistics and the teaching of Intonation", The Modern Language Journal, Vol. LVI, N° 3.

MERCIER, G., 1976 : "Deux Dialogues parlés pour un dialogue intelligent entre l'homme et la machine", Compte-rendu final de la convention S.E.S.O.R.I., N° 74-80.

QUINTON, C., LETY, M., 1977 : "Programme de Cadrage pour l'apprentissage Automatique", Recherches Acoustiques, C.N.E.T, pp. 321-342.

STEVENS, K.N., NICKERSON, R.S., 1972 : "Teaching Speech to the Deaf : Can a Computer help ?", IEEE Trans Audio, Vol AU-21, pp. 445-454.

STEVENS, K.N, KLATT, D.H, 1973 : "V.O.T, Frication, and Aspiration in word-Initial Consonant Clusters", Res. Lab. Electron. Quart. Progr. Rept, M.I.T, pp. 124-136.

VIVES, R., 1979 : "Un programme de Cadrage pour l'adaptation au locuteur en reconnaissance automatique de la parole", Congrès AFCET, Toulouse, pp.343-350.

ZURCHER, F., FRICHOU, M., 1973 : "Vocodeur à canaux ETA", note technique C.N.E.T., TMA/ETA/17.

ANNEXE

EXEMPLES DE RESULTAT DU PROGRAMME.

PHRASE LA PATTE

NOMBRE D'ECHANTILLONS= 50

LOCUTEUR FRANCAIS.

SORTIES VOCODEUR.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	F0	P	PL	E	EL		COU
1	4	2	1	0	0	0	1	1	3	3	1	3	3	1	0	16	0	23	0		0
2	7	4	2	1	1	2	1	0	0	3	1	1	3	1	0	6	32	27	77		0
3	7	5	2	1	1	3	1	0	1	3	0	1	2	0	136	10	23	27	89		0
4	7	5	2	1	2	3	0	0	1	3	2	4	4	1	138	7	46	35	104		0
5	8	5	3	2	2	4	1	0	2	3	2	4	4	2	135	29	45	42	144		0
6	7	8	7	6	5	7	5	2	4	5	3	4	4	2	132	9	49	69	185		0
7	7	9	9	7	6	7	5	4	4	5	1	4	4	2	136	11	23	74	226		0
8	7	9	9	9	7	8	4	4	6	6	3	4	4	3	136	3	22	83	239		0
9	7	9	10	9	7	7	4	4	6	6	3	4	3	3	137	8	28	82	251		0
10	7	9	10	9	8	7	4	4	7	6	4	5	5	1	135	17	87	86	245		0
11	7	10	8	7	8	6	3	3	6	3	2	5	7	2	135	62	87	77	178		0
12	4	4	2	0	3	0	0	0	0	0	0	0	2	0	135	8	81	15	103		0
13	1	4	3	1	2	0	0	0	0	0	0	0	0	0	0	11	22	11	29		0
14	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	15	2	14	****	163
15	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	1	3	****	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	****	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	****	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	87	0	20	****	0
25	5	4	2	2	2	1	0	0	1	1	0	0	1	1	0	67	97	20	107		0
26	8	10	10	9	9	6	4	3	6	5	3	4	6	2	113	10	82	87	202		0
27	8	11	11	8	10	7	4	3	6	5	3	7	8	4	116	5	24	95	278		0
28	8	10	10	9	10	7	4	3	6	5	4	7	8	5	112	9	21	96	278		0
29	8	10	10	8	9	7	4	2	4	5	4	4	6	4	107	7	25	87	269		0
30	8	10	10	8	8	7	4	3	3	4	3	4	6	6	106	9	35	86	254		0
31	8	10	10	7	8	8	4	3	3	4	4	5	4	3	103	19	89	81	231		0
32	8	9	7	7	6	6	3	1	2	5	3	5	1	3	101	61	85	64	148		0
33	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	5	73	3	75	****	185
34	1	2	3	1	1	0	0	0	0	0	0	0	0	0	0	7	13	8	12		0
35	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	1	9	****	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	****	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	****	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	28	0	8	****	0
45	0	0	0	0	0	1	0	0	1	1	1	2	1	1	0	20	52	8	36		0
46	1	3	1	2	2	1	3	0	3	2	2	3	3	2	0	24	52	28	42		0
47	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	8	37	6	49		0
48	0	1	1	1	1	1	0	0	0	0	0	0	2	1	0	5	0	8	0		0
49	0	1	1	1	1	0	0	0	0	1	0	0	0	2	0	6	0	7	0		0
50	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	****	14

REPONSE DU PROGRAMME.

ADRESSE DE DEBUT DU P 25

ADRESSE DE FIN DU P 26

L'ADRESSE EST CELLE DE L'EXPLOSION
BONNE PRONONCIATION DU P SUR LE MOT NUMERO 3

25	5	4	2	2	2	1	0	0	1	1	0	0	1	1	0	67	97	20	107		0 I
26	8	10	10	9	9	6	4	3	6	5	3	4	6	2	113	10	82	87	202		0 I
27	8	11	11	8	10	7	4	3	6	5	3	7	8	4	116	5	24	95	278		0 I
28	8	10	10	9	10	7	4	3	6	5	4	7	8	5	112	9	21	96	278		0 I
29	8	10	10	8	9	7	4	2	4	5	4	4	6	4	107	7	25	87	269		0 I
30	8	10	10	8	8	7	4	3	3	4	3	4	6	6	106	9	35	86	254		0 I
31	8	10	10	7	8	8	4	3	3	4	4	5	4	3	103	19	89	81	231		0 I

PHRASE LE PERE

NOMBRE D'ECHANTILLONS= 63

LOCUTEUR FRANCAIS.

SORTIES VOCODEUR.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	FO	P	PL	E	EL		CO
1	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	7	0		0
2	6	3	0	0	0	3	3	0	0	0	0	0	0	0	135	2	14	15	37		0
3	6	3	1	0	0	3	2	0	0	0	0	0	0	0	135	4	31	15	49		0
4	6	3	1	0	1	3	2	1	0	2	0	0	0	0	135	25	50	19	78		0
5	7	5	3	2	2	3	3	2	5	4	3	4	2	0	132	21	53	44	128		0
6	8	7	4	3	3	4	4	3	6	5	5	7	5	1	133	7	35	65	169		0
7	8	8	4	3	3	4	3	2	6	3	3	7	5	1	130	7	20	60	180		0
8	8	8	4	3	3	5	2	2	5	3	3	6	4	0	129	6	29	55	164		0
9	8	8	4	3	3	4	2	2	3	2	2	6	3	0	129	16	42	49	137		0
10	8	7	4	2	3	4	1	0	1	0	0	2	1	0	142	20	47	33	95		0
11	5	5	1	1	0	0	0	0	0	0	0	1	0	0	132	11	32	13	48		0
12	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	2	16	****	120
13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	3	****	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	****	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	****	0
21	3	2	0	0	1	1	1	0	2	0	0	1	0	0	0	11	62	0	11	****	0
22	6	7	5	4	4	4	4	4	6	5	3	5	4	1	0	51	90	11	73		0
23	8	10	7	6	6	8	7	5	8	7	4	6	6	2	0	28	92	62	163		0
24	8	11	8	6	5	6	7	5	7	6	3	4	4	1	140	13	51	90	233		0
25	8	11	8	6	6	6	9	5	7	6	4	7	6	2	120	10	29	81	262		0
26	8	11	8	6	6	6	9	5	6	6	2	5	5	2	119	6	21	91	257		0
27	8	11	8	6	5	6	8	5	6	6	4	6	5	2	119	5	16	85	262		0
28	8	11	9	6	5	5	8	5	6	6	3	4	4	3	117	5	11	86	256		0
29	8	11	9	6	5	5	8	5	6	6	3	4	5	3	116	1	10	85	257		0
30	8	11	9	6	5	5	8	5	5	6	4	7	6	3	115	4	9	86	259		0
31	8	11	9	6	5	5	8	6	6	7	4	7	6	4	113	4	9	88	266		0
32	8	11	9	6	5	5	8	6	6	7	4	7	6	5	112	1	7	92	273		0
33	8	11	9	6	5	5	8	6	6	7	5	7	6	4	112	2	6	93	278		0
34	8	11	9	6	5	6	8	6	7	7	4	7	6	4	110	3	8	93	280		0
35	8	11	9	6	5	6	8	7	7	7	5	7	6	4	110	3	9	94	284		0
36	8	11	9	6	6	6	8	6	7	8	5	7	6	5	110	3	9	97	289		0
37	8	11	8	7	6	6	9	6	7	8	5	7	6	5	110	3	10	98	294		0
38	8	10	8	7	6	6	9	6	8	7	5	7	7	5	109	4	11	99	296		0
39	8	10	8	7	6	6	8	5	7	6	5	7	7	5	106	4	15	99	293		0
40	8	10	8	7	6	6	6	4	7	5	4	7	7	3	105	7	16	95	282		0
41	7	10	8	8	6	6	6	4	6	5	3	7	6	3	104	5	17	88	268		0
42	7	9	8	8	6	7	5	4	6	5	3	6	5	3	101	5	25	85	255		0
43	7	8	8	8	6	6	3	3	6	4	1	4	2	1	99	23	55	67	199		0
44	7	7	6	6	5	3	2	0	2	0	0	4	2	0	99	17	48	44	142		0
45	6	6	4	3	5	2	0	0	1	1	1	3	0	0	96	8	29	31	106		0
46	5	6	4	6	6	2	0	0	0	0	0	3	0	0	96	4	16	31	93		0
47	5	6	5	7	6	2	0	0	0	0	0	0	0	0	93	4	18	31	97		0
48	5	6	5	7	6	2	0	0	0	2	0	3	0	0	89	10	19	35	91		0
49	4	6	4	5	5	1	0	0	0	0	0	0	0	0	90	5	15	25	80		0
50	4	5	4	4	3	0	0	0	0	0	0	0	0	0	91	0	9	20	65		0
51	4	5	4	4	3	0	0	0	0	0	0	0	0	0	88	4	9	20	56		0
52	4	4	2	4	2	0	0	0	0	0	0	0	0	0	88	5	16	16	49		0
53	4	3	3	2	1	0	0	0	0	0	0	0	0	0	88	7	0	13	0		0
54	2	1	2	1	0	0	0	0	0	0	0	0	0	0	90	6	0	6	0		0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	****	195

REPONSE DU PROGRAMME.

ADRESSE DE DEBUT DU P 21

ADRESSE DE FIN DU P 23

L'ADRESSE EST CELLE DE L'EXPLOSION
BONNE PRONONCIATION DU P SUR LE MOT NUMERO 2

21	3	2	0	0	1	1	1	0	2	0	0	1	0	0	0	51	90	11	73		0 I
22	6	7	5	4	4	4	4	4	6	5	3	5	4	1	0	28	92	62	163		0 I
23	8	10	7	6	6	8	7	5	8	7	4	6	6	2	0	13	51	90	233		0 I
24	8	11	8	6	5	6	7	5	7	6	3	4	4	1	140	10	29	81	262		0 I
25	8	11	8	6	6	6	9	5	7	6	4	7	6	2	120	6	21	91	257		0 I
26	8	11	8	6	6	6	9	5	6	6	2	5	5	2	119	5	16	85	262		0 I
27	8	11	8	6	5	6	8	5	6	6	4	6	5	2	119	5	11	86	256		0 I

LOCUTEUR ANGLAIS.

PHRASE LE PERE

NOMBRE D'ECHANTILLONS= 54

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	FO	R	RL	E	EL		CO
1	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0		0
2	5	4	2	0	0	0	0	0	0	0	0	0	0	0	0	8	9	11	30		0
3	5	4	2	1	1	0	0	0	0	0	0	0	0	0	0	8	5	13	39		0
4	5	4	2	2	2	0	0	0	0	0	0	0	0	0	0	3	6	15	44		0
5	6	4	2	2	2	0	0	0	0	0	0	0	0	0	0	5	12	16	50		0
6	6	5	3	2	2	1	0	0	0	0	0	0	0	0	0	8	15	19	62		0
7	7	6	5	3	3	2	0	0	1	0	0	0	0	0	102	0	51	27	77		0
8	7	7	6	4	3	2	1	0	1	0	0	0	0	0	98	39	52	31	128		0
9	8	9	7	7	6	5	4	3	3	2	2	4	6	4	98	9	53	70	178		0
10	7	9	7	7	8	5	4	3	3	4	3	4	7	6	95	5	16	77	227		0
11	7	9	8	7	8	5	4	4	4	4	2	5	7	6	93	8	8	80	235		0
12	7	9	8	7	8	5	4	4	4	3	2	5	7	9	95	0	7	78	237		0
13	7	9	8	7	8	5	4	5	4	3	2	5	7	5	99	4	13	79	232		0
14	7	8	8	7	8	5	4	5	3	2	2	5	6	5	97	8	83	75	231		0
15	8	9	8	8	8	5	4	3	3	3	3	5	6	4	98	13	61	77	222		0
16	8	8	7	7	8	5	5	3	4	2	2	5	4	2	99	48	79	70	175		0
17	6	5	5	5	5	1	1	0	0	0	0	1	1	0	0	20	74	88	100		0
18	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	36	2	34	****	148
19	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4	15	4	10		0
20	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	5	16	4	15		0
21	1	3	0	2	1	0	0	0	0	0	0	0	0	0	0	7	15	7	15		0
22	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	5	12	4	12		0
23	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	1	6	****	0
24	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	3	1	2	****	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	****	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	****	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	46	0	37	****	0
28	3	2	3	2	2	2	3	2	4	3	1	2	4	4	0	9	38	37	65		0
29	2	2	2	2	2	2	3	2	3	1	1	2	1	3	0	18	28	28	91		0
30	4	4	3	2	2	1	0	2	2	1	0	2	1	2	0	57	86	26	137		0
31	7	8	8	6	6	6	7	6	7	6	3	4	4	5	0	17	86	83	209		0
32	8	10	9	0	7	7	9	7	8	7	3	5	6	6	81	12	34	100	275		0
33	7	10	9	0	5	6	8	5	7	6	3	7	6	5	79	5	24	92	283		0
34	7	10	8	0	6	6	8	5	7	6	3	5	7	5	77	9	15	91	277		0
35	8	10	8	0	6	5	7	6	7	6	3	7	7	6	76	3	15	94	280		0
36	8	10	9	7	6	5	7	6	7	5	3	7	7	6	77	3	12	95	283		0
37	8	10	9	7	6	5	7	6	6	4	4	7	8	7	77	4	15	94	287		0
38	9	10	9	7	6	5	8	6	7	4	3	7	8	7	78	0	16	98	290		0
39	9	11	9	7	6	6	8	6	6	5	4	6	8	7	80	0	18	98	292		0
40	9	11	9	7	6	6	9	6	5	5	3	5	8	5	81	0	18	96	290		0
41	9	11	8	7	6	6	9	6	6	5	3	6	8	5	83	0	18	96	294		0
42	9	11	8	0	6	6	7	7	6	5	4	6	7	4	85	0	20	92	298		0
43	9	11	8	5	6	6	8	8	5	4	3	6	7	4	91	8	32	90	266		0
44	8	9	8	5	4	6	8	8	5	4	4	6	6	3	98	18	39	84	246		0
45	8	8	6	5	5	7	8	5	4	4	2	4	6	4	106	13	44	72	237		0
46	8	9	7	5	6	7	8	6	3	3	3	5	6	5	112	13	32	81	225		0
47	7	10	7	5	6	7	8	4	3	4	3	4	4	2	120	8	25	72	219		0
48	7	9	6	4	6	7	8	4	3	3	2	3	4	2	126	6	24	66	204		0
49	6	8	6	4	5	7	7	4	3	4	2	3	4	3	128	9	19	86	191		0
50	5	7	5	4	6	7	8	3	3	2	2	3	4	3	132	4	27	59	186		0
51	6	7	5	4	6	7	8	3	3	3	2	4	4	2	0	16	84	61	167		0
52	5	7	4	4	5	6	8	3	1	1	2	3	2	2	0	8	22	47	159	****	188
53	5	7	4	4	5	6	8	2	2	2	2	3	4	2	0	8	11	51	149		0
54	5	7	4	4	6	6	8	2	2	2	2	3	4	2	0	8	11	51	149		0

REPONSE DU PROGRAMME.

VOUS AVEZ ASPIRE LE M SUR LE MOT NUMERO 2

28	3	2	3	2	2	2	3	2	4	3	1	2	4	4	0	9	38	37	65		0 %
29	2	2	2	2	2	2	3	2	3	1	1	2	1	3	0	18	28	28	91		0 %
30	4	4	3	2	2	1	0	2	2	1	0	2	1	2	0	57	86	26	137		0 %
31	7	8	8	6	6	6	7	6	7	6	3	4	4	5	0	17	86	83	209		0 %
32	8	10	9	0	7	7	9	7	8	7	3	5	6	6	81	12	34	100	275		0 %
33	7	10	9	0	5	6	8	5	7	6	3	7	6	5	79	5	24	92	283		0 %
34	7	10	8	0	6	6	8	5	7	6	3	5	7	5	77	9	15	91	277		0 %

PHRASE LA PATTE

NOMBRE D'ECHANTILLONS= 50

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	FO	M	RL	E	EL	
1	5	3	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	9	0	
2	4	4	2	1	2	0	0	0	0	0	0	0	0	0	0	13	20	13	46	
3	6	6	3	1	3	2	0	0	0	0	0	0	1	2	0	9	50	24	68	
4	7	6	5	4	3	2	1	0	0	0	0	0	2	1	0	30	57	31	116	
5	8	8	7	6	5	4	3	4	3	1	2	4	3	3	99	18	59	61	171	
6	9	10	8	7	6	5	4	4	5	4	4	5	5	3	99	13	57	79	216	
7	8	9	9	8	8	5	4	4	4	4	2	4	4	3	98	8	26	96	235	
8	8	8	9	7	8	5	5	4	4	4	3	5	6	4	100	7	19	80	239	
9	8	9	10	7	8	5	4	5	4	4	3	6	7	3	100	4	12	83	246	
10	8	9	10	7	8	5	5	5	4	4	3	5	6	4	100	3	10	83	250	
11	8	9	10	7	8	5	5	5	4	4	4	5	6	4	100	3	12	84	252	
12	8	9	10	7	8	5	5	6	5	3	3	6	6	4	101	3	24	85	250	
13	8	9	10	7	7	5	5	4	5	4	3	5	5	4	102	13	83	81	234	
14	8	9	9	7	6	5	5	3	4	2	2	3	4	3	106	60	91	68	153	
15	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	88	4	86	**** 16
16	2	3	4	1	2	2	0	0	0	0	0	0	0	0	0	10	26	14	22	
17	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	2	16	4	20	
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	2	8	****
19	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3	7	2	7	****
20	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	3	3	8	****
21	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	9	****
22	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	8	****
23	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	4	2	5	****
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0	3	****
25	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	26	1	4	****
26	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	24	40	3	28	****
27	2	1	2	3	2	1	0	0	2	0	2	5	2	2	0	15	50	24	82	
28	2	3	4	3	3	2	1	1	3	1	1	6	4	1	0	14	94	35	82	
29	3	3	3	3	2	1	1	1	2	0	0	2	2	0	0	63	88	23	146	
30	7	9	10	10	9	6	8	5	6	4	4	5	5	3	0	9	79	88	208	
31	8	10	11	10	9	7	9	5	7	4	4	6	7	4	74	3	19	97	281	
32	8	10	10	10	9	7	8	4	6	4	4	6	7	6	73	3	14	96	290	
33	8	10	10	10	9	7	8	5	6	5	5	6	6	5	73	6	11	97	292	
34	8	10	10	10	9	7	6	6	6	6	5	6	6	4	74	8	14	99	293	
35	8	10	10	10	9	7	6	5	6	6	8	5	6	4	76	8	16	97	289	
36	8	10	10	10	10	7	9	4	6	5	3	4	7	4	78	4	20	93	285	
37	8	10	10	10	10	7	6	4	6	5	5	6	6	4	84	8	20	95	281	
38	8	9	10	10	9	7	5	5	5	5	8	7	7	3	90	8	23	93	279	
39	8	8	10	10	10	7	8	4	6	5	3	5	6	4	97	9	22	91	272	
40	8	8	10	8	9	7	4	4	5	5	4	6	6	4	101	9	25	88	260	
41	8	7	9	8	9	5	4	4	5	5	3	4	6	4	0	13	34	81	239	
42	7	7	9	6	6	5	4	4	4	3	3	4	5	3	0	16	81	70	207	
43	7	7	8	5	6	5	3	2	2	2	1	1	3	4	0	54	78	36	128	
44	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	66	2	66	**** 18
45	1	0	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	8	0	
46	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	8	0	

LOCUTEUR ANGLAIS.

DUREE DU VOT 4

VOUS AVEZ ASPIRE LE H SUR LE MOT NUMERO 3

27	2	1	2	3	2	1	0	0	2	0	2	5	2	2	0	15	50	24	82	
28	2	3	4	3	3	2	1	1	3	1	1	6	4	1	0	14	94	35	82	
29	3	3	3	3	2	1	1	1	2	0	0	2	2	0	0	63	88	23	146	
30	7	9	10	10	9	6	8	5	6	4	4	5	5	3	0	9	79	88	208	
31	8	10	11	10	9	7	9	5	7	4	4	6	7	4	74	3	19	97	281	
32	8	10	10	10	9	7	8	4	6	4	4	6	7	6	73	3	14	96	290	
33	8	10	10	10	9	7	8	5	6	5	5	6	6	5	73	6	11	97	292	

RESULTATS du PROGRAMME.

ANNEXE

· CORPUS

a. Mots isolés.

La pipe
Le père
La patte
La poule
Le port
La praire
Le plan
Le pied
Le poids
Le puit
La copie
Le compère
Le compas
Le rapport
La pipe
La guêpe
La nappe
La soupe
Le stop
Le sport

b. Phrases

Papa est parti à Paris.
Ce petit port de pêche est pittoresque.
Le parapet du pont de pierre est tombé.
Nous ne pouvons pas vous le promettre.
Pierre a acheté des pommes.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

INDICES ACOUSTIQUES SECONDAIRES DE LIEU ET CONTRAINTES DISTRIBUTIONNELLES

Denis RAMASSE

Institut de Phonétique d'Aix-en-Provence
et Université de Trèves

.RESUME

Les indices secondaires de lieu d'articulation des consonnes portés par les formants des voyelles adjacentes présentent, dans bien des cas, un manque de distinctivité qui les rend difficiles à utiliser, notamment en reconnaissance automatique de la parole. Nous montrons ici qu'une des causes de ce "flou" dans les indices doit être imputée à une influence des contraintes distributionnelles venant elles-mêmes instaurer une redondance due au contexte et rendant dans certains cas inutile un indiçage précis des lieux d'articulation des consonnes occlusives, choisies ici comme cadre d'étude.

PHONOLOGICAL CONSTRAINTS AND SECONDARY ACOUSTIC CUES FOR PLACE OF ARTICULATION

. ABSTRACT

Acoustic cues brought by vocalic formants for place of articulation of stop consonants often show a lackness of distinctivity that, for one part, must be atttributed to an influence of phonological contraints which institute a redundancy.

Phonological constraints are, at the phonological level and according to Trubetzkoy, very near to neutralisation. As the latter is known to have an acoustic reality, it is not so surprizing to see, as this study shows, that there is also a projection of phonological constraints on the acoustic level. This projection is important both for theoretical and practical purposes, and it must be taken into account, for example, in an automatic speech recognition device.

BLUMSTEIN et STEVENS (1979) établissent une distinction, dans la description acoustique des lieux d'articulation des consonnes occlusives, entre les caractéristiques principales qui sont invariantes et contenues dans les spectres de bruit et d'implosion, et les caractéristiques secondaires, créées par la coarticulation entre ces consonnes et les voyelles contiguës, et portées par les formants de celles-ci. Les propriétés acoustiques secondaires, bien qu'elles soient toujours contextuelles, qu'elles varient en fonction de chaque voyelle, sont néanmoins importantes dans le codage des informations concernant le lieu de la consonne par l'apport d'une redondance, parfois indispensable.

Leur variabilité, qui les rend difficiles à cerner, pourrait être accrue par des influences non plus phonétiques mais phonologiques, et en particulier par celle des contraintes distributionnelles. Cette hypothèse d'une pression des contraintes sur les indices acoustiques est d'ailleurs rendue d'autant plus plausible que la neutralisation -qui est la paralysie d'une opposition bilatérale alors qu'une contrainte distributionnelle correspond à celle d'une opposition multilatérale (TROUBETZKOY, 1964, p.262 et p.70) - présente une grande analogie avec celles-ci sur le plan phonologique et a, quant à elle, une existence phonétique.

Celle des contraintes devrait se traduire, dans l'onde de parole, par des écarts dans les indices acoustiques car, s'il y a réduction d'une opposition, la manifestation acoustique d'un phonème, par un principe d'économie, n'est plus tenue d'être nette et bien marquée, puisque l'information distributionnelle vient, dans ce cas, supplanter l'information acoustique. Il devrait donc y avoir une corrélation entre la présence de contraintes phonologiques et celles d'indices déviants par rapport à une distribution normale des indices, qu'il convient auparavant de définir.

LES INDICES ACOUSTIQUES

Les oppositions de lieu d'articulation sont, pour les consonnes occlusives du français, multilatérales, donc susceptibles de subir la pression des contraintes distributionnelles. C'est pourquoi, dans le souci de réduire le cadre d'étude et pour éviter toute influence parasite due aux différences d'accentuation, nous nous sommes intéressé uniquement aux consonnes occlusives des syllabes accentuées de type CVC du français standard et nous avons fait enregistrer par deux locuteurs un corpus de mots isolés présentant, dans leur syllabe finale, toutes les combinaisone requises (1).

L'analyse acoustique a été faite sur sonagraphe avec un filtre large. Nous avons mesuré la fréquence des formants au début, à la fin, au quart et aux trois-quarts de la voyelle finale (paramètre HFiPj) et évalué la pente

---

(1). La syllabe finale est, en français, toujours accentuée dans un mot prononcé isolément.

L'enregistrement a été réalisé en chambre anchoïque avec un magnétophone Nagra III. Les mots étaient lus sur des fiches de carton séparées et présentées dans le désordre. Les locuteurs avaient une prononciation "standard".

des formants (paramètre FiQj) en calculant, sur chaque part de la voyelle, leur variation moyenne en Hz/5ms (1). Une partie de ces paramètres, concernant la première moitié de la voyelle, a servi à définir le lieu d'articulation de la consonne prévocalique, tandis que l'autre partie, portant sur la seconde moitié de la voyelle a été attribuée à la consonne finale (cf. les figures 1 et 2 rappelant et illustrant les paramètres utilisés).

Pour une recherche des indices de lieu, une segmentation plus précise des transitions n'était pas indispensable, leur direction étant dans ce cas beaucoup plus utile que leur degré de pente, lequel est surtout un indice de mode d'articulation (LIBERMAN et al. 1956). Cette manière de découper la voyelle rendait d'autre part les valeurs beaucoup plus aisément comparables entre elles.

Nous avons ensuite opéré, en fonction de chaque noyau syllabique, une classification des mesures ainsi obtenues de façon à découvrir des frontières acoustiques entre les consonnes analysées, en faisant un traitement séparé pour les sourdes et pour les sonores, de même que pour les consonnes prévocaliques et postvocaliques. Ceci nous a permis de ne retenir que les valeurs les plus distinctes ou plus exactement les distributions de valeurs les plus distinctes, celles-ci ayant été groupées dans les histogrammes établis en fonction de chaque consonne, chaque position et chaque contexte vocalique étudiés.

Notons que les paramètres permettant d'opérer la meilleure discrimination entre les consonnes se sont révélés être ceux utilisant le F2, et surtout sa pente initiale (F2Q1), sa fréquence initiale (HF2P1), sa fréquence finale (HF2P4) et sa pente dans le troisième quart de la voyelle (F2Q3). Les pentes du F3 (paramètre F3Qj) ont apporté aussi une information très utile.

Il subsistait cependant, à l'issue de cette classification, des valeurs marginales que franchissaient les limites acoustiques entre les consonnes, en créant ainsi des zones de confusion entre les distributions ou du moins en élargissant ces zones, car il est rare qu'un seul indice suffise à établir la séparation totale entre deux classes de consonnes laquelle requiert le plus souvent plusieurs indices (4 au maximum). Des tests statistiques avec le $\underline{t}$ de Student ont permis d'éliminer de ces valeurs marginales, celles qui n'étaient pas significativement distinctes des distributions desquelles elles paraissaient s'éloigner.

LES CONTRAINTES DISTRIBUTIONNELLES

Nous avons établi un inventaire des contraintes distributionnelles intervenant dans les combinaisons CVC à la finale des mots du français standard de façon à étudier quelle part elles prennent dans l'écart, précédemment mis en évidence, de certains indices.

---

(1). Nous nous sommes servi de la formule suivante :

$$p = \frac{20\ (HFiPn - HFiPm)}{d}$$

où p est la pente, HFiPm la fréquence du formant i au point m et HFiPn celle au point de mesure suivant n, d étant la durée (le coefficient 20 s'explique par le fait que la durée est divisée par 4, puisque la pente est calculée sur un quart de la durée vocalique, et qu'il est nécessaire de multiplier la moyenne par 5, celle-ci étant calculée sur 5 ms).

Il est nécessaire de distinguer deux degrés de contrainte, l'opposition de lieu pour les consonnes occlusives du français étant trilatérale puisqu'il y a trois consonnes dans chaque corrélation de sonorité, /p/, /t/, /k/, ou /b/, /d/, /g/. Un premier degré de contrainte correspond à l'exclusion de deux consonnes d'une position déterminée, ce qui n'en laisse qu'une seule possible - c'est le cas pour le /d/ de /dig/ ("digue") puisque ni */big/, ni */gig/ ne sont possibles en français standard -, et il y a contrainte de deuxième degré quand un seul phonème est exclu - le /p/ final de */bap/, par exemple, laissant alors deux consonnes finales possibles, objet d'une contrainte de second degré, le /t/ de /bat/ ("batte") et le /k/ de /bak/("bac")-. Nous avons donc représenté les contraintes dans des tableaux à deux dimensions (cf. figure 3) par le seul phonème possible pour les contraintes de premier degré et le seul phonème exclu (précédé d'un astérisque dans les tableaux) pour les contraintes de second degré (nous avons en outre indiqué par un astérisque isolé, les cas où aucune des trois consonnes de la même corrélation de sonorité ne peut apparaître). Le deuxième tableau de la figure 2 indique par exemple qu'avant /ig/ seule /d/ est possible comme consonne occlusive sonore (6ème ligne, 4ème colonne) et le troisième tableau (4ème ligne, 1ère colonne) que /b/ est impossible devant /ap/.

Soulignons que les contraintes distributionnelles ainsi décrites et déterminées à partir de mots isolés semblent être également valables en parole continue bien qu'elles soient pour l'instant limitées aux seules syllabes accentuées du français standard. Pour les syllabes inaccentuées, le nombre des structures syllabiques est peut-être plus important en parole continue que pour les mots isolés à cause de phénomènes de liaison ou d'élision. Mais la formulation des règles de contraintes dans le cadre de la syllabe permet de réduire l'écart entre les contraintes distributionnelles qui sont propres au lexique et celles qui n'apparaissent que dans la parole continue. Ainsi, dans "un habit gris" /œ̃nabigʁi /(1), la combinaison /big/ ne forme pas une syllabe puisque /b/ + /i/ constituent la syllabe pénultième (d'ailleurs non accentuée), tandis que /g/ fait partie de la syllabe finale. De même, il n'y a pas de syllabe /bap/ dans le syntagme "un combat prodigieux" / œ̃kɔ̃bapʁɔdiʒjø/ car /ba/ et /p/ n'appartiennent pas à la même syllabe. Le cadre offert par la syllabe semble donc approprié à une étude des contraintes en tant qu'elle permet "d'exprimer de la façon la plus économique les latitudes combinatoires des voyelles et des consonnes dans un langage déterminé" (ARNOLD 1965, pp.273-274).

RESULTATS ET DISCUSSION

Il s'agissait ensuite de mettre en relation écarts acoustiques et contraintes, rapprochement dont le tableau 1 montre le résultat en indiquant le nombre et la proportion d'écarts pour chaque condition de contrainte (absence, 1er degré, 2ème degré) et chaque position (prévocalique et postvocalique). Il apparaît, et ceci est confirmé par des χ^2, que seules les contraintes de premier degré ont une projection acoustique marquée (degré de signification à .01) tandis qu'il n'y a aucune différence statistiquement significative entre contraintes de deuxième degré et absence de contrainte. On peut voir, d'autre part, que les contraintes de premier degré se manifestent presque uniquement en position prévocalique et que, près d'un quart des consonnes prévocaliques subissant une contrainte de premier degré (23,75 %) est influencé.

Ainsi, c'est seulement quand une consonne est totalement contrainte et

---

(1). Pour répondre à une objection de J.P.Tubach en reprenant ses contre-exemples.

que le contexte ne laisse plus subsister aucune hésitation sur le lieu d'articulation de cette consonne, que les indices acoustiques peuvent devenir plus libres, qu'une économie est apportée dans la manifestation acoustique d'un lieu d'articulation.

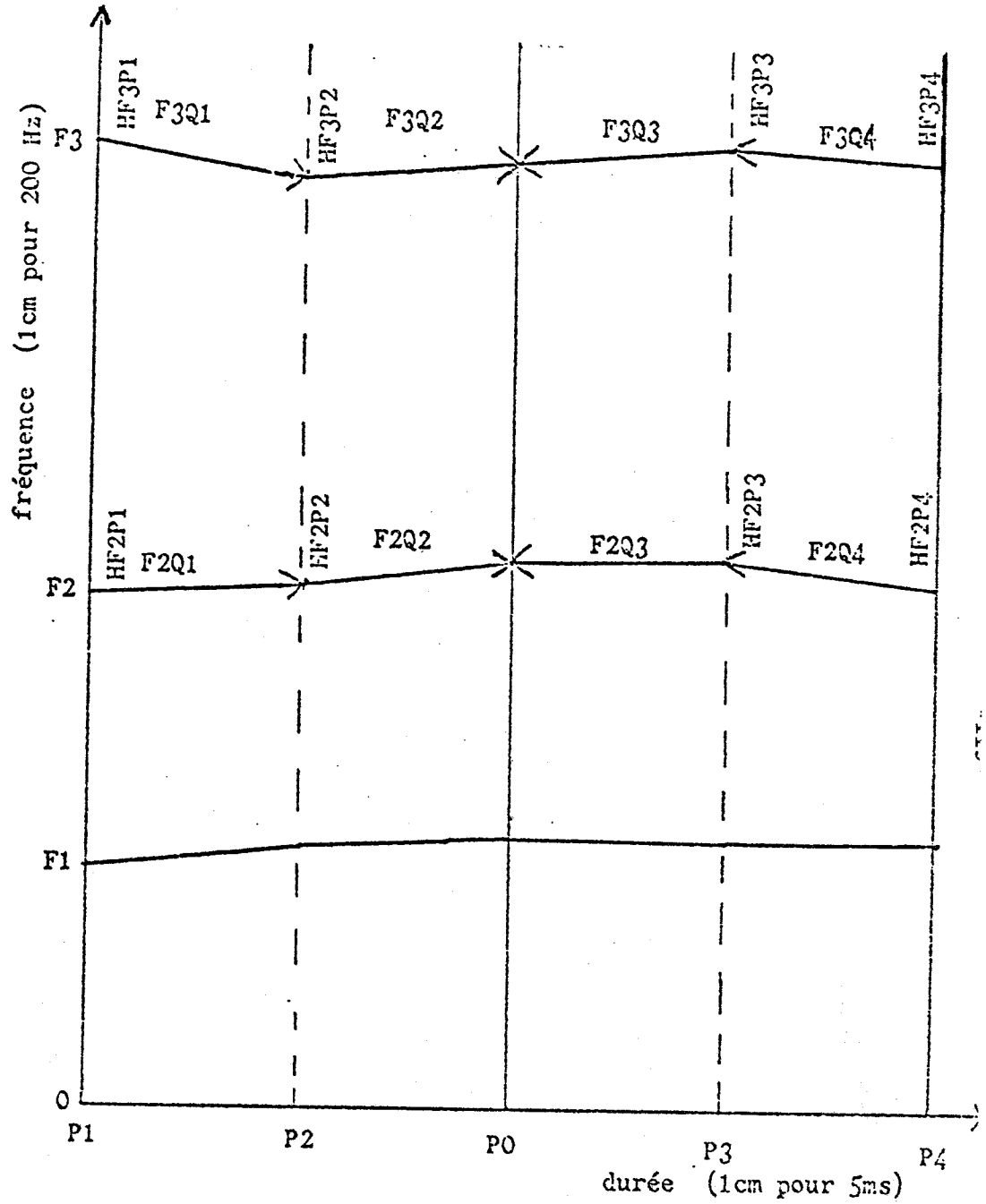
La figure 4 présente deux réalisations du même mot "banque" / bɑ̃k/ dont le /b/ initial est sous une contrainte de premier degré (/d/ et /g/ ne pouvant apparaître en français dans ce même contexte), par deux locuteurs (appelés locuteur A et locuteur B). Or, pour ces locuteurs, deux indices suffisent à la caractérisation du lieu d'articulation d'un /b/ initial d'une syllabe CVC accentuée comportant / ɑ̃/ comme noyau syllabique : il s'agit, d'une part, d'une pente descendante (en direction du centre vocalique) ou nulle de F2 dans le deuxième quart de la voyelle (paramètre F2Q2) et, d'autre part, d'une fréquence initiale de F2 inférieure à 1400 Hz (paramètre HF2P1). Si les deux réalisations répondent à la deuxième condition, les fréquences initiales de F2 étant respectivement pour les locuteurs A et B de 1240 Hz et de 1262 Hz, en revanche elles diffèrent pour le premier paramètre. Pour le locuteur A, en effet, la pente du F2 dans le deuxième quart de la voyelle est nulle, la fréquence de ce formant restant à 1300 Hz dans cette portion de la voyelle. Le caractère labial de la consonne initiale de cette syllabe peut donc être identifié sans aucune ambiguïté puisque la pente n'est pas positive. Chez le locuteur B, la fréquence du F2 évolue dans le deuxième quart de la voyelle de 1148 Hz à 1199 Hz; la pente de ce formant est donc positive (dans la direction consonne → centre vocalique) et sa vitesse d'évolution sur ce segment est de 5Hz/5ms (la durée totale de la voyelle étant de 194,5 ms). Le /b/ du locuteur B est donc mal indicé et il se distingue significativement pour ce dernier indice des /b/ figurant dans un contexte analogue ([—— voyelle cs occlusive]) mais ne subissant pas l'influence des contraintes ($\underline{t}$ de Student significatif à .025). Cet écart (après une comparaison avec d'autres écarts de même type pour d'autres consonnes et d'autres contextes) peut donc être imputé à une influence d'une contrainte distributionnelle de premier degré.

Il semble, par ailleurs, d'après les observations de TSEMEL' et KRINOV (1974), que les transitions finales soient toujours moins nettes que les transitions initiales. Manifestant de façon plus distincte les indices acoustiques de lieu, celles-ci pourraient également en rendre plus accentués les écarts.

Une influence des contraintes distributionnelles sur les indices secondaires de lieu revêt donc, pour l'instant, une importance non négligeable en reconnaissance automatique de la parole où il devrait être tenu compte de celles-ci dans des algorithmes d'identification des consonnes occlusives en attendant que l'information apportée par le burst et l'implosion puisse être, si cela est réalisable, exclusivement utilisée.

paramètre	éléments concernés			
	consonne	point de mesure sur la voyelle	segment de la voyelle	formant
F2Q1	prévoc.		1er quart	F2
F2Q2	prévoc.		2ème quart	F2
F2Q3	postvoc.		3ème quart	F2
F2Q4	postvoc.		4ème quart	F2
F3Q1	prévoc.		1er quart	F3
F3Q2	prévoc.		2ème quart	F3
F3Q3	postvoc.		3ème quart	F3
F3Q4	postvoc.		4ème quart	F3
HF2P1	prévoc.	début		F2
HF2P2	prévoc.	quart		F2
HF2P3	postvoc.	trois quarts		F2
HF2P4	postvoc.	fin		F2
HF3P1	prévoc.	début		F3
HF3P2	prévoc.	quart		F3
HF3P3	postvoc.	trois quarts		F3
HF3P4	postvoc.	fin		F3

(a)

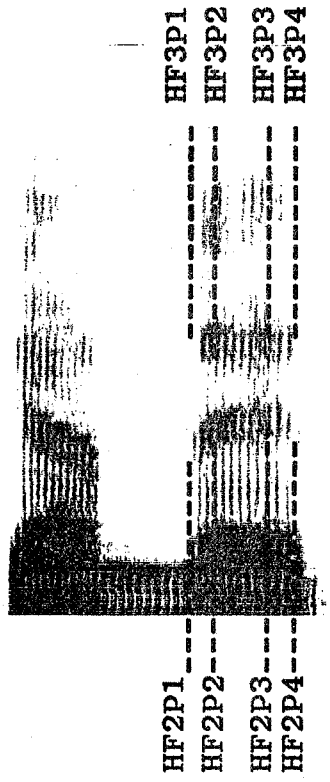


(b)

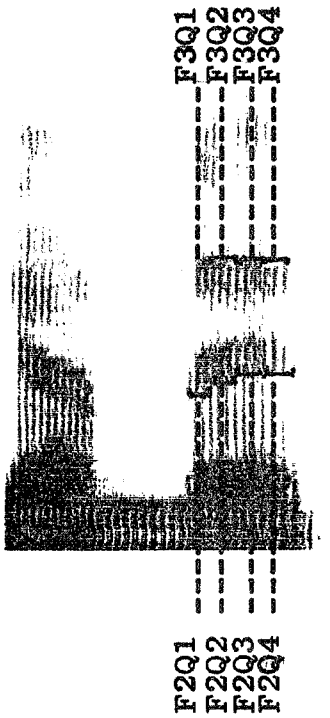
Définition (a) et illustration (b) des paramètres utilisés dans cette étude. La figure b (à droite) est une représentation schématisée du sonagramme de la voyelle /a/ du mot "soupape" prononcé par un des locuteurs, donnant un exemple de localisation des points et des segments mesurés.

Figure 1

mesures ponctuelles



mesures dynamiques



Exemple de localisation des indices utilisés sur la voyelle /ɛ/ du mot "diabète" /djabɛtə/. La segmentation est faite en fonction du F2. Dans le cas de lacune d'un indice statique - c'est ce qui arrive ici pour HF3P1- l'indice dynamique dans le calcul duquel il doit entrer, en l'occurrence F3Q1, est également absent.

Figure 2

- 1 -

contraintes sur les cs

occlusives sourdes initiales

	*	*p		t	*	t		*k		k	*	p
	*k					k			*p	p	*t	t
	p	t	*k		k	*		k	*	t	*	k
k	*	*	*	*	*	*	*	*p	*	*p	*	b
	*			p	*	*p			k	t	*	d
*	*	t	t	*	t	*	*	*k	*	*	*	g

ə a ɑ̃ i ɛ ɛ̃ o ɔ ɔ̃ u y ø

- 3 -

contraintes sur les cs

occlusives sourdes finales

	*p	t		*p	t	*		*k	p	t	t
	t				t	p		*k	*k	k	*
	*	*k	*k	*p	*p	t		*p	*k	p	t
*p	t	*p	*p	*p	t	t		*	*p	t	*
*p	*	t	*p	*p	*	*		*p	t	*t	*
*k	t	t	*	*k	p	*	t	*	t	*	*

ə a ɑ̃ i ɛ ɛ̃ o ɔ ɔ̃ u y ø

g	*	*	*	*	g	*	*g	*	*	d	*	p
	*d		*g		b	b		d		b	*	t
*g	*	b	*g	*g	*	*	*g	d	b	d	*	k
b	*	g	b	*	*	d	g	b	*	*	*	b
	*			*	*b	*g	g	*d	b	*	*	d
	*	g	d	b	d	*	*b	g	*	*	*	g

contraintes sur les cs

occlusives sonores initiales

- 2 -

d	*	d	d	d	*	*	d	*b	*	*	*	p
d	*	*b	*b	*	g	d	d		*	*g	*	t
*g	*	d	d	*	*	d	d	*g	d	b	*	k
	*	d	*g	g	*	d	*	*g	d	*	*	b
*b	*	d	*b	*	*b	*g	g	*	*	*	*	d
*b	*		d	*	d	*		*b	*	*	*	g

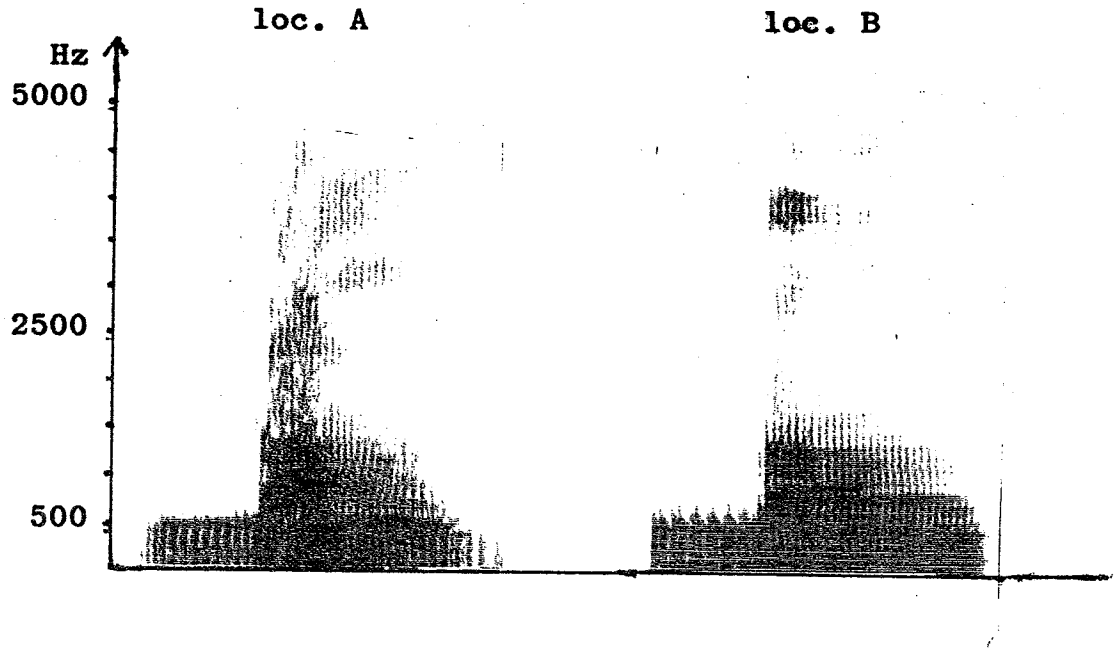
contraintes sur les cs

occlusives sonores finales

- 4 -

Tableaux représentant les contraintes distributionnelles sur les consonnes occlusives dans les syllabes accentuées du français de la forme cs occlusive - voyelle - cs occlusive. Les consonnes occlusives placées sur la ligne médiane verticale entre les tableaux représentent pour ceux de gauche les consonnes finales de syllabe, et dans ce cas les contraintes portent sur les initiales, et pour ceux de droite les initiales, et dans ce cas les contraintes portent sur les finales. Dans les cases des tableaux, les consonnes sans astérisque indiquent les contraintes du premier degré, celles avec astérisque les contraintes du second degré, et les astérisques seuls les contraintes absolues tandis que les blancs signifient des absences de contrainte.

Figure 3



Exemple de manifestation (à droite) d'une contrainte de premier degré portant sur un /b/ prévocalique dans le deuxième quart de la voyelle /ã/ du mot "banque" /bãk/ pour le paramètre F2Q2). La réalisation du locuteur A (à gauche), servant de référence, ne présente pas cet écart.

Figure 4

contraintes	consonne prévocalique			consonne postvocalique		
	N	n	%	N	n	%
premier degré	80	19	23,75	90	6	6,67
deuxième degré	82	5	6,1	148	2	1,35
absence	159	6	3,77	83	4	4,82

Nombre (n) de consonnes occlusives de syllabes CVC du français standard manifestant des écarts dans leurs indices acoustiques, et pourcentage de ces écarts pour chaque catégorie et par rapport au nombre total (N) de consonnes occlusives du français standard analysées.

Tableau 1

REFERENCES

ARNOLD (G.F.), 1965, A Phonological Approach to Vowel, Consonant and Syllable, Lingua 5, 253-287.

BLUMSTEIN (S.E.), STEVENS (K.N.), 1979, Acoustic Invariance in Speech Production : Evidence from Measurements of Spectral Characteristics of Stop Consonants, JASA 66(4), 1001-1017.

BROAD (D.J.), FERTIG (R.H.), 1970, Formant Frequency Trajectories in Selected CVC syllabes, JASA 47(6), 1572-1582.

HOUSE (A.S.), FAIRBANKS (G.), 1967, The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels, in Readings in Acoustic Phonetics, publié par Ilse Lehiste, M.I.T. Press, Cambridge (Massachusetts) Londres, 128-136 (publié auparavant dans JASA, vol. 25 (1953), 105-113).

LIBERMAN (A.M.), DELATTRE (P.), GERSTMAN (L.J.), COOPER (F.S.), 1956, Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds, J.Exp.Psych. 52, 127-137.

LYUBLINSKAYA (V.V.), 1966, Recognition of Articulation Cues in Stop Consonants in transition from Vowel to Consonant, Soviet Physics-Acoustics, vol.12 n°2, 185-192.

POLS (L.C.W.), SCHOUTEN (M.E.H.), 1979, Spectral and Perceptual Aspects of Vowel Coarticulation, Proc.9th Int. Congr. Phon. Sc., Copenhague.

RAMASSE (D.), 1980, La Projection acoustique des contraintes distributionnelles - contribution à la reconnaissance automatique de la parole, Thèse de troisième cycle, Aix-en-Provence, Décembre 1980.

ROSSI (M.), 1975, Les Contraintes phonologiques dans un système de reconnaissance de la parole, 6èmes Journées d'Etude sur la Parole, Toulouse, 160-190.

SEARLE (C.L.), ZACHARY JACOBSON (J.), RAYMENT (S.G.), 1979, Stop Consonant Discrimination Based on Human Audition, J.Acoust.Soc.Am., 65(3), 799-809.

STANLEY (R.), 1967, Redundancy Rules in Phonology, Language 43/2, 393-436.

TROUBETZKOY (N.S.), 1964, Principes de phonologie, traduction française par J.Cantineau de Grundzüge der Phonologie (TLCP VII, Prague,1939), Klincksieck, Paris.

TSEMEL' (G.I.), KRINOV (S.N.), 1974, On the Initial and Final Transitional Segments of Vowels, in Proc. Speech Com.Sem., Stockholm, 137-140.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

RECONNAISSANCE PHONEMIQUE PAR PREDICTION LINEAIRE

L'EXPERIENCE DE L'E.N.S.T.

MM. L. MICLET
Y. GRENIER
J. LE ROUX

ECOLE NATIONALE SUPERIEURE DES TELECOMMUNICATIONS
Département Systèmes et Communications
46 rue Barrault - 75634 - PARIS Cedex 13

.**RESUME**

The introduction of identification methods, stemming from System Theory and in particular the Linear Prediction has recently given rise to the development of new methodologies in signal analysis, and especially in speech. The underlying idea behind these methods consider the vocal system as a filter belonging to a certain family, described by the variation of parameters. According to this approach, the examination of speech signals enables, inside the established mathematical framework, the identification of the characteristic parameters of this signal. The latter, during a short period, is thus represented by the numerical value of these paramaters. The inversion of the model enables to check the correctness of the hypotheses made, by producing a synthetic signal, the quality of which measures the validity of the model.
For the past ten years, big developments have been made in the techniques of speech modelisation, this being due especially to the facility of designing corresponding algorithms on a computer.

Apart from the applications in analysis, compression and synthesis, the speech signal study by identification methods provides the instruments for its recognition : if the parameters extracted are pertinent, one should be able to use them to compare a speech instant thus described with learned prototypes, and with the help of an adapted distance, choose the sound which it most resembles.

The recognition of speech signals based on their analysis by linear prediction has been developed at ENST for more than 5 years. The purpose of this article is to recall the different studies which have been made, and to propose certain subjects, the study of which seeming indispensable if one wishes to progress in this domain. After the introduction of the basic notions, the principle of "channel adapted vocoder" is recalled : this being a bunch of inversed

filters, each one being matched with a phoneme (or sub-phoneme) of the speaker. The passage of the signal in the "vocoder" designates as the phoneme articulated at a certain moment the one which corresponds to the channel where the error of prediction is lowest. This system of recognition works directly on the autocorrelation of the entrance signal, at a cadence of 100 HZ.

The main problem of such a system is of course the learning of each channel model. A procedure based on a method of clustering has been developed at ENST, which enables one to determine in a semi-automatic fashion, a coherent series of channels.

Another aspect is that of adaptation to the speaker. We are introducing a method, stemming from the Canonical analysis of correlations, which automatically enables, starting with the channel adapted vocoder of a speaker, the calculation of that of another speaker. The projection function is learned from a series of key words pronounced by the two speakers : it is thus a question of completly automatic adaptation to the speaker.

So the reconstruction of phonemes can be explained based on recognition factors stemming from the "Vocoder". Automata associated with the channels make a local smoothing, and the series of segments is consequently treated in order to reach a classic "phonemic lattice".

The qualities and faults of such a system of phonemic recognition can be analysed, and certain research axes are hence proposed in order to improve the performances of this system. The latter are already promising.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

RECONNAISSANCE PHONEMIQUE PAR PREDICTION LINEAIRE

L'EXPERIENCE DE L'E.N.S.T.

MM. L. MICLET
Y. GRENIER
J. LE ROUX

ECOLE NATIONALE SUPERIEURE DES TELECOMMUNICATIONS
Département Systèmes et Communications
46 rue Barrault - 75634 - PARIS Cedex 13

1 - ANALYSE ET RECONNAISSANCE DE LA PAROLE PAR PREDICTION LINEAIRE

L'introduction des méthodes d'identification, issues de la Théorie des Systèmes, et en particulier la Prédiction Linéaire, a permis récemment de développer de nouvelles méthodologies en analyse du signal, tout particulièrement sur la parole. Ces méthodes ont pour idée de base de considérer le canal vocal comme un filtre, appartenant à une certaine famille décrite par la variation de paramètres. Sous cette approche, l'examen du signal de parole permet, à l'intérieur du cadre mathématique ainsi fixé, d'identifier les paramètres caractéristiques de ce signal. Celui-ci, pendant une courte durée, est donc représenté par la valeur numérique de ces paramètres. L'inversion du modèle permet de vérifier la justesse des hypothèses faites, en produisant un signal synthétique dont la qualité mesure la validité du modèle (LE ROUX, MICLET, 1979). Depuis une dizaine d'années, les techniques de modélisation de la parole par prédiction linéaire ont connu un grand développement, en particulier dû à la facilité d'implantation sur calculateur des algorithmes correspondants. Les références (GUEGUEN, 1976), (MAKHOUL, 1975) et (MARKEL, GRAY, 1976) développent le sujet et présentent les travaux de base dans le domaine. La qualité de l'analyse réalisée par ces techniques laisse espérer un développement encore supérieur à celui que l'on peut déjà constater - Voir par exemple, la carte de synthèse de TI (TM 990-306), réalisant la compression par prédiction linéaire de quelques centaines de mots courts sous un faible encombrement.

Outre les applications en analyse, compression et synthèse, l'étude du signal de parole par les méthodes d'identification fournit des outils pour sa reconnaissance : si les paramètres extraits sont en effet pertinents, on doit pouvoir les utiliser pour comparer un instant de parole ainsi décrit à des prototypes appris, et, à l'aide d'une distance adaptée, choisir le son duquel il se rapproche le plus.

On se retrouve ici dans un problème de Reconnaissance des Formes dans l'espace des paramètres de Prédiction Linéaire. D'ailleurs, plus généralement, les méthodes d'identification peuvent être utilisées en Reconnaissance des Formes (GUEGUEN, 1978). Mais la nature du signal analysé et celle des traits caractéristiques extraits en font un cas très spécifique. On se propose dans cet article de décrire les méthodes de reconnaissance phonémiques utilisées à l'ENST, sur les bases de l'analyse du signal de parole par prédiction linéaire. On commencera par introduire ici les notations indispensables.

On notera S(t) le signal à étudier ; par échantillonnage à cadence fixe de période T (typiquement, T = 0.1 ms), on obtient la suite de valeurs :

$$S_n = S(n\,T)$$

L'évolution du signal S_n est représentée par le modèle linéaire autoregressif :

$$e_n = a_o S_n + a_1 S_{n-1} + \ldots + a_p S_{n-p} \qquad (1)$$

p est l'ordre du modèle. Les coefficients a_i sont les coefficients de prédiction linéaire, la relation (1) pouvant s'écrire sous la forme :

$$s_n = \frac{1}{a_o} \cdot (e_n - \sum_{i=1}^{p} a_i S_{n-i}) \qquad (2)$$

Sous cette forme S_n est prédit à partir de l'entrée e_n et des p échantillons précédents. On impose la condition $a_o = 1$. Pour une entrée nulle $e_n = 0$, le signal S_n est estimé par $\hat{S}_n$:

$$\hat{S}_n = - \sum_{i=1}^{p} a_i S_{n-i} \qquad (3)$$

La différence e_n entre le signal réel S_n et son estimation $\hat{S}_n$ est l'erreur de prédiction.

L'identification des paramètres a_i, qui réalise l'ajustement du modèle au signal étudié, peut se faire de différentes façons, sensiblement équivalentes. Rappelons ici l'approche par les "moindres carrés".

$$\text{Soit } E : \sum_n (a_o S_n + \sum_{i=1}^{p} a_i S_{n-i})^2 \qquad (4)$$

le critère des moindres carrés sur l'erreur de prédiction e_n. La minimisation de ce critère relativement aux coefficients a_i permet d'écrire le système de p équations :

$$\frac{\partial E}{\partial a_i} = 0 \qquad \forall\, i \in 1, \ldots, p \qquad (5)$$

Ces équations sont les équations normales de Yule-Walker

$$\sum_{k=1}^{p} a_k \sum_n S_{n-i}\, S_{n-k} = - \sum_n S_n\, S_{n-i} \qquad (6)$$

Elles sont résolues par l'algorithme de Durbin.

On peut donner à cette méthode d'analyse une interprétation spectrale, qui permet de se rapprocher des techniques classiques.

Le modèle $\{a_i\}$ a pour fonction de transfert en z :

$$\frac{S(z)}{E(z)} = \frac{1}{A(z)}$$

avec $A(z) = 1 + a_1 z^{-1} + \ldots + a_p z^{-p}$

Alors $E(z) = A(z)\ S(z)$ et $E(\omega)^2 = p(\omega) . A(\omega)^2$, où p est le spectre du signal.

Le théorème de Parseval permet de donner une interprétation spectrale du critère E.

$$E = \sum_n e_n^2 = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} E(\omega)^2 \, d\omega = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} p(\omega) \cdot A(w)^2 \, d\omega \qquad (7)$$

En définissant le spectre $\hat{p}(\omega)$ du modèle comme étant le spectre de la sortie de ce modèle pour une entrée en bruit blanc :

$$\hat{p}(\omega) = \frac{1}{A(\omega)^2}$$

La relation (7) sécrit alors :

$$E = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \frac{p(\omega)}{\hat{p}(\omega)} \, d\omega \qquad (8)$$

Le critère quadratique j est donc la moyenne du rapport des spectres du modèle et du signal.

Une autre relation spectrale peut être définie. En effet, si l'on définit le cepstre d'un signal comme la transformée de Fourier inverse du logarithme du spectre du signal, on peut interpréter les coefficients cepstraux du signal à partir du spectre du modèle :

$$C_n = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log p(\omega) \ e^{jn\omega} d\omega$$

On démontre alors que les $\{a_i\}$ et les $\{C_i\}$ sont liés par les relations :

$$C_1 = a_1$$

$$j\, C_j = j\, a_j - \sum_{k=1}^{j-1} (k\, C_k)\, a_{j-k} \qquad \text{pour } j = 2 \ldots p$$

Tout ceci permet donc de retrouver spectre et cepstre, estimés par prédiction linéaire, et autorise la comparaison directe avec les méthodes fréquencielles classiques.

2 - LE VOCODEUR A CANAUX-ADAPTES

2.1 Principe

Les méthodes de modélisation par prédiction linéaire suggèrent une méthode de reconnaissance par des discriminateurs linéaires dans l'espace du signal à reconnaître.

Supposons en effet qu'on ait à classer un signal inconnu dans l'une des C classes, chacune d'entre elles étant définie par les coefficients (a_i^j) $i = 1,p$ $j = 1,C$ d'un modèle linéaire autorégressif. Il est possible à chaque instant d'échantillonnage du signal de calculer les erreurs de prédiction du signal par rapport à chacun des modèles :

$$e_n^j = - \sum_{i=1}^{p} a_i^j \; S_{n-i} \qquad (9)$$

Les méthodes d'ajustement présentées au paragraphe précédent assurent que le signal d'entrée sera d'autant plus adéquat au modèle que l'erreur de prédiction correspondante sera la plus faible. Ceci correspond à faire un filtrage inverse du signal par tous les modèles, et à chercher quel modèle filtre au mieux le signal d'entrée.

On obtient une interprétation spectrale de l'adéquation du signal à un modèle à partir de l'équation (8). La sortie E^j du $j^{ième}$ canal est :

$$E^j = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \frac{p(\omega)}{\hat{p}^j(\omega)} \qquad (10)$$

où $\hat{p}^j(\omega)$ est le spectre du modèle j.

E^j sera d'autant plus petit que le spectre du signal à reconnaître et le spectre du modèle j sont plus proches en moyenne.

Ce type de comparaison entre un signal et un modèle a été utilisé couramment depuis peu : on peut par exemple consulter les références (ITAKURA,1975), (WAKITA, 1979) pour plus de détails et d'expériences.

2.2 Simulation

En pratique, les calculs ne se feront pas à chaque nouvel échantillon de signal, mais en moyenne sur une fenêtre temporelle de longueur N adaptée à la rapidité de l'évolution du signal à reconnaître. L'énergie en sortie du filtre j vaut donc :

$$E^j = (A^j)^T \, V(A^j)$$

Avec : $$A^j = \begin{bmatrix} a_o^j \\ a_1^j \\ \\ a_p^j \end{bmatrix}$$

$$V = \begin{bmatrix} V_o & V_1 & \cdots & V_p \\ V_1 & & & . \\ . & & & . \\ . & & & V_1 \\ V_p & \cdots & V_1 & V_o \end{bmatrix}$$

où l'on a : $V_i = \sum_{n=p}^{N} S_n S_{n-i} = V_{-i}$

Ceci peut s'écrire encore :

$$E_j = (\alpha^j)^T V \qquad (11)$$

avec $v^T = (V_{+p}\ V_{p-1} \cdots V_1\ V_o\ V_1 \cdots V_p)$

$$\alpha^T = (\alpha^j_{-p}\ \alpha^j_{-p+1} \cdots \alpha^j_1\ \alpha j_o\ \alpha j_1 \cdots \alpha j_p)$$

Les coefficients α_i^j étant la corrélation des coefficients de prédiction linéaire :

$$\alpha_i^j = \sum_{k=0}^{k=p-|i|} a_k^j\ a_{k+1}^j$$

Cette formule fournit une méthode de simulation simple d'un système de reconnaissance à "canaux adaptés" : les modèles des canaux seront alors les ensembles de paramètres :

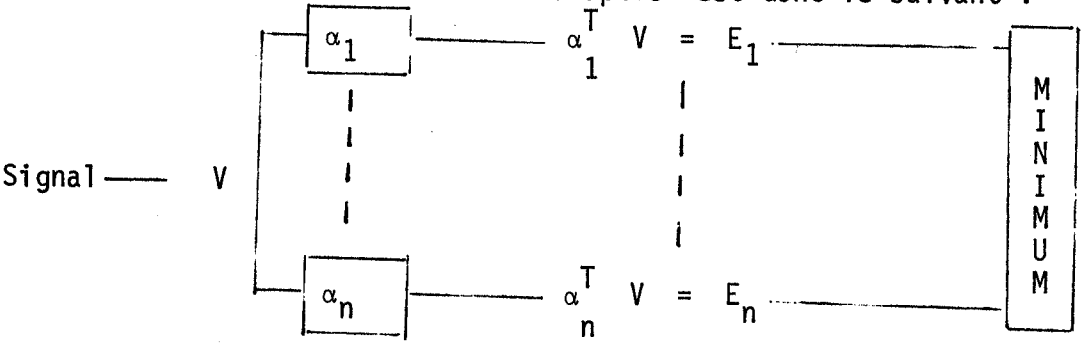
$$(\alpha_i^j) \quad j = 1, C$$
$$i = 1, p$$

Ils joueront le rôle de discriminateurs linéaires dans l'espace de l'autocorrélation du signal à reconnaître. On cherchera en effet à trouver le modèle (canal) j tel que :

$$\forall\, k \neq j \qquad E^k > E^j \qquad k \in 1,C$$

c'est-à-dire celui pour lequel le critère E^j est minimum.

Le schéma d'un "vocodeur à canaux adaptés" est donc le suivant :



Il fournit donc une décision de reconnaissance à chaque instant où a été calculée la corrélation V du signal S (typiquement, 10 ms) (voir la figure 1).

Ce type de distance utilise d'une façon simple et directe l'analyse par prédiction linéaire. On peut la comparer à d'autres distances (calculées à partir de paramètres de prédiction linéaire) sur des bases de données d'apprentissage, afin de mesurer ses propriétés ; en particulier à la distance euclidienne entre les coefficients cepstraux dont la qualité est reconnue (MICLET, NEHME, 1979), (MARKEL, GRAY, 1976). Une conclusion que nous avons tirée de ces comparaisons est que le vocodeur à canaux adaptés représente un bon discriminateur dans l'espace des paramètres de prédiction linéaire, et que l'économie de calculs qu'il représente couvre la faible dégradation, par comparaison avec la distance spectrale précédemment citée .

3 - UTILISATION DE METHODES DE CLASSIFICATION AUTOMATIQUE POUR AIDE A L'APPRENTISSAGE

3.1 Apprentissage standard

Comme dans tout problème de Reconnaissance de Formes, la question de la qualité de l'appentissage est ici cruciale et doit être reliée au choix de la distance utilisée en décision. La définition même du Vocodeur à canaux adaptés fournit une méthode simple d'appentissage.

On suppose d'abord avoir constitué une base d'apprentissage avec professeur, c'est-à-dire connaître un ensemble de morceaux de la parole à la fois sous leur forme de signal et sous leur étiquette phonémique. En se plaçant dans les hypothèses de la prédiction linéaire, il faut supposer que les éléments de la base d'apprentissage sont stationnaires, et donc correspondent à des parties stables de phonèmes. Cette hypothèse de travail est raisonnable pour les sons-voisés et les fricatives, et très discutable pour les plosives non-voi voisées. Cette constation se retrouve évidemment dans les résultats de reconnaissance. En s'en tenant à ce cadre, on va donc travailler au niveau du phonème, supposé stable, et non pas à celui des transitions (diphonèmes). L'application du vocodeur à canaux adaptés à la parole continue suppose un traitement ultérieur des transitions, que l'on abordera au paragraphe 5. Sous ces hypothèses, on a donc défini une base d'apprentissage comportant des morceaux de signal étiquetés. Ceci est réalisé en pratique par un examen du signal sur une console de visualisation graphique interactive, qui permet sa segmentation et son archivage en parties étiquetées par un label phonémique.

L'apprentissage standard consiste à calculer un canal (et un seul) du vocodeur à canaux adaptés pour chaque classe phonémique donnée a priori. En pratique, ce nombre est fixé à 32. Le calcul des paramètres du canal est effectué en prenant pour critère de minimiser la moyenne des sorties du canal pour tous les points étiquetés conformément à ce canal. Autrement dit, si la classe phonémique n° i possède M_i "points", c'est-à-dire M_i instants de 10 ms dont on a calculé la corrélation, on calcul α_i de façon à minimiser :

$$\overline{E_i} = \frac{1}{M_i} \sum_{k=1}^{M_i} (\alpha_i)^T V_k$$

où les (V_k) $k = 1, M_i$ sont tous les points de la base d'apprentissage étiquetés par le phonème n° i.

On a, par simple développement :

$$\overline{E_i} = (\alpha_i)^T V_{ci}$$

avec

$$V_{ci} = \frac{1}{M_i} \sum_{k=1}^{M_i} V_k$$

On calcule donc dans ce procédé d'apprentissage les paramètres α_i du canal par application des algorithmes de prédiction linéaire au centre de gravité des vecteurs de corrélation des points de la base de même étiquette phonémique que le canal.

Ce type d'apprentissage suppose par principe que les classes phonémiques imposées a priori sont bien celles du locuteur étudié ; d'autre part que les erreurs d'étiquetage, inévitables au moment de la constitution de la base, soient rares : le déplacement d'un centre de gravité par des points aberrants peut en effet induire des erreurs importantes dans la discrimination entre les classes phonémiques. Malgré ces faiblesses, auxquelles tente de remédier l'aide semi-automatique exposée à la suite, cette méthode d'apprentissage fournit un bon support pour évaluer la qualité du vocoder à canaux adaptés. Donnons-en un exemple ici ; la base considérée est composée de morceaux de signaux extraits de mots prononcés par un seul locuteur. Elle comporte 1 700 points (soit donc 17 secondes de parole au total), et elle est divisée en 32 classes phonémiques.

On détermine les 32 canaux du vocodeur à canaux adaptés correspondant par la méthode des centres de gravité. On repasse ensuite toute la base sur ce vocodeur ; on mesure pour un point donné, à quel rang apparaît (par erreurs de prédictions croissantes) le canal qui porte la même étiquette phonémique que ce point (l'idéal est bien sûr qu'il soit au premier rang). Sur l'ensemble des points de la base, on obtient typiquement (MICLET, NEHME, 1979) pour l'ordre de prédiction 15 :

Pourcentage des points reconnus au premier rang	66.0
Pourcentage des points reconnus dans les quatre premiers rangs	88.6

Ces résultats sont assez satisfaisants, et semblent valider les hypothèses faites, d'autant plus que la majorité des erreurs provient des plosives non voisées et des sons essentiellement transitoires (certains phonèmes /R/, ou les semi-voyelles).

3.2 Amélioration

Une façon de remédier au côté "aveugle" de la méthode précédente est d'étudier les classes phonémiques les moins bien reconnues dans l'optique de les modéliser par deux ou plusieurs canaux ; ou même de créer un canal dont l'étiquette comportera deux phonèmes, indiquant par là qu'il modélise un ensemble de réalisations intermédiaires au sens de l'analyse effectuée. Si on descend ainsi dans une description plus fine, on a besoin de méthodes d'apprentissage sans professeur, c'est-à-dire de classification automatique. L'expérience nous a en effet prouvé que l'examen direct des nuages de points de la base, par analyse factorielle, était très biaisée du fait de la distance utilisée en reconnaissance par la méthode du vocoder à canaux. Par contre,les méthodes sans professeur, comme les Nuées Dynamiques, permettent d'utiliser en apprentissage la même métrique qu'à la reconnaissance.

On veut dans ce cas chercher dans un ensemble de points, que l'on peut tous supposer étiquetés par le même phonème n° i, les n_i sous-canaux optimaux. La valeur de n_i doit être fixée a priori ; les sous-canaux sont donnés par la

partition de l'ensemble C_i de points qui minimise le critère

$$J = \sum_{j=i}^{n_i} \alpha_j \sum_{k \in C'_j} V_k$$

où C'_j désigne la $j^{ième}$ sous-classe.

L'algorithme des Nuées Dynamiques fait intervenir pour "noyaux" (DIDAY, 1973) les sous-canaux adaptés (α_1 , ..., α_{ni}). La partition est donnée par la distance

$$E(V,\alpha) = \alpha^T V$$

permettant d'attribuer tout point de C_i à la classe la plus proche.

La partition trouvée, le sous-canal est défini à l'aide du centre de gravité de chaque classe trouvée dans la partition.

La méthode des nuées dynamiques consiste donc ici à effectuer alternativement une reconnaissance par les canaux adaptés (c'est-à-dire avec la "distance" de la prédiction linéaire) sur une population non étiquetée, puis un apprentissage par le centre de gravité de chaque classe ainsi séparée.

La convergence en un nombre fini de pas est assurée par les propriétés de la métrique ainsi choisie.

En pratique, la méthode d'apprentissage est semi-automatique, et se déroule ainsi : on calcule les canaux standard pour un locuteur ; on passe la base d'apprentissage sur ces canaux, et on regarde les canaux les plus confondus, les moins discriminants à la reconnaissance. On isole les points de la base d'apprentissage correspondant à ces canaux, et on effectue une classification automatique par l'algorithme des Nuées Dynamiques, avec un nombre de canaux supérieur à deux.

Un certain nombre de procédures interactives permettent à ce stade d'éliminer les points aberrants, très gênants en classification automatique, et de chercher le nombre de classes le plus adapté. On repasse ensuite la base sur le vocodeur à canaux adaptés ainsi modifié, et on évalue l'amélioration apportée. On continue ensuite à chercher des sous-canaux dans les zones phonémiques mélangées. A titre d'exemple, pour un autre locuteur, donnons ici les mêmes chiffres que précédemment (ordre de prédiction : 15)

	Premier rang	4 premiers rangs
Vocodeur à canaux adaptés standard (32 canaux)	60.1	85.6
Vocodeur à canaux adaptés modifié (73 canaux)	75.6	87.2

On constate donc une amélioration sensible de la qualité de la reconnaissance - vite stabilisée si l'on retient les quatre meilleurs canaux. Cette méthode d'apprentissage donne en fait une description plus fine de l'espace d'apprentissage en définissant des sous-phonèmes qui "collent" mieux à la réalité de l'espace d'apprentissage étudié. L'aspect interactif et semi-automatique de l'apprentissage ainsi réalisé permet une manipulation élaborée des données, et conduit à une définition précise d'un vocoder à canaux adaptés associé à un locuteur.

4 - ADAPTATION AU LOCUTEUR

Une des caractéristiques du vocodeur à canaux adaptés - et d'une façon générale des méthodes de reconnaissance de parole par prédiction linéaire - est sa grande sensibilité au locuteur. La reconnaissance d'une base de données issue d'un locuteur L_1 sur le vocodeur à canaux adaptés appris sur un locuteur L_2 est en effet très dégradée. On peut citer par exemple les chiffres suivants (GRENIER, MAURIN, 1979) toujours sous les mêmes conventions :

Pourcentage de reconnaissance (ordre de prédiction : 15) dans les quatre premiers rangs.

Locuteur d'apprentissage	L_1	L_2
Locuteur de reconnaissance L_1	97.1	70.0
L_2	55.0	93.1

Cette dégradation n'est pas forcément un handicap pour construire un système multi-locuteurs : elle traduit simplement la forte adéquation entre un vocodeur à canaux adaptés et l'espace phonémique du locuteur dont il est issu ; s'il est possible de trouver des règles de passage d'un locuteur à un autre, traduites par une transformation à faire subir aux canaux, elle peut au contraire être un avantage : plutôt que de rechercher un vocodeur à canaux adaptés reflétant un locuteur "moyen", on s'attachera à trouver une expression mathématique de la transformation d'un locuteur à un autre, à partir des bases d'apprentissage ; mieux encore, on trouvera cette transformation en examinant une série de mots ou de phrases identiques prononcés par les deux locuteurs, qui permettra de mettre en correspondance les réalisations différentes des mêmes phonèmes.
La méthode qui a été retenue relève de l'Analyse des Données, où elle est connue sous le nom d'Analyse des Corrélations Canonique. Il s'agit d'un procédé statistique destiné à mettre en correspondance la plus étroite possible les deux "univers phonémiques" des locuteurs, en calculant une double projection de l'un et de l'autre. Une première expérience a consisté à calculer ces opérations de projections à partir des bases d'apprentissage des locuteurs (GRENIER, MAURIN, 1979). On obtient alors, pour les mêmes locuteurs :

	L_1	L_2
L_1	97.1	94.3
L_2	90.4	93.1

Ces chiffres permettent de vérifier la validité théorique de la méthode. Cependant, une véritable expérience d'adaptation doit se faire automatiquement : la seconde expérience a consisté à extraire les projections, non pas des bases d'apprentissage, mais des mots identiques prononcés par les deux locuteurs. Le "calage" entre les deux locuteurs, nécessaire pour faire correspondre leurs phonèmes identiques, est réalisé par comparaison dynamique sur l'énergie du signal. Ensuite, le vocodeur à canaux adaptés du second locuteur est calculé à partir de celui du premier, celui-ci étant déterminé à partir d'une base d'apprentissage. Le second locuteur est donc paramétrisé de façon entièrement automatique. Dans ces conditions, on trouve, pour les mêmes locuteurs (MICHEL, 1979) :

	L_1	L_2
L_1	97.1	89.4
L_2	81.6	93.1

Les expériences faites en adaptation automatique semblent assez encourageantes, et c'est un point capital pour la réalisation d'une unité phonémique de qualité dans un système de reconnaissance de la parole.

5 - DECODAGE EN PHONEMES

Le système de reconnaissance par vocodeur à canaux adaptés fournit (cf figure 1) une décision à cadence fixe, chaque instant de parole de 10 ms étant décodé sous la liste des phonèmes les plus proches ; la segmentation en phonèmes et le passage à un treillis phonémique plus classique peut donc se faire à ce stade, c'est-à-dire que la segmentation se fait ici postérieurement à la reconnaissance (On pourrait naturellement imaginer le processus inverse, qui est plus classique). L'intérêt de cette approche est de fournir une description fine du signal dans l'espace phonémique du vocodeur à canaux adaptés à travers lequel il a été analysé. Pour extraire les phonèmes, l'idée est de "lisser" les phonèmes reconnus dans le temps, et de s'appuyer sur l'énergie du signal pour aider cette ségmentation sur les données de reconnaissance. Le procédé choisi est d'associer à chaque canal adapté un automate, dont le rôle est de suivre le comportement de ce canal (relativement aux autres) dans la reconnaissance. Un bon rang dans l'ordre des canaux (déduit des erreurs de prédiction associées) pendant une durée suffisante assure que le phonème associé au canal doit être "proche" pendant cette durée du phonème vraiment émis. Une automate indique donc les débuts et fins des segments où le phonème associé est présent, et lui attribue un score calculé à partir des erreurs de prédiction correspondantes.

L'ensemble des résultats ainsi obtenus est ensuite analysé hors-ligne et les indices de segmentation produits par les automates sont confrontés avec ceux d'un autre automate, qui segmente prioritairement conformément à l'énergie. Les frontières de segmentation sont "agglutinées" pour marquer les limites des phonèmes, et les segments ainsi définis constituent le treillis phonétique cherché (voir la figure 2).

Un tel système permet une grande souplesse de réalisation puisque l'on peut associer à chaque phonème un automate différent (par exemple pour une plosive, le délai de décision doit être plus court que pour une fricative). On peut aussi adapter ces automates aux caractéristiques globales du locuteur (lenteur d'articulation, ou coarticulations systématiques).

Mais le réglage de ces paramètres demande un apprentissage sur des masses très importantes de données ; leur détermination optimale pourrait, là encore, être aidée par des méthodes d'apprentissage automatique. Dans l'état actuel, les caractéristiques du décodeur phonémique ont été fixées de façon simple et indépendantes du locuteur - il n'a pas encore été traité assez de données pour avoir une idée précise sur un réglage plus fin de ses paramètres.

6 - AVANTAGES, DEFAUTS ET PROJETS

L'expérience de plusieurs années d'analyse et de reconnaissance de la parole par prédiction linéaire nous a appris à la fois les grandes possibilités de ces techniques, mais aussi leur "fragilité", et le soin avec lequel il faut mener l'apprentissage avant la synthèse ou la reconnaissance. Sur le plan des avantages, un atout important des méthodes de prédiction linéaire est leur simplicité algorithmique - tant en simulation qu'en réalisation matérielle. La "réalisation" d'un vocodeurà canaux adaptés est en effet très aisée, et on peut en tester l'efficacité à très faible coût informatique. Le principe même de cet organe de reconnaissance fait pencher vers un traitement "en ligne" du signal - et si possible en temps réel grâce à un corrélateur rapide.

Les avantages théoriques de ces méthodes de modélisation sont désormais bien connus. Sur le plan pratique, notons la qualité de la reconnaissance de la plupart des phonèmes (moyennant un apprentissage soigneux - voir par la suite), l'adéquation étroite à un locuteur de son vocodeur à canaux adaptés - et la possibilité non théorique de réaliser une adaptation par locuteur.

Il faut cependant observer des précautions à l'analyse : une préemphase (filtrage passe-haut) du signal est nécessaire - nous la réalisons par une simple double-dérivation de la corrélation. D'autre part, celle-ci est calculée toutes les 10 ms, avec une fenêtre de 25,6 ms - ceci peut induire parfois une "résonance" sur le pitch du locuteur.

Au plan des faiblesses de la méthode, un défaut évident est l'impossibilité de reconnaître avec précision les phonèmes plosifs non voisés. Ils sont en effet caractérisés par des phénomènes temporels brefs, et leur reconnaissance par vocodeur à canaux adaptés est rare, et peu significative. On doit semble-t-il envisager pour ces phonèmes une détection, puis une reconnaissance par d'autres méthodes. Par contre, les voyelles nasales, qui en théorie sont mal décrites par des modèles "tout-pôles", sont en pratique bien reconnues. Le point le plus délicat dans la mise au point d'un système de reconnaissance par vocodeur à canaux adaptés semble en définitive celui de l'apprentissage des canaux. On constate en effet une grande sensibilité des résultats pour des variations relativement légères des paramètres des canaux. Ceci tient au fait que la décision effectuée est prise sur un critère de "distance" que l'on contrôle mal : l'espace des paramètres est de dimension élevée (typiquement, 15), et les méthodes d'analyse de données et projections dans des sous-espaces n'en donnent pas une représentation correcte. Des erreurs légères à l'apprentissage peuvent induire des biais non négligeables ; c'est la raison pour laquelle la méthode exposée au paragraphe 3, ou toute autre approche permettant de vérifier la validité de la phase d'apprentissage, semble nécessaire.

Sur le plan des projets, et directions de recherche dans ce domaine, citons les quelques axes suivants :

- étude et comparaison de distances spectrales issues de la Prédiction Linéaire (généralisation des vocodeurs à canaux adaptés) ;

- détermination variable de l'ordre ;
- étude d'un système de transmission à faible débit par reconnaissance sur un vocoder à grand nombre de canaux adaptés ;
- essai de bases de données importantes pour fixer les paramètres du système.

En particulier sur ce dernier point, beaucoup de travaux restent à effectuer (dont certains grâce à la collaboration avec les équipes du CNET-Lannion). L'expérience déjà acquise nous permet d'espérer d'autres progrès dans ces domaines.

Phonème correspondant au canal adapté reconnu en premier, en second...

Energie du signal

E	Y	J-	EE	I	ON	AI	R	.010133	
E	Y	L	EU	EE	AI	N	AU	.015160	
E	Y	J-	EU	I	EE	AI	R	.018555	chaque ligne
Y	E	J-	I	EU	EE	AI	R	.010079	représente
Y	E	J-	EE	N	I	EU	AI	.002033	10 ms
Y	N	EE	E	J-	EU	I	AI	.002538	
L	U-	Y	N	EE	E	EU	AI	.002367	
L	N	U-	EU	E	Y	AU	AI	.003892	
L	N	M	U-	EU	AI	AU	Y	.006950	
L	N	EU	U-	AI	AU	A	Y	.011926	
L	EU	N	AU	AI	E	A	Y	.018637	
L	EU	W-	EE	AU	O	N	AI	.018803	
W-	O	OU	L	ON	EE	EU	AU	.016706	
W-	OU	O	ON	L	R	EE	N	.014269	
OU	W-	O	ON	R	L	EE	N	.012131	Phrase prononcée :
W-	OU	O	ON	R	EE	L	N	.010858	"Eloignez-vous"...
W-	OU	O	ON	R	B	G	EE	.011307	
W-	OU	O	ON	R	B	EE	G	.013187	
W-	O	OU	ON	R	B	EE	G	.017232	
W-	O	OU	ON	R	AU	EE	L	.025201	/E/L/W-/A/N/...
O	W-	ON	R	OU	AU	EE	L	.041955	
W-	O	R	AU	ON	OU	EU	EE	.062543	
W-	AU	R	EU	AN	ON	AI	EE	.056222	
EU	AU	EE	AI	IN	A	L	AN	.041285	
EU	AI	AU	A	L	E	N	Y	.046832	
AI	EU	E	L	A	N	U-	Y	.049599	
E	AI	EU	L	U-	A	Y	N	.043428	
E	AI	Y	L	U-	EU	A	N	.034469	
E	Y	AI	U-	L	EU	N	A	.026383	
E	Y	L	U-	AI	EU	N	AU	.012976	
L	N	M	U-	W-	O	EU	E	.012364	
L	N	M	U-	EE	W-	O	EU	.014129	
L	N	M	U-	W-	OU	O	EU	.015281	
N	L	M	U-	W-	O	EU	AU	.015642	
N	M	L	U-	OU	W-	O	EU	.015855	
M	N	L	U-	OU	W-	EE	O	.016129	
U-	OU	EE	W-	M	G	N	ON	.015419	
OU	V	EE	R	J-	ON	I	E	.014983	
U-	EE	J-	V	R	ON	I	G	.015370	

Figure 1 Exemple de reconnaissance par vocodeur à canaux adaptés.

534	E	Y	J-	EE	I	ON	AI	R	.010133
535	E	Y	L	EU	EE	AI	N	AU	.015160
536	E	Y	J-	EU	I	EE	AI	R	.018555
537	Y	E	J-	I	EU	EE	AI	R	.010079
538	Y	E	J-	EE	N	I	EU	AI	.002033
539	Y	N	EE	E	J-	EU	I	AI	.002538
540	L	U-	Y	N	EE	E	EU	AI	.002367

Phonème le mieux reconnu pour le segment

FIN DU SEGMENT 45**** Y E

541	L	N	U-	EU	E	Y	AU	AI	.003892
542	L	N	M	U-	EU	AI	AU	Y	.006950
543	L	N	EU	U-	AI	AU	A	Y	.011926
544	L	EU	N	AU	AI	E	A	Y	.018637
545	L	EU	W-	EE	AU	O	N	AI	.018803
546	W-	O	OU	L	ON	EE	EU	AU	.016706

FIN DU SEGMENT 46**** L EU N

547	W-	OU	O	ON	L	R	EE	N	.014269
548	OU	W-	O	ON	R	L	EE	N	.012131
549	W-	OU	O	ON	R	EE	L	N	.010858
550	W-	OU	O	ON	R	B	G	EE	.011307
551	W-	OU	O	ON	R	B	EE	G	.013187
552	W-	O	OU	ON	R	B	EE	G	.017232
553	W-	O	OU	ON	R	AU	EE	L	.025201
554	O	W-	ON	R	OU	AU	EE	L	.041955
555	W-	O	R	AU	ON	OU	EU	EE	.062543
556	W-	AU	R	EU	AN	ON	AI	EE	.056222

FIN DU SEGMENT 47**** W- OU O

557	EU	AU	EE	AI	IN	A	L	AN	.041285
558	EU	AI	AU	A	L	E	N	Y	.046832
559	AI	EU	E	L	A	N	U-	Y	.049599
560	E	AI	EU	L	U-	A	Y	N	.043428
561	E	AI	Y	L	U-	EU	A	N	.034469
562	E	Y	AI	U-	L	EU	N	A	.026383
563	E	Y	L	U-	AI	EU	N	AU	.012976

FIN DU SEGMENT 48**** E EU AI

564	L	N	M	U-	W-	O	EU	E	.012364
565	L	N	M	U-	EE	W-	O	EU	.014129
566	L	N	M	U-	W-	OU	O	EU	.015281
567	N	L	M	U-	W-	O	EU	AU	.015642
568	N	M	L	U-	OU	W-	O	EU	.015855
569	M	N	L	U-	OU	W-	EE	O	.016129
570	U-	OU	EE	W-	M	G	N	ON	.015419
571	OU	V	EE	R	J-	ON	I	E	.014983
572	U-	EE	J-	V	R	ON	I	G	.015370

FIN DU SEGMENT 49**** N L M

Figure 2

Exemple de décodage phonémique

BIBLIOGRAPHIE

1. DIDAY (E.) 1973
Analyse Factorielle Typologique
Rapport Laboria n° 27
INRIA (Rocquencourt)

2. GRENIER (Y.), MICLET (L.) 1978
Reconnaissance de signaux par prédiction linéaire

Congrès AFCET/IRIA
Traitement d'images et Reconnaissance des Formes
Chatenay-Malabry Février 1978

3. GRENIER (Y.), MAURIN (J.C.) 1979
Adaptation au locuteur par analyse canonique des corrélations

2° Congrès AFCET/IRIA
Reconnaissance des Formes et Intelligence Artificielle
Toulouse Septembre 79

4. GRENIER (Y.), MICLET (L.), MAURIN (J.C.), MICHEL (H.) 1981
Speaker adaptation for phoneme recognition
ICASSP 1981 Atlanta (Georgia) USA

5. GUEGUEN (C.), CARAYANNIS (G.), FARJAUDON (T.), LE CHEVALIER (F.) 1975
Un vocodeur à canaux adaptés
L'Onde Electrique Vol 55 n° 7 (1975)

6. GUEGUEN (C.) 1976
Introduction à l'analyse de la parole
7° journées d'Etude sur la Parole
(GACF). Nancy Mai 1976

7. GUEGUEN (C.) 1978
Sur une approche "système" de la Reconnaissance des Formes

Congrès AFCET/IRIA
Reconnaissance des Formes et Traitement des Images
Chatenay-Malabry Février 1978

8. ITAKURA (F.) 1975
Minimum prediction residual principle applied
to speech recognition
IEEE on ASSP Vol 23 1975

9. LE ROUX (J.), MICLET (L.) 1979
Transcription orthographique-phonétique et synthèse en temps réel de la parole par prédiction linéaire

10° journées d'Etudes sur la Parole du GALF
Grenoble Mai 1979

10. MAKHOUL (J.) 1975
Linear Prediction : A tutorial review
Proc. of I E E E Vol 63 n° 4 Apr 75

11. MARKEL (J.), GRAY (A.) 1976
Linear Prediction of Speech
Springer-Verlag 1976

12. MARKEL (J.), GRAY (A.) 1976
Distance Measures for Speech Processing
IEEE on ASSP Vol 24 n° 5 Oct 1976

13. MICLET (L.), NEHME (A.) 1979
Expériences en Reconnaissance de la Parole par Prédiction Linéaire

2° Congrès AFCET/IRIA
Reconnaissance des Formes et Traitement des Images
Toulouse 12-14 Sept 1979

14. MICHEL (H.) 1979
Adaptation d'un système de reconnaissance
Rapport de stage projet de 3° Année
ENST 1979

15. WAKITA (H.) 1979
Residual energy of Linear Prediction applied to vowel and speaker recognition
IEEE on ASSP Vol Juin 1976

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

CLASSIFICATION HIERARCHIQUE AUTOMATIQUE POUR LA CREATION DE PHONES OU DE MOTS DE REFERENCE EN RECONNAISSANCE DE LA PAROLE.

Alain LELIEVRE C.N.E.T. LANNION

.RESUME

Cette étude a été réalisée dans le cadre d'une Thèse de 3ème Cycle ayant pour sujet l'application des méthodes de Classification Hiérarchique automatique aux problèmes posés par la préhension des données en phonétique.

On va envisager la classification de deux genres d'éléments phonétiques : d'une part, les phonèmes et, d'autre part, les mots isolés. On utilise la méthode de Classification Ascendante Hiérarchique de I.C. LERMAN (1970,1976), fondée sur l'algorithme de la vraisemblance du lien (I.C. LERMAN, 1976).

L'étude des phonèmes nous a permis de retrouver la typologie des voyelles et des consonnes établie par les phonéticiens (P. DELATTRE, 1965), ainsi que de définir des taux de reconnaissance pour chacun des phonèmes de la langue française .

En ce qui concerne les mots, nous nous sommes orientés vers la création, à partir d'un grand nombre d'enregistrements, d'un vocabulaire de référence de taille raisonnable, destiné à la reconnaissance multi-locuteurs d'un ensemble de 19 mots isolés.

. ABSTRACT

The study presented in this paper was conducted for a Doctorate of Science Thesis on the application of hierarchical classification methods to problems related to the data acquisition in phonetics.

Two kinds of phonetic elements are considered : phonemes and isolated words. We use the hierarchical classification method (I.C. LERMAN, 1970, 1976) based on the Link Likelihood Algorithm (I.C. LERMAN, 1976).

The study of phonemes enabled to find out the typology of vocalic and consonantal features (P. DELATTRE, 1965), and to define recognition rates for all the phonemes of the French Language.

For isolated words, we tried to establish a reference vocabulary of a reasonable size, from a large number of records. This vocabulary is intended for the recognition of a list of 19 isolated words by different speakers.

I. - L'OBTENTION DES DONNEES.

1) Le matériel d'analyse.

Afin d'obtenir des données utilisables par l'ordinateur, il a été nécessaire de transposer le signal sonore sous forme numérique. Cette opération a été réalisée à l'aide d'un vocodeur à 14 canaux (300 - 4200 Hz) avec une période d'échantillonnage de 13,3 ms.

Couplé à un détecteur de mélodie, le vocodeur fournit ainsi, toutes les 13,3 ms, un vecteur de 15 composantes, l'une de ces composantes représentant la fréquence fondamentale F_o, et les 14 autres donnant l'énergie du signal dans chaque bande de fréquence.

2) Les données relatives aux phonèmes.

On a utilisé ici un texte de 61 phrases relativement courtes, destiné à la Conception Assistée par Ordinateur. Ce texte a été prononcé une fois par trois locuteurs masculins.

Ayant été traité par le vocodeur à canaux, le signal numérisé a ensuite été segmenté en phonèmes. Cette opération (G. MERCIER, 1978) a consisté :

a) à déterminer les limites extrêmes de la parole en éliminant les zones de bruit dues aux conditions d'enregistrement du signal ;

b) à segmenter la parole en syllabes en utilisant la courbe de l'énergie du signal ;

c) à décomposer les syllabes en phonèmes en partant de la voyelle centrale de chaque syllabe.

La fréquence fondamentale F_o est un des critères utilisés pour la décomposition en syllabes et en phonèmes. La donnée de F_o étant de nature tout à fait différente de celle des énergies dans les 14 bandes de fréquence, nous l'avons ensuite écartée. Chaque phonème est alors représenté par une suite d'échantillons de 14 nombres entiers, le nombre de ces échantillons variant, en général, entre 4 et 12. Les phonèmes sont ainsi représentés par des tableaux rectangulaires de nombres entiers positifs dont l'une des dimensions est variable : on peut définir ces suites d'échantillons comme les "parties stables des représentations des phonèmes".

Dans le corpus de données étudié, on a distingué :

- 12 sortes de voyelles : I(/i/), EI(/e/), AI(/ɛ/), A(/a/), O(/o/), OU(/u/), U(/y/), E(/ø/), OE(/œ/), IN(/ɛ̃/), AN(/ɑ̃/), ON(/ɔ̃/).

- 16 types de consonnes,

dont 6 occlusives : P, T, K, B, D, G,

et

10 consonnes non occlusives : F, S, CH(/ʃ/), V, Z, J, L, R, M et N.

.../...

3) Les mots isolés.

On a traité grâce au vocodeur à canaux, une liste de 19 mots (ou phrases) isolés, prononcés 12 fois chacun, et séparément, par dix locuteurs, dont six du sexe masculin et quatre du sexe féminin.

La liste des 19 "mots isolés" est la suivante :

- zéro
- un
- deux
- trois
- quatre
- cinq
- six
- sept
- huit
- neuf
- plus
- moins
- multiplier
- diviser
- parenthèse
- fermer la parenthèse
- virgule
- égale
- annuler

Chaque mot ayant été prononcé douze fois par dix locuteurs, nous nous sommes trouvés en présence de 120 représentants de 19 mots ; chaque référence étant représentée sous la forme d'une suite d'échantillons donnant les énergies correspondant à chaque canal.

II. - LA METHODE DE CLASSIFICATION UTILISEE.

1) La classification automatique.

La classification est si souvent confondue avec le classement qu'il ne nous paraît pas inutile de rappeler la distinction entre ces deux concepts.

Le classement est une opération fondée sur un système de classes déjà conçu. Si l'on dispose d'un certain nombre d'éléments à classer, le seul problème sera alors de choisir la classe à laquelle on affectera chacun d'entre eux.

La classification s'avère être une opération beaucoup plus complexe. En effet, partant d'un ensemble fini d'éléments I décrits par un ensemble fini de caractéristiques, il ne s'agit plus de ranger ces éléments dans un système de classes déjà connu ; mais, au contraire, de définir un ensemble de classes en utilisant uniquement les caractéristiques des éléments de I.

Généralement, ce système de classes est construit, soit sous la forme d'une partition, c'est-à-dire d'une décomposition de l'ensemble I en un certain nombre de parties non vides et disjointes deux à deux, soit sous la forme d'une hiérarchie de classes emboîtées : l'ensemble I est divisé en un certain nombre de classes, elles-mêmes divisées en classes, et cela jusqu'à ce que chacune des classes enfin obtenues ne contienne qu'un élément unique.

La classification automatique désigne simplement tout processus de construction, à partir d'un ensemble fini d'éléments, d'un ensemble de classes empiétantes ou non, hiérarchisées ou non, ce processus étant formalisé pour un traitement de données par l'ordinateur.

.../...

Les techniques de classification automatique constituent donc l'une des composantes essentielles de l'Analyse des Données.

Tout traitement par une méthode de classification automatique doit être précédé d'un travail de préparation que l'on peut décomposer en trois étapes :

1°) Le choix des descripteurs, c'est-à-dire des variables et des unités d'observation ;

2°) La construction du tableau de données brutes, pas forcément à double entrée (par ex. : un tableau d'expériences en physique, les mesures variant dans le temps) ;

3°) Le codage des données, lié aux particularités des objets et des variables.

Pour toute méthode de classification, comme d'ailleurs pour toute méthode d'Analyse des Données, se pose alors le problème du choix d'un tableau de données à traiter. Ce choix, qui donne une signification mathématique à ce tableau, est appelé codage.

Le codage joue un rôle d'autant plus essentiel qu'il a souvent une influence directe sur la qualité des résultats obtenus.

2) La méthode de Classification Ascendante Hiérarchique de I.C. LERMAN. -

Nous présentons brièvement cette méthode de Classification hiérarchique.

Elle est orientée vers la condensation de l'information d'un tableau de données sous la forme d'un arbre de classification. Essentiellement combinatoire, elle procède de l'analyse des corrélations entre structures finies de même type permettant d'approcher le langage du spécialiste des données.

a) L'Algorithme de la Vraisemblance du Lien (A.V.L.)

La notion de proximité entre objets ou entre classes d'objets est fondée sur le principe statistique suivant : cette proximité ne doit tenir compte que de ce qui peut être significatif par rapport à l'hypothèse d'absence de lien entre les sous-ensembles mis en jeu.

Pour établir l'arbre des classifications sur un ensemble I d'objets, on étend la notion de proximité entre objets à celle entre deux classes C et D d'objets ; et ceci grâce à l'Algorithme de la Vraisemblance du Lien (A.V.L.). Soit U un indice de proximité centré réduit sur I (c'est- à-dire se référant à une loi de probabilité centrée réduite) ; dans l'hypothèse d'absence de lien entre deux individus c et d de I, u = U (c,d) est la réalisation d'une variable aléatoire normale centrée et réduite N (0, 1).

.../...

Et donc :

$$P(c,d) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-x^2/2).\, dx,$$

est une réalisation d'une variable aléatoire uniformément répartie entre 0 et 1.

Pour l'A.V.L., la base de la constitution de l'indice de proximité entre C et D sera :

$$p(C,D) = \max \left\{ P(c,d) \; ; \; (c,d) \in C \times D \right\}.$$

Soit 1 et m les cardinaux des classes C et D. Par association de variables aléatoires de même distribution sous l'hypothèse d'absence de lien, on aboutit à l'indice suivant :

$$P(C,D) = [p(C, D)]^{lm}$$

b) Construction de l'arbre.

L'Algorithme de la Vraisemblance du Lien permet à partir de cette mesure de proximité, de procéder à l'agrégation des classes de l'arbre de classification : Card I étant le cardinal de l'ensemble des objets à classifier, au niveau 1, on agrège les deux objets les plus proches ; on obtient ainsi une partition constituée de (Card I - 1) classes. Au niveau suivant, on réunit le couple de classe possédant l'indice de proximité le plus élevé. On obtient une nouvelle partition, et ainsi de suite jusqu'au niveau (Card I-1), où s'agrègent les deux classes de la partition précédente.

On obtient ainsi un arbre de (Card I - 1) partitions emboîtées .

c) Les statistiques de signification des niveaux.

Le nombre de niveaux de l'arbre étant presque aussi grand que le cardinal de l'ensemble I à classifier, il en résulte la nécessité d'un guidage statistique accompagnant la formation des partitions et de la condensation de l'arbre à ses niveaux les plus significatifs.

A cette fin, on attache à chacun des niveaux de l'arbre détaillé des classifications, deux statistiques :

1°) Une statistique globale dont nous allons préciser la notion.

Nous avons déjà calculé les similarités de tous les couples d'éléments de I. Nous ordonnons de manière croissante ces similarités ; c'est ce que nous appelons l'ordonnance des similarités.

Une partition sera bonne si les classes qui la composent sont bien distinctes (c'est-à-dire les similarités entre classes faibles) et bien cohérentes (c'est-à-dire les éléments d'une même classe sont très voisins). Pour mesurer la qualité d'une partition, on utilise une formule statistique introduite par I.C. LERMAN (1976) tenant compte de cette ordonnance des similarités.

.../...

2°) Une statistique locale.

Celle-ci mesure au niveau étudié la qualité de la classe qui vient d'être constituée (c'est-à-dire sa cohérence). Cette classe est formée par la réunion des deux classes les plus proches du niveau précédent. Pour déterminer cette valeur, on utilise à nouveau une formule statistique de même type que la précédente mais restreinte aux deux classes qui viennent de s'agréger, d'où la notion de statistique locale par rapport à celle de statistique globale (I.C. LERMAN, 1976).

La distribution de cette statistique le long de la suite des niveaux est telle que sa valeur augmente lorsqu'une classe en cours de formation se confirme et décroît sensiblement devant l'arrêt de constitution d'une classe au profit de la naissance de l'embryon d'une autre classe.

Les niveaux associés aux maximums locaux de cette distribution correspondent par conséquent à des niveaux d'achèvement des classes. Ils sont appelés niveaux significatifs.

III. - CLASSIFICATION DES PHONEMES.

1) L'indice de proximité utilisé : l'indice COSINUS.

Il s'est tout d'abord posé le problème du choix d'un indice de proximité entre échantillons constitutifs des tableaux représentatifs des phonèmes. Après avoir testé plusieurs indices, tels que l'indice de corrélation ou l'indice COSINUS, nous avons opté pour ce dernier, qui nous a fourni les meilleurs résultats.

Soit I un ensemble d'items à classifier.

Soit V l'ensemble des variables définissant ces objets.

L'indice de proximité entre deux items $x = \{ v(x) ; v \in V \}$ et $y = \{ v(y) ; v \in V \}$ sera défini de la manière suivante :

$$S(x, y) = \cos(x,y) = \frac{\sum_{v \in V} v(x)\, v(y)}{\sqrt{\left[\sum_{v \in V} v^2(x)\right]\left[\sum_{v \in V} v^2(y)\right]}}$$

Si on considère la représentation du nuage des objets dans RP, où un point x a pour coordonnées $\{ v(x) ; v \in V \}$, où $v(x)$ est la mesure de la variable v sur l'individu x, alors l'indice $S(x, y)$ mesure le Cosinus de l'angle des deux rayons-vecteurs représentant x et y.

L'indice COSINUS présente l'avantage d'être normalisé, propriété qui ici revient sensiblement à diminuer l'effet du niveau d'énergie auquel a été prononcé le phonème.

.../...

Nous avons utilisé, pour le calcul des proximités entre phonèmes, deux méthodes différentes : d'abord, nous avons réduit chaque tableau représentatif d'un phonème, sous la forme d'un vecteur, et appliqué directement l'indice COSINUS ; ensuite, nous avons appliqué la programmation dynamique à la comparaison entre ces tableaux, l'indice de proximité entre échantillons étant toujours l'indice COSINUS.

On a étudié séparément les voyelles et les consonnes.

2) Classification des voyelles.

En appliquant la méthode de Classification Ascendante Hiérarchique des échantillons d'un même tableau représentatif d'une voyelle du français, il est facile de distinguer trois zones qui se distinguent nettement les unes des autres :

- une zone instable initiale, qui correspond, en général, à l'articulation avec le phonème précédent ;
- une zone stable centrale, qui comprend les échantillons médians ;
- une zone terminale, instable, correspondant souvent à l'articulation avec le phonème suivant.

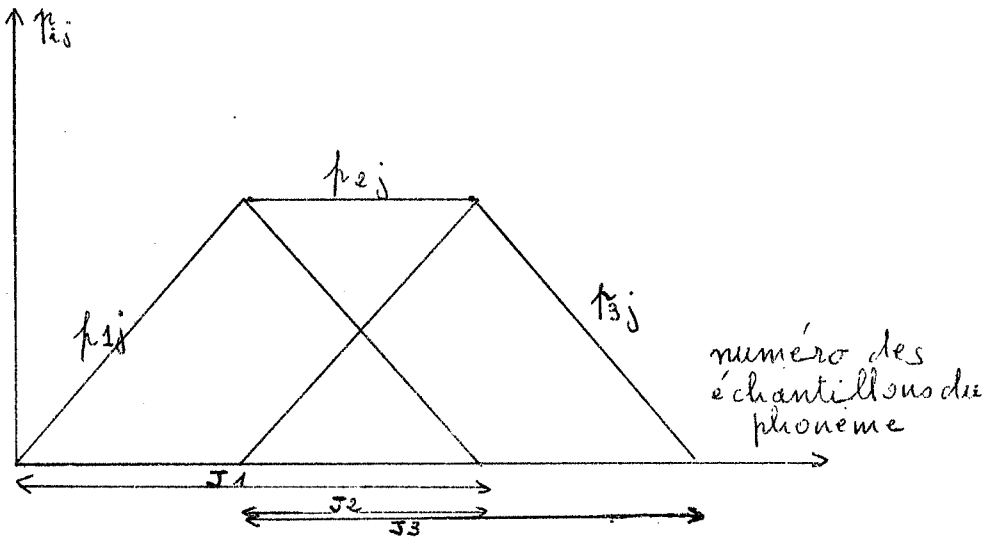
D'où l'idée de coder les tableaux représentatifs des voyelles à partir du processus suivant :

a) Division du bloc phonème en trois zones.

b) Elaboration d'un système de pondérations de ces trois zones ; on obtient ainsi un tableau de 3 vecteurs à 14 composantes.

d) Constitution, à partir de ces trois nouveaux échantillons, d'un vecteur à 42 composantes.

Nous avons ainsi testé divers sytèmes de pondérations sur trois zones représentatives du phonème. Nous avons obtenu les meilleurs résultats à partir d'un système de pondérations empiétantes sur trois zones J_1, J_2 et J_3, que l'on peut représenter de la manière suivante :



Les pondérations $p_1 j$, $p_2 j$ et $p_3 j$ vérifient l'égalité suivante :

$$\sum_{j \in J_1} p_{1j} = \sum_{j \in J_2} p_{2j} = \sum_{j \in J_3} p_{3j} = 1.$$

Pour chacun des trois locuteurs, la classification des voyelles (environ 160 représentants) nous a permis de bien distinguer, aux plus hauts niveaux, les voyelles antérieures (I, U, EI, AI) des voyelles postérieures (OU, ON, AN, IN, ON). Ensuite, aux niveaux légèrement plus bas, aparaissent les séparations entre voyelles fermées et voyelles ouvertes. On peut noter, jusqu'à un niveau très élevé, la séparation des voyelles A en deux classes : on retrouve ainsi la distinction entre le A fermé et le A ouvert du français.

On a considéré alors comme partition optimale, celle correspondant à un noeud significatif (cf. § II) et dont le nombre de classes est le plus proche de 12, c'est-à-dire du nombre de voyelles. Chacune des classes de cette partition a alors été considérée comme représentant la ou les deux voyelles les plus présentes au sein de cette classe, ces voyelles sont alors considérées comme reconnues par la classification.

On a ainsi construit un taux de reconnaissance, égal au rapport du nombre de phonèmes reconnus sur le nombre de phonèmes total. On obtient les résultats suivants pour chacun des trois locuteurs et sur l'ensemble d'apprentissage :

Taux de reconnaissance	Locuteur 1	Locuteur 2	Locuteur 3
1ère position	57,0%	69,5 %	55,9 %
2ème position	85,4 %	84,4 %	77,6 %

On a ensuite utilisé la même méthode, mais sans coder les données et en utilisant une version de la programmation dynamique (H. SAKOE and S. CHIBA, 1978), que nous avons adaptée à la classification hiérarchique. Comme indice de proximité entre échantillons, nous avons adopté le Cosinus entre vecteurs représentant les différences entre canaux consécutifs.

On a obtenu des résultats sensiblement identiques aux précédents, mais légèrement meilleurs. Les taux de reconnaissance ont été les suivants:

Taux de reconnaissance	Locuteur 1	Locuteur 2	Locuteur 3
1ère position	67,7 %	69,5 %	67,7 %
2ème position	88,0 %	89,2 %	89,4 %

Ces résultats demandent à être vérifiés sur un ensemble test différent de l'ensemble d'apprentissage.

.../...

3) Classification des consonnes.

La classification des consonnes s'est avérée beaucoup plus délicate, en particulier en raison de la présence des occlusives.

Nous avons tout d'abord utilisé une méthode de codage des blocs représentatifs des phonèmes. Cette méthode suit le même algorithme que celui exposé plus haut, mais le codage optimal obtenu a consisté à affecter des pondérations uniformes aux échantillons, et cela sur deux zones contigües d'égale longueur, et constitutives des blocs phonèmes. L'existence de ces deux zones a également été validée par la classification hiérarchique.

On a également utilisé, de la même manière que pour les voyelles, l'indice de programmation dynamique.

On a obtenu les taux de reconnaissance suivants sur l'ensemble d'apprentissage :

PONDER. UNIFORMES	1er locuteur	2ème locuteur	3ème locuteur
1ère position	49,0 %	55,6 %	63,7 %
2ème position	69,7 %	72,9 %	78,1 %

PROGR. DYNAMIQUE	1er locuteur	2ème locuteur	3ème locuteur
1ère position	49,7 %	51,4 %	61,0 %
2ème position	71,7 %	67,4 %	79,5 %

Comme pour les voyelles, cela demande à être vérifié sur l'ensemble test.

L'arbre des classifications, dans les deux cas, nous permet de distinguer assez bien les occlusives et les consonnes liquides. Les principales difficultés semblent provenir de la séparation des fricatives voisées (V, J, Z), des fricatives non voisées (F, S, CH) et des nasales (M, N).

Les taux de reconnaissance ne sont pas très élevés ; ils sont plus faibles que pour les voyelles. On retrouve ici les difficultés inhérentes à la reconnaissance des consonnes et peut-être à des erreurs de segmentation.

Que ce soit pour les voyelles ou pour les consonnes, l'indice obtenu à partir d'une version adaptée de la programmation dynamique semble fournir des résultats légèrement meilleurs. Mais cet indice présente l'inconvénient de nécessiter un temps calcul beaucoup plus long que celui issu d'un précodage des données.

IV. - LA RECONNAISSANCE DES MOTS ISOLES.

A partir des données citées plus haut, nous avons créé deux corpus: le premier, comprenant la moitié des références, soit six représentants de chacun des mots prononcés par chaque locuteur, a été destiné à la reconnaissance ; l'autre corpus a été utilisé afin de créer un vocabulaire de taille relativement réduite, qui servira de référence en vue de la reconnaissance multi-locuteurs.

Dans ce but, on a réalisé, pour chacun des types de mots ou phrases prononcés, une classification hiérarchique de ces 60 représentants retenus, en utilisant la programmation dynamique, appliquée à l'indice COSINUS, comme mode de calcul des proximités.

Puis, dans chaque arbre de classification, on a retenu la partition optimale (c'est-à-dire correspondant à un noeud significatif) dont le nombre de classes est le plus proche de 6. On a alors calculé l'élément-type de chacune des classes retenues, en définissant cet élément-type comme le représentant dont la somme des proximités aux autres éléments de sa classe est maximale.

Ainsi, on a réduit l'ensemble des 60 représentants de chacun des 19 mots isolés, à un groupe d'environ 6 éléments résumant au mieux l'ensemble initial.

On a donc obtenu un nouveau vocabulaire de référence multi-locuteurs, comprenant en tout 105 références.

Nous avons alors appliqué notre programme de reconnaissance aux représentants constituant l'autre partie du corpus des données : pour chacune des références, le mot reconnu est celui possédant la proximité la plus élevée avec la référence considérée. Les proximités ont été, ici aussi, calculées grâce à l'algorithme de programmation dynamique.

On a ainsi obtenu les taux de reconnaissance suivants :

HOMMES						
	Loc.1	Loc.2	Loc.3	Loc.4	Loc.5	Loc.6
1ère position	94,7 %	89,5 %	84,2 %	94,7 %	84,2 %	68,4 %
2ème position	100 %	89,5 %	89,5 %	100 %	89,5 %	68,4 %

FEMMES				
	Loc.7	Loc.8	Loc.9	Loc.10
1ère position	78,9 %	94,7 %	68,4 %	84,2 %
2ème position	89,5 %	94,7 %	84,2 %	89,5 %

.../...

Les taux de reconnaissance sont, en général, supérieurs à 2/3 en première position et à 80 % en 2ème position. Le 6ème locuteur constitue une exception ; ses moins bons résultats proviennent vraisemblablement d'une mauvaise séparation bruit-parole.

On constate une faible dégradation des résultats par rapport à une reconnaissance mono-locuteur, pour laquelle les taux de reconnaissance varient entre 90 et 95 %.

CONCLUSION :

Bien que la classification ait déjà été utilisée dans ce domaine, la méthode de Classification Ascendante Hiérarchique semble être une nouvelle approche intéressante de la reconnaissance de la parole. Il serait pertinent d'utiliser d'autres méthodes et de comparer les résultats. De même, la classification hiérarchique pourrait, sans doute, avoir des applications dans d'autres domaines de la phonétique.

En ce qui concerne la reconnaissance des mots isolés, nous avons été limités par le temps. Nous envisageons d'utiliser le vocabulaire de référence pour la reconnaissance de mots prononcés par des locuteurs qui ne sont pas intervenus dans l'élaboration de ce vocabulaire.

N.B. : Je tiens à remercier Mr le Professeur I.C. LERMAN, de l'Université de Rennes I, ainsi que Mmes N. VALETTE, M.T. ROTH, et Mrs C. GAGNOULET, D. LE CLERC et G. MERCIER, du CNET-LANNION, pour leurs précieux conseils tout au long de ce travail.

BIBLIOGRAPHIE

BENZECRI (J.P.), 1976 : L'Analyse des Données, Tomes 1 et 2 ; Dunod, Paris.

DELATTRE (P.), 1965 : Comparing the phonetic features of english, german, spanish and french ; Julius Groos Verlag, Heidelberg.

JAMBU (M.), 1978-1979 : Classification Automatique pour l'Analyse des Données, Tomes 1 et 2, Dunod, Paris.

LEBART (L.), FENELON (J.P.), 1971 : Statistique et Informatique appliquées, Dunod, Paris.

LERMAN (I.C.), 1970 : Les bases de la Classification Automatique ; Gauthier-Villars, "Collection Programmation" Paris.

LERMAN (I.C.), 1976 : Reconnaissance et Classification des structures finies en Analyse des Données ; Rapport I.R.I.S.A. n° 70, Université de Rennes I.

MERCIER (G.), 1978 : Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL, 9è J.E.P., p 321 à 342, Lannion 1978.

RABINER (L.R.), 1978 : On creating reference templates for speaker independant recognition on isolated words ; I.E.E.E. Transactions on A.S.S. Processing, vol. A.S.S.P.26, n° 1, febr. 1978.

ROTH (M.T.), 1979 : Variations acoustiques de la zone stable de la voyelle sous l'influence des consonnes dans les monosyllabes en français ; Note Technique DAS/SST/23, CNET-LANNION, Oct. 1979.

SAKOE (H.) and CHIBA (S.), 1978 : Dynamic Programming Algorithm Optimization for spoken word Recognition ; I.E.E.E. Transactions on A.S.S. Processing, vol. A.S.S. P. 26, n° 1, febr. 1978.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

L'ANALYSE ACOUSTIQUE, L'ANALYSE PHONETIQUE ET L'APPRENTISSAGE DANS LE SYSTEME DE RECONNAISSANCE DE LA PAROLE CONTINUE KEAL II.

GUY MERCIER — CNET - LANNION A - TSS/RCP

MARTINE GERARD — CNET - LANNION A - TSS/RCP

.**RESUME**

Ce papier décrit les modules liés à l'analyse acoustique-phonétique et à l'apprentissage des paramètres phonétiques.

Dans le premier paragraphe on présente très brièvement la structure générale de ces modules et leur enchaînement.

Dans le paragraphe suivant, on décrit les différentes représentations du signal et les paramètres acoustiques choisis pour identifier les phonèmes. Dans la troisième partie on détaille les étapes successives de l'analyse phonétique ainsi que le module d'adaptation à un nouveau locuteur.

ACOUSTIC, ANALYSIS, PHONETIC ANALYSIS AND TRAINING IN THE CONTINUOUS SPEECH RECOGNITION SYSTEM : KEAL II

. ABSTRACT

This paper describes the acoustic-phonetic recognizer which performs the early stages of analysis in the KEAL speech understanding system. The objective of this module is to transform the continuous speech signal representing the uttered sentence into a string of phoneme-like units.

The KEAL acoustic phonetic recognizer consists of components for carrying out three main tasks : Acoustic analysis, labelling, training (fig 1). These components are themselves composed of modules, each corresponding to an important step of analysis. Each labelling module proceeds from the data or the hypotheses constructed during previous steps, refining them or building new ones in order to advance the phonemic recognition process.

The labelling procedure we have carried out consists in applying both an ordered sequence of logical rules (structural approach) and a set of decision functions (statistical discriminant approach) in order to distinguish phonemes.

- The logical rules are used for segmentation and for recognizing rough-phonetic categories.

- The discriminant approach we use consists in representing the units to be recognized by points in a R^n space and then in separating them by means of hyperplanes. This method is well-adapted for recognizing stationary acoustic-segments like vowels.

The parameters of the discriminant functions are computed automatically during a training phase. They can be easily turned to other speakers or to new recording conditions.

- The end result is a phonetic lattice displaying alternative possibilities for segmentation and labelling.

L'ANALYSE ACOUSTIQUE, L'ANALYSE PHONETIQUE ET L'APPRENTISSAGE DANS LE SYSTEME DE RECONNAISSANCE DE LA PAROLE CONTINUE KEAL II

MERCIER G. GERARD M. CNET LANNION

INTRODUCTION

La reconnaissance de la parole continue est une tâche très complexe qui nécessite l'étude et la mise au point de nombreux modules liés à différentes sources de connaissance (phonétique, syllabique, lexicale, prosodique, syntaxique, sémantique etc...). Par l'intermédiaire de ces modules de reconnaissance,un système peut émettre des hypothèses successives de plus en plus "sûres" sur des phonèmes, syllabes, mots, phrases etc....

De nombreuses équipes françaises et étrangères travaillent depuis plusieurs années à l'étude de tels systèmes (CARRÉ et al.1979).

Au CNET une première version en temps réel du système KEAL de dialogue homme-machine, admettant la parole continue comme support, a fonctionné en 1977. Deux années d'expériences nous ont permis d'évaluer les performances des différents modules composant la partie reconnaissance et de définir une seconde version (KEAL II). Ce système, destiné à la reconnaissance multi-locuteurs, comporte également des modules d'adaptation automatique (ou tout au moins semi-automatique dans une première étape) à tout nouveau locuteur. Ce système, en cours de mise au point sur un IRIS80, va être implanté sur un ensemble de mini-calculateurs.

Nous nous limitons dans cette communication à la description des modules liés à l'analyse acoustique-phonétique et à l'apprentissage des paramètres phonétiques.

Dans le premier paragraphe nous présentons très brièvement la structure générale de ces modules et leur enchaînement. Dans le paragraphe suivant, nous décrirons les différents moyens dont on dispose pour représenter le signal et les paramètres acoustiques qui seront effectivement choisis pour identifier les phonèmes. Dans la troisième partie, nous détaillerons les étapes successives de l'analyse phonétique et enfin nous présenterons le module d'apprentissage des paramètres spécifiques d'un locuteur et d'adaptation à un nouveau locuteur.

I ORGANISATION GENERALE

L'analyseur acoustique phonétique est constitué d'un ensemble de modules correspondant chacun à une étape importante de l'analyse : traitement acoustique du signal, extraction de paramètres, segmentation, hypothèses sur les traits etc.... Chaque module a pour fonction de faire progresser la reconnaissance phonétique à partir des hypothèses émises à l'étape précédente.

La procèdure adoptée consiste à appliquer soit une hierarchie de règles logiques, soit un ensemble de fonctions de décision permettant de distinguer des classes dans un espace R^n. La règle suivante est un exemple de règles utilisées : si le rapport entre l'énergie dans les hautes fréquences et l'énergie dans les basses fréquences est supérieur à un seuil fixé à l'avance pendant un temps suffisamment long, on a détecté un segment fricatif avec une probabilité plus ou moins grande.

Ces règles et ces fonctions représentent la source de connaissance phonétique du système. Les régles sont en général fixées au départ ; par contre les paramètres des fonctions de décision et les seuils sont, soit calculés par apprentissage, soit éventuellement modifiés en cours de traitement pour les adapter à un nouveau locuteur ou à de nouvelles conditions d'enregistrement.

La figure 1 montre la structure générale de cet analyseur. Les flèches symbolisent les transferts de données et d'hypothèses entre les divers modules et indiquent la stratégie générale adoptée pour aboutir au treillis phonétique final.

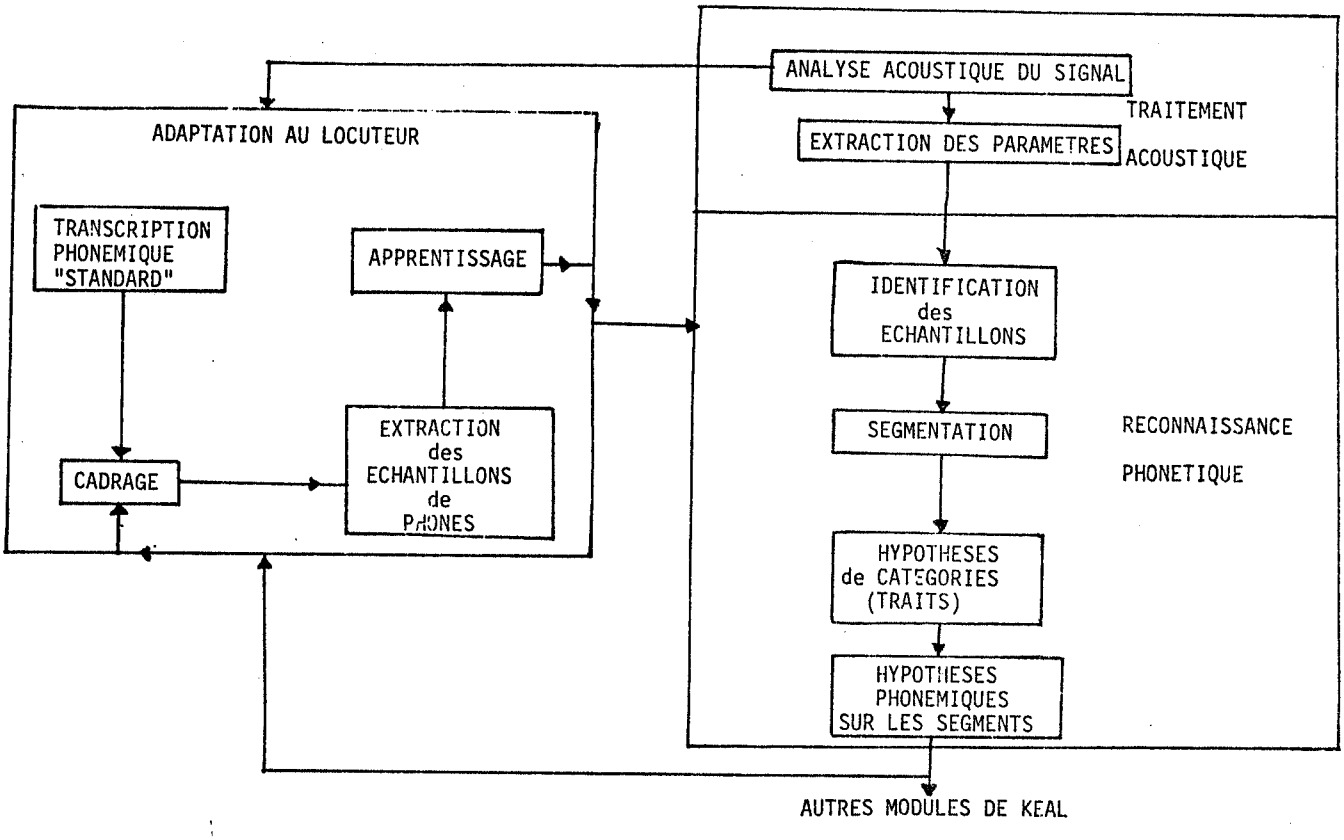


Fig. 1 - STRUCTURE GENERALE DE L'ANALYSEUR PHONETIQUE ET DU MODULE D'APPRENTISSAGE

II L'ANALYSE DU SIGNAL ET L'EXTRACTION DES PARAMETRES

Différentes représentations sont disponibles sur le système (fig.2) ; la représentation utilisée dans la première version de KEAL était une représentation spectrale obtenue à la sortie d'un vocodeur à canaux analogique (14 canaux couvrant la bande 250-4200 Hz) ; nous disposons maintenant d'une représentation identique obtenue à la sortie d'un vocodeur numérique dont le nombre et la répartition des filtres ainsi que la période d'analyse spectrale(échantillonnage) sont programmables. Un deuxième vocodeur à canaux analogique doit être relié au système ; la répartition et le nombre des filtres sont différents (16 canaux couvrant la bande 250-5000 Hz).

Al'aide d'un codeur analogique-numérique, branché sur le système, une représentation différée du signal sous forme de F.F.T., de coefficients de prédiction linéaire ou de cepstre est possible ainsi qu'une représentation par un modèle de cochlée (CAELEN,1979).

Un mélographe (ZURCHER,1977) fonctionnant en temps réel détecte et mesure le fondamental dans la gamme 60-500 Hz.

L'adaptation du programme d'analyse phonétique à ces représentations diverses est en cours ; la comparaison de ces méthodes d'analyse dans le cadre de la reconnaissance de mots isolés par programmation dynamique sur les paramètres acoustiques a déja donné quelques résultats(STELLA,1980 ;COUVRAT, GAGNOULET,1981)

En plus des paramètres fournis par l'une ou l'autre de ces analyses, on calcule un certain nombre de paramètres de base : les fréquences et amplitudes des pics d'énergie(DUBOIS,1980 ; ROTH,1981) ou des formants obtenus soit à partir du banc de filtre et d'un programme simple de restitution des formants (ROSSI,1978) soit à partir de la prédiction linéaire ou du cepstre, l'énergie totale toutes les centisecondes, le centre de gravité fréquentiel, les énergies et les centres de gravité fréquentiels dans les bandes suivantes :250-850 hz, 850-1600 Hz , 1600-2800 Hz , 2800-4200 Hz , la dérivée du spectre (pente entre les spectres de deux échantillons consécutifs), le relief (variation de l'énergie entre les canaux) et l'évolution des formants ou les déplacements des centres de gravité sont autant d'indices intéressants pour la segmentation et la détection des phonèmes.

III L'ANALYSE PHONETIQUE

La sortie du module d'analyse phonétique correspond à une transcription du signal de parole sous la forme d'un treillis de symboles phonétiques. Grâce à une série de tests sur les paramètres extraits précédemment et sur les données acoustiques, on élabore des hypothèses de segmentation et de classification ; l'application de nouvelles règles permet d'affiner progressivement ces hypothèses.

Les phases principales sont les suivantes (figure 3) :

- la détection des limites de laphrase, la distinction voyelle-consonne, l'étiquetage des échantillons toutes les centisecondes, la segmentation en pseudo-syllabes, la détection des zones stables et instables séparant les noyaux vocaliques, l'identification phonétique de ces zones, la constitution du treillis des hypothèses phonémiques.

III-1 La détection des limites extrêmes de la phrase.

Une première détection est obtenue à partir des trois variables suivantes mesurées toutes les centisecondes :
l'énergie, la dérivée du spectre et la répartition de l'énergie en fréquence ; une fonction linéaire de ces trois paramètres évaluée sur un intervalle de temps suffisant (40 à 80 ms) et une fonction de décision déterministe nous donnent une première approximation du début et de la fin de la phrase. Par la suite, l'étiquetage phonétique toutes les centisecondes nous donne la possibilité d'éliminer certains échantillons isolés et d'ajuster au mieux les limites.
L'introduction d'un algorithme de reconnaissance des formes syntaxiques, pour repérer les pics d'énergie, et de l'algèbre floue, pour évaluer la possibilité pour ces pics de correspondre à des extrémités du signal de parole, a permis d'améliorer cette détection (R.DE MORI, G. MERCIER, 1981). Grâce à ces différents filtres, les vraies limites de la phrase sont obtenues dans presque tous les cas.

III-2 La distinction voyelle-consonne.

Dans cette seconde étape, on affecte la classe voyelle ou consonne à chacun des échantillons spectraux de la phrase. On introduit les paramètres acoustiques suivants : les énergies et les centres de gravité de la distribution d'énergie dans les bandes de fréquences basses, centrales, hautes et totales. On applique quelques tests simples sur ces paramètres et, en fonction de la valeur vraie ou fausse on prend la décision : voyelle ou consonne. Les divers paramètres et seuils sont ajustés par rapport à la valeur maximale de l'énergie des formants des voyelles de l'enregistrement. L'utilisation de l'algèbre floue en cas de doute (semi-voyelles par exemple) , permet d'envisager une prise de décision moins brutale.

III-3 L'étiquetage phonétique des échantillons.

A chaque échantillon spectral est affectée une étiquette phonétique plus précise, à l'aide de fonctions linéaires dont on a calculé les coefficients par apprentissage. Ces fonctions sont obtenues de la façon suivante : on transforme chaque échantillon à l'instant t en définissant un vecteur X(t), dont les premières coordonnées correspondent aux pentes de l'énergie entre deux canaux (bandes de fréquence) consécutifs et les coordonnées suivantes

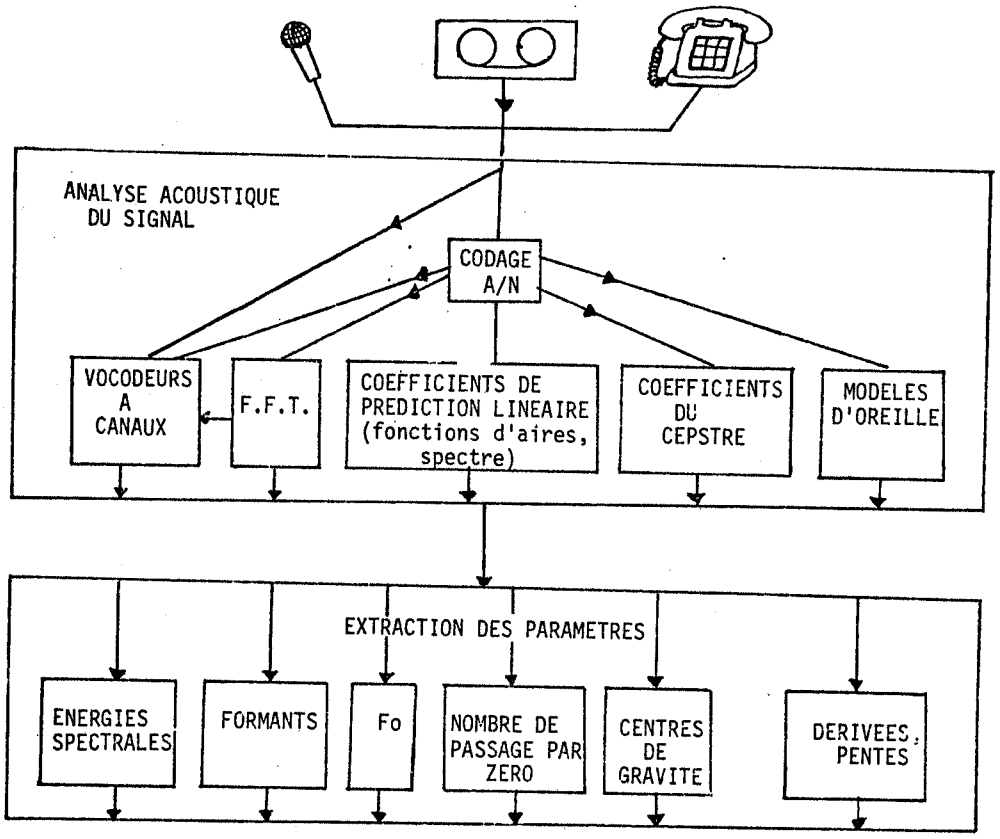


Fig. 2 - Analyse du signal et paramètres disponibles

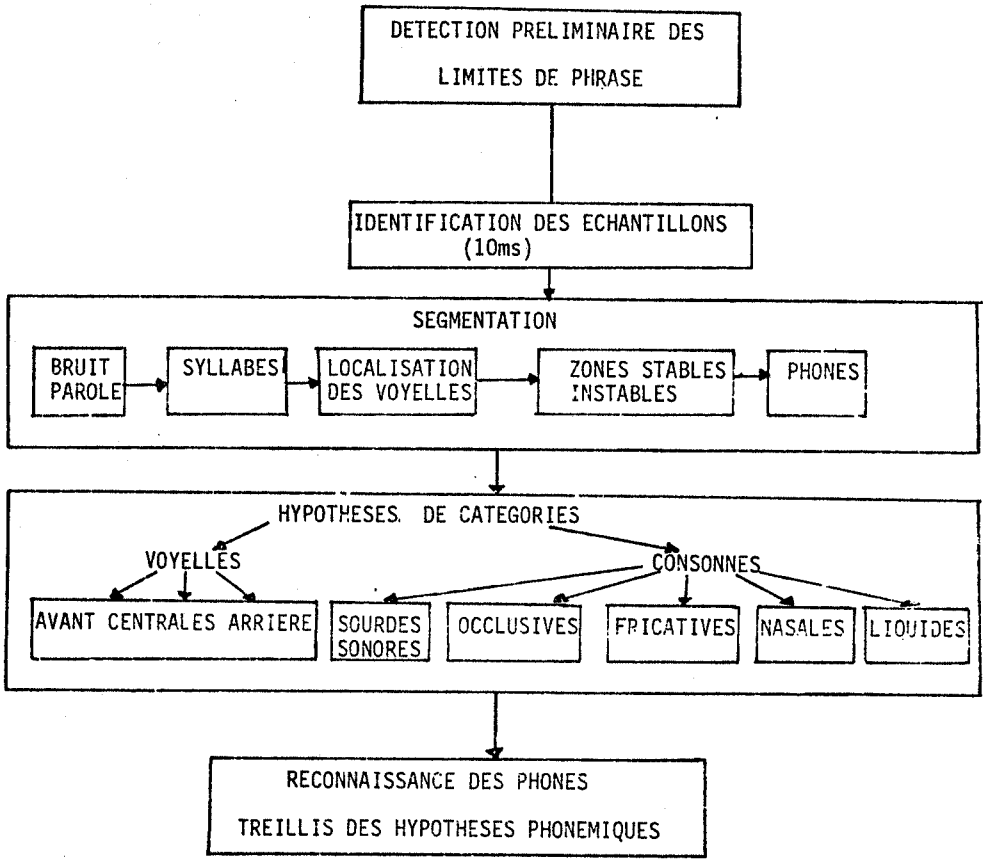


Fig. 3 - Schéma général de l'analyse phonétique

correspondent respectivement à l'énergie du spectre à l'instant t, au centre de gravité fréquentiel du spectre et aux énergies et centres de gravité fréquentiels dans les 4 zones de fréquence définies dans le paragraphe II. Dans la version KEAL II, ces coordonnées seront paramètrées et pour représenter les échantillons, à l'instant t, on pourra disposer aussi bien des coefficients du cepstre, de la prédiction linéaire, du modèle d'oreille que d'un banc de filtres (dont le nombre et la répartition des canaux sont eux aussi paramètrables).

On se définit ensuite trois classes : la classe des voyelles, la classe des consonnes et la classe silence qui, elle, subit un traitement particulier. Dans chacune des classes (voyelles ou consonnes) on a un nombre de variétés allophoniques, c'est-à-dire de "phones" différents; ce nombre ayant été précisé ou déterminé pendant la phase d'apprentissage correspond approximativement au nombre d'allophones qu'on peut distinguer dans chaque classe.

Soit l'échantillon $X(t) = (X_1^t, \ldots, X_j^t, \ldots, X_n^t, X_{n+1}^t)$
avec n = nombre de coordonnées retenues et $X_{n+1}^t = 1$
on calcule les m produits scalaires suivants

$$F^t(i) = \sum_{j=1}^{n+1} v(i,j)\, X_j^t \qquad i = 1, m$$

où v(i,j) correspond aux coefficients calculés pendant l'apprentissage (paragraphe IV et annexe I).

On compare les m produits scalaires F(i) les uns par rapport aux autres. On garde en mémoire les 3 phones correspondant aux trois produits scalaires les plus élevés : ce qui nous permet d'étiqueter chaque échantillon. L'étiquette silence sera affectée à l'échantillon si celui-ci possède une énergie faible et un relief très plat.

III-4 La segmentation en pseudo-syllabes.

Le principe de base de cette segmentation est la recherche du noyau vocalique de la syllabe ; elle s'effectue en trois étapes.

Dans la première étape, la courbe représentant la variation de l'énergie E (t) au cours du temps est décomposée en segments consécutifs, par localisation des maxima d'énergie. Au cours de la deuxième étape, chacun des segments précédents est réexaminé afin de savoir s'il contient un noyau vocalique. On fait appel à 9 paramètres acoustiques représentant des énergies, des positions de formants, des proportions d'échantillons non-voisés ou vocaliques contenus dans le segment.

Au cours de la troisième étape on détermine les frontières des pseudo-syllabes qui, habituellement, coïncident avec le début de la zone stable contenant le mininum d'énergie situé entre les deux noyaux vocaliques des deux syllabes adjacentes.

L'adaptation de certains paramètres à la vitesse de l'élocution devrait conduire à une meilleure segmentation dans le cas de personnes ayant un debit rapide.

III-5 La segmentation en phonèmes.

Les voyelles sont approximativement localisées dans la zone stable entourant le maximum d'énergie de la syllabe.

Le paramètre le plus important pour repérer les parties stables et transitoires des consonnes est la pente P (t) du spectre ; cependant les résultats de l'identification hiérarchisée des consonnes peuvent, par la suite, entraîner une modification de cette segmentation.

III-6 L'identification des segments.

III-6.a La reconnaissance des voyelles.

Un premier étiquetage des échantillons de la voyelle ayant été effectué dans l'étape III-3, on calcule sur l'ensemble des échantillons de la voyelle une valeur moyenne pour chaque fonction linéaire mesurant la ressemblance aux divers phonèmes, et on conserve les réponses phonétiques correspondant aux valeurs les plus élevées de G(i)

avec $G(i) = \frac{1}{m} \cdot \sum_{t=1}^{m} F^t(i)$

avec $F^t(i)$ défini au paragraphe III -3

et m = nombre d'échantillons de la voyelle

Un programme parallèle de reconnaissance des voyelles utilisant la notion de traits est en cours de mise au point (ROSSI,NISHINUMA, LE MONTREER,1981) et doit par la suite s'intégrer dans le système KEAL II.

III-6.b La reconnaissance des consonnes.

La procédure suivante est utilisée pour reconnaître les segments"consonne".

- Les consonnes voisées sont séparéees des consonnes sourdes à partir des sorties du mélographe, des énergies et différences d'énergie dans les bandes (250-650 hz) et des étiquettes précédemment affectées aux échantillons "consonne". A l'avenir la décision finale voisée-non voisée utilisera des règles contextuelles.
- Une "plosive" (ou plutôt le trait "interrompu") est détectée si le segment est composé d'une zone de silence ou de faible énergie (barre de voisement) dans les basses fréquences (250-450 Hz) suivie d'une forte explosion et d'une variation brusque du spectre, c'est-à-dire de l'énergie et des centres de gravité.
- Une constrictive (fricative) sera repérée si le segment contient une proportion suffisante d'échantillons dont le centre de gravité de la distribution d'énergie dans le spectre est supérieur à un seuil fixé à l'avance.
- Si, après une plosive, on repère sur la courbe d'énergie un maximum, en présence d'une énergie suffisante dans les premiers canaux, on affecte au segment la classe "liquide" : /l/ ou /r/.
- Un module de reconnaissance de la place d'articulation des plosives est en cours de mise au point ; il utilise les centres de gravité, dans les zones de fréquence basse, hautes et totales, moyennés sur le segment correspondant à l'explosion ("burst"), et les transitions des formants vers la voyelle suivante ; les algorithmes de décision prennent en compte la place d'articulation de la voyelle suivante (DE MORI,MERCIER 1981)

Un module de reconnaissance des consonnes nasales est en cours d'étude. L'algorithme, basé sur les valeurs et l'évolution des fréquences des formants F_1, F_2, F_3 ainsi que sur la position du premier antiformant, tiendra compte du contexte vocalique entourant la nasale (DE MORI, GUBRYNOWICZ, LAFACE,1979; CAELEN,1979).

NUMERO DU SEGMENT		1	2	3		4	5	6	7	8
POIDS DU SEGMENT		1	1	1		1	0.5	1.	1.	1.
N° DE LA SYLLABE		1	2	2		2	3	3	4	4
ADRESSE DE DEBUT DU SEGMENT		4	18	24		37	49	53	72	77
ADRESSE DE FIN DU SEGMENT		14	22	36		47	53	57	77	85
SORTIES PHONEMIQUES		ɔ̃ o	z k j	e y		b d n g	k	e ɛ	ʒ t j	ɔ̃
TRANSCRIPTION "STANDARD"	P	o	z	e	y	n	k	ɛ	S j	ɔ̃
ADRESSE DE DEBUT SUGGEREE	NON CADRE	4	18	24		37	49	53	NON CADRE 72	77
ADRESSE DE FIN SUGGEREE	NON CADRE	14	22	36		47	53	57	CADRE 77	85

RESULTAT DE L'ANALYSE PHONETIQUE

RESULTAT DU CADRAGE

Fig.4 :Sorties partielles du treillis phonémique et du programme de cadrage

- L'identification finale du segment se fait à partir de cette identification hiérarchisée et des étiquettes affectées à chacun des échantillons du segment à l'étape III-3.

III-7 La constitution du treillis des hypothèses phonémigues - résultats

La sortie de l'analyseur phonétique se présente sous la forme d'une suite de segments partageant les informations suivantes : numéro d'ordre dans la suite, probabilité d'existence, frontières, suite de phonèmes candidats rangés en fonction du degré de ressemblance, numéro de la syllabe à laquelle appartient le segment, affecté d'une probabilité (fig.4).

Les résultats obtenus dans la première version de KEAL sur la parole continue sur une dizaine de locuteurs après apprentissage sur deux voix, varient entre 40 et 50% de reconnaissance exacte pour le premier candidat, entre 75 et85% pour les trois premiers candidats et entre 85 et 90% pour l'ensemble des phonèmes candidats dans le treillis phonétique ; ces résultats tiennent compte des erreurs de segmentation. L'introduction des modifications et des nouveaux algorithmes précédemment décrits ainsi que l'adjonction du programme d'adaptation automatique devrait améliorer sensiblement ces resultats.

III-8 L'apprentissage.

Pour l'apprentissage des coefficients des fonctions de séparation des classes on procède comme suit (fig.5) :

On extrait automatiquement les segments phonétiques d'apprentissage à partir d'un programme (VIVES,1980) permettant le cadrage du treillis phonétique sur la transcription phonétique idéale des syllabes, mots ou phrases utilisés comme corpus de base(fig.4). Pour obtenir les bons échantillons, on corrige manuellement les frontières des segments en cas d'erreur de cadrage. Un programme d'approximation stochastique minimisant un critère des moindres carrés sur cet ensemble de segments d'apprentissage détermine les coefficients optimaux. L'annexe I décrit plus précisément le calcul de ces coefficients. Le nombre m de phones à distinguer dans une classe donnée (voyelle ou consonne) correspondait au départ au nombre de voyelles et consonnes du français. Cependant nous avons légèrement augmenté ce nombre pour obtenir des ensembles homogènes et compacts dans l'espace représentatif des voyelles et des consonnes :

Actuellement nous pouvons distinguer 17 phones dans la catégorie des voyelles et 26 phones dans la catégorie des consonnes ; un programme de classification automatique nous permet également de calculer de façon automatique ce nombre de phones (LELIEVRE,1981). Les apprentissages ont été réalisés pour deux voix masculines et deux voix féminines.

Dans la version définitive de KEAL II nous prévoyons d'appliquer cette technique de reconnaissance paramètrique et d'apprentissage à d'autres catégories telles que les nasales, les liquides, les plosives et les fricatives. D'autre part, après amélioration des programmes de cadrage du treillis phonétique sur la transcription phonétique idéale, nous pensons rendre complétement automatique cet apprentissage et ainsi adapter automatiquement les coefficients à tout nouveau locuteur.

CONCLUSION

Dans la première version de KEAL, certaines lacunes et faiblesses de l'analyseur phonétique abaissaient sensiblement les résultats sur la reconnaissance de mots ou de phrases ; nous espérons améliorer ces performances grâce aux diverses modifications en cours : extension de la bande de fréquence d'analyse de la parole, utilisation et comparaison de différentes méthodes d'analyse, introduction de nouveaux paramètres spectraux et temporels, écriture de modules de reconnaissance de plosives et de nasales, utilisation de l'algebre floue pour éviter les décisions trop brutales, adaptation automatique au locuteur, etc....

On peut ainsi rêver de voir un jour le système de reconnaissance utilisant l'analyse phonétique se rapprocher peu à peu des taux de reconnaissance obtenus par les opérateurs humains sur les logatomes.

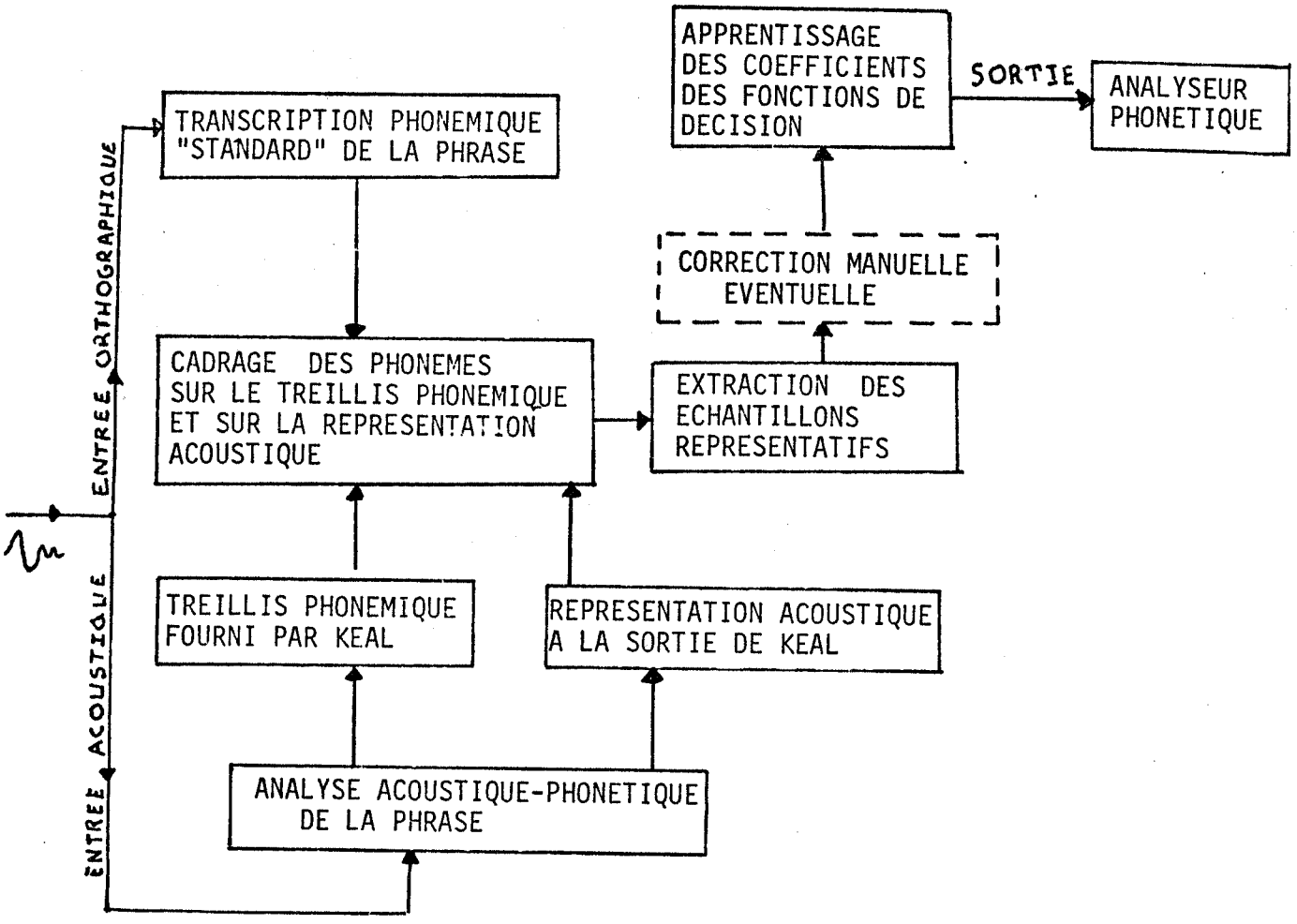


Fig.5 Schéma du module d'apprentissage

Locuteur	F.E	M.G	L.B	P.Q
Nombre de phones différents	17	12	17	12
Nombre total de représentants	106	151	244	78
Pourcentage cumulé de reconnaissance en 1ère position	66.98	65.56	82.79	74.36
Pourcentage cumulé de reconnaissance en 2ème position	83.96	79.4	93.03	89.74
Pourcentage cumulé de reconnaissance en 3ème position	93.40	89.4	95.49	93.59

Tableau I.a

Reconnaissance monolocuteur sur les voyelles après apprentissage.

Locuteur	F.E	M.G	L.B	P.Q
Nombre de phones	26	20	26	12
Nombre de représentants	127	191	405	66
Pourcentage cumulé de reconnaissance Choix 1	55.91	39.2	53.09	65.15
Choix 2	77.95	58.1	71.60	84.85
Choix 3	87.40	70.1	77.04	93.94
Choix 4	92.91	78.5	80.49	96.97

Tableau I.b

Reconnaissance sur les consonnes.

ANNEXE I : Optimisation des coefficients des fonctions linéaires de séparation des phones.

- Soit m le nombre de phones d'une catégorie donnée (voyelle ou consonne)

- On dispose d'un corpus d'apprentissage de N échantillons X_{ij}
  i variant de 1 à m et j de 1 à p(i) : nombre d'échantillons pour l'ensemble des représentants Phone (i)

$$N = \sum_{i=1}^{m} p(i)$$

$$X_{ij} = (X_{ij}^1, \ldots, X_{ij}^n, 1) \in \mathbb{R}^{n+1}$$

et n etant la dimension de l'espace de représentation des échantillons de parole.

- On se donne une application de séparation "théorique" f :

$$f : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^{m-1} \qquad X_{ij} \longmapsto f(X_{ij}) \qquad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, p(i) \end{matrix}$$

telle que $f(X_{ij}) = a_i \quad \forall (i,j)$

Les a_i sont les sommets d'un polyèdre équilatéral de $\mathbb{R}^{m-1}$ défini par les relations suivantes :

$$a_i = (a_i^1, \ldots, a_i^{m-1}) \text{ et } a_i^h = 0 \text{ si } i < h$$

$$a_i^h = \left(\frac{m}{m-1} \cdot \frac{m-h}{m-h+1}\right)^{\frac{1}{2}} \text{ si } i = h$$

$$a_i^h = -\frac{1}{m-h}\left(\frac{m}{m-1} \cdot \frac{m-h}{m-h+1}\right) \text{ si } i > h$$

Cette fonction f permet de séparer les m classes.

- On cherche à approximer f par une application linéaire

$$H^* : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^{m-1} \qquad X_{ij} \longmapsto W^{*T} X_{ij}$$

(W^* matrice de l'application H^*)

qui minimise l'erreur quadratique moyenne suivante :

$$E[e^2] = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{p(i)} (e_{ij})^2$$

où e représente la "distance" entre les applications f et H^*

et $$e_{ij} = f(X_{ij}) - H^*(X_{ij}) = a_i - W^{*T} X_{ij} \qquad \begin{matrix} i = 1, \ldots, m \\ j = 1, \ldots, p(i) \end{matrix}$$

On détermine W^* par une méthode itérative de type "gradient" (GLADYSHEV, 1965 ; MERCIER, 1970) définie par la relation (3) suivante :

Soit k l'indice d'itération du gradient
Soit $\varphi : \mathbb{N} \longrightarrow \mathbb{N}^2 \qquad p \longmapsto (i,j)$

Posons $(X_p)_{p=1,\dots,N} = (X_{11}, \dots, X_{m\,p(m)})$

$(e_j)_{p=1,\dots,N} = (e_{11}, \dots, e_{1\,p(1)}, \dots, e_{m\,p(m)})$

alors :

$$W_k(p+1) = W_k(p) + DW_k(p) \qquad (3)$$

avec $DW_k(p) = -\varepsilon \nabla_{W_k} [E(e^2)] = -2\varepsilon\, e_p \cdot X_p \quad p=1,..N$

et $W_1(1) = 0$

$W_k(1) = W_{k-1}(N)$

On arrête le processus itératif quand

$$\| W^*_{k+1}(N) - W_{k*}(N) \| < \lambda$$

dans ces conditions : $W^* = W_{k*}(N)$

Cet algorithme a été modifié pour tenir compte du passé de l'apprentissage et pour conserver la direction moyenne de l'accroissement des coefficients W survenu lors de l'apparition de tous les représentants X_{ij} de la classe i dans l'itération k. Pour cela on remplace dans (3) $DW_k(p)$ par $DW'_k(p) = \dfrac{DW_k(p)}{\| DW_k(p) \|}$

La direction moyenne de changement pour une classe phonétique i donnée et une itération k est alors définie par :

$$DWPASS_{k\,i} = \frac{1}{p(i)} \sum_{j=1}^{p(i)} DW'_k(l)$$

$$\text{avec} \quad l = \Phi^{-1}(i,j) \qquad (5)$$

La modification (4) permet de se limiter à une seule itération de l'algorithme du gradient : $k^* = 1$

La modification (5) présente les avantages suivants :

1 -Si au cours d'un apprentissage un phone n'a plus de représentants c'est la matrice DWPASS qui remplace la matrice DW' .

2 -Si on ne dispose pas de représentants de tel ou tel phone lors de l'adaptation à un nouveau locuteur on utilise les accroissements DWPASS correspondants.

3 -Si on veut tenir compte du passé au cours de l'adaptation, on peut affecter un coefficient de pondération aux accroissements DWPASS et les combiner aux accroissements actuels DW' dans l'itération (3).

REFERENCES

CAELEN, J. ,1979, Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique. Thèse d'état Toulouse.

CHAPLIN W.G. , LEVADI V.S. ,1967, A généralization of the linear threshold decision algorithme to multiple classes. Computer and Information Sciences II, Tou(A.P.) New-york London.

COUVRAT M. , GAGNOULET C. ,1981 , Optimisation en reconnaissance par mots. The Fourth F.A.S.E. symposium on acoustics and speech.

DE MORI R. , MERCIER G. , 1981 , The phonetic and phonemic knowledge of the KEAL speech understanding system ;à paraître.

DE MORI R. , GUBRYNOWICZ R. , LAFACE P. , 1979, Inference of a knowledge source for the recognition of nasals in continuous speech. I.E.E.E. Transactions on acoustics, speech and signal processing vol. ASSP -27 n°5 p 538-549.

EL MALLAWANY I. , 1977 , Area function extraction over the closed glottis interval, Symposium sur les modèlisations articulatoires GRENOBLE p65-76.

GLADYSHEV E.G. , 1965 , On stochastic approximation, theory of probability and its application, U.S.A., 10, p. 275-278.

GRESSER J.Y. , MERCIER G. , 1975 , Automatic segmentation of speech into syllabic and phonemic units : application to french words and utterances. Auditory analysis and perception of speech. G. FANT and M.A.A. TATHAM editors, A.P. London p359-382.

LECORRE C. , VIVES R. , 1979 , Un programme de cadrage automatique pour l'adaptation au locuteur en recherche automatique de la parole 2ème Congrès AFCET - IRIA - RFIA Toulouse Tome II p 343-350.

LELIEVRE A. , 1981, Classification hierarchique automatique pour la création de phones ou de mots de référence en reconnaissance de la parole. 12èmes J.E.P. Montréal.

MERCIER G. , 1970 , Approximation stochastique et reconnaissance acoustique d'un vocabulaire limité. Annales des Télécom. Tome25 p.207-216.

MERCIER G. , QUINTON P. , VIVES R. , 1978 a , KEAL, un système pour un dialogue avec une machine. Théorie et technique de l'information Actes du congrès de l'AFCET 13-15 nov. p.304-314.

MERCIER G. , 1978 b , Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL, 9ème J.E.P. Lannion p.321-342.

QUINTON P. , 1980 , Contribution à la reconnaissance de la parole : utilisation des méthodes heuristiques pour la reconnaissance de phrases, Thèse d'état, Université de Rennes.

RAMASSE D. , 1981 , La projection acoustique des contraintes distributionnelles, contribution à la reconnaissance automatique de la parole, Thèse 3ème cycle Université d'Aix Marseille I.

ROSSI M. , LECORRE C. , MERCIER G. ,1978 , Indice de détection de formants sur analyse spectrale par canaux, 9èmes J.E.P. Lannion P.105-114.

ROTH M.T. , 1981 , Contribution à l'étude acoustique des voyelles dans les les monosyllabes en français, Thèse de 3ème cycle, Université de Rennes, U.E.R. de linguistique.

SANCHEZ C. , MESSENET G. , HATON J.P. , 1978, Etude et utilisation des indices acoustiques et des traits pour la segmentation et la reconnaissance phonémique de la parole 9èmes J.E.P. Lannion p.271 -276.

VAISSIERE J. , 1980 , Speech recognition programs as models of speech perception. In the cognitive representation of speech, MYERS T. LAVER J. and ANDERSON J. Editeurs, North-Holland Publishing Company, inpress.

WOLF J. , WOODS W. , 1980 , The HWIM Speech understanding systèm. Trends in speech recognition, LEA editor Prentice-Hall signal processing series p.316-339.

ZUE V. , SCHWART R. , 1980 , Acoustic processing and phonetic analysis. Trends in speech recognition, LEA W. Prentice-Hall signal processing series p.101-124.

ZURCHER E. , 1977 , La mesure du fondamental par la détection des crêtes 8èmes J.E.P. GALF, Aix en Provence p.120-126.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

SYNTAXE, SEMANTIQUE ET DIALOGUE DANS LE SYSTEME KEAL APPLICATION AU RENSEIGNEMENT AUTOMATIQUE

DOMINIQUE GILLET	CNET - LANNION A - TSS/RCP
ANDREE NOUHEN-BELLEC	IRISA, IUT DE LANNION
PATRICE QUINTON	IRISA, LA 227
JACQUES SIROUX	IRISA, IUT DE LANNION

.RESUME

On décrit dans cet article la façon dont les contraintes syntaxiques, sémantiques et pragmatiques sont mises en oeuvre dans le système de reconnaissance de la parole continue Keal, actuellement en cours de développement au Centre National d'Etudes des Télécommunications (CNET) à Lannion. L'utilisation de ces informations présente un double avantage : d'une part, on réduit partiellement l'indéterminisme de l'analyse phonétique et lexicale ; d'autre part, on obtient un dialogue plus souple et plus naturel entre l'homme et la machine. L'application qui a servi de support à ces recherches est un centre de renseignements téléphoniques automatique.

On présente d'abord l'application choisie et l'organisation générale des niveaux syntaxique, sémantique et pragmatique dans le système Keal. Puis on décrit successivement les trois étapes suivantes :

- la reconnaissance de phrase, qui permet de construire l'arbre syntaxique de la phrase prononcée à partir des mots détectés dans le signal de parole ;

- l'interprétation de la phrase, dont le rôle est d'extraire l'information sémantique à partir de l'arbre syntaxique ;

- la gestion du dialogue.

Enfin, on souligne les principales caractéristiques de cette réalisation, et on indique en conclusion les principaux développements envisagés pour cette étude.

SYNTAX, SEMANTICS AND DIALOG IN THE KEAL SYSTEM

APPLICATION TO THE AUTOMATIC DIRECTORY ASSISTANCE

. ABSTRACT

The object of this paper is to describe how the syntactic, semantic and pragmatic constraints are implemented in the Keal continuous speech recognition system ; this system is currently under development at the french Telecommunication Research Center (CNET) in Lannion. The use of these informations provides two types of advantages : on the one hand, it allows to partially reduce the non-determinism of the phonetic and lexical analysis ; on the other hand, it provides more flexibility and more naturalness in the man-machine dialog. The application which has been used for developping this research is an automatic directory assistance service.

The selected application and the general organization of the system at the syntactic, semantic and pragmatic levels are first presented. The three following steps are then successively described :

- the sentence recognition, which allows to build the parse-tree of the uttered sentence from the detected words in the speech signal ;

- the sentence interpretation, whose role is to extract the semantic information from the parse-tree ;

- the dialog management.

Finally, the main features of this realization are underlined, and the main improvements of this study are indicated as conclusion.

SYNTAXE, SEMANTIQUE ET DIALOGUE DANS LE SYSTEME KEAL APPLICATION AU RENSEIGNEMENT AUTOMATIQUE.

GILLET D. — CNET LANNION
NOUHEN-BELLEC A. — IRISA, IUT LANNION
QUINTON P. — IRISA, LA 227
SIROUX J. — IRISA, IUT LANNION.

1. INTRODUCTION

Depuis quelques années, un grand nombre de travaux ont porté sur l'utilisation de la syntaxe et de la sémantique pour reconnaître la parole continue (WOODS et al., 1976 ; WALKER, 1976 ; PIERREL et HATON, 1980 ; MERCIER et al., 1980 ; MELONI et GUIZOL, 1980 ; GROC et TUFFELLI, 1980 ; SAITTA, 1978 ; SHIGENAGA, 1978). Deux idées sont à l'origine de ces recherches. D'une part, essayer de surmonter les difficultés rencontrées au niveau de la reconnaissance des phonèmes et des mots en faisant appel aux contraintes apportées par "ce dont on parle", autrement dit faire de la compréhension de la parole plutôt que de la reconnaissance. D'autre part, réaliser des systèmes ayant la capacité de dialoguer réellement avec des êtres humains. Cette approche fait appel à des techniques dites d'intelligence artificielle : il faut en effet modéliser le thème du dialogue et donner au système une certaine intelligence simulant le comportement d'un interlocuteur humain.

L'objet de cet article est de présenter la façon dont ces problèmes ont été abordés dans le système Kéal. L'application qui a servi de support aux recherches décrites ici est le renseignement téléphonique automatique, pour lequel le comportement de Kéal doit simuler celui d'une opératrice d'un centre de renseignements. Cette application présente certaines difficultés qui font qu'elle reste du domaine des recherches à long terme : citons par exemple le problème de l'adaptation au locuteur et celui de la reconnaissance d'un nom (de ville, de personne ou de société) parmi une grande liste. Le choix de cette application présente cependant l'avantage de mettre clairement en évidence ce que l'on peut attendre de l'introduction des contraintes syntaxiques et sémantiques dans la reconnaissance, et au contraire ce qui ressort exclusivement des étapes antérieures de la reconnaissance allant du signal de parole à l'identification des mots.

Cet article comporte sept parties. Dans la seconde, on décrit l'application choisie et l'organisation générale des niveaux syntaxique, sémantique et pragmatique de Kéal. La troisième partie concerne l'étape reconnaissance de phrase, c'est-à-dire celle qui permet de reconstituer la phrase prononcée par l'interlocuteur du système à partir des mots détectés dans le signal de parole. Dans la quatrième partie, on décrit l'étape d'interprétation d'une phrase, dont l'objet est d'extraire de la phrase prononcée son contenu sémantique. La gestion du dialogue fait l'objet de la cinquième partie. Enfin, dans la sixième partie on souligne les principales caractéristiques de cette réalisation, et la septième partie indique pour conclure les principales directions envisagées pour le développement ultérieur de cette étude.

2. CARACTERISTIQUES DE L'APPLICATION ET ORGANISATION GENERALE DU SYSTEME

Dans cette partie, nous indiquons tout d'abord les principales caractéristiques du dialogue du centre de renseignements, puis l'organisation générale des traitements syntaxico-sémantiques dans Kéal.

2.1. Caractéristiques de l'application

Dans le cadre de l'application au centre de renseignements téléphoniques, nous nous limitons aux trois types de demandes suivats :

- numéro d'un abonné (particulier ou société),
- concordance pour un numéro modifié,
- acheminement (indicatif d'un département français ou d'un pays étranger).

L'examen de 315 conversations réelles montre que ces trois types représentent environ 95% des demandes. Cette restriction, si elle limite le domaine sémantique de l'application, n'en supprime cependant pas la difficulté essentielle : les demandes étant formulées par les usagers en langue naturelle, le langage de communication est donné a priori, et on ne peut qu'essayer de l'approcher le mieux possible.

Dans une première approche, on a délibérément ignoré un certain nombre de problèmes posés par cette application. On suppose que :

- l'utilisateur s'exprime de la même façon qu'il le fait actuellement en face d'une opératrice d'un centre de renseignements. Cela signifie que l'on peut, pour définir le langage de communication, s'appuyer sur les exemples de dialogues de renseignements recueillis. Une expérience décrite dans QUINTON et al. (1975) donne à penser que cette supposition n'est pas absurde ; toutefois, il faudra attendre la réalisation d'un système complet pour valider totalement cette approche.
- l'utilisateur du système formule sa demande sans hésitation ni reprise. Si la syntaxe du dialogue prévoit un grand nombre de formulations possibles définies par l'analyse des demandes réelles, elle ne permet pas de traiter les formes "dégradées" que l'on rencontre pourtant communément dans une conversation courante.
- on possède les caractéristiques vocales de l'utilisateur du système. Il est clair que l'application réaliste de Kéal au renseignement automatique nécessite la mise au point de méthodes d'adaptation au locuteur. Des recherches récentes (ROSENBERG et al., 1980) montrent qu'il est sans doute possible de résoudre partiellement ce problème par l'utilisation de méthodes de classification des locuteurs.

Ces hypothèses de travail, qui peuvent paraître dénaturer le problème, ont l'avantage de faire porter l'effort sur d'autres difficultés du dialogue vocal, à savoir :

- l'incertitude des données acoustico-phonétiques que l'on cherche à lever à l'aide des contraintes linguistiques,
- la mise au point d'un générateur de dialogue cohérent, capable de simuler correctement le rôle d'une opératrice.

2.2. Principe du dialogue

Le principe du dialogue retenu (NOUHEN et SIROUX, 1979) est de laisser l'usager formuler sa question, puis de lui demander de préciser éventuellement les informations qui manquent pour que le renseignement puisse être déterminé. La figure 1 représente le diagramme fonctionnel d'un cycle question-réponse.

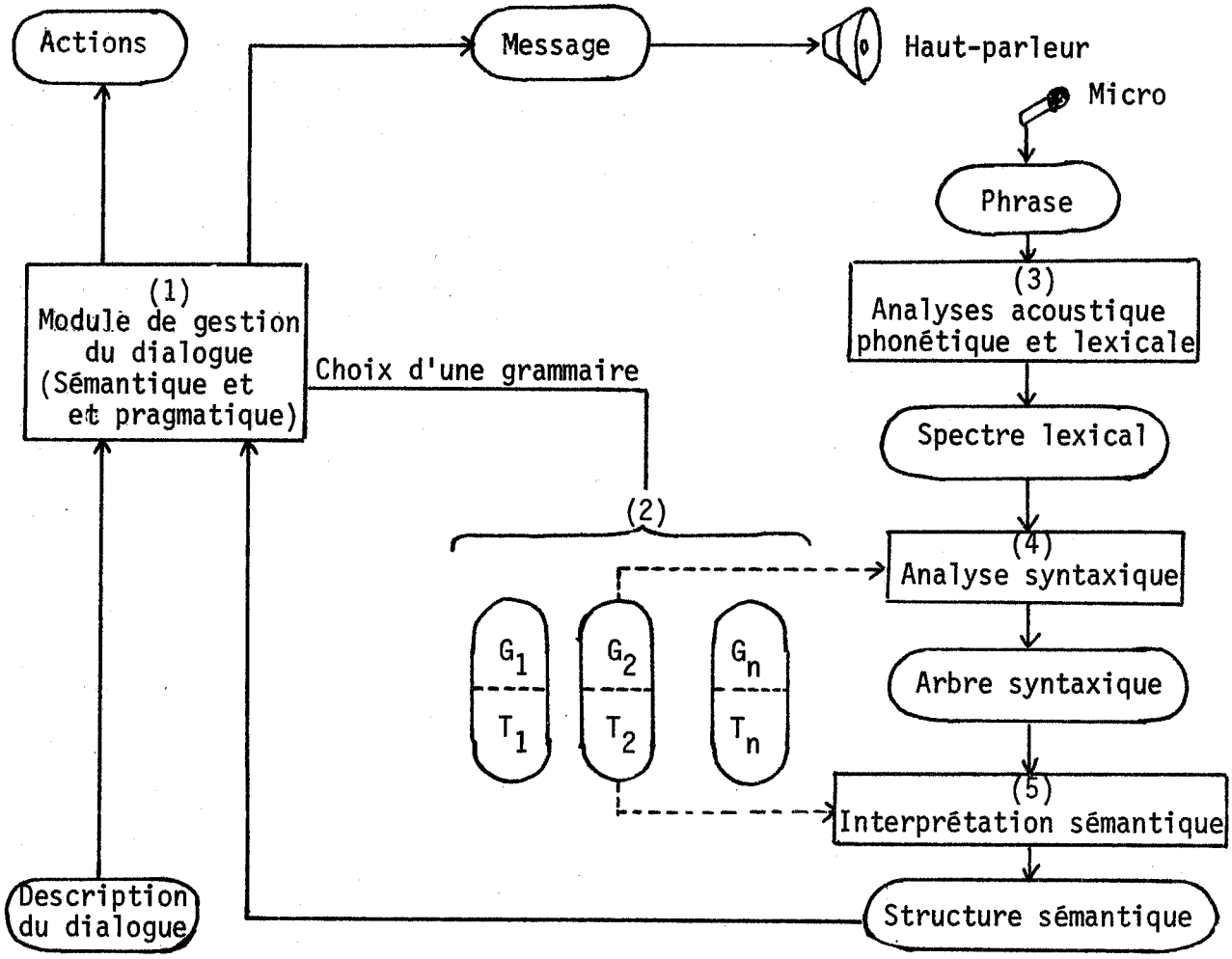


Figure 1 : Diagramme d'un cycle question-réponse dans le dialogue d'un centre de renseignements automatique.

Le module de gestion du dialogue (1) choisit la grammaire (2) qui décrit toutes les réponses possibles de l'usager, et il active les programmes de reconnaissance proprement dits (3) et (4), dont l'analyseur syntaxique (4) qui fournit l'arbre syntaxique de la meilleure phrase-solution. Cet arbre est interprété (5) à l'aide des transformations associées à la grammaire choisie, et l'on obtient ainsi la structure sémantique de la phrase qui sert d'entrée au module de gestion du dialogue pour un nouveau cycle question-réponse.

Ce qui suit décrit avec plus de détails la reconnaissance de phrases, l'interprétation sémantique et la gestion du dialogue.

3. LA RECONNAISSANCE DES PHRASES

3.1. Le problème

Le rôle de cette étape est, après chaque intervention de l'interlocuteur, de reconstituer la phrase qu'il a prononcée en réponse à une question du système. On suppose que cette phrase appartient à un langage que l'on peut décrire au moyen d'une grammaire hors-contexte G, dont le vocabulaire terminal V_T est un sous-ensemble du vocabulaire de toute l'application. Cette grammaire est sélectionnée par le programme de gestion du dialogue en fonction du contexte courant de l'intervention ; de cette façon, on restreint à la fois la tâche du reconnaisseur de mots et du reconnaisseur de phrases.

Les données d'entrée du reconnaisseur de phrases sont constituées par le spectre lexical construit par le reconnaisseur de mots. Celui-ci compare chaque mot de V_T à la phrase telle qu'elle résulte de l'analyse phonétique (VIVES, 1978 . GRESSER et MERCIER, 1975). Chacun des mots donne lieu à une ou plusieurs détections à différents endroits de la phrase. Le spectre lexical contient donc à la fois des vraies détections, qui correspondent aux mots effectivement prononcés par l'interlocuteur, et des fausses détections dues à l'incertitude des résultats phonétiques. Pour un vocabulaire d'une centaine de mots, on trouve en moyenne vingt fausses détections pour une vraie détection, ce qui montre que la reconnaissance de phrase est un problème très combinatoire. Sa résolution passe à la fois par l'utilisation des contraintes apportées par la grammaire et par la mise en oeuvre de techniques heuristiques évitant d'effectuer toute la combinatoire. Ce sont ces deux aspects que nous abordons successivement.

3.2. Expression des contraintes syntaxico-sémantiques

Le modèle utilisé pour décrire le langage de réponse d'un interlocuteur est une grammaire hors-contexte notée $G = (V_N , V_T , S , P)$ où :

- V_N est le vocabulaire non-terminal, dont les éléments sont aussi appelés classes syntaxiques ;
- V_T est le vocabulaire terminal, composé ici de classes lexicales groupant les mots de même comportement syntaxique ;
- S est l'axiome appartenant à V_N ;
- P est un ensemble fini de productions de la forme $X \rightarrow \alpha$ où $X \in V_N$ et $\alpha \in (V_N \cup V_T)^*$.

A titre d'exemple, la grammaire suivante permet de décrire un sous-ensemble des questions initiales possibles du dialogue du centre de renseignements automatique :

V_N = { <demande> , <numéro> , <expression> , <contenu> , <partie numéro> , <partie correspondant> , <titre> , <nom de personne> }

V_T = { 'je voudrais' , 'le numéro' , 'de' , 'Monsieur' , 'Madame' , 'Mademoiselle' , 'Dupont' , 'Durand' , ... }

Axiome = <demande>

Productions =

<demande> → <numéro>

<numéro> → <expression> <contenu>

<expression> → 'je voudrais'

<contenu> → <partie numéro> <partie correspondant>

<partie numéro> → 'le numéro'

<partie correspondant> → 'de' <titre> <nom de personne>

<titre> → 'Monsieur' *ou* 'Madame' *ou* 'Mademoiselle'

<nom de personne> → 'Dupont' *ou* 'Durand' *ou* ...

Cette grammaire engendre les phrases de la forme :

$$\text{Je voudrais le numéro de } \left\{ \begin{array}{l} \text{Monsieur} \\ \text{Madame} \\ \text{Mademoiselle} \end{array} \right\} \left\{ \begin{array}{l} \text{Dupont} \\ \text{Durand} \\ \ldots \end{array} \right\}$$

Dans la grammaire complète, chacune des classes syntaxiques décrit d'autres types de syntagmes. Par exemple, la classe <expression> peut aussi engendrer :

Est-ce-que je pourrais avoir
Pourrais-je avoir
Puis-je avoir
...

qui sont toutes compatibles avec les séquences décrites par la classe <contenu>.

Comme on le constate sur l'exemple précédent, le modèle utilisé permet de représenter à la fois des contraintes syntaxiques et sémantiques. En effet, les classes syntaxiques choisies ne sont pas les "parties du discours" (substantif, verbe, etc...) que l'on utilise normalement pour la description des contraintes syntaxiques. Cette façon de faire conduirait à décrire un ensemble de phrases syntaxiquement correctes, mais sémantiquement inacceptables : il serait alors nécessaire d'ajouter d'une façon ou d'une autre des vérifications sémantiques au cours de l'analyse syntaxique, problème toujours difficile à résoudre techniquement. Ce modèle présente cependant l'inconvénient de ne pouvoir être généralisable à des applications comportant de grands vocabulaires, le nombre de productions devenant alors trop important. De plus, un changement d'application conduit nécessairement à réécrire la grammaire. Néanmoins, à notre sens, cette approche s'applique à la plupart des applications envisageables actuellement pour la reconnaissance de la parole continue.

3.3. Méthode de reconnaissance

La reconnaissance de la phrase consiste à explorer le spectre lexical de la gauche vers la droite en construisant de proche en proche des séquences ω de détections adjacentes de plus en plus longues, et en vérifiant que ces séquences sont compatibles avec les contraintes syntaxico-sémantiques exprimées par la grammaire. Par ce procédé, on peut espérer trouver la séquence ω^* correspondant à la phrase effectivement prononcée, si toutefois tous les mots prononcés ont été détectés avec des scores suffisamment élevés.

Le choix des séquences ω que l'on considère est déterminé par une stratégie de recherche qui utilise le score des détections de ω comme fonction heuristique : il n'est en effet pas pensable d'examiner toutes les possibilités qui se présentent.

La stratégie de recherche retenue pour Kéal est la recherche dite "en faisceau" inspirée du système Harpy (LOWERRE, 1976). Ce choix a été fait de façon expérimentale en comparant cette stratégie à deux autres stratégies de recherche

classiques : le décodage séquentiel et la recherche à retour-arrière (QUINTON, 1980).

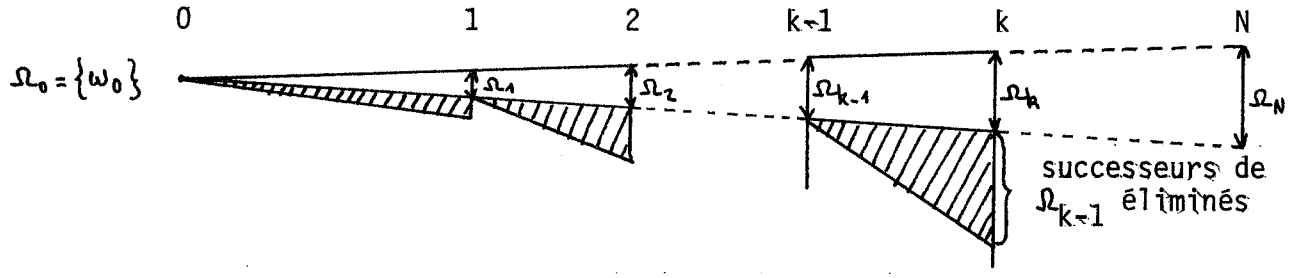


Figure 2 : Schéma de la recherche en faisceau

Le principe de la recherche en faisceau peut être schématisé de la façon suivante. Remarquons tout d'abord que chaque séquence ω de n mots se déduit d'une séquence unique ω' de n-1 mots par adjonction d'un mot. L'ensemble des séquences ω de longueurs croissantes forme donc un arbre dont la racine est la suite vide ω_0. La recherche en faisceau consiste à parcourir ect arbre en considérant successivement les séquences ω d'extrémité 1, 2, ..., k, ..., N et en ne gardant que celles dont le score moyen est suffisamment élevé. Schématiquement (figure 2), si l'on note Ω_{k-1} l'ensemble des séquences conservées après le pas k-1, l'ensemble Ω_k s'en déduit en construisant tous les successeurs des éléments de Ω_{k-1} dans l'arbre, puis en effectuant une sélection sur les séquences obtenues ainsi. La trace de l'algorithme dans l'arbre se présente comme un faisceau de largeur constante, d'où son nom.

La solution fournie par la recherche est la meilleure séquence de l'ensemble Ω_N. L'arbre syntaxique de cette séquence est construit à partir de la trace d'exécution de l'algorithme.

4. INTERPRETATION SEMANTIQUE

Le rôle de cette étape est d'extraire d'une phrase l'information qui est pertinente pour le dialogue. On vient de voir qu'à l'issue de la reconnaissance, on dispose de l'arbre syntaxique de la phrase prononcée. L'interprétation consiste en un parcours de cet arbre syntaxique, au cours duquel sont produites des relations (n-uplets) qui représentent la structure sémantique de la phrase.

4.1. Représentation de la structure sémantique d'une phrase

On entend ici par "structure sémantique" l'ensemble des informations apportées par une phrase qui permettent au dialogue de progresser. Les seules informations qui nous intéressent sont donc celles qui ont relatives à l'univers du dialogue. Ces informations sont codées à l'aide d'un modèle unique appelé n-uplet de la forme :

Relateur (actant 1, ..., actant n)

qui exprime une relation mettant en jeu les actants.

A titre d'exemple, la phrase : "Je voudrais le numéro de Monsieur Dupont" que prononce l'interlocuteur de Kéal est représentée par les n-uplets suivants :

1 Vouloir (je, numéro)

2 Appartenir (numéro, correspondant)

3 Titre de (Monsieur, correspondant)

4 Nommer (Dupont, correspondant)

Le n-uplet 1 indique qu'il s'agit d'une demande de numéro par opposition au n-uplet *Vouloir (je, acheminement)* qui aurait été produit dans le cas d'une demande d'acheminement. Le n-uplet 2 indique que le correspondant a été effectivement cité dans la demande ; il serait absent si l'usager avait dit, par exemple : "Je voudrais un numéro de téléphone". Le n-uplet 3 indique que le correspondant est un particulier et non une société. Enfin, le n-uplet 4 donne le nom du correspondant.

La liste des n-uplets transmis au module de gestion du dialogue dépend évidemment de la quantité et de la nature des informations fournies par l'usager. La demande peut être très vague ; ainsi la phrase : "Je voudrais un numéro de téléphone" ne permettra d'obtenir qu'un n-uplet *vouloir (je, numéro)*. Elle peut au contraire être très précise : il arrive que dès la première phrase on soit en possession de toutes les informations nécessaires à la recherche du renseignement ; par exemple, si l'usager dit : "Je voudrais le numéro de téléphone de Monsieur Dupont à Lannion dans les Côtes-du-Nord", on obtiendra les n-uplets :

Vouloir (je, numéro)

Appartenir (numéro, correspondant)

Titre de (Monsieur, correspondant)

Nommer (Dupont, correspondant)

Nommer (Lannion, ville)

Nommer (Côtes-du-Nord, département)

Deux remarques éclairent le choix qui a été fait ici. On notera tout d'abord que toutes les informations sémantiques sont représentées de la même façon. Ceci traduit le désir d'utiliser un mécanisme unique pour traiter ces informations. Notons par ailleurs que les n-uplets n'ont pas de signification en eux-mêmes : ils ne prennent de sens que par la façon dont ils sont intégrés par le programme de gestion du dialogue dont le rôle, comme on le verra plus loin, est d'expliciter toutes les informations que l'on peut déduire de la présence d'un n-uplet. Le même n-uplet, apparaissant à des stades différents du dialogue, peut donc avoir plusieurs interprétations différentes.

4.2. Mécanisme d'interprétation

Le principe de l'interprétation est analogue à celui d'une grammaire d'attributs (KNUTH, 1968 ; LORHO, 1974 ; BOCHMANN, 1976) dans laquelle on n'emploie que des attributs synthétisés. A chaque production de la grammaire G définissant le langage de réponse correspondant à une étape du dialogue, on associe une liste de fonctions (appelées transformations) dont le rôle est de construire les n-uplets représentant la structure sémantique de la phrase. Le détail du mécanisme d'interprétation peut être trouvé dans GILLET et QUINTON (1980).

La figure 3 représente l'arbre syntaxique de la phrase : "Je voudrais le numéro de Monsieur Dupont" et schématise l'ordre dans lequel sont produits les quatre n-uplets qui correspondent à la structure sémantique de cette phrase. Le n-uplet 1 *Vouloir (je, numéro)* est obtenu au niveau de la production ⟨*demande*⟩ → ⟨*numéro*⟩ dont la présence résume le fait que la phrase est une demande de numéro et non une demande d'acheminement ou de concordance. Le n-uplet 2 est est engendré au niveau de la production ⟨*partie correspondant*⟩ → ... . Les n-uplets 3 et 4 sont construits en deux étapes : leur cadre (*Titre de (— , correspondant)* et *Nommer (— , correspondant)*) est généré à la rencontre de la

production ⟨*partie correspondant*⟩ → ..., et les actants manquants sont ajoutés au vu respectivement des productions ⟨*titre*⟩ → *'Monsieur'* et ⟨*nom de personne*⟩ → *'Dupont'*.

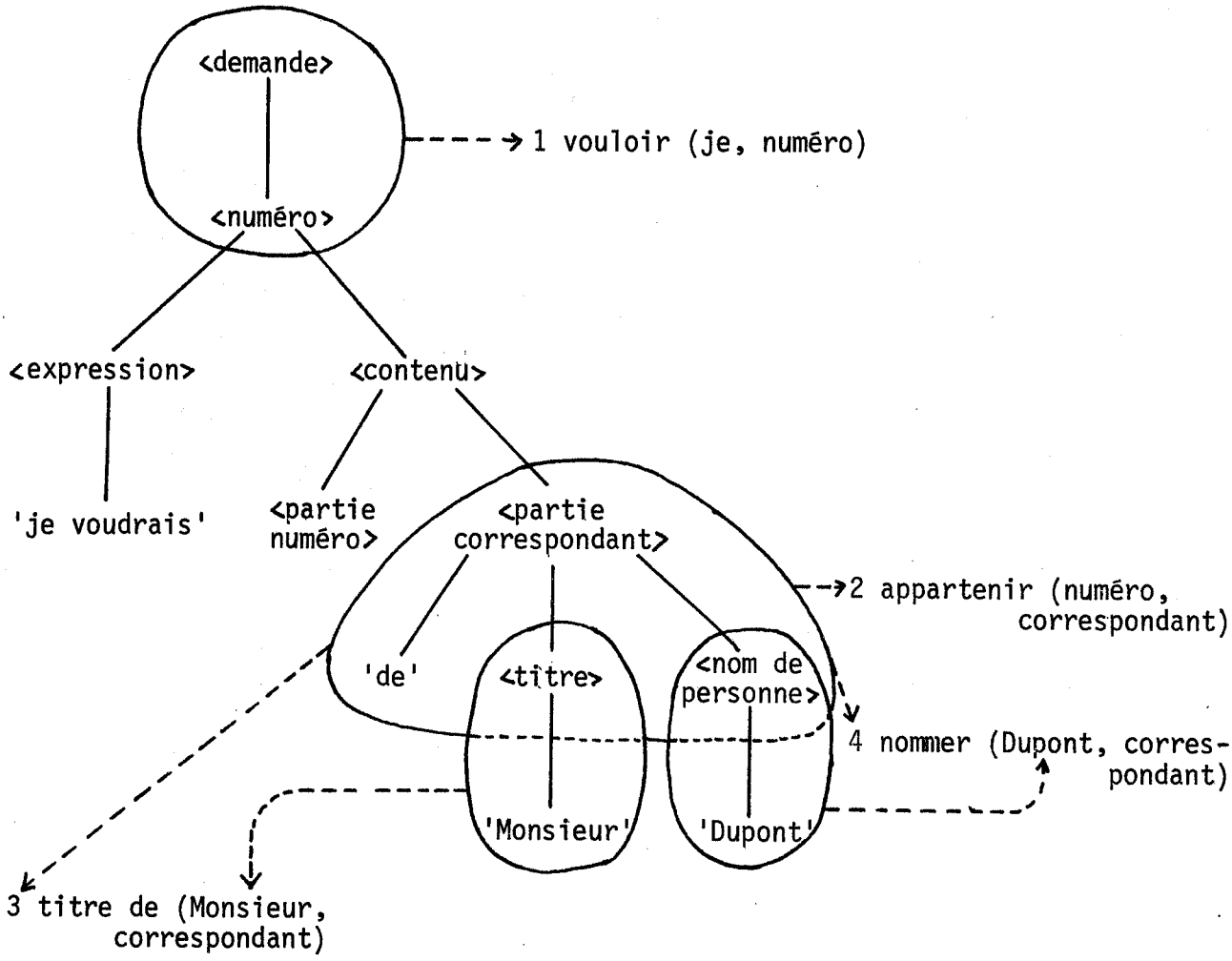


Figure 3 : Schéma d'interprétation de la phrase : "Je voudrais le numéro de Monsiëur Dupont". Les productions concernées par la génération des n-uplets sont entourées. Notons que les n-uplets 3 et 4 sont produits au vu de deux productions différentes.

5. GESTION DU DIALOGUE

Le module de gestion du dialogue (MGD dans la suite) a deux fonctions principales :

- assimiler les informations qui lui sont données par l'utilisateur en les intégrant à ses propres connaissances ;
- susciter par des questions appropriées les données qui lui manquent pour exécuter la tâche confiée au système.

Nous montrons dans cette partie comment le dialogue est modélisé, et nous indiquons brièvement son fonctionnement à partir d'un exemple.

5.1. Modèle de dialogue

Le dialogue est conçu comme une hiérarchie de phases appelée graphe de dialogue, chaque phase représentant l'acquisition d'un certain groupe d'informations

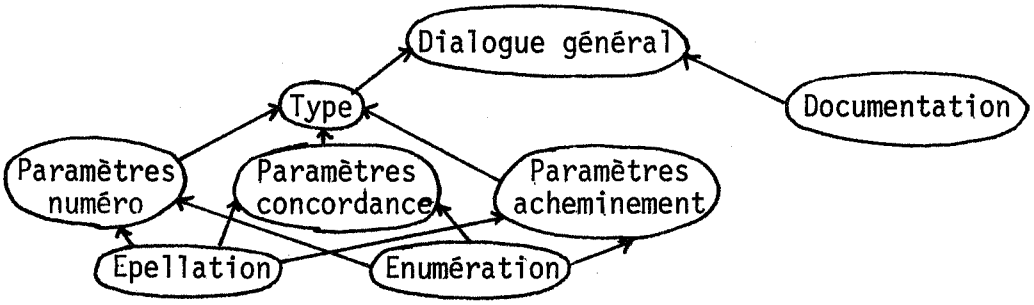


Figure 4 : Graphe simplifié du dialogue de renseignement

A titre d'exemple, la figure 4 symbolise la hiérarchie des phases par lesquelles il faut passer dans le cas du dialogue du centre de renseignements. Dialogue général est la phase qui permet de lancer le dialogue et de boucler sur les demandes successives formulées par un même utilisateur ; Type a pour fonction d'acquérir l'information permettant de distinguer les trois demandes envisageables (demande de numéro, demande de concordance et demande d'acheminement). Les phases Paramètres-numéro, Paramètres-concordance et Paramètres-acheminement ont pour fonction de rassembler les paramètres de chaque type de demande avant d'effectuer un accès à la documentation. Chacune de ces trois phases peut faire appel à Epellation et Enumération, la première pour reconstituer un nom propre par épellation, la seconde pour reconstituer un numéro par énumération. Enfin, la phase Documentation effectue l'accès à la documentation et en fournit les résultats à l'utilisateur.

Le graphe de dialogue représente de façon potentielle ce qui peut se produire au cours d'un dialogue. Suivant les informations que reçoit le MGD, telle ou telle phase sera activée. Ainsi, dès que le type de la demande peut être identifié, une seule des trois phases Paramètres-numéro, Paramètres-concordance ou Paramètres-acheminement est activée.

Chaque phase est elle-même une structure arborescente dont les noeuds, appelés concepts, représentent des informations à acquérir. L'arbre modélise la dépendance entre ces informations. La figure 5 schématise la phase Paramètres-numéro destinée à obtenir les coordonnées du correspondant dont l'utilisateur désire le numéro.

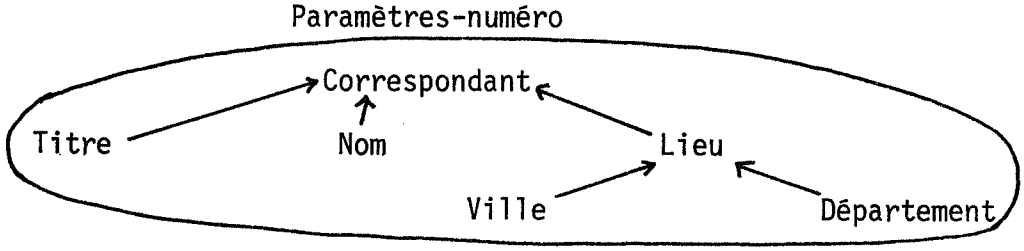


Figure 5 : Structure de la phase Paramètres-numéro

Chaque concept est représenté par les informations suivantes :

- un statut,
- un état,
- une forme,

- un ensemble de procédures associées à des valeurs d'état.

Le statut d'un concept peut prendre les valeurs obligatoire ou facultatif. Ceci traduit le fait que le concept est indispensable ou non à la recherche documentaire.

L'état d'un concept définit son mode d'activité à un instant donné du dialogue. Il peut prendre les valeurs :

- en attente d'information, lorsque le concept qui lui correspond n'a pas encore été rencontré au cours du dialogue ;
- à confirmer, lorsque le concept a été rencontré mais n'est pas sûr (i.e. le score acoustique qui est associé à son expression n'est pas bon) ;
- satisfait, lorsque le concept a été trouvé et confirmé ;
- dormant, c'est-à-dire hors du contexte actuel du dialogue ;
- détruit, c'est-à-dire hors dialogue.

La forme associée à un concept désigne l'information qui peut déclencher l'activité du noeud. Cette information est représentée par un n-uplet (ou plus généralement une expression logique dont les termes sont des n-uplets) avec des actants partiellement ou totalement fixés. Ainsi, la forme associée au concept *ville* serait le n-uplet *Nommer (— , ville)* où le seul actant libre est le nom de la ville.

Enfin, les procédures associées à chaque valeur d'état expriment des actions à effectuer lorsque le concept passe dans l'état correspondant. Nous indiquons les principales actions possibles en décrivant le mécanisme du dialogue.

5.2. Mécanisme du dialogue

Globalement, le mécanisme du dialogue consiste à remplir le graphe de dialogue au moyen des structures sémantiques obtenues après reconnaissance des phrases de l'utilisateur. Lorsque le graphe de dialogue contient tous les paramètres d'une demande, un accès à la documentation est effectué, et les résultats correspondants sont transmis à l'utilisateur. Ce qui suit détaille le fonctionnement du MGD.

Plaçons-nous après la reconnaissance d'une phrase dont la structure sémantique est un certain ensemble de n-uplets. L'état du dialogue est représenté par le graphe de dialogue, qui contient déjà un certain nombre d'informations acquises ; une phase seulement du graphe de dialogue est active, et parmi les concepts de cette phase, un certain nombre sont accessibles (ceux qui ne sont ni dormants, ni détruits, ni satisfaits). Le MGD examine un à un les concepts accessibles de la phase active ; pour chaque concept, il cherche dans la structure sémantique des n-uplets qui remplissent la forme du concept, et exécute ensuite une procédure attachée à l'état courant du concept, procédure qui dépend du succès de la tentative de remplissage de la forme.

Une procédure est une suite d'actions qui peuvent être immédiatement exécutables ou différées. Parmi les actions immédiatement exécutables, citons par e exemple la modification de l'état ou du statut d'un concept. Les actions différées sont mémorisées dans une liste appelée le bloc-note, et leur exécution est reportée à la fin du parcours de la phase par le MGD. Deux types d'actions différées ont un rôle fondamental :

- l'activation d'une nouvelle phase, dont le rôle est de déplacer le contexte du dialogue selon les informations reçues dans la phase courante. Ainsi, c'est par ce mécanisme que le MGD peut choisir entre les phases Paramètres-numéro, Paramètres-concordance ou Paramètres-acheminement, qui sont bien entendu exclusives.
- la construction d'une question à poser à l'utilisateur, qui a pour effet de permettre au dialogue de se poursuivre tant que toutes les informations nécessaires à la recherche documentaire ne sont pas acquises. A chaque question est associée la grammaire décrivant les réponses attendues de l'utilisateur, ce qui permet, après que la question a été posée, de fixer les paramètres des programmes de reconnaissance pour l'analyse de la réponse de l'utilisateur.

A titre d'exemple, considérons le cas du concept *ville* de la phase Paramètres-numéro, auquel est attachée la forme *Nommer (— , ville)*, ainsi qu'on l'a vu au paragraphe 5.1. A l'état *en attente d'information* de ce concept sont attachées deux procédures : l'une, exécutée lorsque la forme est remplie par un n-uplet, aura pour effet de modifier l'état du concept *ville*, qui deviendra *à confirmer* si le score acoustique du n-uplet trouvé est médiocre, ou *satisfait* si le score est bon ; l'autre procédure, exécutée lorsque la forme ne peut être remplie, a pour conséquence de construire la question : "Dans quelle ville habite votre correspondant ?", à laquelle est attachée une grammaire décrivant les réponses possibles de l'utilisateur.

Le parcours de la phase active par le MGD a donc pour conséquence des modifications du graphe de dialogue (par les actions immédiates et le remplissage des formes) et la création d'une liste d'actions différées dans le bloc-note. A la fin du parcours de la phase active, le MGD choisit dans le bloc-note une action à exécuter : s'il s'agit d'une activation de phase, cette nouvelle phase est parcourue ; s'il s'agit d'une question, la question est posée et le MGD donne le contrôle aux programmes de reconnaissance.

6. DISCUSSION

Les programmes réalisant les trois étapes que nous venons de décrire ont été mis au point séparément et sont actuellement en cours d'intégration. La description du dialogue, les grammaires et les règles d'interprétation sont des paramètres compilés, ce qui permet de définir et de mettre au point rapidement un dialogue. L'ensemble de ces programmes constitue donc un bon outil pour expérimenter des dialogues vocaux, tâche qui, à notre sens, est essentielle pour définir de façon précise les conditions d'utilisation de la parole comme moyen de communication avec une machine.

Dans ce qui suit, nous insistons sur deux points fondamentaux de cette étude : le premier point concerne le modèle syntaxico-sémantique choisi, et le second a trait à la place des contraintes syntaxico-sémantiques dans la reconnaissance de la parole.

6.1. Modèle syntaxico-sémantique

Par rapport à d'autres travaux ayant trait au même sujet, le modèle syntaxico-sémantique de Kéal présente quelques originalités qu'il convient de souligner.

Tout d'abord, le fait que la reconnaissance d'une phrase et sa compréhension sont deux processus indépendants (dans la plupart des systèmes existants, reconnaissance et compréhension se font simultanément). Deux raisons nous ont conduits à adopter cette solution. D'une part, le fait que la reconnaissance de

la phrase prononcée est un processus indertéministe au cours duquel de nombreuses tentatives inutiles sont effectuées. Il est donc préférable de n'effectuer à ce stade que les traîtements indispensables à la reconnaissance proprement dite, et par conséquent de différer l'interprétation sémantique et l'intégration de la phrase au contexte du dialogue, étapes dont le rôle n'est pas déterminant pour la reconnaissance de la phrase prononcée. La seconde raison de ce choix est qu'il permet de décrire indépendamment le langage des réponses de l'utilisateur et leur contenu sémantique. En ce sens, la phase d'interprétation sémantique joue un rôle fondamental, puisqu'elle permet le passage de l'arbre syntaxique d'une phrase à sa structure sémantique.

Une autre caractéristique du modèle choisi ici est qu'il autorise des stratégies de dialogue très variées. D'une façon générale, lorsqu'une question est posée, le MGD s'attend à une forme privilégiée de réponse dont le contenu sémantique est lié à la phase active du dialogue. Cependant rien n'empêche, lorsque la réponse contient des informations non directement liées au contexte immédiat du dialogue, d'intégrer ces informations en changeant de phase active sans pour autant que le dialogue ne "dérive" trop. Cette possibilité, qui doit bien entendu être utilisée de façon contrôlée, est indispensable si l'on veut pouvoir faire autre chose que des dialogues dirigés.

6.2. Place des contraintes syntaxico-sémantiques dans la reconnaissance de la parole

L'idée d'utiliser les contraintes syntaxico-sémantiques au cours de la reconnaissance de la parole a été appliquée à de nombreux systèmes classés sous le nom générique de systèmes de compréhension de la parole. Cette idée a été et est toujours controversée. Pourtant, les performances atteintes par ces systèmes montrent clairement l'intérêt de cette approche, puisqu'en dépit de taux de reconnaissance phonétique assez faibles, ces systèmes atteignent un bon taux de compréhension sur des langages de plusieurs centaines de mots.

L'étude que nous venons de présenter est en grande partie fondée sur cette idée. Cela ne signifie pas que nous considérions comme inutiles des progrès au niveau acoustico-phonétique, bien au contraire ; l'expérience montre qu'une amélioration, même faible, des performances à ce stade a un effet très significatif sur le taux de compréhension global d'un système. Par ailleurs, il est évident que la syntaxe et la sémantique ne sont pas d'un grand secours pour l'identification d'un nom propre dans une grande liste, problème qui se présente dans l'application que nous avons choisie. A défaut, elles peuvent donner des indications permettant de savoir quand un problème se présente, et de le traiter de façon différente des fragments de phrases dont l'identification est plus aisée.

7. CONCLUSION ET PERSPECTIVES

On a présenté la façon dont la syntaxe et la sémantique sont utilisées dans le système Kéal, en décrivant les mécanismes utilisés sur l'exemple du dialogue des renseignements automatiques. A court terme, l'intégration des programmes que nous avons décrits doit permettre des expériences de dialogues vocaux, afin de mettre en évidence les avantages et les inconvénients du modèle choisi. A plus long terme, on étudiera diverses extensions souhaitables de ce modèle :

- au niveau de la reconnaissance des phrases, il serait souhaitable de pouvoir effectuer dans certains cas une analyse partielle de la phrase, reportant ultérieurement l'analyse de certains fragments difficiles (par exemple, le nom du correspondant qui est plus facile à reconnaître lorsqu'on en connait l'adres-

se). L'une des approches possibles pour ce problème serait de s'aider des marques prosodiques que l'on peut repérer sur la phrase prononcée.

- au niveau de la gestion du dialogue, de nombreuses extensions sont possibles : ainsi, par exemple, l'introduction de quantificateurs dans la représentation des formes et l'ajout de mécanismes logiques permettant de construire dynamiquement certaines informations déductibles des énoncés de l'utilisateur.

REFERENCES

BOCHMANN, G.V., 1976 : "Semantic Evaluation from Left to Right", Communications of the ACM, Vol. 19, N° 2, pp. 55-62.

GILLET, D., QUINTON, P., 1980 : "Utilisation de transformations d'arborescences pour l'extraction de l'information sémantique dans le système Kéal", Séminaire GALF-AFCET Syntaxe et Sémantique, Paimpont, et Recherches Acoustiques CNET, Vol. 6.

GRESSER, J.Y., MERCIER, G., 1975 : "Automatic Segmentation of Speech into Syllabic and Phonemic Units : Application to French Words and Utterances", in "Auditory Analysis and Perception of Speech", G. Fant and M.A.A. Editors, Academic Press, London.

GROC, B., TUFFELLI, D., 1980 : "A Continuous Speech Recognition System for Data Base Consultation", IEEE ICASSP, Denver, Vol. 3, pp. 896-899.

KNUTH, D.E., 1968 : "Semantics of Context-Free Languages", Math. Systems Theory, Vol. 2, N° 2, pp. 127-145.

LORHO, B., 1974 : "De la définition à la traduction des langages de programmation : méthode des attributs sémantiques", Thèse d'Etat, Université Paul Sabatier, Toulouse.

LOWERRE, B.T.,1976 : "The Harpy Speech Recognition System", Ph. D. Thesis, Carnegie Mellon University, Pittsburg.

MELONI, H., GUIZOL, J., 1980 : "Un système de compréhension automatique de la parole continue", Congrès AFCET Informatique, Nancy, pp. 505-516.

MERCIER, G., NOUHEN, A., QUINTON, P., SIROUX, J., 1980 : "The Keal Speech Understanding System", in "Spoken Language Generation and Understanding" (Proceedings of the NATO ASI), J.C. Simon Editor, Reidel Publishing Company, Dordrecht.

NOUHEN, A., SIROUX, J., 1979 : "Sémantique et pragmatique dans le système Kéal", Congrès AFCET-IRIA Reconnaissance des Formes et Intelligence Artificielle, Toulouse, Vol. 1, pp. 83-91.

NOUHEN-BELLEC, A., SIROUX, J., 1981 : "CADI : Constructeur Automatique de Dialogue Intelligent intégré dans un système de reconnaissance et de compréhension de la parole", Thèse de 3ème Cycle, Université de Rennes, à paraître.

PIERREL, J.M., HATON, J.P., 1980 : "Syntactic-Semantic Interpretation of Sentences in the Myrtille II Speech Understanding System", IEEE ICASSP, Denver, Vol. 3, pp. 892-895.

QUINTON, P., VIVES, R., GRESSER, J.Y., 1975 : "Dialogues avec un robot", 6èmes Journées d'Etudes sur la Parole, Toulouse, Vol. 1, pp. 411-421.

QUINTON, P., 1980 : "Contribution à la reconnaissance de la parole. Utilisation de méthodes heuristiques pour la reconnaissance des phrases", Thèse d'Etat, Université de Rennes.

ROSENBERG, A.E., RABINER, L.R., WILPON, J.G., 1980 : "Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independant Templates", The Bell System Technical Journal, Vol. 59, N° 4, pp. 571-592.

SAITTA, L., 1978 : "Fuzzy Semantic Network for a Speech Understanding System. An Experimental Study", IEEE ICASSP, Tulsa, pp. 433-436.

SHIGENAGA, M., SEKIGUCHI, Y., 1978 : "Speech Recognition of Spoken Fortran Programs", 4th. Int. Joint Conf. on Pattern Recognition, Kyoto, pp. 990-994.

VIVES, R., 1978 : "Utilisation de l'information phonémique et syllabique pour la reconnaissance de mots prononcés isolément ou dans des phrases", 10èmes Journées d'Etudes sur la Parole, Grenoble, pp. 375-384.

WALKER, D.E., 1976 : "Speech Understanding through Syntactic and Semantic Analysis", IEEE Trans. on Computers, Vol. C-25, N° 4, pp. 432-439.

WOODS, W.A., et al., 1976 : "Speech Understanding System. Final Report", BBN Report N° 3438, Boston.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

RECHERCHE DE SEGMENTS STABLES DANS LES COMPTINES ENFANTINES PARLEES

N.TIRANDAZ — Laboratoire de Physique Electronique
Université Claude Bernard Lyon
et Ecole Normale Supérieure de Casablanca (Maroc)

C.BERGER-VACHON — Laboratoire de Physique Electronique
Université Claude Bernard Lyon

.RESUME

L'étude du langage enfantin à travers les comptines est une étape intéressante que les auteurs se proposent de développer.

Les comptines (langage très scandé) sont prononcées en continu devant le vocodeur du CNET Lannion et elles sont soumises au prétraitement développé par l'équipe de reconnaissance de la parole du CNET. Il en résulte un treillis phonémique dans lequel il faut essayer d'identifier automatiquement la comptine prononcée.

Tout d'abord des mots du vocabulaire sont repérés dans le continuum. Ensuite, on construit des suites de mots balayant l'ensemble de la prononciation. Ceci conduit à un nombre très important de solutions ; on va essayer de réduire ce nombre en détectant des zones stables qui serviront de point d'ancrage à la reconnaissance en éliminant beaucoup de suites.

On se propose de rechercher des zones stables par la considération :

- de la psychologie de l'enfant ; elle introduit des contraintes syntaxico-sémantiques
- de la limitation des possibilités de succession des mots en français
- de l'impossibilité d'avoir des suites trop longues de mots mono ou diphonémiques.

L'analyse actuelle montre que les hypothèses introduites sont utiles pour réduire l'explosion combinatoire, mais qu'elles sont encore insuffisantes pour aboutir à une solution correcte en un temps de calcul raisonnable sur ordinateur.

RESEARCH OF RELIABLE STRINGS IN SPOKEN CHILD DITTIES

N.Tirandaz et C.Berger-Vachon

. ABSTRACT

The study of child language through child ditties is an interesting step developed by the authors.

Ditties are a language having a strong beat. They are pronounced continuously and analysed by the CNET Lannion's vocoder. Then, the pretraitement developed by the speech recognition group of the CNET Lannion is applied.It produces a matrix of possible phonemes in which we try to identify automatically the spoken ditties.

Firstly, words of the vocabulary are spotted in the matrix. Then successions of words are constructed, covering the whole pronunciation. This lead to a huge number of successions. We try to reduce this number by the detection of reliable strings which will be used as starting points for future recognition.

Following considerations help the research of reliable strings :
- child psychology introduces syntaxico-semantic constraints
- the possibilities of words succession in french language are limitated
- long strings of mono or diphonemic words are very unlikely.

The present study shows that introduced hypothesis are useful to break down the combinative explosion, but are yet insuffisant to produce a correct solution in a reasonable computing time.

I INTRODUCTION

Le langage humain est quelque chose de très difficile à reconnaitre et de nombreuses équipes dans le monde recherchent des invariants du langage pouvant servir de base à un système de reconnaissance automatique de la parole prononcée en continu. On peut penser que l'enfant qui utilise un langage rudimentaire, présente une étape intéressante à explorer.

Bien que séduisante, cette idée se heurte à la complexité de la formalisation du comportement enfantin (PIAGET 1964 ; MAC NEIL 1970) et l'exploration des caractéristiques de la parole de l'enfant est loin d'être évidente (TIRANDAZ 1977,1979). On sait que l'apprentissage de la parole par l'enfant obéit à un système cybernétique associant la phonation et l'audition (LEGROS 1978) et que la rupture de la boucle conduit à des difficultés lourdement handicapantes ; ceci souligne l'aspect global du problème.

Pour espérer reconnaître un sous-ensemble du langage enfantin constitué par les comptines (BRAY-CLAUSARD 1978), on se propose de prendre en compte les différentes phases habituellement abordées dans de telles démarches (MARIANI 1979) avec le niveau acoustique puis une analyse des structures syntaxiques (LEWINSON 1978) et sémantiques. L'emploi de syntaxes rigides facilite la reconnaissance en limitant à chaque instant le nombre de candidats-mots possibles ; ceci permet de pallier aux insuffisances de l'étage acoustique (BERGER-VACHON 1978 ; PIERREL 1975). Il semble plus difficile de faire intervenir les données sémantiques, mais l'expérience de T.WINOGRAD (1970) a montré qu'on pouvait, à l'aide de la robotique, se servir de ces considérations.

L'étude que nous avons effectuée constitue une première analyse dans la reconstitution des comptines enfantines parlées ; elle s'appuie sur les résultats donnés par le vocodeur du CNET Lannion et sur le prétraitement habituellement effectué par l'équipe de reconnaissance automatique de la parole (MERCIER 1978). Ce prétraitement conduit à un treillis phonémique dans lequel il est classique de localiser les mots par des méthodes matricielles (VIVES 1976) ; mais on peut aussi choisir en première approximation des méthodes plus simples, lorsqu'on s'attache à l'étude d'autres aspects du langage.

Les comptines sont un langage créé par les adultes pour aider les enfants à compter. Deux notions de base (au moins) ont été retenues lors de leur élaboration :

- l'importance du rythme (responsable de leur aspect scandé) ; on peut étudier l'influence de ce trait lors de la segmentation automatique
- la"simplicité" du langage employé qui doit être compris par l'enfant. Tout mot ambigu ou inconnu doit en principe être suivi par sa définition. Par exemple, le prénom "TARA" est immédiatement précisé par la séquence "le petit rat".

On indique sur l'annexe I un exemple de comptines.
Le but de cette étude est une meilleure compréhension des liens qui existent entre la psychologie de l'enfant et l'acquisition du langage. La reconnaissance automatique des comptines parlées permet d'évoquer certaines hypothèses et de vérifier leur validité. Nous souhaitons que ces hypothèses soient une aide pour la détection des comptines dans le continuum vocal. Pour commencer, nous nous sommes proposés de rechercher des traits simplifiant l'analyse syntaxico-sémantique du langage.

L'exemple traité est ici très simple ; en utilisant les mots de cinq comptines, nous avons construit un vocabulaire de 160 mots. Le treillis phonémique correspondant à la prononciation de la comptine "Tara, le

petit rat, s'en va au Canada avec Sacha, le petit chat" (cf. Annexe II) a été soumis à l'analyse.

Nous avons recherché dans le continuum du langage tous les mots du vocabulaire qu'il est possible de trouver à l'aide d'un algorithme de comparaison simple. Ensuite, on a tenté de reconstituer la comptine en alignant ces mots-candidats. La solution théorique amène à considérer environ 1600 milliards de solutions. L'emploi de règles simples a permis de former rapidement des suites de mots qui serviront de point d'ancrage dans nos études futures pour reformer les comptines parlées.

II ANALYSE DU PROBLEME

2-1 - Notion de sémantème et phonatème

Nous appellerons "sémantème" un mot considéré dans son acceptation sémantique. Le sémantème est représenté dans le dictionnaire par un code. Ce code porte la signification du mot, il comprend 6 bits : abcdef.

Les bits "a" et "b" permettent la distinction entre trois grandes catégories syntaxiques : les noms communs (00), les verbes (01 et 11) et tous les autres mots (10).

Les quatre autres bits : cdef, portent des informations plus précises sur le mot à l'intérieur de sa catégorie. Il est donc possible d'augmenter la précision de l'information, en augmentant le nombre des bits, sans pour autant revoir tout le système de codification.

Indiquons quelques exemples de codes :

abcdef		
000000	:	nom commun féminin abstrait (ex : bonté)
001000	:	nom commun masculin abstrait (ex : voyage)
100111	:	nom propre
010011	:	verbe infinitif transitif direct et indirect(ex :dire)
100000	:	article masculin

Nous appellerons "phonatème" la représentation phonémique du mot. Ainsi, au mot chat sera associé le sémantème 001100 (qui signifie "nom commun masculin animé non humain petit"). Il aura deux phonatèmes : /ʃa/ et /ʃɑ̃/. En effet, le phonème /a/ sera parfois analysé selon le contexte ou le locuteur comme un /ɑ̃/. Il en est de même par exemple pour le mot "avec" qui, pour des raisons de "sandhi" ou d'assimilation, peut perdre son /k/ final et être perçu comme /avɛ/ au lieu de /avɛk/ (MALMBERG 1975).

2-2 - Localisation des mots dans le continuum vocal

Le prétraitement de l'onde acoustique par l'équipe de reconnaissance automatique de la parole du CNET, procède à une segmentation de l'onde. Pour chaque segment, une liste de phonèmes possibles est proposée. Nous disposons à la sortie de ce prétraitement d'un treillis représenté partiellement par le tableau 1 (correspondant à l'annexe II).

La localisation des mots dans le treillis phonémique revient à la construction de tous les phonatèmes possibles. Dans la concaténation des phonatèmes, nous dépendons de la segmentation du prétraitement acoustique. Ce prétraitement basé sur les caractéristiques physiques de l'onde acoustique (énergie, pente, fondamental ...) peut avoir introduit soit une compression phonémique (le même segment porte deux phonèmes qui normalement devraient se suivre), soit une segmentation supplémentaire.

Pour tenir compte de cet aspect des choses, nous sommes amenés à accepter des décalages à l'intérieur des mots jusqu'au tiers de leur longueur phonémique. Ce décalage sera d'une colonne pour les mots diphonémi-

ques. Cet algorithme de concaténation nous donnera une matrice ayant pour colonne les segments donnés par le prétraitement phonémique et en ligne les mots phonétiques ou phonatèmes pouvant commencer par cette colonne. Cette concaténation a pour objet d'élaborer le maximum de phonatèmes possibles et n'introduit aucune règle restrictive (voir tableaux 2 et 3).

Nº de la colonne (ou du segment)	1	2	3	4	5	6	7
phonèmes candidats	t k p g d b ʃ	a ɛ̃ a	f s ʃ r	a o õ ã	n ʒ l z j v	o ø œ	

Tableau 1 : Extrait du treillis phonémique ; chaque colonne correspond à un segment et ses éléments sont les candidats proposés par l'étage acoustique.

normal	t	a	r	a			pour /tara/
avec 1 décalage		a	r	.	n	o	pour /arno/ phonatème de Arnaud
avec compression			f	o a			pour /foa/ phonatème pour fois, foie et foi

Tableau 2 : exemples d'analyse de prononciation de mots.

Nº de colonne	1	2	3
	tã dã tara kã tro pa dɛ̃	a ɛ̃ arno ã	fã ra ʃã sõ ʃo ʃa

Tableau 3 : extrait de la matrice des phonatèmes : il s'agit de la représentation phonétique de mots reconnus à partir de la colonne indiquée sur la première ligne du tableau (cf : annexe III).

2-3 - Détection d'éléments stables dans la comptine

Par une stratégie d'alignement et selon un graphe arborescent, on balaie toutes les suites de phonatèmes qui par enchaînement peuvent être constituées à partir des phonatèmes candidats. Dans cette construction, on est amené à tenir compte de quelques règles de la phonétique combinatoire et de la linguistique.

2-3-a - Introduction de quelques règles de la phonétique combinatoire

En prononçant les sons du langage, l'homme a tendance à appliquer la règle de l'effort minimum. Cela entraîne une modification des sons en contact avec d'autres sons. Certaines de ces modifications ont été prises en compte en introduisant plusieurs phonatèmes pour un même mot. Mais d'autres modifications interviennent non pas à l'intérieur des mots mais à la frontière de ceux-ci. Par exemple, l'expression "tête de veau" peut être reconnue non pas comme deux plosives qui se suivent (/t/ et /d/) mais d'une seule allongée. Il en est de même dans "cette table" qui peut être prononcée /sɛtabl/. Ces phénomènes nous amènent à envisager la possibilité de chevauchement des phonatèmes. Si un mot finit en colonne j, le mot suivant peut commencer :
- en colonne j , il y a alors chevauchement
- en colonne j+1, cas normal et le plus fréquent
- en colonne j+2, soit qu'un allongement du dernier phonème ait entrainé une segmentation supplémentaire, soit qu'une segmentation fortuite se soit produite due, par exemple, au bruit. Ainsi, dans notre exemple de comptine (t a ra lə pəti ra sɑ̃ wa o k a n a d a avɛk s a ʃ a lə pəti ʃa) le phonatème /pəti/ finit en colonne 10, le phonatème /ʃa/ commence en colonne 11 (succession classique), tandis que le phonatème /sɑ̃wa/ ne commence qu'en colonne 14 (j+2). Compte tenu de cette remarque, nous ne détecterons que les silences durant au moins deux segments (ex : colonnes 26-27-28 sur l'annexe II).

7	8	9	10	11	12	13	14	15	16	17
p	ə	t	i	ʃ	a	...	s	ã	w	a

saut d'1 colonne

Tableau 4 : Exemple de segmentation introduisant un phonème supplémentaire par rapport à une décomposition idéale. On voit aussi que le phonème /ʃ/ peut être substitué au /r/ non reconnu.

L'explosion combinatoire qui conduit à environ 1600 milliards de solutions lorsqu'on aligne simplement à la suite les uns des autres les phonatèmes va atteindre des limites astronomiques avec ces nouvelles possibilités. Nous devons dès lors édicter quelques règles simples qui limitent tant soit peu cette explosion et éliminent un grand nombre de solutions triviales. Ces règles porteront sur les chevauchements : par exemple, on n'admet que les chevauchements des phonèmes de même nature. De plus, en cas de chevauchement /ã/ ne peut précéder /a/, car en français, dans une telle éventualité il y a un /t/ ou un /z/ qui se glisse entre les deux. Par l'introduction de ces règles, nous entamons une nouvelle phase de notre démarche : après avoir constitué le maximum de phonatèmes et avoir envisagé toutes les successions théoriques possibles, nous cherchons désormais à éliminer le maximum de suites possibles.

2-3-b - Quelques règles linguistiques

1 - Longueur des comptines en mots

L'exemple ci-dessus nous montre une autre source de combinaisons triviales : les mots mono et/ou diphonémiques. Tout en admettant la pos-

sibilité de chevauchement pour les phonatèmes, on doit éliminer les chevauchements sur une colonne des mots monophonémiques. D'autre part, plus de 50% des textes écrits sont constitués des mots outils qui sont à prononciations courtes (exemple : "de", "à", "ou", "en", "le", "la", etc...) Or de tels mots sont détectés à presque toutes les colonnes. Il faut éviter des enchaînements trop longs de tels mots. On introduit la notion de longueur des comptines en mots. Le rapport λ du nombre de segments détectés par le prétraitement acoustique sur le nombre des mots d'une comptine varie de 2,4 à 3,9. L'étude de l'évolution de ce rapport le long de la comptine (fig.1) montre que ce rapport est toujours supérieur à 2. Toutes les solutions qui ne respectent pas cette inégalité dès le 3ème mot seront éliminées.

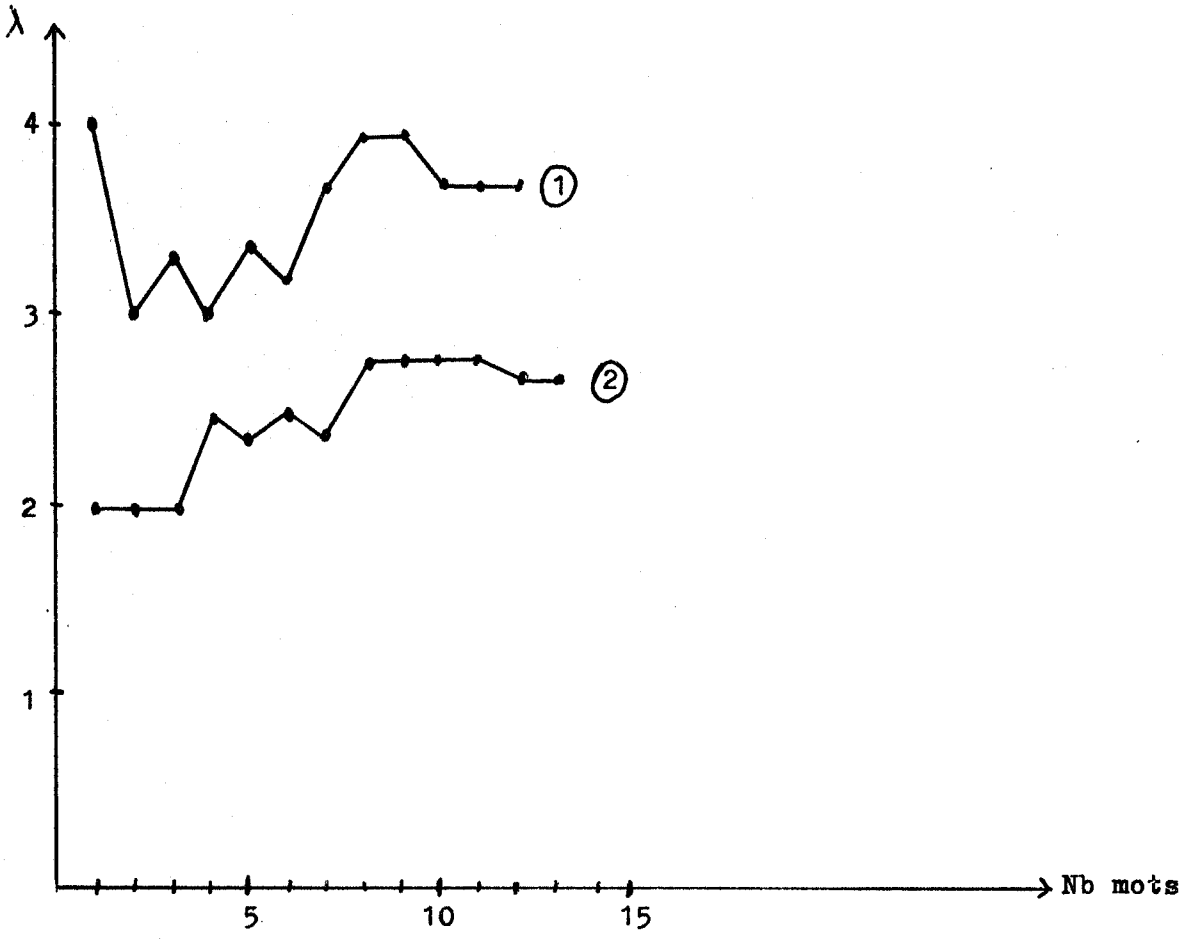


Figure 1 - Evolution de λ le long de deux comptines
1 : cas de notre exemple
2 : cas le plus défavorable.

2 - Des règles syntaxiques

Les considérations précédentes ont pour objet de casser l'évolution explosive de notre stratégie d'alignement. On peut y adjoindre quelques règles syntaxiques. Elles sont prohibitives et rigides. Elles sont destinées à réduire le nombre des enchaînements et non à construire par concaténation des bribes de phrases. On ne peut donc parler de grammaire à proprement dit.

* Règles de phonétique combinatoire pour les chevauchements et les successions

- une sifflante ne peut chevaucher qu'une sifflante
- une plosive ne peut chevaucher qu'une plosive
- on ne peut avoir les successions suivantes :

/ã/ /a/ /a/ /y/ /ɛ̃/ /a/

* Règles syntaxiques

- Un nom commun est toujours précédé d'un des éléments suivants :
  - . un article
  - . un nombre
  - . un adjectif
  - . un possessif
  - . une préposition
- Un verbe non infinitif ne peut être précédé d'une préposition
- Une conjonction ne peut être précédée d'une préposition.

3 - Une structure syntaxico-sémantique et phonologique

Il est intéressant de considérer la structure propre du langage des comptines. Henri Wallon nous apprend (1945) que la distinction des objets parait subordonnée à leur identification. Nous avons donc recherché, dans les comptines, les mots qui étaient définis dans le texte. Cette recherche a été rarement positive. Si on suppose que tous les mots utilisés sont compris par l'enfant, nous avons déjà affaire à un langage très évolué. Nous pensons que la procédure de définition a été insuffisamment utilisée dans les comptines.

La procédure de désignation la plus fréquemment employée concerne la définition d'un nom propre (tableau 5).

TARA	le petit rat
SACHA	le petit chat
MIMI	le ouistiti
ZOZO	le souriceau
BUBU	la tortue
AMEDEE	le jardinier
LUCIEN	le magicien
FLAVIEN	le musicien
PAMPAM	le faisan
PIMPIM	le dauphin
POMPOM	le hérisson

Tableau 5 : Noms propres suivis d'une identification (conforme à la règle étudiée dans ce paragraphe) dans les comptines.

Ceci peut être formalisé en utilisant les éléments de la théorie des langages :

$$\left\{\text{Nom propre}\right\} \underset{R_2}{\uparrow} \left\{\left\{\text{le}\right\} \left\{\text{la}\right\}\right\} \underset{R_2}{\uparrow} \left\{\left[\left\{\text{petit}\right\}\right]\right\} \left\{\text{nom commun}\right\} \quad R_1$$

Cette écriture indique qu'un nom propre doit être suivi de "le" ou "la", puis de "petit" qui peut être facultatif et enfin d'un nom

commun. Bien évidemment, "petit" pourrait être remplacé par une liste d'adjectifs dans une généralisation de cette analyse. Cette structure obéit de plus à un certain nombre de contraintes phonologiques liées à sa présence dans les comptines :

R_1 : le rythme impose au dernier phonème du nom propre d'être identique au dernier du nom commun

R_2 - (règle de limite nette) : "le" ou "la" commencent immédiatement après le phonème (une voyelle) terminant le nom propre ; il n'y a ni chevauchement, ni décalage (ceci, vérifié expérimentalement, est lié à l'aspect scandé de la comptine)

- le nom commun ou l'adjectif "petit" commencent immédiatement après "le" ou "la".

La procédure d'identification de cette structure élémentaire se déduit de ce qui vient d'être exposé ; lorsqu'un nom propre est détecté dans le treillis, la logique essaie de construire la structure élémentaire. La réponse est positive dans les cas indiqués ci-dessus (tableau 5) et négative ailleurs, pour les premiers exemples que nous avons analysés.

Comme il est nécessaire de reconnaître les "zones stables", nous avons dû introduire un équivalent de /lə/ qui est /yə/ et un de /pəti/ qui est /pəsi/. De plus, il est important que les noms propres soient correctement détectés, ce qui n'est pas toujours possible lorsqu'ils sont choisis au hasard. Nous n'avons eu qu'un seul échec dans la liste indiquée tableau 5 (pour BUBU les /b/ n'ont pas été détectés, le premier étant omis et le second remplacé par un /t/) ; il est évident qu'un accroissement anarchique du vocabulaire comporte un risque important de voir diverger cette analyse.

En ce qui concerne notre comptine étudiée (particulièrement favorable puisqu'elle contient deux fois une telle structure), on remarque que les structures "Tara le petit chat" (chat est reconnu à la place de rat) et "Sacha le pesi chat" sont correctement identifiées.

III RESULTATS ET DISCUSSION

La segmentation automatique de la comptine "Tara le petit rat s'en va au Canada avec Sacha le petit chat" conduit à 44 segments (cf. annexe II). Les mots détectés dans le continuum sont indiqués annexe III. Si on exclue les parties correctement identifiées par la procédure syntaxico-sémantique et phonologique ("Tara le petit rat" et "Sacha le petit chat") il ne reste que 19 segments. Cette réduction est très favorable compte tenu de l'analyse arborescente utilisée, mais il y a encore plusieurs milliers de possibilités pour retrouver la suite "s'en va au Canada avec". Nos travaux actuels essaient d'utiliser de nouveaux traits du langage enfantin pour aboutir plus vite à une solution correcte.

Le but ultime de cette étape étant la reconnaissance des comptines en continu, nous avons limité le temps total de lecture et de l'exécution du programme de concaténation à une minute. La logique de reconnaissance employée est schématisée sur la figure 2. On essaie d'aligner les mots détectés dans le continuum de la parole pour recouvrir l'ensemble de la prononciation de la comptine. Lorsqu'une suite a été trouvée, on remplace le n^e mot par une autre possibilité au n^{ieme} niveau. Il peut y avoir un nombre variable de mots pour recouvrir les 44 colonnes du continuum. Bien entendu, la colonne de début d'un niveau n'est pas fixe ; elle dépend de la longueur des mots employés. Le nombre n est aussi variable.

Une fois que toutes les possibilités au n^{ieme} niveau ont été épuisées, on prend le deuxième mot au niveau n-1, et ainsi de suite. Cette logique a été programmée en Fortran sur IRIS 80 ; nous avons limité le temps d'unité centrale à une minute.

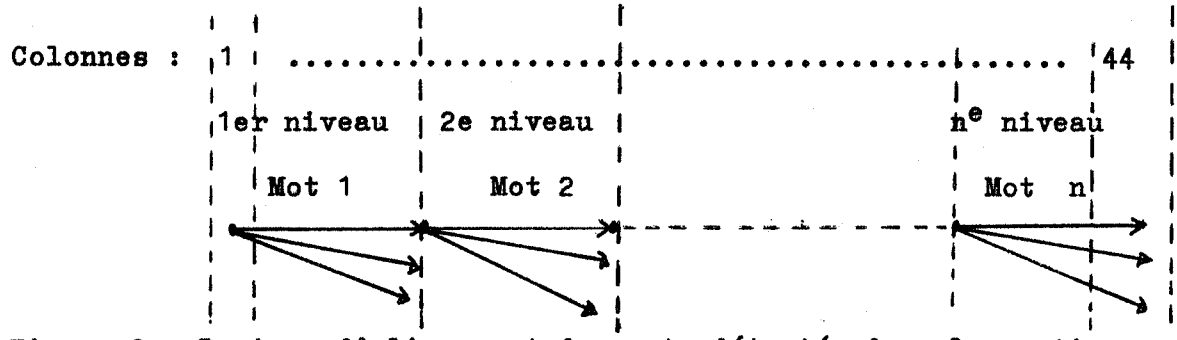


Figure 2 - Logique d'alignement des mots détectés dans le continuum de la parole pour concerner les 44 segments.

Un premier essai sur les 44 colonnes, sans aucune règle restrictive autre que l'interdiction des doubles chevauchements, permet de constater que la machine donne en 0.77 minutes, 671 solutions. Ces solutions correspondent à un balayage entre la 35ème colonne et la 43ème colonne. Aucune de ces solutions, comme on pouvait s'y attendre, ne contient une bribe de la solution finale. Dans le même laps de temps, avec un algorithme contenant les règles syntaxiques et de longueur précitées, on balaie à partir de la 33ème colonne et la machine donne 329 enchaînements. Dans les solutions construites par l'IRIS 80, on constate que la limitation des combinaisons n'est pas trop brutale. Sur le plan qualitatif cependant, les résultats sont dignes d'intérêts : dès la première suite, on constate l'existence des bribes de la comptine. Dès la 230ème, nous retrouvons 8 mots sur les 12 de la comptine groupés en deux éléments stables. Ces deux éléments stables correspondent à Tara le petit rat et Sacha le petit chat.

Les sorties de l'ordinateur donnent des suites phonémiques du type suivant :

Tara	le	petit	cha	s'en va	au canada		avec	Sacha	le
dɛ̃ra	lə	pəti	ʃa	dãpa	p r e e l a də	s e	a v e	saʃa	lə

petit	chant
pəti	ʃɑ̃

L'utilisation de l'indice de stabilité λ sur des ensembles de segments phonétiques renforce les résultats obtenus après application de la recherche de la structure syntaxico-sémantique et phonologique. On pourra s'en servir pour accroître la longueur des zones stables identifiées.

Exemple :

comptine à reconnaître : Tara lə pəti ra sãva o kanada avɛk saʃa lə pəti ʃa

reconnue dãra lə pəti ʃa avɛ saʃa lə pəsi ʃa

1er élément stable(*) 2ème élément stable(*)

(* : stabilité obtenue en considérant principalement l'indice λ)

Les structures syntaxico-sémantiques et phonologiques sont détectées avec une sécurité voisine de 100% (ceci est possible quand on a affaire à un vocabulaire limité, à une prononciation coopérative, et à un prétraitement efficace tel que celui effectué par l'équipe du CNET Lannion). Cette détection réduit le nombre de phonèmes à analyser et introduit des données contextuelles qu'il sera souhaitable d'utiliser dans la suite. L'analyse de la comptine réduite (par l'exclusion des segments "Tara le petit rat" et "Sacha le petit chat") permet de ne soumettre à la logique arborescente indiquée sur la figure 2 que la portion comprise entre les colonnes 13 et 31 (extrémités comprises). Cette étude est actuellement en cours.

Nous recherchons aussi d'autres structures stables dans les comptines, à partir des données de la psychologie infantile, pour que plus de comptines soient concernées et pour limiter l'explosion combinatoire. L'élaboration d'un score de reconnaissance permettra ensuite de classer les suites formées par l'ordinateur.

Comme il semble peu raisonnable de vouloir reconnaître toutes les comptines (trop proches du langage naturel) nous devrons probablement, dans le futur, nous limiter à un sous-ensemble qu'il sera possible d'agrandir en modifiant certaines comptines initialement abandonnées.

IV CONCLUSION

La méthode présentée et analysée ici n'est qu'une partie d'un processus destiné à analyser et si possible à reconnaître des comptines prononcées en continu. Elle a pour objet la détection d'éléments pouvant servir de point d'ancrage. Elle utilise une stratégie arborescente. Elle trouve sa limitation en cela même : il faut l'abandonner dès que possible pour gagner du temps. Elle devra être suivie d'une stratégie constructive (par opposition à celle de l'investigation caractérisant les méthodes arborescentes), c'est à dire la construction par l'utilisation d'une grammaire appropriée autour des points d'ancrage. Cette phase que l'on peut qualifier d'initiale doit donc satisfaire aux conditions suivantes :

a) construire le maximum de mots à partir du treillis phonémique
b) rester suffisamment ouvert pour ne pas éliminer la bonne solution dès le départ
c) arrêter l'explosion combinatoire et éliminer suffisamment de solutions triviales pour accélérer le balayage arborescent
d) reconnaître quelques éléments stables pouvant servir d'ilots de confiance et permettre l'abandon de la stratégie arborescente.

Dans cette phase, on n'a pas épuisé les informations des sémantèmes : la plus grande partie des informations contenues dans les codes est encore inexploitée. D'autre part, les algorithmes mis en oeuvre dans cette phase n'ont établi aucune démarche hiérarchique dans le choix des termes non éliminés. Nous espérons et croyons que la mise en oeuvre des stratégies tenant compte des certaines priorités sémantiques et syntaxiques d'une part, psychologiques d'autre part, sont de nature à accélérer la recherche de la bonne solution.

REFERENCES

BERGER-VACHON C, MESNARD G., 1978, Application des décisions markoviennes dans un problème de reconnaissance séquentielle de la parole ; -Annales des Télécommunications ; T.33 N°3-4.

BRAY-CLAUSARD, 1978, Apprendre à écouter, comptines ; -OCDL Paris.

LEGROS C., 1978, Reconnaissance des voyelles et application à la rééducation de la parole ; Revue Acoustique ; Vol 11, N°44 , pp 25-29.

LEWINSON S.E., ROSEMBERG A.E., FLANAGAN J.L., 1978, Evolution of a word recognition system using syntax analysis ; -Bell System, Tech. Journal (USA) T.57 N°5 , pp 1619-1626.

MAC NEIL D., 1970, The development of language ; -P.E. MUNSEN (ed) Carmichael Manual of Child Psychology, 2° ed - New York.

MAC NEIL D., 1970, The acquisition of language, the study of developmental linguistics, New York, Harper Row.

MALMBERG B., 1975, La Phonétique ; - 11° Ed. PUF - Paris.

MARIANI J.J., 1979, Formalisation du lexique et des règles phonologiques sur le système Esope 0 ; -10èmes Journées d'étude sur la parole. Proc. pp 224-234 - Grenoble.

MERCIER G., 1978, Evolution des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL ; -9èmes journées d'étude sur la parole. Proc. pp 321-342 - Lannion.

PIAGET J., 1964, Six études de psychologie ; -Denoel Gontier.

PIERREL J.M., 1975, Contribution à la compréhension automatique du discours continu ; -Thèse de spécialité - Université de Nancy1

TIRANDAZ N., BERGER-VACHON C., 1977, La place des éléments voisés dans l'acquisition du langage ; -8èmes journées d'étude sur la parole. Proc. pp 255-260 - Aix en Provence.

TIRANDAZ N., BERGER-VACHON C., 1979, Eléments d'un lexique pour la reconnaissance automatique des comptines enfantines parlées ; -10èmes journées d'étude sur la parole. Proc. pp 364-374 - Grenoble.

VIVES R., 1976, L'analyse lexicale dans le système Keal pour la reconnaissance de la parole ; -7èmes journées d'étude sur la parole. Proc. pp 115-127 - Nancy.

WALLON H., 1945, Les origines de la pensée chez l'enfant ; -P.U.F. Paris

WINOGRAD T., 1970, Procedures as a representation for data in a computer for understanding natural language ; -Ph. D. Thesis ; Department of mathematics MIT USA.

ANNEXE I : exemple de comptines

- La souris a mis du riz dans le nid de ses petits
- Tara le petit rat s'en va au Canada avec Sacha le petit chat
- Mimi le ouistiti a choisi à midi un ananas pour son repas
- Zozo le souriceau joue au cerceau sur le piano avec un anneau de rideau
- Dans la rue j'ai vu Bubu la tortue qui m'a vendu de la laitue
- Un oiseau, il a froid dans la rue, il a chaud dans son nid
- Amédée le jardinier m'a donné une clef toute dorée
- L'escargot et le crapaud sont venus chez la tortue, ils ont mangé des artichauts des abricots et des gâteaux
- Au mois de mai, avec mon poney, je fais des bouquets de muguet, de genêts et d'oeillets
- La chatte a mal aux pattes, la belette a mal à la tête, l'hirondelle a mal aux ailes, un lapin médecin est arrivé pour les soigner
- Henri a un colibri qui chante en si, Bruno a un loriot qui chante en do, Nicole a un rossignol qui chante en sol.

.....

Nous disposons au total d'une cinquantaine de comptines de ce type.

ANNEXE II : Treillis phonémique de la comptine étudiée : "Tara le petit rat s'en va au Canada avec Sacha le petit chat".

On indique le numéro du segment et l'ensemble des phonèmes-candidats par segment.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
t	a	f	a	n	o	t	o	t	i	f	a	b	s	ã	w	ã	r	r	t
k	a	s	a	ʒ	ə	p	ə	k	ɛ	s	ã	d	ʃ	o	p	õ	p	o	k
p	ɛ̃	ʃ	ã	i	œ	k	œ	g	y	k	ɔ̃	g	f	õ	r	o	w	ã	d
d	e	ʁ	o	z	ʁ	d	a	p		ʃ	o	w				a			p
g			õ	j		b		b			ɛ̃	r				a			b
b				v		g		d											g
ʃ						s		j											ʃ
						ʒ		u											

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
a	n	a	b	a	w	f	t	ə	p	ɛ	s	a	ʃ	a	n	o	p	o	s
ɛ	z	ɛ̃	d	a	r	s	p	ɛ̃	d	e	ʃ	ɛ̃	ʒ	a	ʒ	ə	d	o	ʃ
ɛ̃	l	ɛ	g	ɛ̃	v		d	a	n	a	f	a	s	e	l	œ	k	œ	f
		a	t		n		k	o	v		t	ɛ	j	õ	z		t	a	
			p				g	a	z		ʒ	e	u	ɛ̃			g		
			k				b							o			b		
							f												

41	42	43	44
i	ʃ	p	õ
y	s	ʒ	ã
	j	ʃ	o
	ʒ	l	a
	f	n	
		a	

ANNEXE III : Mots reconnus dans le treillis phonémique (on indique la colonne de début) ; ∧ = blanc (ou silence) ;

1 - Tara ; dã ; kã ; dɛ̃ ; tro ; pa ; tã ;
2 - a ; ɛ̃ ; arno ; ã ;
3 - ra ; ʃa ; ʃã ; sõ ; ʃo ; fã ;
4 - a ; o ; ano ; õ ; ã ;
5 - lə ; lo ; ʒə ; vœ ;
6 - o ;
7 - də ; pəti ; bo ; ʒə ; sə ;
8 - o ;
9 - dy ; ki ; ty ; di ;
10 -
11 - ʃa ; ʃã ; sõ ; kã ; ʃo ; fã ;
12 - õ ; o ; a ; ã ;
13 - ∧ ; dã ; bo ; bã ;
14 - sãva ; ʃã ; sõ ; fwa ; ʃo ; fã ;
15 - o ; õ ; ã ;
16 - ra ; pã ; pa ;
17 - ã ; õ ; o ; a ;
18 - pre ; pã ;
19 - ã ; o ;
20 - se ; kanada ; ʃa ; kle ; ʃe ; de ; dɛ̃ ; gaetã ; pa ;
21 - e ; el ; ɛ̃ ; a ;
22 - la ; le ; letã ; ne ;
23 - a ; ɛ̃ ; e ;
24 - də ; dɛ̃ ; pa ; tã ; kã ; pã ; bã ; dã ;
25 - ã ; ɛ̃ ; a ;
26 - ∧ ;
27 - sə ;
28 - də ; bo ; twa ; fwa ; done ; pone ; dɛ̃ ; bwa ; pa ;
29 - ɛ̃ ; ave ; o ; a ;
30 - va ; de ; pa ; ne ;
31 - e ; a ;
32 - sə ; ʃa ; saʃa ;
33 - ɛ̃ ; a ;
34 - sə ; ʃã ; ʃo ; ʒə ; sõ ; ʃa ;
35 - ã ; õ ; ano ; a ; ɛ̃ ; o ;
36 - ʒə ; lo ; lə ;
37 - o ;
38 - pa ; dɛ̃ ; bo ; pəsi ; də ;
39 - ɛ̃ ; o ; a ;
40 -
41 - u ; il ;
42 - ∧ ; ʃã ; sõ ; ʃo ;
43 - pã ; lo ;
44 - õ ; o ; ã ;

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ALGORITHMES ET METHODES POUR LA DETERMINATION DU FONDAMENTAL

Wolfgang Hess

Lehrstuhl für Datenverarbeitung, TU München
Postfach 202420, 8000 München 2, R.F.A.

Résumé. - Dans cette présentation, on groupe les algorithmes et méthodes divers pour la détermination du fondamental en deux classes majeures: les algorithmes qui opèrent dans le domaine temporel, et les algorithmes d'analyse à terme court. Les derniers algorithmes quittent le domaine temporel par une transformation à terme court; ils fournissent une séquence de valeurs d'appréciation moyennes extraites d'intervalles de mesure successifs, l'algorithme individuel étant caractérisé par la transformation à terme court appliquée. Les algorithmes qui opèrent dans le domaine temporel examinent le signal période par période. Isolation et extraction de l'harmonique fondamentale d'un part, et l'examination de la structure temporelle par un algorithme de décision d'autre part représentent les deux extrêmes entre lesquels se trouve la plupart de ces algorithmes. D'après la présentation brève de ces methodes, on discute le rôle de la determination du fondamental pour des applications diverses: phonétique, education, pathologie et communication parlée.

Abstract. - In this paper, the various pitch determination methods and algorithms (PDAs) are grouped into two major classes: time-domain PDAs and short-term analysis PDAs. The short-term analysis PDAs leave the signal domain by a short-term transformation. They supply a sequence of average pitch estimates from consecutive frames. The individual algorithm is characterized by the short-term transform it applies. The time-domain methods, on the other hand, track the signal period by period. Extraction and isolation of the fundamental harmonic, and investigation of the temporal signal structure are the two extremes between which most of these PDAs are found. After a short review of these principles, the paper finally discusses different application-oriented aspects, i.e. the role of the PDA in phonetics, education, phoniatrics, and speech communication systems.

1. Introduction

Le parametre du fondamental (fréquence fondamentale FO et période fondamentale TO) prend une position clef dans le signal acoustique de la parole. L'information prosodique d'une phrase est dominée par ce parametre. L'ouie humain est d'un ordre de magnitude plus sensible aux changements de ce parametre qu'à ceux d'autre parametre quelconque (Flanagan 1972, Flanagan et Saslow 1958). La qualité de mesure du fondamental a des grands effets sur la qualité du signal de sortie d'un vocodeur (Gold 1977). Consécutivement, l'importance de ce parametre exige des methodes bonnes et fiables afin de le determiner.

Au premier coup d'oeil le problème apparaît simple: on ne doit que détecter la fréquence (ou période) fondamentale d'un signal approximativement périodique. Cependant, la condition préalable que le signal soit approximativement périodique se trouve souvent loin de la réalité s'il s'agit d'un signal de parole. Fig.1 en montre un exemple arbitraire (mais caractéristique). De plusieurs raisons, la tâche de détermination du fondamental compte parmi les problèmes les plus difficiles de l'analyse de la parole:

(1) En principe, la parole est un processus dynamique: la position instantanée se peut toujours changer brutalement. Des variations brusques dans la structure temporelle du signal résultent d'un tel changement qui peut se passer même entre des periodes voisines.

(2) Dû à la flexibilité du conduit vocal humain, et dû à la grande diversité des voix, il y a une multitude de structures temporelles possibles. Des formants à bande étroite qui coincident avec des harmoniques d'ordre basse (particulièrement la deuxième ou troisième harmonique) sont une source additionnelle de problèmes.

(3) Dans un signal arbitraire issu d'un locuteur inconnu, la fréquence fondamentale peut varier dans une bande d'à peu près quatre octaves (50 Hz jusqu'à 800 Hz). Particulièrement pour des voix féminines, il arrive souvent

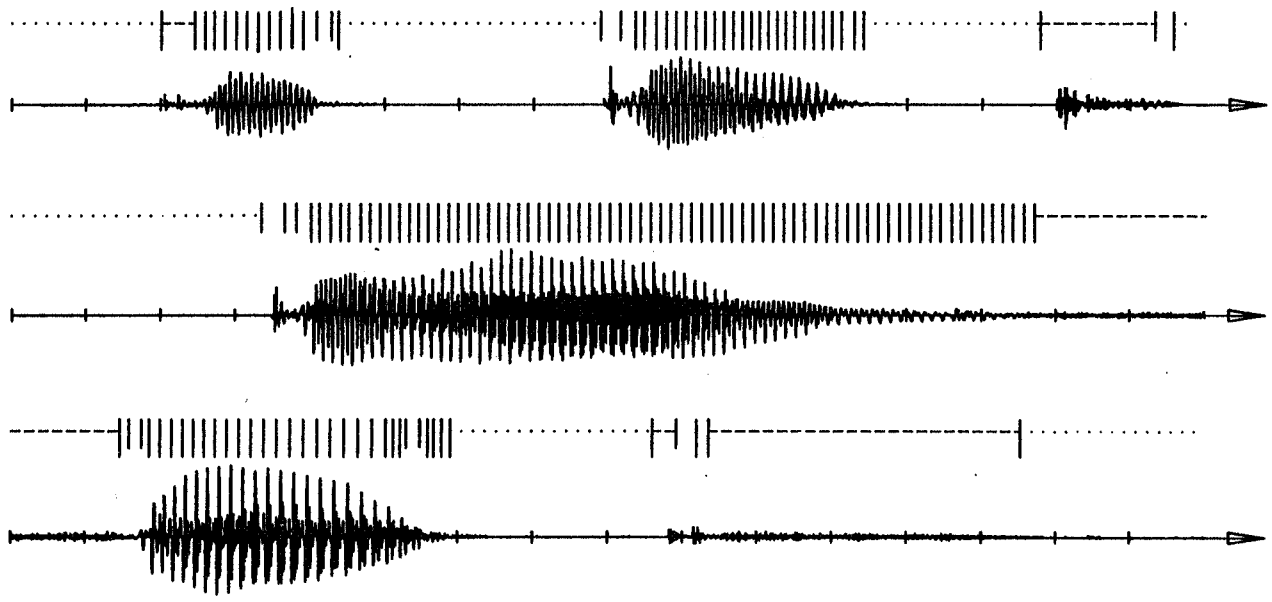


Fig. 1
Exemple d'un signal parole (après la détermination du fondamental): début de la phrase "The big brown fox ...". Locutrice féminine, signal limité à la bande téléphonique. L'analyse a été faite en utilisant l'algorithme de Hess (1979).

------ Non voisé ..... Silence | Marqueurs des périodes

Echelle temporelle: 500 ms par ligne
Marqueurs indiqués par ligne courte ont été trouvés irréguliers.

que FO coincide avec le premier formant (ce dernier varie entre 200 et 1400 Hz environ), ce qui cause de problèmes lorsqu'on emploie la méthode du filtrage inverse.

(4) Le signal excitateur est irrégulier de temps en temps. Même quand la voix est ni enrouée ni malade, le signal glottique présente certaines irrégularités (Fujimura 1968, Lieberman 1963). De plus, la voix peut tomber temporairement dans un mode irrégulier ("vocal fry"), qui est un mode d'excitation non pathologique avec des intervalles irréguliers entre des impulsions excitatrices consécutives.

Littéralement des centaines de méthodes de détection du fondamental ont été developpées. Cette publication donnera une présentation synoptique des principes les plus importants. Quant à la réalisation, on ne va pas distinguer entre un dispositif câblé (soit analogique, soit numérique) et une solution par algorithme: tous les méthodes seront regardées comme algorithmes de détection du fondamental[1)]. De plus, on ne va pas faire une différence par rapport à ce qu'on mesure actuellement: la fréquence fondamentale FO, la durée (moyenne) TO de la période fondamentale, ou la durée (ou position) de périodes fondamentales individuelles. FO et TO sont liés par la rélation

$$FO := 1 / TO \qquad (1)$$

Si l'on mesure des periodes individuelles, TO est defini instantanément comme le temps écoulé du début de la période actuelle jusqu'au début de la période suivante.

La détermination du fondamental est un sous-problème de la tâche d'analyse de la source vocale, l'autre grand sous-problème étant "où l'on trouve les périodes". Cette tâche, qui s'appelle comme "la détection du voisement" ne sera pas discutée ici. Dans la plûpart des analyseurs, ce problème est résolu séparément de la détermination du fondamental. Alors, on va supposer que la détection de voisement a déjà été faite, et qu'il n'y a que de signaux voisés a l'entrée de l'analyseur.

En suivant, on va d'abord catégoriser les méthodes diverses et puis décrire brièvement les catégories individuelles. Enfin, on va discuter quelques domaines d'application et les exigences et conditions préalables liées au domaine d'application particulier.

2. Catégorisation des principes de détermination du fondamental.

L'analyseur ordinaire se compose de trois étages (McKinney 1965): l'étage de prétraitement, l'extracteur, et l'étage de traitement ultérieur. L'extracteur exécute le mesurage propre: il convertit le signal d'entrée en une séquence de valeurs estimatives du fondamental. La tâche de l'étage de prétraitement est d'exécuter une réduction de l'information afin de faciliter le travail de l'extracteur. Enfin, les tâches de l'étage de traitement ultérieur sont plus divers et plus orientées vers l'application respective: la correction d'erreurs, le lissage de la courbe de mélodie et la visualisation des résultats sont des problèmes caractéristiques pour cet étage.

Les principes de détermination du fondamental se laissent diviser en deux grandes catégories selon le domaine du signal d'entrée de l'extracteur. Lorsque ce signal possède la même base temporelle que le signal parole, l'analyseur opère en domaine temporal. Dans tous les autres cas le domaine temporel a été quitté quelque part pendant le prétraitement. Le signal parole étant variable avec le temps, on ne peut quitter le domaine temporel que par une transformation

---

1) A cause de brévité, on appelle toute sorte de dispositif ou d'algorithme de detection du fondamental simplement comme "analyseur".

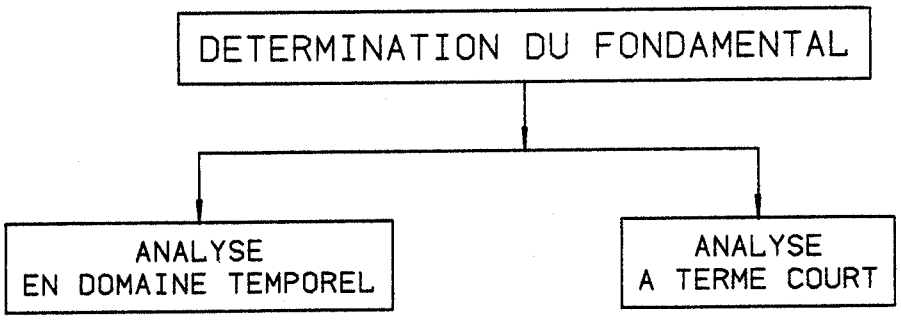
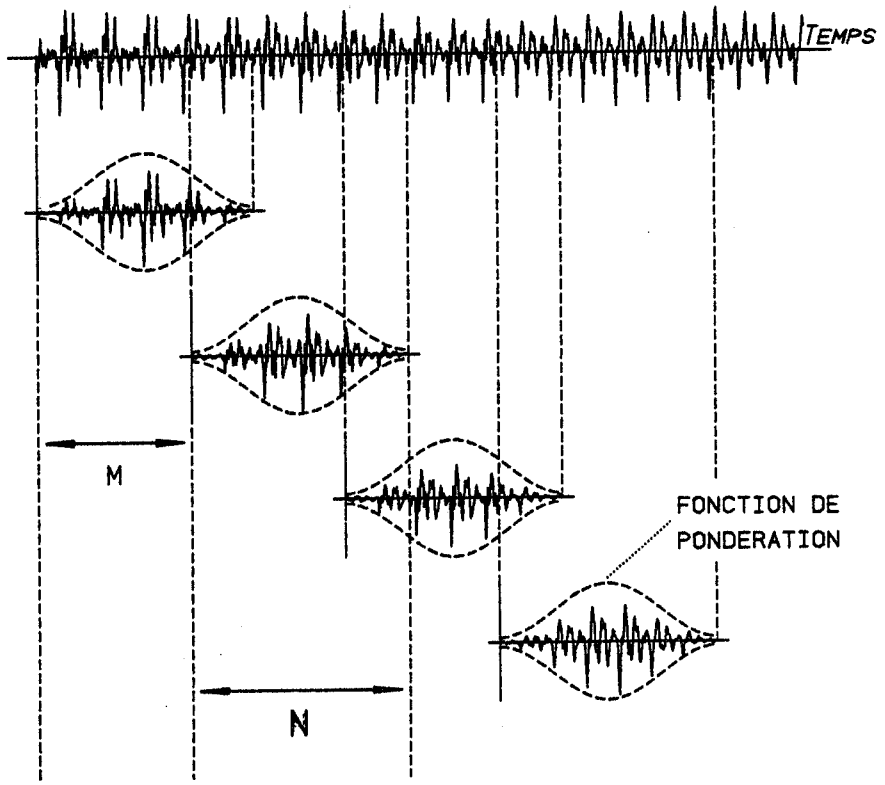


Fig. 2
Catégorisation des principes de détermination du fondamental

à terme court. En consequênce, on a les deux catégories: (a) détermination du fondamental dans le domaine temporel, et (b) détermination du fondamental par analyse à terme court (fig.2). Cette catégorisation sera discutée en detail dans les paragraphes suivants.

3. Détermination du fondamental par analyse à terme court.

3.1 Le principe d'analyse à terme court. - En employant ce principe, l'analyseur exécute une transformation à terme court comme partie de l'étage de prétraitement. Pour ce faire, le signal parole est divisé en des intervalles à terme court successifs (fig.3). On obtient l'intervalle à terme court en prenant un nombre limité d'échantillons consécutifs du signal à partir d'un point de départ, K, jusqu'à un point d'arrêt, (K+N-1). La longueur de cet intervalle, N, est choisie tellement que, d'un part, le parametre à mesurer peut être supposé comme suffisamment constant pendant l'intervalle. D'autre part, la longueur N doit être suffisamment grande afin de garantir que le parametre reste mesurable. Pour la plupart des analyseurs à terme court, cet intervalle donc contient au moins deux ou trois périodes fondamentales complètes. Dans des cas particuliers, quand FO se change brusquement, ou quand le signal est irrégulier, ces deux



M Intervalle d'échantillonage du parametre

N Longueur de la fenêtre d'intervalle à terme court

Fig. 3 - Principe de l'analyse à terme court.

conditions sont contradictoires, ce qui est une source d'erreurs additionnelle. La distance entre deux intervalles consécutifs, M, est fixée tellement que chaque changement signifiant du parametre soit documenté dans les résultats du mesurage. De même que le signal original, la sequénce des valeurs estimatives du parametre mesuré constitue les échantillons d'un signal échantillonné qui obéit à la loi d'échantillonage, bien que la fréquence d'échantillonage du parametre, comparée à celle du signal, puisse être beaucoup plus basse. En conséquence, le mesurage d'un parametre par analyse à terme court est un cas spécial de sous-échantillonage.

Rentrons au problème de détermination du fondamental. Ici, on demande de la transformation à terme court qu'elle se comporte, pour ainsi dire, comme un miroir concave qui focalise l'information sur le fondamental, distribuée sur tout l'intervalle de mesure dans le signal original, en une seule valeur extrême (maximum ou minimum) dans le domaine spectral. Ce pic est puis detecté par un détecteur de pointe, ce qui est la réalisation usuelle de l'extracteur dans ce type d'analyseur. En conséquence, le signal de sortie d'un tel analyseur est donné par une séquence de valeurs estimatives moyennes du fondamental et pas par le tracé du signal période par période. Au cours de la transformation les relations de phase entre le domaine temporel et le domaine spectral sont perdues; en même temps, toutefois, l'algorithme perd sa sensibilité à des déphasages et autres distorsions du signal parole; en particulier, l'algorithme devient insensible aux affaiblissements des fréquences basses comme ils arrivent dans les transmissions téléphoniques.

Malheureusement, la fiabilité augmentée doit être payée par une dépense de calcul plus grande(au moins d'un ordre de magnitude par rapport à un analyseur en domaine temporel). La plupart de la dépense de calcul est causée par l'exécution numérique de la transformation à terme court. Outre la recherche de fiabilité, la recherche des calculs rapides et efficaces est donc toujours un point important. Cependant, le developpement des technologies nouvelles, qui emporte des circuits de calcul de plus en plus complexes et rapides aussi que des dispositifs en hardware de moins en moins chers, est en train de diminuer la signifiance de ce dernier aspect.

3.2 Les algorithmes individuels. - Pas tous les transformations spectrales connues présentent l'effet d'établir un point dans le domaine spectral où toute l'information sur le fondamental sera concentrée. Celles qui le font sont les suivantes (fig.4): méthodes de corrélation, méthodes qui opèrent en domaine fréquentiel, et l'algorithme des moindres carrés dit "Maximum likelihood". Parmi les méthodes de corrélation, on trouve l'algorithme bien connu de l'autocorrélation qui devenait fiable quand il était combiné avec l'écrêtement central (Sondhi 1968, Rabiner 1977). Le calcul de la fonction d'autocorrélation est considérablement simplifié quand on exécute un écrêtement des pics additionnel. Si l'on ne retient comme information qu'un signal qui peut prendre les trois valeurs +1, 0, et -1, on évite tous les multiplications. Tels dispositifs peuvent opérer en temps réel (Dubnowski et al. 1976, Galand et al. 1976).

Le pendant à la fonction de l'autocorrélation est fait par un algorithme, pour ainsi dire, "d'anticorrélation". On calcule la fonction de la valeur absolue moyenne de difference, dit "AMDF" (average magnitude difference function)

$$\mathrm{AMDF}(d) := \sum_{n} | x(n) - x(n+d) | \qquad (2)$$

Lorsque le décalage, d, est égal a la durée TO de la période fondamentale, l'AMDF revèle un minimum distingué. L'AMDF, n'ayant pas besoin de multiplications, s'exécute assez vite. De plus, la definition de l'AMDF ne réquiert pas de stationarité; l'intervalle de mesure donc peut être très bref, c'est-à-dire une période on même moins (Ross et al. 1974, Moorer 1974). Un et Yang (1977) ont appliqué ce principe avec succès au signal résiduel de la prédiction linéaire.

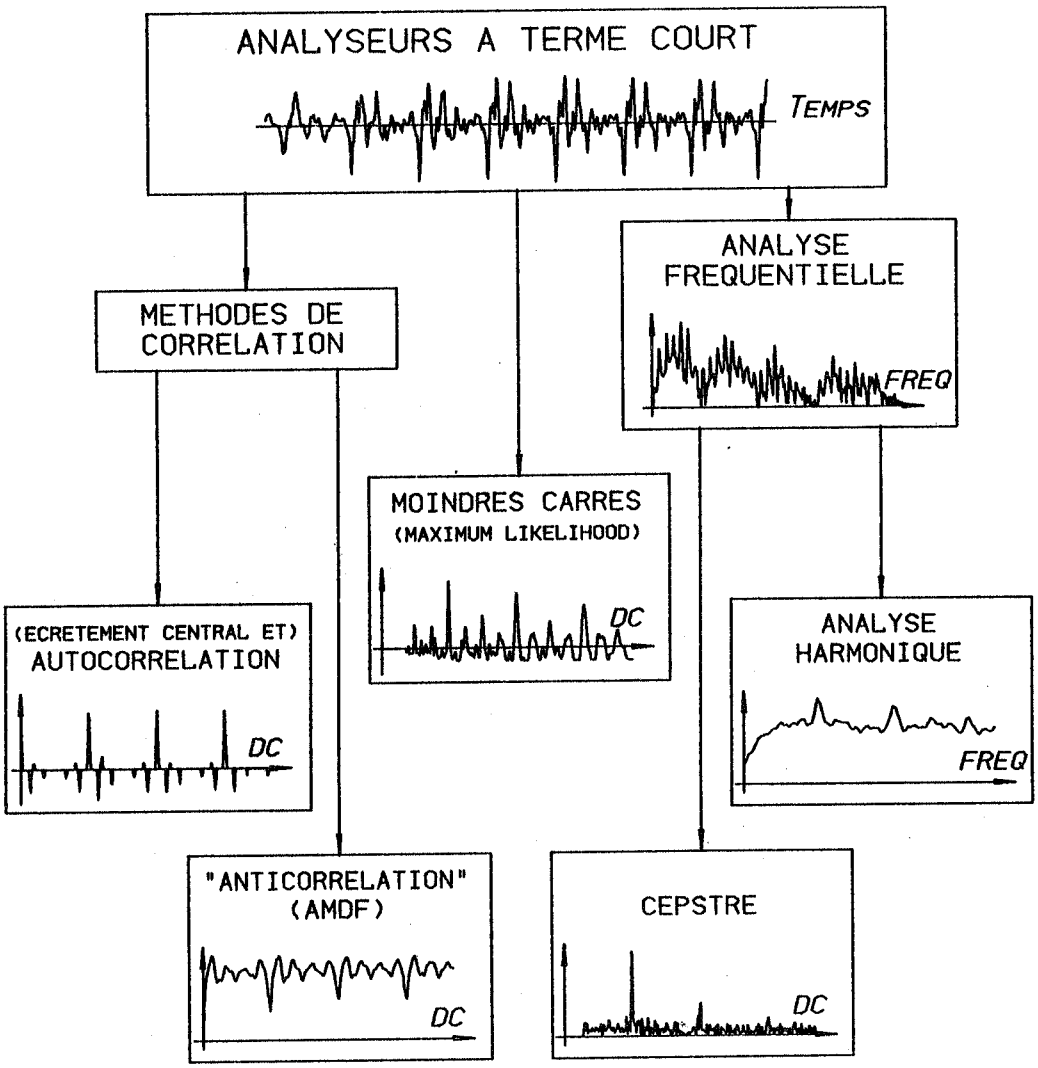


Fig. 4 - Méthodes de détermination du fondamental à terme court

Les méthodes qui opèrent dans le domaine fréquentiel peuvent être divisées en deux catégories. Il n'est pas fiable de déterminer FO comme la location du premier maximum dans le spectre d'amplitude. On donc préfère examiner la structure harmonique du spectre. Une possibilité de ce faire se fond sur la compression spectrale. Le spectre d'amplitude (ou de puissance) est comprimé le long de l'axe fréquentielle par des facteurs de 2, 3 etc. et additionné au spectre primordial. On obtient un maximum fort pour FO ce qui résulte de la contribution cohérente des harmoniques élevées après la compression (Schroeder 1968, Noll 1970, Terhardt 1979). Miller (1970) a construit un dispositif analogique qui applique le même principe en domaine temporel. Un calcul différent qui prend une décision de majorité selon les distances des pics spectraux voisins a été proposé par Seneff (1978). La deuxième méthode d'analyse fréquentielle retourne en domaine temporel. Au lieu de transformer le spectre de puissance (ce qui conduirait à la fonction d'autocorrélation), on transforme le logarithme de spectre de puissance. Ce procédé conduit au "cepstre" bien connu, dont la variable indépendente dans le domaine temporel, pour raison d'uniformisation, est appelé décalage comme la variable indépendente de la fonction d'autocorrélation[2]. Le cepstre présente un maximum distingué qui indique la période fondamentale TO.

Enfin, on doit mentionner la méthode des moindres carrés (dit "maximum likelihood").

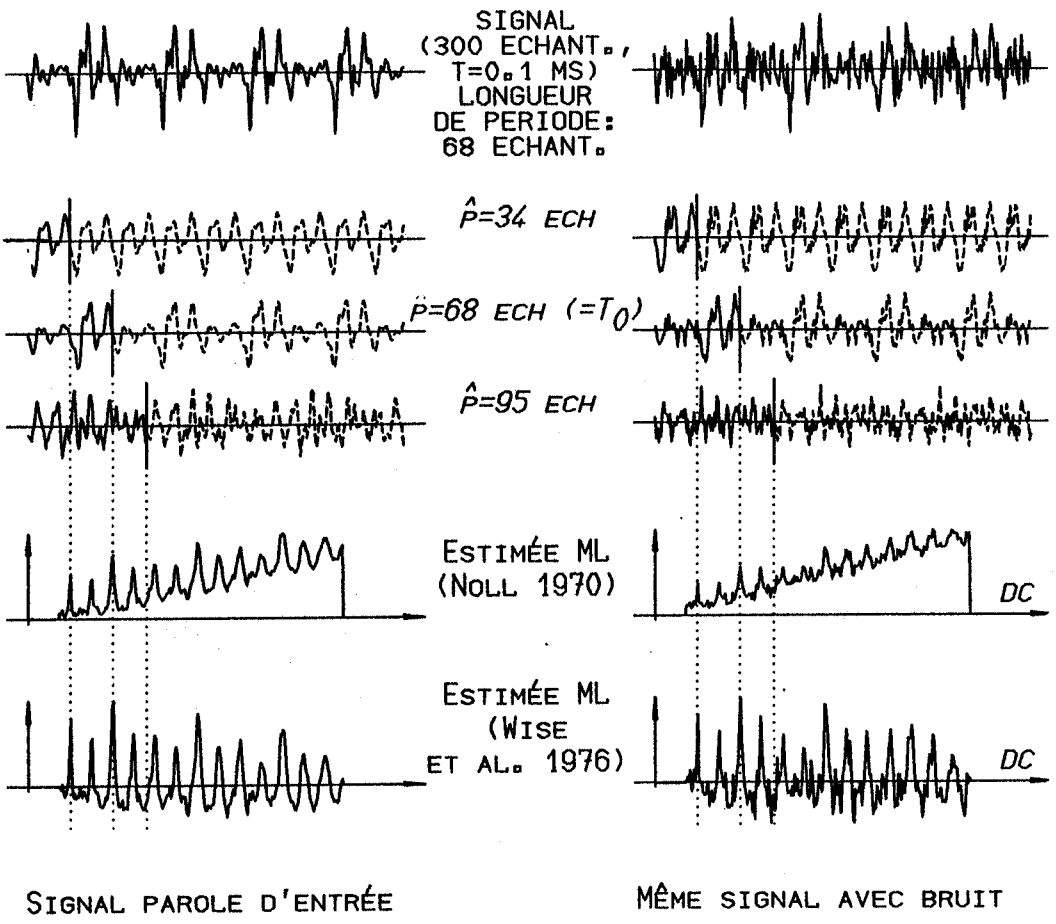


Fig. 5
Exemple de mode d'action d'un analyseur selon le principe des moindres carrés ("Maximum Likelihood")

ech échantillons
ML Maximum Likelihood

C'est originairement un calcul statistique afin de séparer un signal périodique d'une période inconnue d'un signal de bruit à distribution de Gauss (Noll 1970). Cette méthode doit être modifiée car ni le signal parole est exactement périodique ni le bruit (qui aussi contient les parties non periodiques du signal) obéit exactement à la distribution de Gauss (Wise et al. 1976, Friedman 1977, 1978). Même que le principe de compression spectrale, cet analyseur est extrêmement insensible au bruit de fond. Il continue à fonctionner fiablement même quand le rapport signal/bruit prend une valeur de 0 dB (Noll 1970, Irwin 1979, voir fig.5). Avec un tel bruit de fond, la parole elle-même devient inintelligible.

En sommaire, les analyseurs à terme court fournissent une séquence de valeurs estimatives moyennes de TO ou FO. Ils ne sont pas très sensibles au bruit de fond ni à des déphasages ni à l'absence de l'harmonique fondamentale. D'autre

---

2) L'auteur est d'avis que les termes américains concernant le cepstre (comme "quefrency", "liftering" etc.) ne sont pas transférables en autres langues.

part, la dépense de calcul a une tendance d'être assez grand, ce qui peut causer de problèmes quand on envisage une application en temps réel.

4. Détermination du fondamental dans le domaine temporel.

Les analyseurs de cette catégorie sont moins homogènes que les analyseurs à terme court. Une possibilité de les subdiviser encore une fois se fond sur le partage de la reduction d'information entre l'étage de prétraitement et l'extracteur. En ce faisant, on arrive à deux extrêmes entre lequels se trouve la plupart des analyseurs (fig.6):

(1) Toute la réduction d'information est exécutée par l'étage de prétraitement. Dans l'extrême cas, le signal d'entrée de l'extracteur ne consiste que dans la première partielle.

(2) Le réduction d'information est faite exclusivement par l'extracteur. Dans ce cas, l'extracteur doit venit à bout de toute la complexité et la variabilité des structures temporelles du signal parole. Dans l'extrême cas, l'étage de prétraitement peut être omis.

En principe, l'analyseur en domaine temporel est capable de tracer le signal période par période. A la sortie de l'extracteur, on trouve une sequence de points de repère appartenants aux débuts ou à des autres points significatifs des périodes individuelles. Comme les algorithmes prennent l'information locale individuellement de chaque période, ils sont en général plus sensibles à des

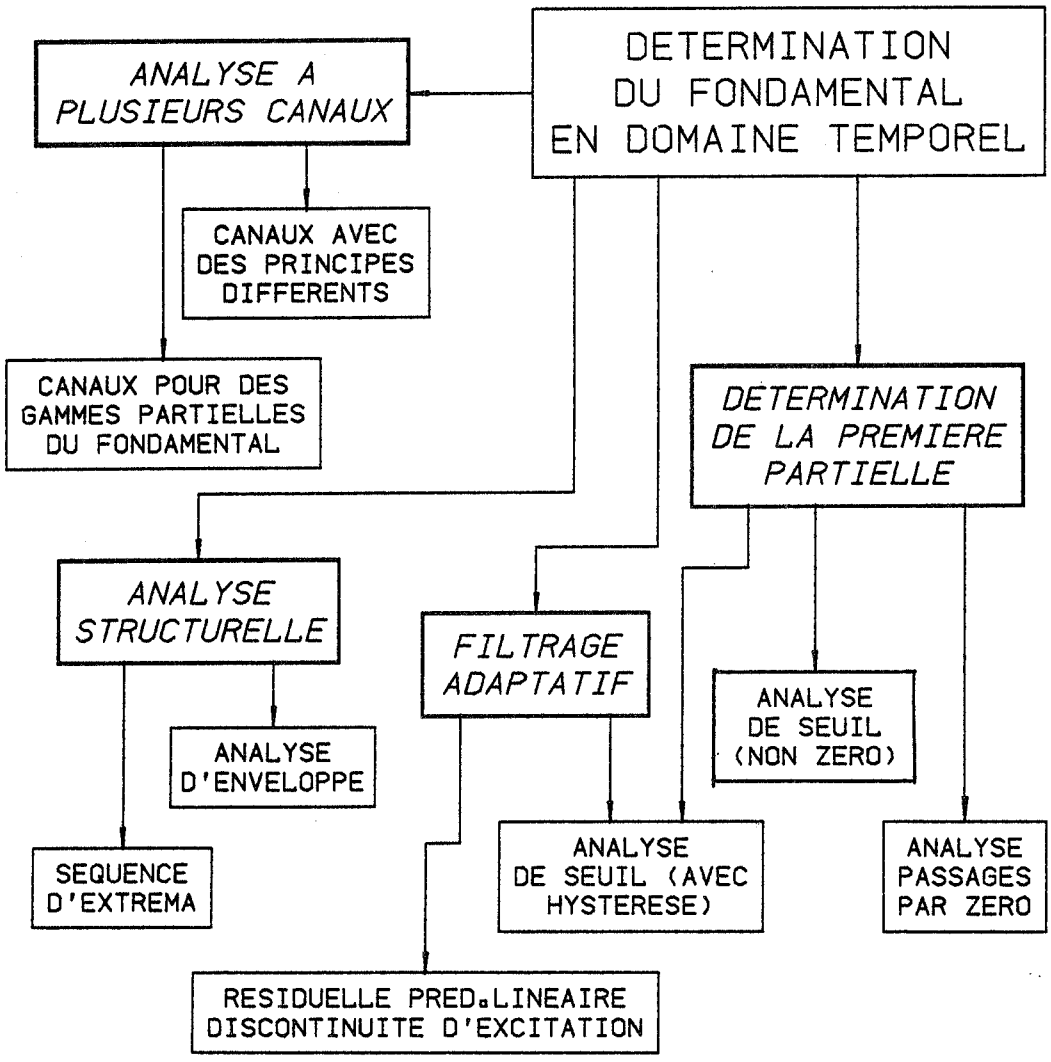


Fig. 6 - Principes de détermination du fondamental en domaine temporel

distorsions locales et donc moins fiables que leurs pendants à terme court. D'autre part, les analyseurs en domaine temporel sont capables - au moins en principe - de fournir des résultats corrects quand le signal apparaît irrégulier dû à une perturbation temporaire de la voix.

4.1 Analyse structurelle. - La période fondamentale est la réponse tronquée du conduit vocal à une impulsion laryngienne individuelle. Parce que le conduit vocal se comporte comme un système linéaire avec pertes, cette réponse se compose de plusieurs oscillations sinusoidales amorties par une fonction exponentielle. Il est donc à supposer que les extrema signifiants prennent les valeurs les plus grandes au début de la periode et pas vers la fin (fig.7), c'est-à-dire, l'inspection appropriée des pics du signal (maxima et/ou minima) conduira à des indices de la périodicité.

Cependant, cette méthode pose de problèmes. Premièrement, les fréquences des oscillations dominantes sont definies par les valeurs instantanées des formants; elles peuvent se changer brusquement. Deuxièmement, l'amortissement du premier formant (quand il se trouve à fréquence basse) est souvent faible; il peut être compensé et dépassé par un changement général du niveau du signal. Troisièmement, si le signal a été distordu de phase, il peut arriver que l'excitation des formants divers se passe à des instants différents. On peut surmonter les problèmes, mais ils conduisent vers des solutions algorithmiques fort compliquées. Comme la plupart des instructions se compose de décisions, ces algorithmes n'ont pas besoin d'une grande dépense de calcul malgré leur complexité. Usuellement, ces algorithmes sont exécutés comme suit (Reddy 1967, Miller 1975, Baudry et Dupeyrat 1976):

(1) Le signal subit un filtrage passe-bas moderé afin d'affaiblir l'influence des hauts formants;
(2) on détermine tous les maxima et minima du signal;

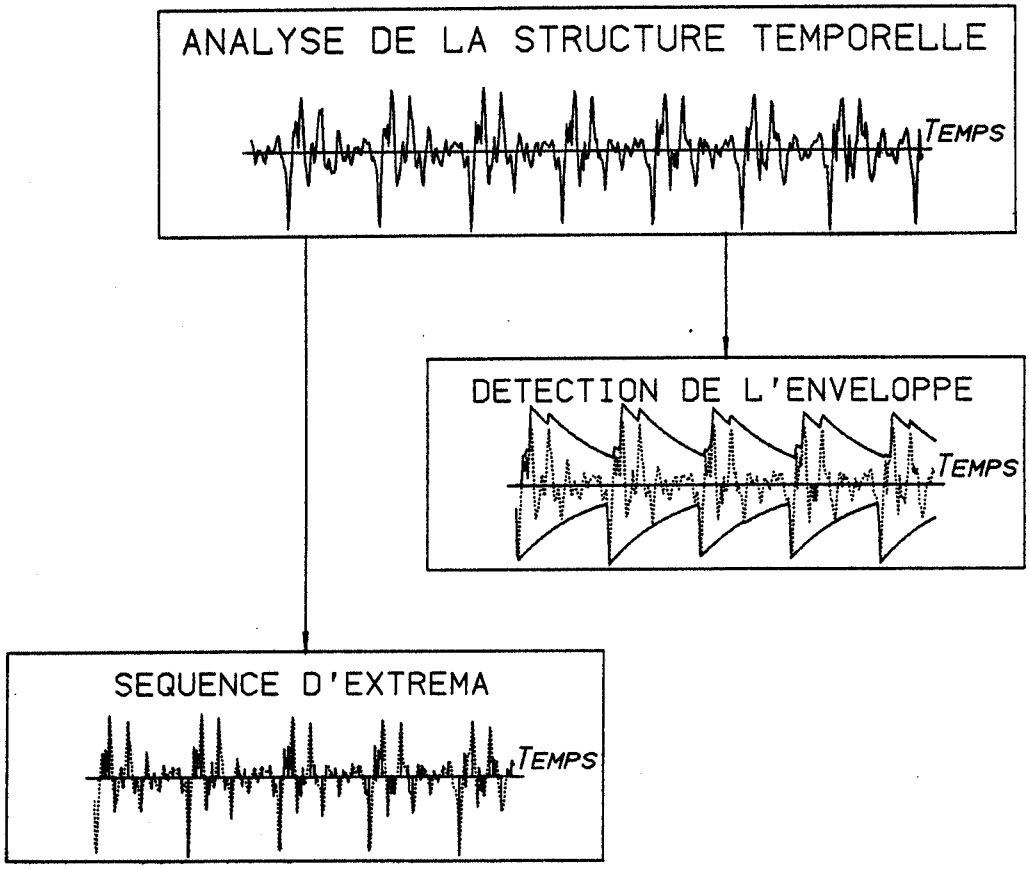


Fig. 7 - Analyse en domaine temporel: examination de la structure temporelle

(3) on exclut ces pics qui se trouvent être insignifiants jusqu'à ce qu'on retienne un seul point de repère par période;
(4) une routine de correction rejette tous les points qui sont evidemment faux.

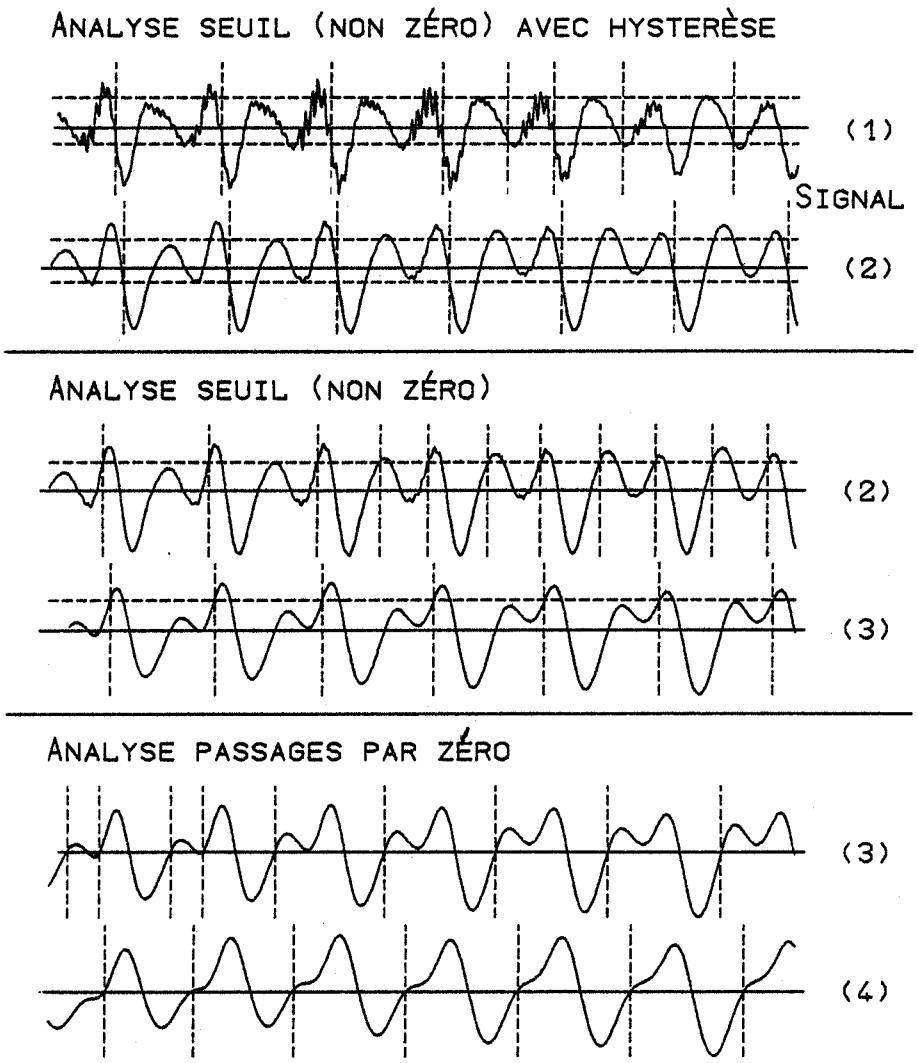
Pour des applications restreintes, ces analyseurs peuvent être considérablement simplfiés (Gold et Rabiner 1969, Baudry et Dupeyrat 1976, Tillmann 1978).

Une solution différente a été proposé en domaine analogique (Dolansky 1955, Winckel 1964, Filip 1969, Zurcher 1972). L'enveloppe de la période est modelée par un circuit diode-resisteur-capaciteur avec une constante de temps de montée très faible et une constante de temps de déscente assez longue. Ce circuit détecte les crêtes du signal: il accentue la pointe principale de la période et supprime tous les autres. Malheureusement, la performance depend fortement de l'ajustage correcte de la constante de temps de déscente; de plus, cet analyseur souffre d'interactions entre l'amortissement local des formants et les changements globaux du niveau du signal. Au moins dans cette réalisation simple, ce dispositif ne fonctionne bien que pour des gammes de fréquence fondamentale restreintes (deux octaves environ).

Le dispositif de Zurcher (1972 et 1977) évite ces problemes en partie. D'abord, il emploie un compresseur de dynamique qui normalise le niveau du signal (ce compresseur opérant avec plusieurs constantes de temps simultanément, son emploi ne change pas le rapport de hauteur des crêtes diverses au cours d'une période individuelle, ce qui autrement compromettrait la performance de la méthode), puis le signal subit un filtrage passe-bas. Le détecteur de crêtes emploie un circuit de rétroaction qui réapplique le signal de sortie à l'entrée avec un certain décalage. A cause de cela les pics non signifiants sont supprimés davantage. Cependant, dû au filtrage passe-bas, cet analyseur présente une certaine sensibilité à l'affaiblissement de la première partielle par filtrage passe-haut, comme il se passe dans la chaîne telephonique.

4.2 Isolation et détection de la première partielle. - On peut détecter et mesurer TO par l'oscillation temporelle de la première partielle. Celle-ci est extraite du signal par un filtrage passe-bas extensif dans l'étage de prétraitement. Fig.8 présente les principes de trois extracteurs: (1) détecteur de passages par zéro, (2) détecteur de passages par une ligne de seuil différénte de zéro, et (3) détecteur de passages par une ligne de seuil avec hysterèse. Le premier détecteur réquiert que le signal aie deux et seulement deux passages par zéro au cours d'une période, ce qui demande un filtrage passe-bas très fort (McKinney 1965). Le détecteur de passages par seuil avec hysterèse note un point de repère aussitôt que le signal excède le seuil; cependant, ce point n'est pas vérifié avant que le signal passe un deuxième seuil en sens inverse. Cet extracteur plus élaboré ne réquiert comme prétraitement qu'un filtrage passe-bas faible.

L'exigence d'un filtrage passe-bas extensif est l'une de deux déficiences sévères de ce principe. Le détecteur de passages par zéro demande un filtre passe-bas avec un affaiblissement des hautes fréquences d'au moins 18 dB par octave au dedans de bande de mesure de la fréquence fondamentale (McKinney 1965, Erb 1974). En conséquence, si la bande de mesure est envisagé de s'étendre sur toute la gamme des fréquences fondamentales, l'extracteur doit venir à bout avec variations du niveau de plus que 50 dB dues à la variation de FO seulement. Cette plage de niveau, (au moins 30 dB au-delà) est trop grande pour que l'analyseur puisse en sortir. Avec ce détecteur primitif, on ne peut donc analyser qu'une bande de fréquences fondamentales restreinte. Pour le détecteur de passages par seuil, ce problème n'est pas aussi grave; cependant, la necessité d'adapter le seuil à l'amplitude globale du signal complique la réalisation d'un tel analyseur. - La deuxième déficience s'ensuit de ce que l'algorithme demande la présence de première partielle. Limitation de la bande vers les fréquences basses fait faillir systématiquement cet analyseur.



Signaux:

(1) après filtrage passe-bas faible 6 dB par octave

(2) après filtrage passe-bas modéré 12 dB par octave

(3) après filtrage passe-bas fort 18 dB par octave

Fig. 8
Exemple de mode d'action d'extracteurs divers pour l'extraction de la première partielle

4.3 Simplification de la structure temporelle. - Les algorithmes de ce principe prennent une position intermédiaire entre les principes d'analyse structurelle et l'extraction de la première partielle. La plupart de ces algorithmes suit deux méthodes de procédé: (1) filtrage inverse, et (2) détection de discontinuités dans la fonction laryngienne ("epoch detection"). Tous les deux méthodes se fondent sur le fait que la structure temporelle du signal excitateur laryngien est beaucoup plus simple et régulier que celle du signal parole. Le filtre inverse a une fonction de transfert réciproque a celle du conduit vocal; il donc reconstruit le signal glottique. Si l'on n'est interessé qu'à l'extraction du fondamental, une approximation crue du filtre inverse (dont le calculation demanderait autrement la détermination de formants complète) va suffire. Une telle approximation est realisée, par exemple, en utilisant un filtre passe-bas fixe pour enlever les hautes formants et un filtre coupe-bande variable en temps afin de supprimer le premier formant. On peut mesurer le premier formant en domaine temporel et prendre un filtre peigne (dont l'implementation est très simple) afin d'affaiblir le formant (Hess 1976). Maissis (1972) et les algorithmes successeurs qui se fondent sur ce principe, affaiblissent le premier formant par un filtre à prédiction linéaire de deuxième ordre.

En cours de chaque cycle fondamental, les cordes vocales se ferment brusquement, ce qui cause une discontinuité de la dérivée du signal glottique. L'instant de

cette discontinuité se laisse détecter sans filtrage inverse, mais à condition que le spectre analysé soit rélativement plain. Dû a des déphasages du signal dans le conduit vocal, on ne peut pas dire si l'instant de discontinuité de la fonction excitatrice coincide avec une valeur extrême ou avec un passage par zéro. Afin d'enlever cette ambiguite, plusieurs algorithmes donc emploient une transformation de Hilbert, ce qui construit le signal avec un déphasage constant de 90°. La racine de la somme des carrés du signal original et de la transformée de Hilbert donne l'amplitude instantanée du signal. Après une ponderation spectrale appropriée, cette valeur indique le début de chaque période, c'est-à-dire l'instant de la fermeture de la glotte.

Comme déjà dit, cet procédé réquiert que le spectre du signal soit plein au dedans de la bande analysée. On a donc la possibilité d'examiner le signal aux hautes fréquences (Ananthapadmanabha et Yegnanarayana 1975), le signal résiduel de la prédiction linéaire (les mêmes auteurs, 1979), ou d'appliquer un banc de filtres afin d'égaliser le spectre (Smith 1954).

Le principe du filtrage inverse a un point dangereux qui prend effet quand le premier formant coincide avec la fréquence fondamentale. Dans ce cas, qui arrive fréquemment à des voix féminines, la première partielle prâtiquement n'est pas amortie, et le filtrage inverse, en éliminant cette partielle, enlève l'oscillation du fondamental et fait faillir l'analyseur. Afin d'éviter cette situation, on doit bloquer le filtre inverse aussitôt qu'elle est reconnue. La même chose se met en évidence pour le principe de détection de discontinuité de la fonction excitatrice. Le premier formant coincidant avec FO, presque toute l'énergie du signal s'y concentre. (Hess 1979). En calculant la somme des carrés du signal et de sa transformée de Hilbert, on obtient pour la première partielle

$$(x(n))^2 = (A \cos an)^2 + (A \sin an)^2 = A^2 = \text{const.} \qquad (3)$$

Dans ce cas-là, mais seulement dans ce cas, on peut aussi mesurer FO directement par la derivée de la fonction de phase instantanée (Berthomier 1979).

4.4 Restitution et intensification de la première partielle. - Quand on veut appliquer le principe de l'extraction de la première partielle à des signaux limités de bande vers les fréquences basses, on doit restituer ou amplifier cette partielle. Ça peut être effectué par une distorsion non linéaire du signal. Par rapport à cela, beaucoup des proposition ont été faites dans la littérature (par exemple, Grützmacher et Lottermoser 1937, Risberg 1960, McKinney 1965). Cependant, on ne trouve pas d'opinion uniforme à l'égard de la question quelle fonction non linéare soit la meilleure pour cette tâche. Afin d'éclaircir ce point, l'étude suivante a été faite (Hess 1979): 48 périodes fondamentales isolées issues de trois locuteurs (une féminine, deux masculins) représentant la plupart des voyelles et consonnes soutenues de la langue allemande ont été distordues par plusieurs fonctions non linéaires: des fonctions impaires, fonctions paires (par exemple, redressement double), fonctions "mixtes" (ni pair ni impair, comme le redressement demi-onde, par exemple) et, enfin, le signal modulé à bande latérale unique (et puis redressé). L'effet de ces fonctions a été décrit à l'aide de l'amplitude relative de la première harmonique

$$AR1: = 20 \log \frac{A(1)}{\max\,(A(i))} \quad ; \quad i = 2, 3, \ldots\ldots \text{ pour des fréquences au-dessous de 1200 Hz} \qquad (4)$$

Dans cette équation, A(k) représente l'amplitude de la k-ème harmonique. Afin d'évaluer la valeur de AR1, les harmoniques élevées ont été comptés jusqu'à des fréquences de 1200 Hz. Fig.9 présente l'histogramme de ce parametre pour les signaux non distordus. On observe une discrépance considérable entre les données pour les voix féminine et masculines.

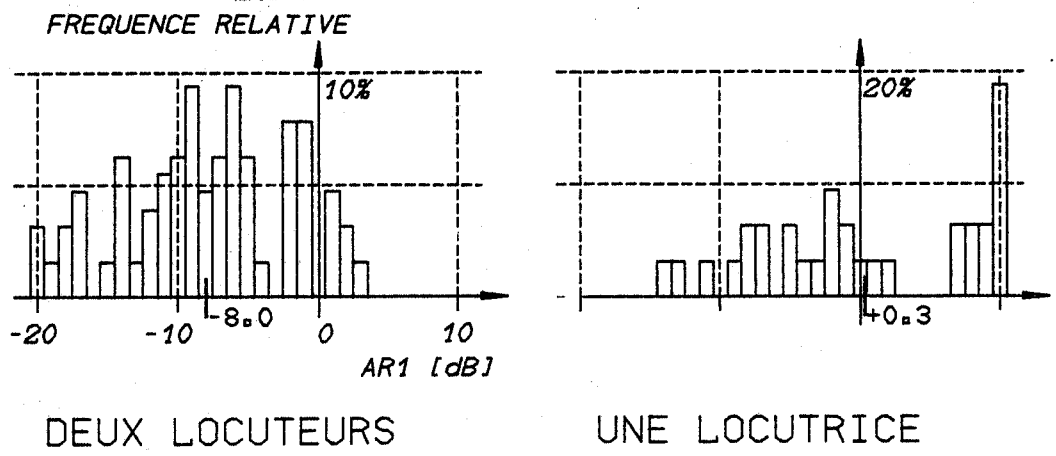


Fig. 9
Distribution de l'amplitude relative de la première partielle, AR1 (selon eq. 5), pour des signaux non distordus.

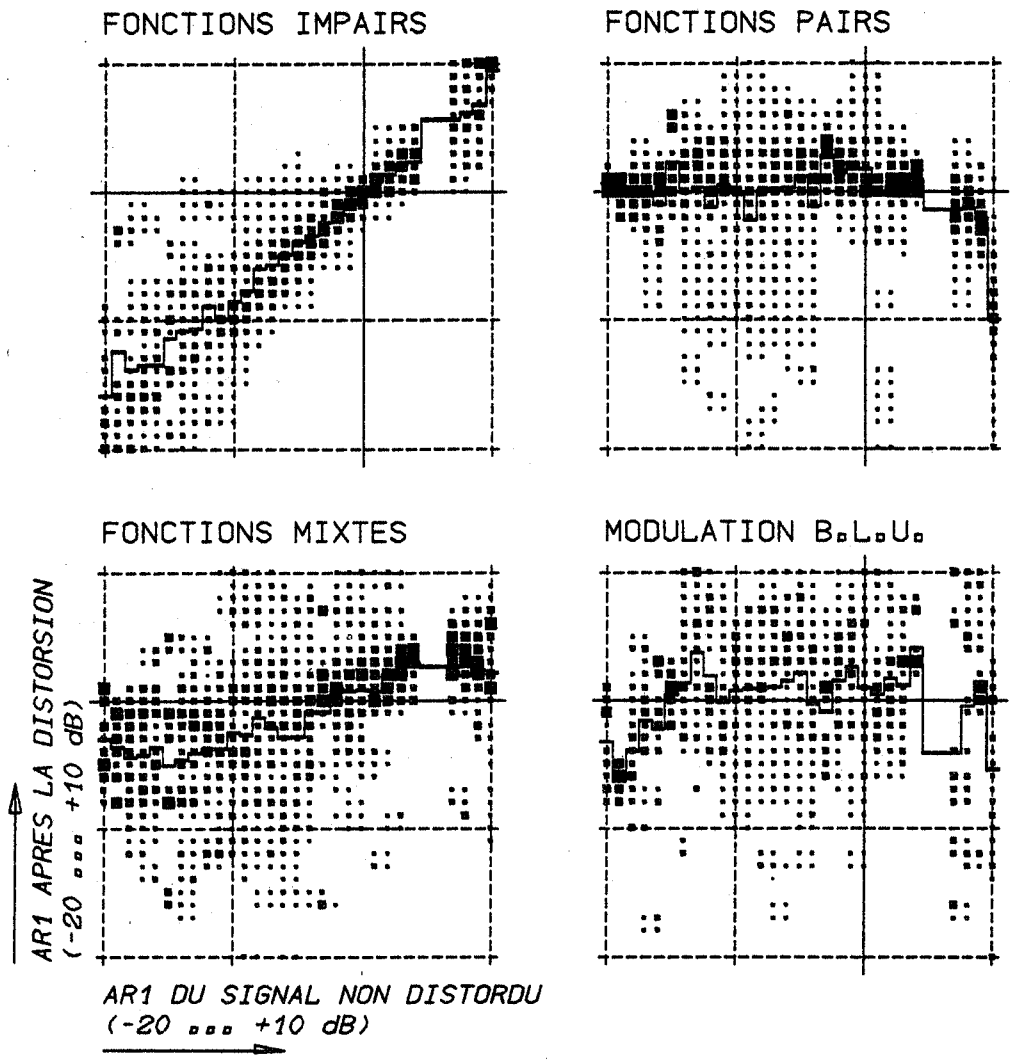


Fig. 10
L'effet de la distorsion non linéaire sur l'amplitude relative de la première partielle, AR1.

Les résultats finals de cette étude sont dépeints dans fig.10. En général, les fonctions impaires n'effectuent pas beaucoup. Les fonctions paires et le signal de sortie de la modulation B.L.U. (voir ci-dessus) intensifient une première harmonique primordialement faible ou manquante, mais elles affaiblissent la première partielle quand elle est dominante. Les fonctions mixtes ne sont pas optimales. On peut donc constater qu'aucune fonction non linéaire ne se comporte optimalement pour toute sorte de signaux d'entrée.

4.5 Dispositifs à plusieurs canaux. - A l'exception d'analyse structurelle, la plupart des algorithmes en domaine temporel est restreinte à l'égard de la plage de FO ou de la sorte de signal d'entrée. Une possibilité de surmonter ces restrictions consiste à realiser plusieurs analyseurs ou plusieurs canaux de prétraitement en parallel et à selectionner puis ce canal qui conduise l'information correcte. Les analyseurs individuels peuvent être identiques; dans ce cas, chacun traite une partie de l'éventail des fréquences fondamentales (Léon et Martin 1969, Dibbern 1972). D'autre part, ils peuvent employer des principes différents pour toute la gamme des fréquences, et selectionner ce canal qui contient l'information propre avec un maximum de probabilité. L'analyseur de Fant (1962, aussi décrit dans Risberg 1960 et ailleurs), par exemple, emploie deux, l'algorithme de Hess (1979) trois fonctions non linéaires afin d'intensifier la première partielle. L'algorithme de DeMori et al. (1977), qui se fond sur le principe de détection de discontinuité (voir section 4.3), prend cinq canaux à bande étroite centrés autour de cinq fréquences différentes (entre 1.2 et 3 kHz). Les critères de sélection entre les canaux sont établis par une certaine hierarchie (Léon et Martin 1969), par contrôle de regularité (Hess 1979), ou par des règles syntaxiques (DeMori et al. 1977). La sélection étant toujours contrôlée, l'analyseur peut changer le canal sélectionné tout le temps.

En sommaire, les analyseurs en domaine temporel ont une tendance à être plus simples et plus vites que les analyseurs à terme court. D'autre part, ils sont moins fiables.Les algorithmes d'analyse structurelle sont sensibles à de déphasages et de signaux parasites brefs, comme, par exemple, de bruits de commutation. Les analyseurs simples souvent ne peuvent qu'analyser une gamme de FO restreinte, ou ils faillissent pour certains types de signaux. Une possibilité de surmonter ce dilemme est d'utiliser un analyseur à plusieurs canaux.

5. Evaluation et application.

Afin d'évaluer et juger le fonctionnement d'un dispositif de mesurage, on devrait avoir à sa disposition un autre instrument avec (au moins) la même précision. Si c'est impossible, on a besoin de critères objectifs afin d'étalonner l'instrument nouveau. Pour la détermination du fondamental, tous ces deux bases de comparaison sont difficiles à vérifier. Aucun analyseur ne fonctionne sans erreurs. De plus, on trouve une telle variété de signaux et de conditions d'environnement, qu'on ne peut guère tester tous les situations possibles en avance. Il est donc à comprendre que les chercheurs, en projetant les algorithmes, ont rarement fourni des données quantitatives sur leur fonctionnement.

Une des études rares sur ce sujet a été faite par Rabiner et al. (1976) et McGonegal et al. (1977) en rapport avec un système de communication parlé. Sept analyseurs (deux en domaine temporel, cinq à terme court) ont été examinés. Les résultats essentiels sont résumés comme suit:

(1) Aucun analyseur ne fonctionne sans erreurs, même si les conditions d'environnement sont bonnes. Chaque analyseur possède ses erreurs "favorites"; néanmoins, une erreur de sorte quelconque se passe chez un analyseur quelconque.

(2) On n'a pas réussi à faire accorder les résultats de l'évaluation objective avec ceux de l'évaluation subjective. Aucune des erreurs mesurables (nombre des erreurs grosses, des petites imprécisions et des erreurs en cours de la détermination du voisement) n'est bien correlée avec les résultats subjectifs. En conséquence, la question quelles erreurs sont les plus ennuyeu-

ses pour l'ouie humaine est encore pendante.

Le champ de la communication parlée (analyse, synthèse, transmission) est un domaine d'application important pour les analyseurs du fondamental. Néanmoins, il y a des autres domaines: (2) la phonétique et linguistique expérimentales, y compris la musicologie, c'est-à-dire le mesurage instrumental des courbes de mélodie comme porteurs d'information prosodique, phonétique, et musicale; (3) éducation: des aides pour les handicappés aussi que des aides pour l'enseignement de langues étrangères; et (4) l'application dans la phoniatrie. Chacune de ces applications a ses exigences intrinsèques aux détecteurs du fondamental. Tableau 1 essaie de compiler et ponderer quelques specifications vues de l'avis de leurs applications respectives. Pour tous les applications, avec l'exception de la communication parlée, la visualisation est un point très important. Pour la phonétique, la fonctionnement en temps réel est insignifiant car le phonéticien peut attendre les résultats (et faire autre chose pendant ce temps), mais indispensable, par exemple, pour l'étudiant d'une langue étrangère. Ce dernier, par contre, n'a pas besoin de résultats fort exacts. Néanmoins, les exigences les plus rigoureuses sont posées dans le domaine de la communication parlée et particulièrement par l'application comme partie d'un vocodeur. Là, l'analyse est faite automatiquement, et l'ouie humaine est le juge le plus critique. En conséquence, on a developpé la plupart des analyseurs pour cette application, et la plupart des évaluations et résumés (y compris celui-ci) a été faite pour cela. Par rapport à cette application, il y a encore de questions pendantes: il n'y a aucun principe qui fonctionne sans erreurs dans toute situation possible. Il semble difficile à croire qu'un principe nouveau et révolutionnaire puisse résoudre le problème complêtement. La combinaison de plusieurs principes à comportement complémentaire, ensemble avec une routine à contrôle et sélection

	CMP	*PHO*	*EDU*	*PAT*
Précision	*2*	*1*	*5*	*3*
Visualisation	*5*	*1*	*1*	*1*
Gamme du fondamental	*2*	*4*	*4*	*2*
Sensibilité à des dégradations du signal	*1*	*3*	*5*	*5*
Necessité de préselection	*1*	*3*	*5*	*5*
Dépenses d'implementation	*3*	*4*	*1*	*5*
Opération en temps réel	*1-5*	*4*	*1*	*5*
Facilité de manoeuvre	*5*	*2*	*2*	*3*

Tableau 1
Aperçu d'exigences des applications respectives aux analyseurs du fondamental.

CMP Communication parlée (vocodeurs etc.)
PHO Phonétique, linguistique, musicologie
EDU Education
PAT Pathologie

1 très important
2 important
3 intéressant
4 marginal
5 insignifiant

L'importance de l'opération en temps réel dans l'application de la communication parlée varie avec les tâches respectives. Pour un vocodeur, par example, l'opération en temps réel est indispensable, pour l'établissement de données de référence (dans un systeme de synthèse par règles ou de reconnaissance), au contraire, elle est insignifiiante.

L'application "pathologie" n'inclut pas une application récente: l'implantation cochléaire afin de fournir de patients sourds avec information sur le fondamental (Fourcin et al. 1978). Les exigences de cette application sont analogues a celles de l'application de la communication parlée.

élaborée, pourrait représenter une conception nouvelle qui permettrait la recherche, en domaine de la détection du fondamental, d'avancer un pas vers la solution générale.

Bibliographie.

ANANTHAPADMANABHA T.V., YEGNANARAYANA B. (1975): Epoch extraction of voiced speech. IEEE Transactions ASSP-23, pp. 562...569

ANANTHAPADMANABHA T.V., YEGNANARAYANA B. (1979): Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Transactions ASSP-27, pp. 309...319

BAUDRY M., DUPEYRAT B. (1976): Analyse du signal vocal - Utilisation des extrema du signal et de leurs amplitudes - détection du fondamental et recherche des formants. Journ.Etud.Parole, vol. 7, pp. 247...257

BERTHOMIER C. (1979): Calcul analogique de la fréquence du fondamental. Dans: Actes du 9me Congrès international des sciences phonétiques, Copenhague 1979, vol. 1, p. 260

DE MORI R., LAFACE P., MAKHONINE V.A., MEZZALAMA M. (1977): A syntactic procedure for the recognition of glottal pulses in continuous speech. Pattern Recognition, vol. 9, pp. 181...189

DIBBERN U. (1972): Grundfrequenzmessung bei der menschlichen Sprache. In: Akustik und Schwingungstechnik. Berlin: VDE-Verlag; pp. 345...348

DOLANSKÝ L.O. (1955): An instantaneous pitch- period indicator. Journ.Acoust. Soc.Am., vol. 27, pp. 67...72

DOLANSKÝ L.O., TJERNLUND P. (1968): On certain irregularities of voiced speech waveforms. IEEE Transactions AU-16, pp. 51...56

DUBNOWSKI J.J., SCHAFER R.W., RABINER L.R. (1976): Real-time digital hardware pitch detector. IEEE Transactions ASSP-24, pp. 2...8

ERB H.J. (1974): Ein Verfahren zur Bestimmung der Sprachgrundfrequenz in Echtzeit. Frequenz, vol. 28, pp. 23...28

FANT G. (1962): Speech analysis and synthesis: Technical (Final) Report. Stockholm: Speech Transmission Lab.; AD 298309.

FILIP M. (1969): Envelope periodicity detection. Journ.Acoust.Soc.Am., vol. 45, pp. 719...732

FLANAGAN J.L. (1972): Speech analysis, synthesis, and perception. Berlin/New York: Springer.

FLANAGAN J.L., SASLOW M.G. (1958): Pitch discrimination for synthetic vowels. Journ. Acoust.Soc.Am., vol. 30, pp. 435...442

FOURCIN A.J., ROSEN S.M., MOORE B.C.J., DOUEK E.E., CLARKE G.P., DODSON H., BANNISTER L.H. (1978): External electrical stimulation of the cochlea. In: Speech and Hearing - Work in progress. London: Dept. of Phonetics and Linguistics, University College; pp. 3...32

FRIEDMAN D.H. (1977): Pseudo-maximum-likelihood pitch extraction. IEEE Transactions ASSP-25, pp. 213...221

FRIEDMAN D.H. (1978): Multidimensional pseudo- maximum-likelihood pitch estimation. IEEE Transactions ASSP-26, pp. 185...196

FUJIMURA O. (1968): An approximation to voice aperiodicity. IEEE Transactions AU-16, pp. 68...72

GALAND C., ESTEBAN D., DUBUS F. (1976): Détection de la mélodie par autocorrélation non linéaire. Journ.Etud.Parole, vol. 7, pp. 333...345

GOLD B. (1962): Computer program for pitch extraction. Journ.Acoust.Soc.Am., vol. 34, pp. 916...921

GOLD B. (1977): Digital Speech Networks. Proceedings of the IEEE, vol. 65, pp. 1636...1658

GOLD B., RABINER L.R. (1969): Parallel processing techniques for estimating pitch periods of speech in the time domain. Journ.Acoust. Soc.Am., vol. 46, pp. 442...448

GRÜTZMACHER M., LOTTERMOSER W. (1937): Über ein Verfahren zur trägheitsfreien Aufzeichnung von Melodiekurven. Akustische Zeitschrift, vol. 2, pp. 242...248

HESS W.J. (1976): A pitch-synchronous digital feature extraction system for phonemic recognition of speech. IEEE Transactions ASSP-24, pp. 14...25

HESS W.J. (1979): Time-domain pitch period extraction of speech signals using three nonlinear digital filters. ICASSP-79, pp. 773...776

HESS W.J. (à paraître): Algorithms and devices for pitch determination of speech signals. Berlin, Heidelberg, New York: Springer.

IRWIN M.J. (1979): Periodicity estimation in the presence of noise. In: Proceedings of the 1979 Autumn Conference. Edinburgh: Institute of Acoustics; pp. 55...58

KUBZDELA H. (1976): An analogue fundamental frequency extractor. In: Speech analysis and synthesis. edited by W. Jassem. Warsaw: Inst. of Fundamental Research, Polish Academy of Sciences; vol. 4, pp. 269...279

LÊON P., MARTIN PH. (1969): Prolégomênes à l'étude des structures intonatives. Paris, Montrêal: Didier; Studia Phonetica, No. 2; 225 pages

LIEBERMAN P. (1963): Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journ.Acoust.Soc.Am., vol. 35, pp. 344...353

MAISSIS A.H. (1972): Methode d'extraction du pitch. Journ.Etud.Parole, vol. 3, pp. 310...315

MARKEL J.D. (1972): The SIFT algorithm for fundamental frequency estimation. IEEE Transactions AU-20, pp. 367-377

McGONEGAL C.A., RABINER L.R., ROSENBERG A.E. (1975): A semiautomatic pitch detector (SAPD). IEEE Transactions ASSP-23, pp. 570...574

McGONEGAL C.A., RABINER L.R., ROSENBERG A.E. (1977): A subjective evaluation of pitch detection methods using LPC synthesized speech. IEEE Transactions ASSP-25, pp. 221...229

McKINNEY N.P. (1965): Laryngeal frequency analysis for linguistic research. Ann Arbor MI: Communic.Sciences Lab., Univ. of Michigan; Res.Rept. No. 14; 340 pages

MILLER N.J. (1975): Pitch detection by data reduction. IEEE Transactions ASSP-23, pp. 72...79

MILLER R.L. (1970): Performance characteristics of an experimental harmonic identification pitch extraction system (HIPEX). Journ. Acoust.Soc.Am., vol. 47, pp. 1593...1601

MOORER J.A. (1974): The optimum comb method of pitch period analysis of continuous digitized speech. IEEE Transactions ASSP-22, pp. 330...338

MOSER M., KITTEL G. (1977): Rechnergestützte Tonhöhenbestimmung. Folia phoniatrica, vol. 29, pp. 119...126

NOLL A.M. (1967): Cepstrum pitch determination. Journ.Acoust.Soc.Am., vol. 41, pp. 293...309

NOLL A.M. (1970): Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In: Symposium on Computer Processing in Communication; edited by Microwave Institute. New York: The University of Brooklyn Press; vol. 19, pp. 779...797

RABINER L.R. (1977): On the use of autocorrelation analysis for pitch detection. IEEE Transactions ASSP-25, pp. 24...33

RABINER L.R., CHENG M.J., ROSENBERG A.E., McGONEGAL A. (1976): A comparative study of several pitch detection algorithms. IEEE Transactions ASSP-24, pp. 399...413

REDDY D.R. (1967): Pitch period determination of speech sounds. Comm.ACM, vol. 10, pp. 343...348

RISBERG A. (1960): Voice fundamental frequency tracking. STL-QPSR, No. 2

RISBERG A., MÖLLER A., FUJISAKI H. (1960): Voice Fundamental Frequency Tracking. STL-QPSR, No. 1, pp. 3...5

ROSS M.J., SHAFFER H.L., COHEN A., FREUDBERG R., MANLEY H.J. (1974): Average magnitude difference function pitch extractor. IEEE Transactions ASSP-22, pp. 353...361

SANCHEZ GONZALEZ F.J. (1977): Application of dissimilarity and a periodicity function to fundamental frequency measure of speech and voiced/unvoiced decision. ICA-9, paper I117, page 523

SCHROEDER M.R. (1968): Period histogram and product spectrum: new methods for fundamental-frequency measurement. Journ.Acoust.Soc.Am., vol. 43, pp. 829...834

SCHULTZ-COULON H.J., ARNDT H.J. (1972): Tonhöhenschreibung in der phoniatrischen Praxis. Folia phoniatrica, vol. 24, pp. 241...258

SENEFF S. (1978): Real-time harmonic pitch detector. IEEE Transactions ASSP-26, pp. 358...364

SMITH C.P. (1954): Device for extracting the excitation function from speech signals. US Patent No. 2,691,137, filed October 5, 1954

SONDHI M.M. (1968): New methods of pitch extraction. IEEE Transactions AU-16, pp. 262...266

TERHARDT E. (1979): Calculating virtual pitch. Hearing Research, vol. 1, pp. 155...182

TILLMANN H.G. (1978): Bestimmung der Stimmperiode in der Zeitfunktion des digitalen Sprachschallsignals. IPSKMünchen, vol. 9, pp. 207...213

TUCKER W.H., BATES R.T.H. (1978): A pitch estimation algorithm for speech and music. IEEE Transactions ASSP-26, pp. 597...604

UN C.K., YANG S.C. (1977): A pitch extraction algorithm based on LPC inverse filtering and AMDF. IEEE Transactions ASSP-25, pp. 565...572

WINCKEL F. (1964): Tonhöhenextraktor für Sprache mit Gleichstromanzeige. Phonetica, vol. 11, pp. 248...256

WISE J.D., CAPRIO J.R., PARKS T.W. (1976): Maximum likelihood pitch estimation. IEEE Transactions ASSP-24, pp. 418...423

ZURCHER J.F. (1972): Dispositif de détermination du fondamental de la parole humaine. Recherches/Acoustique (CNET), vol. 1, pp. 7...16

ZURCHER J.F. (1977): La mesure du fondamental par la détection de crêtes. Techniques employés, résultats. Journ.Etud.Parole, vol. 8, pp. 119...126

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

EXTRACTION DE LA FREQUENCE FONDAMENTALE PAR INTERCORRELATION AVEC UNE FONCTION PEIGNE

Ph. MARTIN

Institut de Phonétique
Université de Provence

.RESUME

On présente une nouvelle méthode de mesure de la fréquence fondamentale du signal de parole, basée sur la recherche d'une structure harmonique dans le spectre d'amplitude. La méthode procède par calcul de l'intercorrélation entre le spectre $|F(\omega)|$ et une fonction "peigne" $P(\omega p, \omega)$, ω étant la fréquence, et ωp un paramètre fixant la distance entre les dents du peigne. Le maximum de la fonction d'intercorrélation

$$I(\omega p) = \int_0^{\omega_s} P(\omega p, \omega) \cdot |F(\omega)| \cdot d\omega$$

est obtenu lorsque les dents du peigne correspondent aux pics harmoniques du spectre analysé, c'est-à-dire lorsque $\omega p = \omega o$, le paramètre d'espacement des dents du peigne est égal à la fréquence fondamentale du signal. On montre que ce résultat est obtenu même si la structure harmonique est réduite à la seule fondamentale ou à la présence de deux harmoniques consécutives quelconques. Les premiers résultats montrent que cette méthode est fiable même dans le cas de la transmission téléphonique.

. ABSTRACT

A new method for fundamental frequency measurement is presented here. This method is based on the crosscorrelation between the power spectrum $|F(\omega)|$ and a comb function $P(\omega p, \omega)$, with ωp being the distance between the comb's teeth.

The maximum of the crosscorrelation function

$$I(\omega p) = \int_0^{\omega_c} P(\omega p, \omega) \cdot |F(\omega)| \cdot d\omega$$

is reached when the comb's teeth correspond to the harmonic peaks of the spectrum, i.e. when $\omega p = \omega o$, the spacing of the comb's teeth is equal to the fundamental frequency of the signal. This remains true when the harmonic structure is reduced to the fundamental frequency component only, or to any two consecutive harmonics. Preliminary results show that the method is very reliable in various noise background, including telephone transmission.

EXTRACTION DE LA FREQUENCE FONDAMENTALE PAR INTERCORRELATION AVEC UNE FONCTION PEIGNE

Ph. MARTIN
Institut de Phonétique
Université de Provence

INTRODUCTION

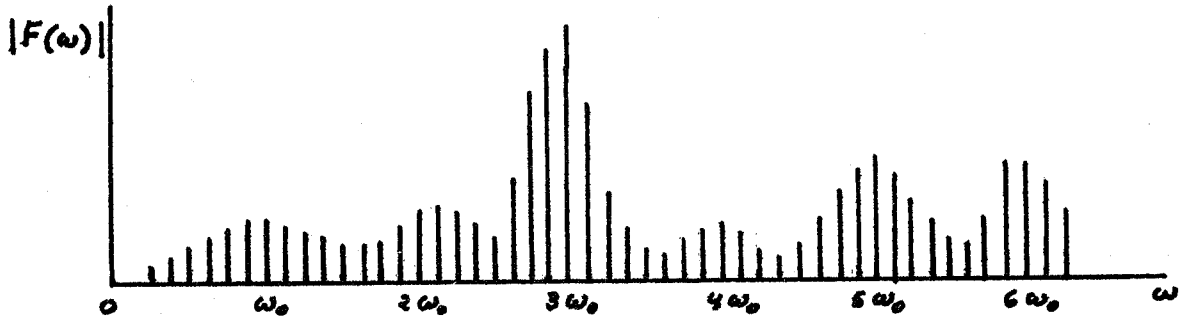
Malgré le très grand nombre de dispositifs d'analyse réalisés à ce jour (cf. McKinney, 1965; Léon et Martin, 1969), la mesure de la fréquence fondamentale du signal de parole reste un problème difficile à résoudre. En effet, l'estimation de la fréquence laryngienne à partir de l'onde acoustique se heurte à la nature essentiellement non périodique de la source, et aux conditions d'interaction continuellement variables entre l'excitation glottale et le conduit vocal. Les durées successives (pseudo-périodes) des cycles ouverture-fermeture des cordes vocales peuvent évoluer dans de grandes proportions à court et à moyen terme, et l'allure de l'onde de parole est continuellement changée par la configuration du conduit vocal. Ces caractéristiques, et plus particulièrement les possibles variations rapides sur plusieurs octaves rendent la mesure de la fondamentale difficile par simple filtrage, et aléatoire si le signal analysé a été transmis en présence de bruit, ou lorsque la fréquence fondamentale est absente du fait d'un filtrage (cas d'une transmission téléphonique, par exemple). Bien que des dispositifs relativement simples permettent une détection convenable de Fo dans une gamme importante (par ex. 70 à 1000 Hz) en opérant par filtrage adaptatif (Martin, 1973), ceux-ci se révèlent inefficaces dans des conditions sévères de bruit ou lorsque la fondamentale n'a pas une amplitude suffisante.

Pour remédier à cette situation, on peut se baser sur une propriété évidente de la fondamentale qui est ... de posséder des harmoniques, et utiliser toutes les informations contenues dans le spectre pour extraire Fo. Les progrès technologiques accomplis ces dernières années permettent d'envisager des dispositifs basés sur l'analyse spectrale du signal et travaillant en temps réel. En utilisant les mesures des fréquences et des amplitudes des harmoniques d'un signal donné pour obtenir une estimation de Fo, on peut espérer réduire fortement les effets du bruit et un fonctionnement satisfaisant même en l'absence de fondamentale. C'est ce principe qui est (indirectement) mis en oeuvre par exemple dans la méthode du cepstre (Noll, 1967), ou dans celle de l'histogramme des périodes (Schroeder, 1968).

On présente ici les principes d'une méthode nouvelle de mesure de Fo, basée sur la détection d'une structure harmonique dans le spectre d'amplitude d'un signal par intercorrélation avec une fonction peigne fréquentielle. Physiquement, cette méthode équivaut à la recherche simultanée des maxima du spectre situés à des fréquences harmoniques les unes des autres.

PRINCIPE DE LA METHODE

La méthode du peigne appartient au domaine fréquentiel, et demande d'effectuer le calcul du spectre d'amplitude du signal à intervalles réguliers, (par exemple tous les 10 ms.), avec une fonction de prélèvement temporelle convenable. En l'absence de bruit, c'est-à-dire en l'absence de tout ce dont un modèle de la phonation ne rend pas compte dans le signal, le spectre d'amplitude $|F(\omega)|$ d'un segment voisé présente par exemple l'allure suivante:



Les pics successifs de $|F(\omega)|$ correspondent à des fréquences qui sont théoriquement multiples entiers les unes des autres, et donc le P G C D est égal à la fréquence fondamentale. On dira dans ce cas que le spectre présente une structure harmonique. Pour qu'il y ait structure harmonique, il n'est pas nécessaire que toutes les composantes multiples de la fondamentale soient présentes. En effet, deux pics harmoniques consécutifs quelconques suffisent pour définir une structure, dont la fréquence fondamentale est alors égale à la différence des fréquences des pics.

Le principe de la méthode du peigne réside dans la recherche de valeurs du spectre situées à des fréquences harmoniques et dont la somme soit maximale pour un intervalle fréquentiel donné. L'intercorrélation spectre-peigne revient au calcul de la somme des composantes spectrales correspondant à une structure harmonique donnée, dont la fondamentale varie dans une gamme prédéterminée (par exemple 70 à 1000 Hz). La fondamentale correspondant à la structure harmonique donnant la somme la plus grande est alors retenue comme fréquence fondamentale du signal, du moins si cette somme est suffisamment différente des valeurs obtenues pour d'autres structures dans le même spectre (cas qui correspond aux segments non voisés).

Formellement, la "meilleure" structure harmonique est liée au maximum de la fonction d'intercorrélation entre le spectre d'amplitude $|F(\omega)|$ et une fonction peigne réelle, ne présentant que des raies d'amplitude A à des intervalles réguliers espacés d'une valeur fréquencielle ωp (le paramètre ωp équivaut à la fréquence fondamentale du peigne).

L'équation du peigne s'écrit donc

$$P(\omega p, \omega) = \sum_{n} An\, \delta(n\,\omega p - \omega)$$

avec An = amplitude de la dent d'ordre n.

La représentation d'une fonction peigne d'amplitude constante A, et de paramètre ωp, est:

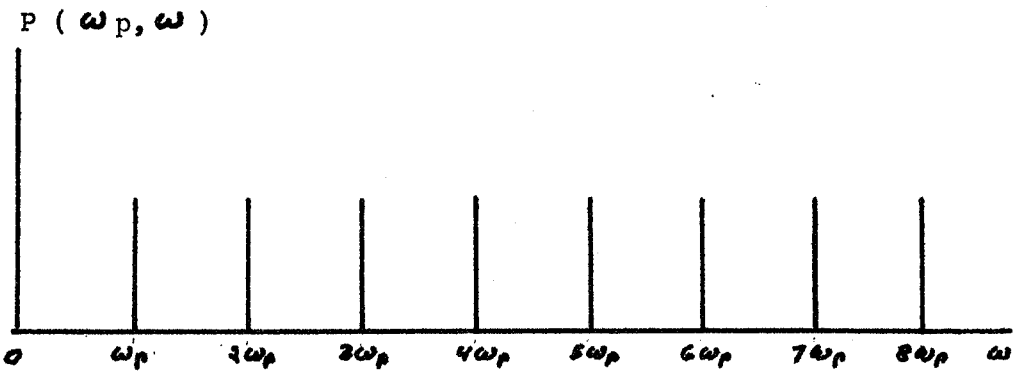


Fig. 1 Fonction peigne P ($\omega p, \omega$) aux dents d'amplitude constante A

La fonction d'intercorrélation spectre-peigne est:

$$I(\omega p) = \int_0^{\infty} P(\omega p, \omega) \quad |F(\omega)| \quad . \, d\omega$$

$|F(\omega)|$ représentant le spectre d'amplitude si l'on désire déterminer la structure harmonique par la somme des composantes, ou le logarithme du spectre d'amplitude si le calcul porte sur le produit des composantes harmoniques.

Si le spectre $|F(\omega)|$ était réduit à un spectre de raies, la fonction d'intercorrélation avec un peigne d'amplitude constante présenterait des maxima chaque fois que les dents du peigne correspondent à l'ensemble des raies (en nombre supposé fini), c'est-à-dire lorsque $\omega p = \omega o$, la fréquence fondamentale du signal, mais aussi chacun de ses sous multiples $\frac{\omega o}{2}$, $\frac{\omega o}{3}$, etc.

Par contre, pour des valeurs multiples de ωo , $\omega p = 2\omega_o$, $3\omega o$, etc., les dents du peigne ne correspondent pas à toutes les raies du spectre, et les valeurs de la fonction d'intercorrélation seront inférieures à celles obtenues pour $\omega p = \omega o$.

Pour une gamme de variation de 70 à 1000 Hz, on peut montrer facilement que le nombre des sous harmoniques détectées par les maxima théoriquement égaux de I (ωp) est au plus de 3.

On peut faire en sorte que seule la fréquence ωo corresponde au maximum de la fonction d'intercorrélation en pondérant différemment les différentes raies du spectre, de manière à pénaliser les harmoniques élevées dans le calcul de I (ωp). Les mêmes pics harmoniques du spectre correspondant à des dents du peigne d'ordre différent pour $\omega p = \frac{\omega o}{2}$, $\frac{\omega o}{3}$, ..., on obtient ce résultat en conférant aux dents du peigne une amplitude décroissante, par exemple selon la loi

$$A_n = (n)^{-p} \qquad \text{avec p entier} > 1$$

Cette amplitude décroissante des dents du peigne est préférable à une normalisation de la fonction d'intercorrélation par le nombre de dents présentes dans la bande de fréquence utilisée. Par cette dernière méthode en effet, on obtient des maxima égaux pour $\omega p = \omega o$ et $\omega p = 2\omega o$ si le spectre ne présente que deux pics harmoniques et que le premier corresponde à la fondamentale.

En résumé, on a donc:

$$I(k\,\omega o) < I(\omega o)$$

$$I\left(\frac{\omega o}{k}\right) < I(\omega o)$$

avec

$$I(\omega p) = \int_0^{\omega_s} \left[\sum_n (n)^{-p}\, \delta(n\,\omega p - \omega)\right] \cdot |F(\omega)|\, d\omega$$

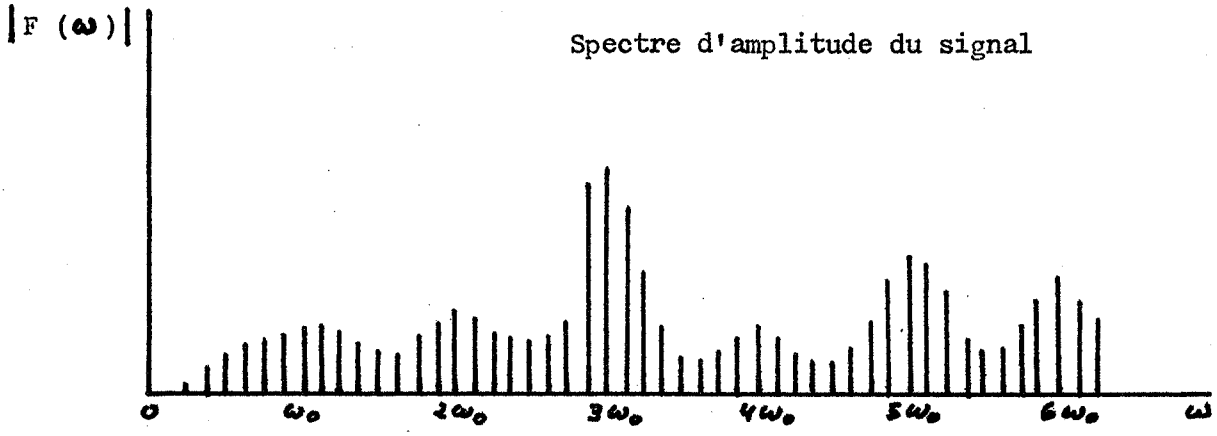
(avec ω_s = borne supérieure du spectre).

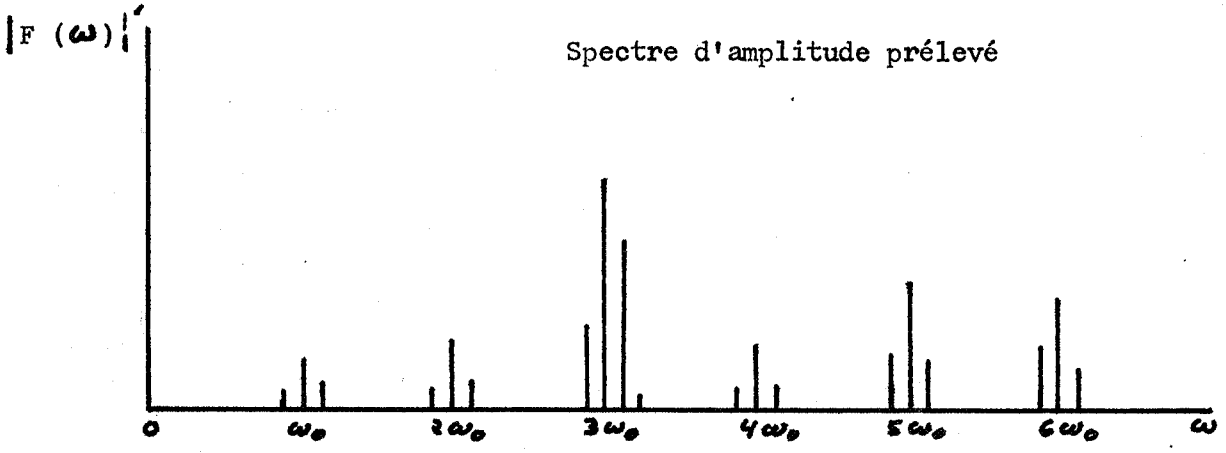
En fait, le spectre d'amplitude n'est pas un spectre de raie et il présente des valeurs non nulles entre ses différents sommets. Malgré une amplitude décroissante des dents du peigne, la valeur de $I(\omega p)$ risque donc d'être supérieure, pour des valeurs de ωp sous harmoniques, à celle obtenue pour $\omega p = \omega o$, puisque des dents intermédiaires du peigne correspondent alors aux creux non nuls entre les pics harmoniques.

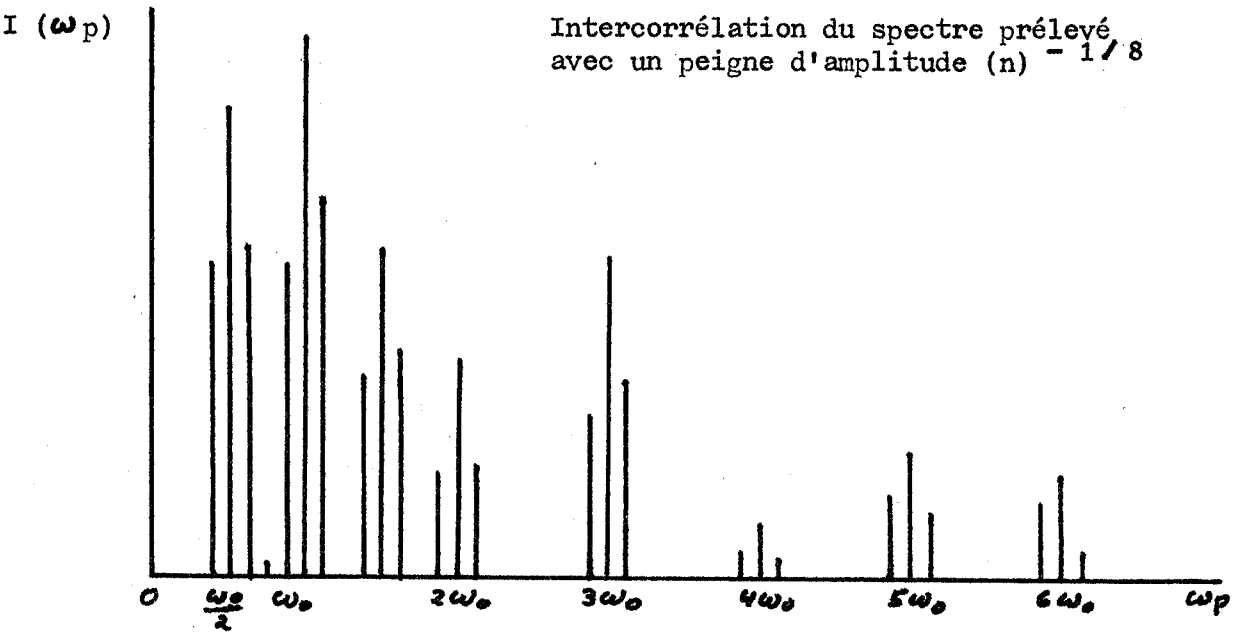
Pour être assuré d'obtenir un maximum de $I(\omega p)$ pour $\omega p = \omega o$, il y a donc lieu d'éliminer l'effet des valeurs spectrales non nulles entre les maxima du spectre, en réalisant un prélèvement fréquentiel de chacun des sommets selon une fonction analogue aux fenêtres utilisées pour le prélèvement temporel du signal (Hamming, Hanning, Gauss, etc.). Ce prélèvement est préférable à une discrétisation réduisant le spectre à ses seuls maxima. En effet, l'imprécision du calcul numérique et les éventuelles distorsions de phase peuvent rendre les fréquences des pics du spectre non harmoniques exacts de la fréquence fondamentale. Il n'y aurait donc plus dans ce cas correspondance exacte des dents du peigne avec chacune des harmoniques. Au contraire, le prélèvement fréquentiel permet d'interpréter le sommet de la fonction d'intercorrélation comme résultat du meilleur ajustement d'un spectre donné avec une contrainte d'harmonicité de fréquence fondamentale ωp.

On voit donc là apparaître l'intérêt de la méthode par rapport au cepstre, pour lequel la mesure de la fondamentale dépend de la détermination d'une périodicité dans le spectre. La fonction d'intercorrélation présentera un maximum pour ωo même si cette fréquence est la seule composante d'un spectre dépourvu de toute harmonique, ou si seules deux harmoniques consécutives sont présentes, alors que ce résultat n'est pas garanti dans le cas du cepstre.

Un critère de voisement peut être obtenu en comparant le maximum de $I(\omega p)$ aux valeurs liées aux peignes définissant une structure dans les segments de bruit du spectre. Un seuil de voisement peut être fixé au dessous duquel le segment analysé est considéré comme non voisé, ce seuil devant correspondre à la somme pondérée moyenne des spectres non voisés.


|F (ω)|
Spectre d'amplitude du signal
0
ω₀
2ω₀
3ω₀
4ω₀
5ω₀
6ω₀
ω


|F (ω)|'
Spectre d'amplitude prélevé
0
ω₀
2ω₀
3ω₀
4ω₀
5ω₀
6ω₀
ω


I (ωp)
Intercorrélation du spectre prélevé
avec un peigne d'amplitude (n) - 1/8
0
ω₀/2
ω₀
2ω₀
3ω₀
4ω₀
5ω₀
6ω₀
ωp

SIMUTATION DE LA METHODE

On a simulé la méthode d'analyse de Fo par une fonction peigne en programme FORTRAN, exécuté sur calculateur T1600. La fréquence d'échantillonnage choisie est de 4 k Hz, fixant la borne supérieure du spectre à 2 k Hz. La fenêtre de prélèvement utilisée est du type Kaiser-Bessel (Harris, 1976), portant sur une durée de 32 ms correspondant à 128 échantillons du signal. L'analyse spectrale est réalisée toutes les 10 ms, ce qui permet d'obtenir 100 valeurs de Fo par seconde.

Le spectre d'amplitude est obtenu par FFT. Des 64 valeurs résultantes, sont retenus les sommets entourés des valeurs immédiatement adjacentes, qui sont ensuite recopiés dans un tableau correspondant au spectre prélevé, et dont la résolution est de 1 Hz. Les valeurs intermédiaires autour des sommets sont obtenues par interpolation parabolique, avec une largeur de bande de chaque pic de 40 Hz ($\pm$ 20 Hz autour du maximum). Les valeurs de $|F(\omega)|$ inférieures au maximum du spectre de plus de 35 dB sont annulées de manière à éliminer les pics de $I(\omega p)$ dus à une structure harmonique de bruit.

La fonction d'intercorrélation est ensuite calculée pour ωp variant de 70 à 1000 Hz, et son maximum est déterminé, qui correspond à la fréquence fondamentale cherchée. La durée de calcul sur un calculateur T1600 est d'environ 200 fois le temps réel.

PREMIERS RESULTATS

Huit locuteurs, 5 hommes et 3 femmes ont été enregistrés dans diverses conditions (studio, salle de calcul, téléphone, transmission télévisée) de manière à faire une première évaluation des performances du système. Les erreurs d'harmoniques établies en comparant les valeurs calculées à celles obtenues par inspection visuelle du signal, ont été relevées pour les segments voisés. (Le caractère voisé d'un segment est déterminé par un critère indépendant élaboré par R. Espesser, utilisant le nombre de passages par zéro, l'intensité, et l'erreur de prédiction linéaire).

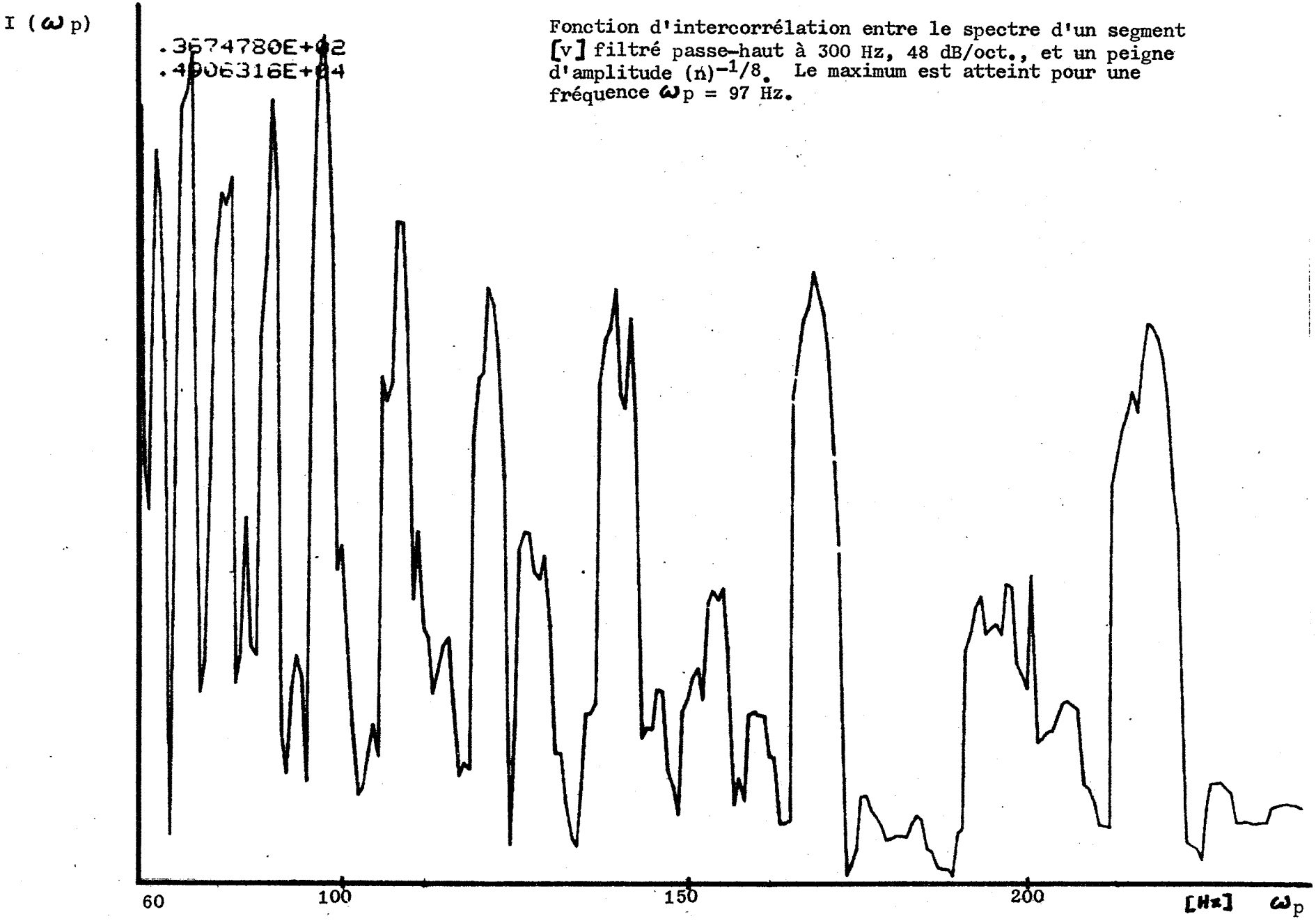
Les résultats sont regroupés dans le tableau suivant;

Locuteur	Condition	Nombre d'erreurs	Nbre total de valeurs	Pourcentage d'erreurs
(M) J E -	Salle de calcul	11	365	3,0 %
(M) R E -	" "	4	340	1,1 %
(M) R B -	" "	2	290	0,7 %
(M) J C -	" "	3	280	1,1 %
(F) M R -	" "	8	325	2,4 %
(F) B R -	" "	3	260	1,1 %
(F) M E -	Studio, voix chantée	0	347	0,0 %
(M) R E -	filtrage passe-haut à 300 Hz, 48 dB/oct	28	210	13,3 %
(M) R E -	transmission téléphonique	12	113	10,6 %
(M) G E -	enregistrement à la télévision	9	71	12,6 %
	Total	88	2727	3,26 %

(M) = voix masculine

(F) = voix féminine

Tableau donnant le pourcentage d'erreurs harmoniques de la méthode pour diverses conditions d'enregistrement.



Fonction d'intercorrélation entre le spectre d'un segment [v] filtré passe-haut à 300 Hz, 48 dB/oct., et un peigne d'amplitude $(n)^{-1/8}$. Le maximum est atteint pour une fréquence ω_p = 97 Hz.

CONCLUSIONS

La méthode proposée se révèle donc fiable même dans des conditions difficiles (transmissions téléphoniques). Bien que les conditions ne soient pas les mêmes, le pourcentage d'erreurs se compare favorablement à celui rapporté dans Rabiner et al., 1976, pour 7 autres méthodes de détection de Fo. Une version cablée, en cours de réalisation, permet un fonctionnement en temps réel dans les mêmes conditions que celles de la simulation, mais en déterminant les maxima du spectre par itération des valeurs obtenues par DFT et non plus par FFT. L'interpolation parabolique n'est alors plus nécessaire, les valeurs prélevées autour des sommets étant calculées directement. Comparé au cepstre, le nombre de multiplications est de 235 x 8 = 1880 pour le calcul de la fonction d'intercorrélation à partir du spectre, alors que le cepstre par FFT est habituellement effectué à partir d'un spectre de 256 points et demande donc 256 x 8 = 2048 multiplications pour une même résolution de 4 Hz dans une gamme de 1000 Hz.

L'avantage principal de la méthode proposée ne vient donc pas d'un gain de temps de calcul (quoiqu'elle se prête à une réalisation cablée plus simple que celle du cepstre), mais plutôt de ce que la détection de Fo se base sur une méthode dont l'interprétation physique est plus satisfaisante parce que basée sur la recherche de la meilleure structure harmonique pouvant se trouver dans le spectre, structure qui peut ne correspondre qu'à la seule fondamentale ou qu'à 2 harmoniques consécutives quelconques, alors que le cepstre est lié à la périodicité des harmoniques dans le spectre.

REMERCIEMENTS: à Robert ESPESSER qui a écrit les programmes de simulation.

REFERENCES

HARRIS, F.J. (1976) "Trigonometric Transforms; A Unique Introduction to the FFT", Spectral Dynamic Technical Publication, DSP-005 6/76.

LEON, P.R. et MARTIN, PH. (1969) "Prolégomènes à l'étude des structures intonatives", Studia Phonetica II, Didier, Montréal.

MARTIN, Ph. (1973) "Réalisation d'un analyseur de mélodie fonctionnant en temps réel en vue de la recherche en phonétique et l'enseignement", Thèse Doctorat en Sciences, ULB. Bruxelles (non publié).

McKINNEY, N.P. (1965) "Laryngeal Frequency Analysis for Linguistic Research", Commun. Sci. Lab., Univ. of Michigan, Ann Arbor, Rept 14.

NOLL, A.M. (1967) "Cepstrum Pitch Determination", JASA, (40) 293-309.

RABINER, L.R. et al. (1976), "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE trans. ASSP-24, 399-418.

SCHROEDER, M.R. (1968) "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurements" JASA 43 (4) 829-834.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

COUPLAGE CONDUIT VOCAL-SOURCE VOCALE - CONSEQUENCES ET ANALYSE DE PAROLE

R. CARRE

Laboratoire de la Communication Parlée
Equipe de Recherche Associée au C.N.R.S.

E.N.S.E.R.G.
23, rue des Martyrs - 38031 GRENOBLE CEDEX

.RESUME

On étudie ici les effets du couplage source vocale-conduit vocal lors de la production de voyelles. Le couplage provoque des déplacements de fréquences des formants, des augmentations des bandes passantes. Par ailleurs, il est à l'origine d'un phénomène de synchronisation qui se caractérise principalement par une absorption spectrale pour des fréquences légérement supérieures à celle du formant considéré réputé sensible au couplage.

Les conséquences de ces phénomènes en analyse de parole sont discutées.

. ABSTRACT

The vocal source-vocal tract coupling effects are studied. Formant frequency displacements, bandwidth frequency increases, are the main results of such a coupling. More, a kind of synchonisation effect is observed, mainly characterized by a spectral zero near the formants which are affected by the coupling effect.

INTRODUCTION

FANT a montré, dès *1960*, l'importance de l'impédance de la source vocale sur les fréquences et les bandes passantes des formants. L'excitation principale du conduit vocal correspond à la fermeture de la glotte *(HOLMES - 1976)* ce qui montre le rôle important de l'amortissement dû aux pertes par la source quand les cordes vocales sont ouvertes.

Ces pertes sont généralement calculées pour une valeur moyenne de la glotte *(FLANAGAN - 1972)*, *(FLANAGAN, ISHIZAKA et SHIPLEY - 1975)*.

Plus récemment, *WAKITA* et *FANT (1978)* ont calculé les fréquences et les bandes passantes des formants pour deux ouvertures de la glotte.

Par ailleurs, *FANT (1980)* a proposé une analyse originale du signal de source.

Enfin, l'impédance moyenne de la source (qui évolue dans le temps au cours d'une période de façon non linéaire) a été étudiée par simulation *(CARRE, GUERIN - 1980)*.

On se propose ici de vérifier sur des voyelles naturelles les résultats obtenus en simulation lorsque les effets de couplage source-conduit sont maximums. Ces résultats sont les suivants :

- augmentation des fréquences et des bandes passantes des formants,

- apparition d'une absorption dans le spectre, légèrement au delà du formant sensible au couplage, absoption pouvant correspondre à un phénomène de "synchronisation" entre la source et le conduit.

L'intéraction source-conduit vocal est maximale lorsque l'impédance d'entrée du conduit vocal vue coté source est élevée.

Les valeurs des parties résistives de l'impédance données tableau I pour 11 voyelles avec les valeurs des fréquences et des bandes passantes des formants correspondants ont été calculées par *MRAYATI, GUERIN, BOE (1976)* à partir d'un modèle du conduit. Les pertes par vibrations des parois, viscosité, chaleur et radiation ont été prises en compte.

D'après cette simulation, les effets maximums de couplage doivent intervenir, pour F_1 et F_2 de la voyelle [*a*], pour F_1 et F_3 de la voyelle [a], pour F_2 de la voyelle [i] etc...

Ils doivent aussi intervenir lorsque les cordes vocales fonctionnent dans des conditions correspondant à un maximum de la durée de la phase d'ouverture c'est à dire lorsque la fréquence fondamentale est élevée ou bien lorsque l'intensité est faible.

Nous allons décrire et commenter les résultats d'analyse obtenus pour 3 voyelles.

Quelques conséquences concernant l'analyse du signal de parole seront discutées.

	[u]	[o]	[ɔ]	[ɑ]	[a]	[ɛ]	[e]	[i]	[y]	[ø]	[œ]
a F_1(Hz)	296	402	540	676	678	458	418	244	280	384	535
a F_2	783	797	1007	1192	1300	1881	2111	2250	1846	1592	1411
a F_3	2286	2417	2480	2208	2583	2584	2665	3082	2230	2461	2569
b B_1(Hz)	75	43	38	36	35	34	37	72	69	48	35
b B_2	20	29	45	42	63	72	93	19	55	86	53
b B_3	18	34	65	49	80	162	269	272	29	54	67
c R_1(Ohms)	50	128	275	330	265	102	95	38	50	98	170
c R_2	175	37	51	350	125	107	93	348	62	16	118
c R_3	186	20	40	210	450	130	70	45	330	300	320

TABLEAU I : Fréquences des formants (a), bandes passantes (b), et parties résistives de l'impédance d'entrée du conduit vocale correspondantes (c) pour 11 voyelles différentes.

I - ANALYSE DE VOYELLES NATURELLES

Trois voyelles [ɑ], [i], [u] ont été prononcées à fréquence fondamentale croissante. On a pris soin de conserver un conduit vocal et une position du larynx stables.

La figure 1 représente le spectre (F.F.T.) de la voyelle [ɑ] pour deux valeurs de la fréquence fondamentale. On constate une absorption caractéristique (spectre de droite) immédiatement au dessus du 1er formant. Par ailleurs, le 2ème formant est nettement affaibli (spectre de gauche).

Les figures 2, 3 et 4 représentent les évolutions des amplitudes des différents harmoniques, calculées par F.F.T., pour les deux premiers formants des voyelles [ɑ] (en fait intermédiaire entre [ɑ] et [ɔ]), [i] et [u].

On constate, figure 3, en particulier pour l'harmonique 6 mais aussi pour l'harmonique 7, des variations d'amplitude, qui pourraient correspondre à l'effet de synchronisation constaté en simulation.

De même, on constate des "instabilités" pour les 2es formants, des trois voyelles ce qui correspond bien à des impédances d'entrée du conduit vocal élevées (voir tableau I).

Les variations pouvant découler d'une instabilité de la forme du conduit vocal, nous avons calculé les formants, par la méthode de covariance en codage prédictif, dans les parties correspondant à la glotte fermée. On calcule donc les résonances du conduit vocal sans le couplage avec la source.

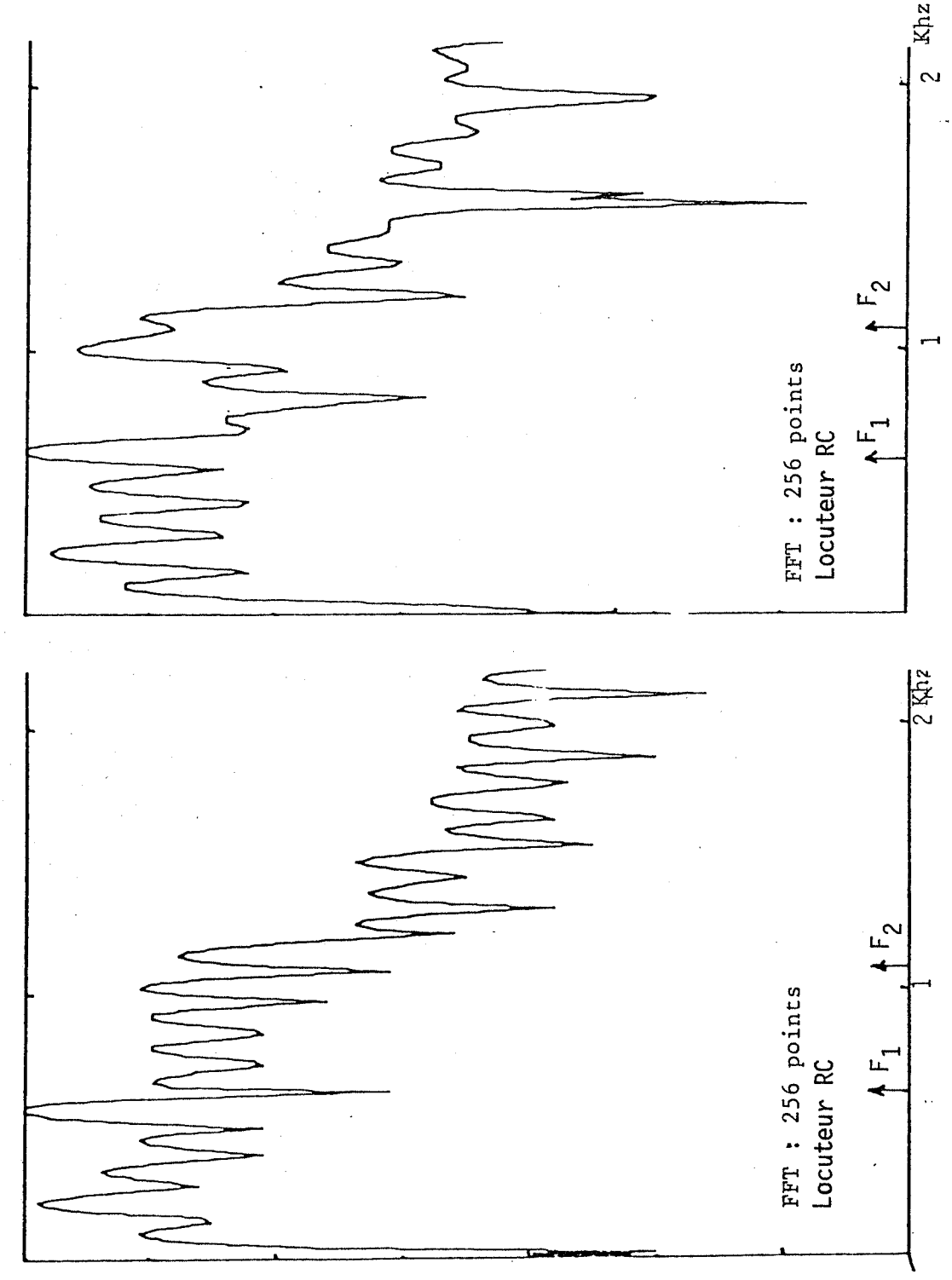
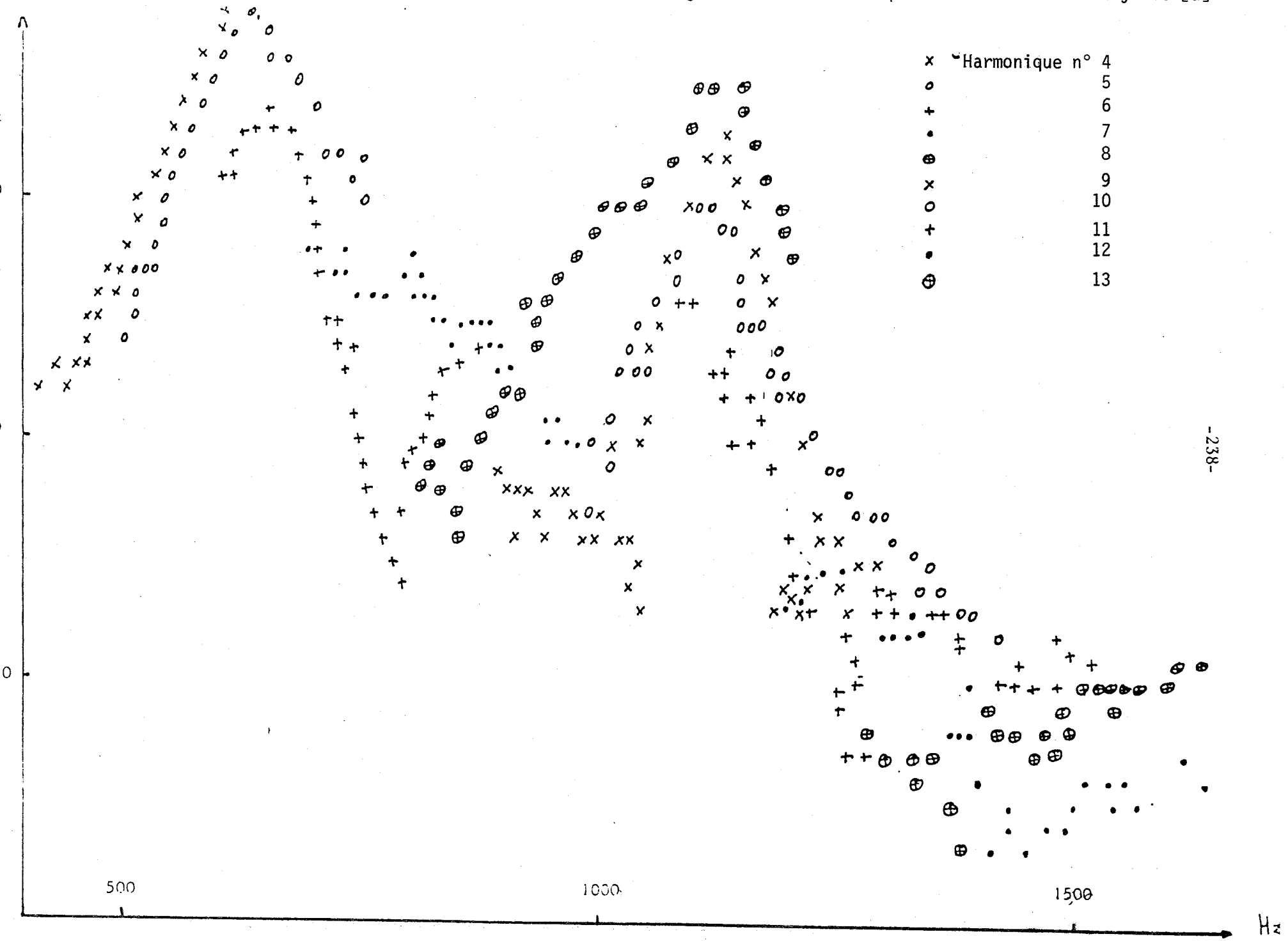


FIGURE 1 Spectres de la voyelle [a] pour deux valeurs de la fréquence fondamentale

FIGURE 2 Variations des amplitudes des harmoniques lors d'une augmentation de la fréquence fondamentale - Voyelle [a]



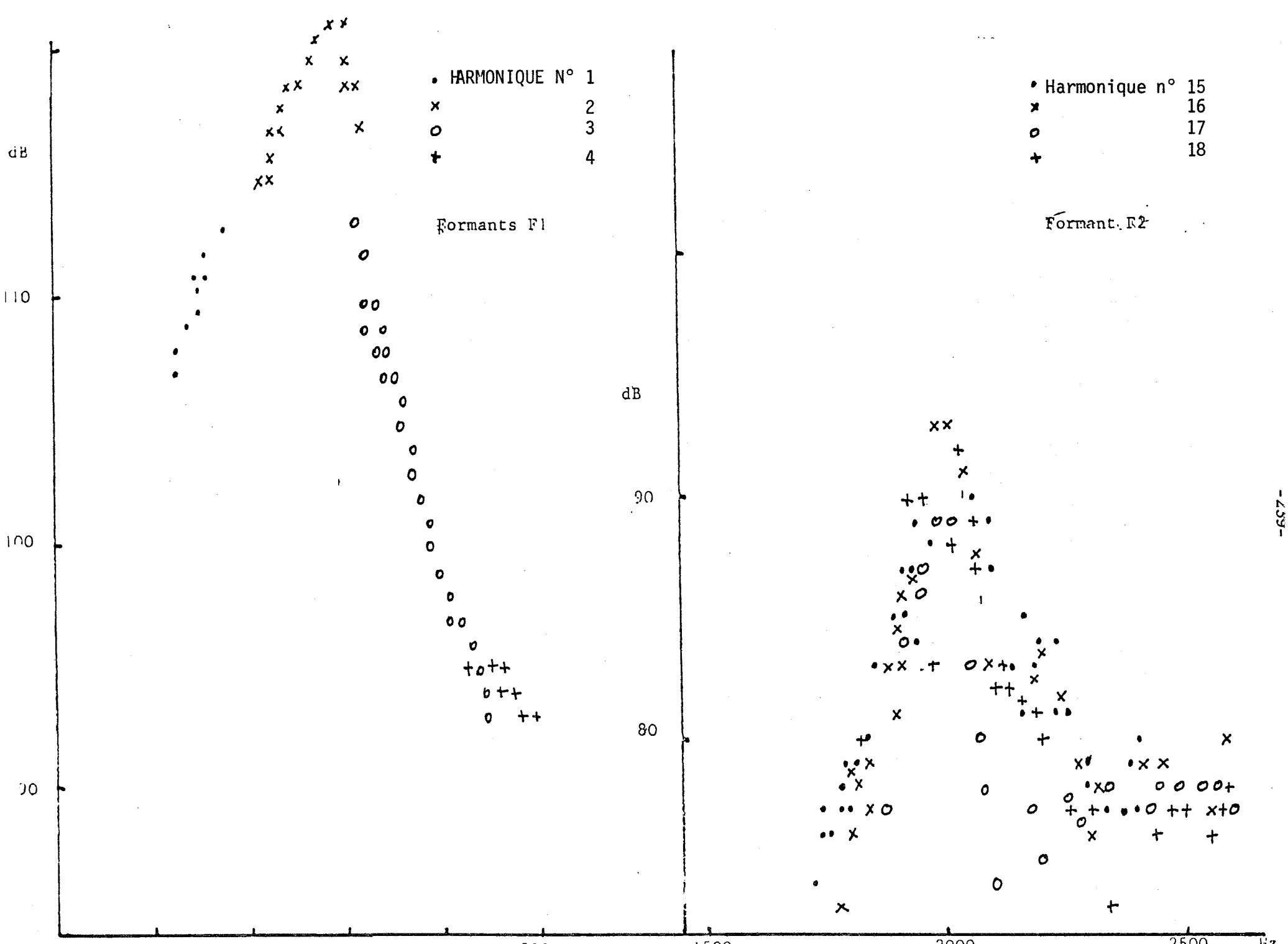
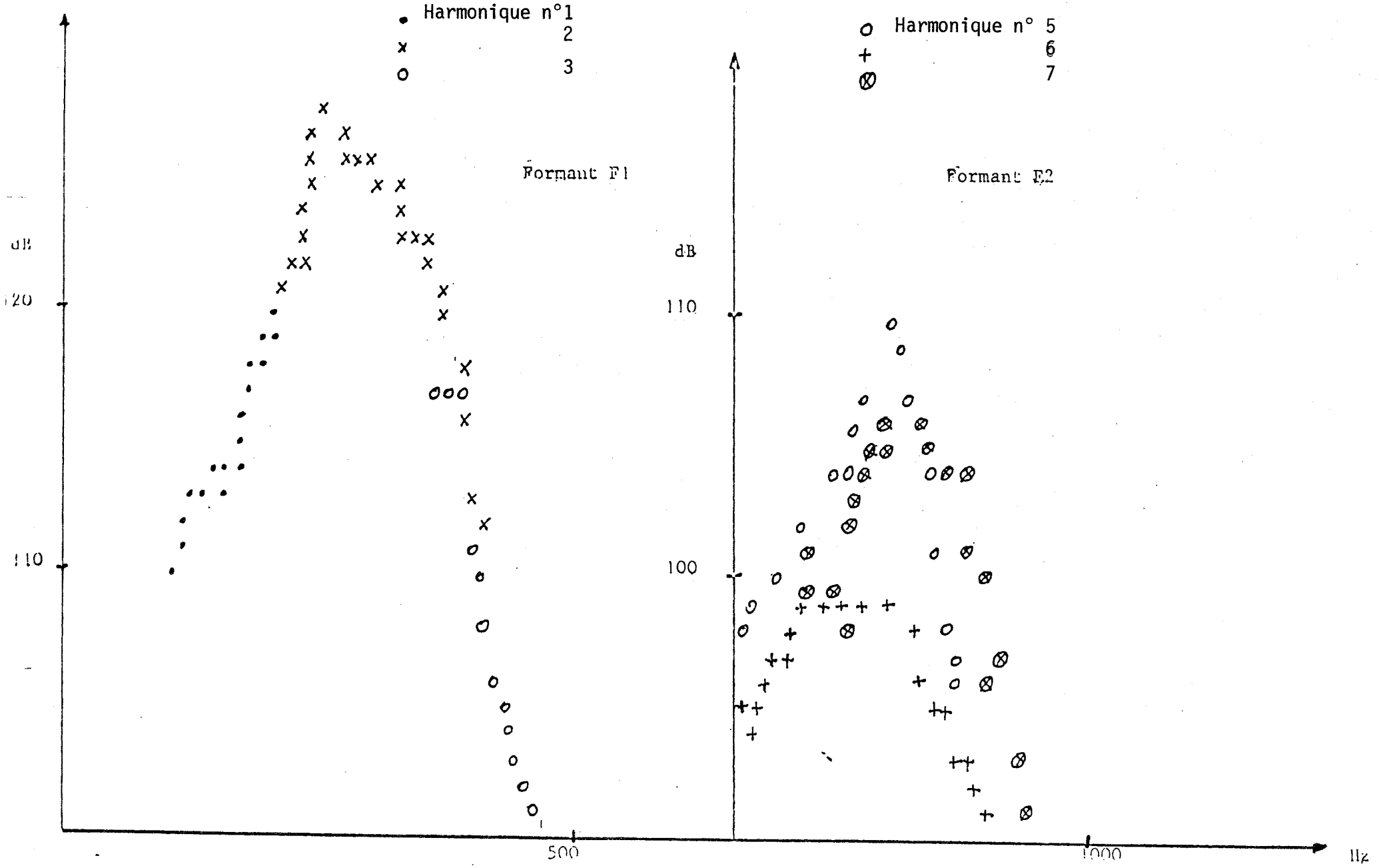


FIGURE 3 Variations des amplitudes des harmoniques lors d'une augmentation de la fréquence fondamentale - voyelle [i]

FIGURE 4 Variations des amplitudes des harmoniques lors d'une augmentation de la fréquence fondamentale.
Voyelle [u]



Comme le montre les résultats des tableaux II et III, les formants ne varient pas de manière notable.

F_0 (Hz)	F_1 (ΔF_1)	F_2 (ΔF_2)	F_3 (ΔF_3)	F'_1 ($\Delta F'_1$)	F'_2 ($\Delta F'_2$)	F'_3 ($\Delta F'_3$)
114	584 (45)	1070 (37)	2984(160)	572 (188)	1077(239)	2900(153)
116	594 (28)	1079 (48)	2984(152)	573 (187)	1080(212)	2945(171)
119	592 (40)	1080 (44)	2980(175)	585 (200)	1082(155)	2915(118)
121	590 (41)	1083 (38)	2976(186)	595 (191)	1090(114)	1894(171)
121	588 (39)	1082 (44)	2989(197)	591 (167)	1093(104)	2929(156)
125	596 (43)	1090 (60)	2972(167)	596 (163)	1098(157)	2959(145)
128	587 (47)	1082 (42)	2963(193)	592 (184)	1081(136)	2992(236)
129	579 (65)	1071 (63)	2968(181)	568 (173)	1075(133)	2944(102)
133	586 (40)	1080 (42)	2984(216)	584 (248)	1080(121)	2887(125)
136	593 (41)	1083 (45)	2983(159)	559 (177)	1092 (73)	2967(119)
142	584 (43)	1086 (38)	2979(189)	567 (161)	1135(155)	2960(172)
144	583 (56)	1089 (46)	2974(258)	570 (141)	1102(179)	2943(135)
144	587 (48)	1092 (47)	2959(183)	564 (130)	1112(126)	2940(148)

TABLEAU II : Fréquences et bandes passantes des trois premiers formants calculées par prédiction-covariance 12 coefficients (F_1, F_2, F_3) et autocorrélation 10 coefficients (F'_1, F'_2, F'_3) à différentes positions correspondant à différentes fréquences fondamentales. Voyelle [a].

Analyse	F_1 (Hz)	ΔF_1 (Hz)	F_2 (Hz)	ΔF_2 (Hz)
Voyelle [a]	588 (5)	44 (9)	1082 (6)	46 (8)
Voyelle [i]	269 (4)	64 (13)	2002 (24)	101 (31)
Voyelle [u]	280 (8)	52 (16)	807 (8)	44 (14)

TABLEAU III : Fréquences moyennes et bandes passantes avec les déviations standard pour les 2 premiers formants - Voyelles [a], [i], [u] - Analyse par covariance (12 coef.).

Les valeurs des bandes passantes obtenues par covariance correspondent généralement avec celles obtenues par simulation du conduit vocal (tableau I).

Par ailleurs, à partir des différentes amplitudes des composantes harmoniques on peut tracer une courbe de réponse moyenne et mesurer F et ΔF.

On retrouve bien alors les résultats obtenus par simulation du couplage source-conduit *(CARRE, GUERIN 1980)* c'est à dire : déplacement de fréquence centrale du formant vers les fréquences élevées, augmentation de la bande passante, effet de synchronisation.

II CONSEQUENCES EN ANALYSE DE PAROLE

Le phénomène de "synchronisation" constaté, qui dépend de la fréquence fondamentale, peut perturber notablement les résultats.

On peut alors souhaiter comparer nos résultats avec ceux obtenus, par analyse prédictive à autocorrélation, technique généralement exploitée pour l'obtention de données ou pour la commande d'un synthétiseur.

Le tableau II facilite la comparaison point par point des résultats.

Le tableau IV donne les valeurs moyennes des formants et des bandes passantes pour la voyelle [*a*] en fonction de la fréquence fondamentale.

F_0 (Hz)	F_1 (Hz)	ΔF_1 (Hz)	F_2 (Hz)	ΔF_2 (Hz)	F_3 (Hz)	ΔF_3 (Hz)	Nbr de Mesures
<110	583,5(10)	183 (28)	1084 (14)	190 (48)	2977 (28)	126(29)	49
110-120	578 (10)	214 (27)	1084 (12)	259 (66)	2923 (23)	143(30)	34
120-130	590 (10)	177 (21)	1091 (10)	142 (27)	2917 (28)	158(33)	23
130-140	570 (25)	211 (18)	1087 (12)	109 (32)	2912 (23)	133(27)	12
>140	565 (9)	148 (23)	1115 (12)	142 (22)	2959 (18)	154(34)	21

Tableau IV : Fréquences moyennes et bandes passantes des 3 premiers formants avec les déviations standard pour différentes plages de fréquences fondamentales. On a indiqué dans la colonne de droite le nombre de mesures effectuées - Voyelle [a].

On a constaté les écarts maximun suivants :

- $554 < F_1 < 617$ Hz pour $F_\emptyset$ correspondant respectivement à 125 Hz et 145 Hz.

- $120 < \Delta F_1 < 300$ Hz pour $F_\emptyset$ correspondant respectivement à 145 Hz et 115 Hz.

Les résultats obtenus sont donc très peu stables comparés à ceux obtenus avec la méthode de covariance. Ces variations pouvent être dues au couplage source vocale-conduit ou à la méthode d'analyse. Par ailleurs, avec un nombre de coefficients de prédiction différents, les résultats sont notablement différents.

Enfin, aux instants de synchronisation notable $121 \text{ Hz} < F_0 < 128$ Hz il est curieux de constater une augmentation de la fréquence du 1er formant. Une diminution aurait pu être attendue. Il est aussi étonnant de constater que la fréquence du formant obtenu par la méthode de l'autocorrélation est inférieure à celle obtenue par la méthode de la covariance. Là encore, le résultat inverse est attendu car avec la méthode de l'autocorrélation, on effectue une analyse sur 20 ms de signal donc on intègre les effets inductifs de la glotte.

En fait, on peut vérifier ici que la méthode de prédiction linéaire par autocorrélation propose un modèle qui a tendance à "suivre" l'harmonique le plus élevé à l'intérieur d'un formant plutot que de se "caler" sur le formant lui-même.

CONCLUSIONS :

Il est difficile de formuler des conclusions définitives. Toutefois, différents résultats concernant le couplage conduit vocal-source vocale peuvent être retenus.

Le couplage conduit vocal-source vocale peut être étudié, dans une première approche, au moyen d'un modèle de source à deux masses. Une mesure de l'impédance interne du modèle de source peut être facilement effectuée *(Voir CARRE, GUERIN - 1980)*.

Un phénomène de synchronisation a été relevé sur le modèle, phénomène que l'on retrouve dans la réalité et qui se caractérise alors par une absorption près du formant sensible au couplage.

Cette absoption peut correspondre aux zéros généralement observé vers 800 Hz sur les spectres moyens de parole.

L'analyse par la méthode de la covariance qui permet d'obtenir les résonances du conduit vocal à glotte fermée donc sans pertes glottiques parait sûre mais elle est longue à mettre à oeuvre car il est nécessaire de déterminer visuellement la phase de fermeture de la glotte.

L'analyse par F.F.T est sûre mais elle ne permet pas l'obtention de fréquences des formants.

L'analyse par la méthode de l'autocorrélation, facile à mettre en oeuvre et automatique, donne des résultats discutables (détection de l'harmonique le plus élevé à l'intérieur d'un formant).

On peut donc se poser de nombreuses questions sur les fréquences de formants généralement attribuées au seul conduit vocal alors que le couplage source-conduit et la fréquence fondamentale peuvent jouer un rôle non négligeable et modifier notablement et de façon imprévisible les caractéristiques.

BIBLIOGRAPHIE :

. CARRE R., GUERIN B. (1980)

Glottal impedance of a two mass model of the vocal cords
J. of Acoust. Soc. of Am. S67, 95

. FANT G. (1960)

Acoustic Theory of Speech Production
Mouton - The Hague

. FANT G. (1979)

Temporal fine structure of formant damping and excitation
50 th Anniversary Meeting of the Acoustical Society of America, BOSTON

. FANT G. (1980)

Voice source dynamics
STL - QPSR, 2-3, 17-37

. FLANAGAN J.L., ISHIZAKA K., SHIPLEY K.L. (1975)

Synthesis of Speech from a dynamic model of the vocal cords and vocal tract
B.S.T.J., 54, 485-506

. GUERIN B., BOE L.J. (1977)

A two mass model of the vocal cords : determination of control parmeters and their respective consequences
IEEE Int. Conf. on Acoust., Speech and Signal Proc., Hartford, 583-586

. HOLMES J.N (1976)

Formant excitation before and after glottal source
IEEE Conf. on Acoustics, Speech and Signal Processing, PHILADELPHIA

. 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

. MRAYATI M., GUERIN B., BOE L.J. (1976)

Etude de l'impédance d'entrée du conduit vocal - Couplage source-conduit vocal
Acustica 35, 330-340

. WAKITA H., FANT G. (1978)

Toward a better vocal tract model
STL - QPSR, 1, 9-29

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

VARIATIONS INTRA ET INTER LOCUTEUR DE F_0 : PROCEDURES DE NORMALISATION.

Etude préliminaire.

Guy JOUANNELLE - Louis-Jean BOË - Gaston BELLET - Institut de Phonétique de Grenoble

RESUME

Dans le cadre d'une étude générale des variations intra- et inter-locuteur de F_0, ce travail préliminaire définit une procédure de normalisation temporelle et/ou fréquentielle de l'évolution du fondamental utilisant le principe de la programmation dynamique.

Diverses variantes (tenant essentiellement aux contraintes de parcours) de cet algorithme ont été testées.

Les caractéristiques de la procédure choisie (largeur et forme de fenêtre, taille de cellule) semblent confirmer une absence de corrélation suprasegmentale entre deux sons voisins d'une phrase donnée puisque les meilleurs résultats sont obtenus pour une fenêtre correspondant à 50 ms.

Au niveau linguistique, des tests perceptifs sont envisagés afin de valider (ou rejeter) cette procédure de normalisation.

. ABSTRACT

Within the frame of a systematic research of inter- and intra- speaker variations of the fundamental frequency of speech (F_0) this paper presents a preliminary study of time - and / or frequency normalization. We use a well known dynamic programming (DP) algorithm generally used for spoken word recognition.

To determine an optimum algorithm for D.P. matching we have tested the symetric and asymetric form and different window shapes and sizes.

The best normalization is obtained for a symetric form and a 50 ms window. This result confirms the fact that there is no correlation, for variations of F_0, for a reference size of 50 ms.

The next step on the linguistic level will consist in verifying the validity of this preliminary study by administering perception tests.

VARIATIONS INTRA ET INTER LOCUTEUR DE F_0 : PROCEDURES DE NORMALISATION

Etude préliminaire.

Guy JOUANNELLE, Louis-Jean BOË, Gaston BELLET

Institut de Phonétique de Grenoble

I - INTRODUCTION

Dans le cadre d'analyses linguistiques ou d'études sur les caractéristiques individuelles, on se trouve confronté à deux types de difficultés lorsque l'on est amené à comparer objectivement deux séries de valeurs de la fréquence fondamentale (F_0) correspondant à la réalisation d'une séquence de parole donnée prononcée par un même locuteur ou par des locuteurs différents :

- Les valeurs moyennes ($\overline{F_0}$) sont décalées d'un locuteur à un autre. Entre adultes les différences les plus importantes sont liées au sexe : $\overline{F_0}$ = 118 Hz pour les hommes et $\overline{F_0}$ = 210 Hz pour les femmes. Pour les enfants, elles sont encore plus marquées et varient avec l'âge avant de se stabiliser après la puberté. A l'intérieur d'un même groupe homme/femme les écarts types sur la moyenne sont assez importants, 18 à 20 Hz. Ces valeurs, relevées pour le français, sont tout à fait comparables à celles qui ont été notées par d'autres communautés linguistiques et il semble bien que l'on puisse les considérer comme des constantes anatomiques et physiologiques (BOË & RAKOTOFIRINGA, 1975).
  Par contre, il apparaît que la fréquence fondamentale moyenne, toutes choses égales par ailleurs, situation de communication (monologue, dialogue, échange téléphonique), contenu de la communication, état nerveux et affectif, etc., reste relativement stable pour un même locuteur (BOË & al., 1975); l'âge ne commence à la modifier, et d'ailleurs que peu, qu'après la soixantaine.

- La durée de réalisation d'un même énoncé peut présenter de notables différences interlocuteur mais aussi intralocuteur. Celles-ci sont dues essentiellement :

- au débit d'élocution,
- à la place et à l'importance de l'accent qui affecte notablement la durée de la partie "stable" des voyelles.

Une observation des réalisations d'une même phrase permet de constater qu'une simple homothétie ne peut permettre de réduire cette distorsion temporelle, celle-ci n'étant pas linéaire.

Dans ce travail, nous avons développé et testé l'efficacité de procédures permettant de normaliser temporellement et/ou fréquentiellement les variations du fondamental correspondant à une même phrase réalisée par un même locuteur ou par des locuteurs différents.

II - PROCEDURE

Pour opérer une transformation non linéaire permettant d'ajuster ainsi les contours de F_0, nous avons choisi la programmation dynamique. Cette technique, développée initialement en recherche opérationnelle, a été introduite à la fin des années 60 par SAKOE & CHIBA dans le domaine de la reconnaissance de la parole. Le "Dynamic Time Warping" a été amélioré depuis par de nombreux travaux (VELICHKO & ZAGORUYKO, 1970; ITAKURA, 1975; WHITE & NEELY, 1975; RABINER & al., 1978; SAKOE & CHIBA, 1979). Les différences entre ces méthodes sont dues pour l'essentiel :

- aux définitions des variables de décision,
- aux contraintes (fenêtres),
- à la distance choisie.

Il existe également des procédures plus simples nécessitant moins de calculs mais donnant des résultats à peu près équivalents. Elles consistent à ne retenir du treillis qu'une cellule de quelques éléments et à n'effectuer la comparaison dynamique que sur celle-ci. Au lieu de faire directement une évaluation globale, on progresse ainsi, pas à pas (schéma 1).

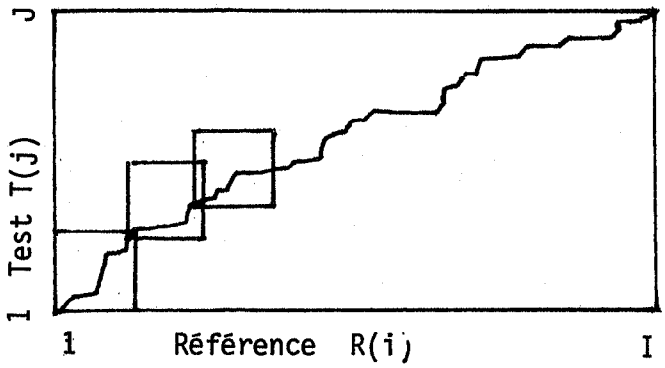


Schéma 1 - Progression pas à pas.

Etant donnés deux contours de F_0 :

. l'un de référence $R(i)$ $1 < i < I$
. l'autre test $T(j)$ $1 < j < J$

il s'agit de trouver une suite de points $c(k) = (i(k), j(k))$ de façon à ce

que les écarts temporels entre R(i(k)) et T(j(k)) soient effacés.

La suite de points c(k) doit satisfaire aux conditions et contraintes suivantes :

- Conditions limites :

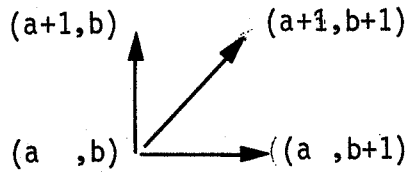
$$i(1) = 1 \qquad i(K) = I$$
$$\text{et}$$
$$j(1) = 1 \qquad j(K) = J$$

- Conditions de monotonicité (pour respecter le sens de l'"écoulement du temps") :

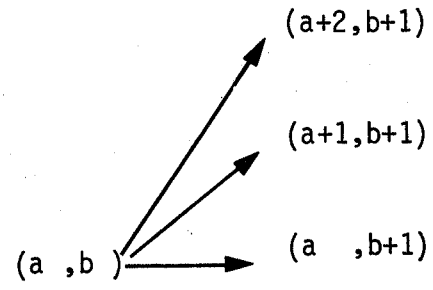
$$i(k-1) \leqslant i(k)$$
$$j(k-1) \leqslant j(k)$$

- Conditions de continuité :
  La continuité est définie par le mode de progression choisi. Les plus usité sont :

  . le mode symétrique (SAKOE & CHIBA)



  . le mode assymétrique (RABINER & al.)



- Contraintes de parcours
  Etant donné que les écarts temporels entre deux réalisations ne sont pas excessifs, il faut limiter le chemin dans une "fenêtre temporelle" autour de la diagonale.

  Ainsi SAKOE & CHIBA ont défini une fenêtre rectangle droite (schéma 2a).

  On peut noter que la nature même de la progression assymétrique définie par RABINER & al. impose au chemin d'appartenir au parallèlogramme hachuré (schéma 2b).

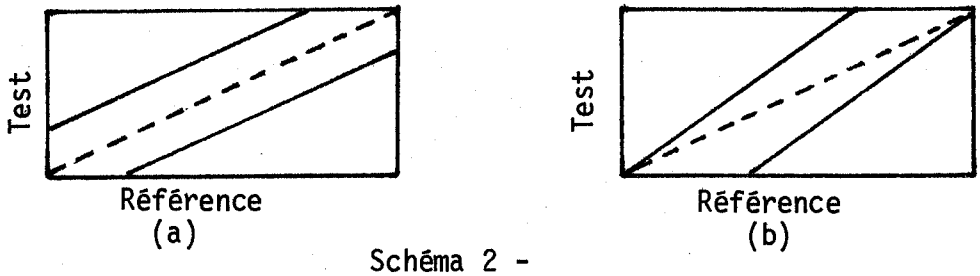


Schéma 2 -

Types de fenêtre : (a) rectangle droite
(b) parallèlogramme

- Contraintes locales

La contrainte locale sur la pente consiste en la restriction sur les possibilités de relation entre plusieurs points consécutifs. En effet une pente trop raide par exemple causerait une correspondance irréelle entre un court segment de l'axe i, et un long segment de l'axe j. On limite donc le nombre de translations successives verticales ou horizontales par exemple.

La distance totale est définie par :

$$D(T, R) = \sum_{k=1}^{k} d(c(k)) \cdot w(k) \Big/ \sum_{k=1}^{k} (w(k))$$

où $d(c(k)) = \| R(i(k)) - T(j(k)) \|$

et $w(k)$ est un coefficient de pondération associé à chaque point $c(k)$ en fonction de son antécédent.

Pour pouvoir comparer les chemins et effectivement utiliser la programmation dynamique, il faut que :

$N = \sum_{k=1}^{k} w(k)$ soit constant quel que soit le chemin suivi

. Pour SAKOE & CHIBA, ce coefficient est défini par :

$w(k) = (i(k) - i(k-1) + (j(k) - j(k-1))$

d'où $N = I + J$.

. Pour RABINER & al., la définition de la progression fait que si

$w(k) = 1$

$N = J$

Donc, on voit que, dans le premier cas, le calcul de la distance est intégré sur les deux échantillons (référence et test) alors que le second privilégie la référence.

Dans ce travail, nous avons testé l'efficacité de ces méthodes en vue d'une normalisation temporelle en fonction :

- des dimensions de la cellule (de 2 à 20 échantillons),
- de la forme de la fenêtre (schéma 3)
  - . rectangle : permet de rattraper également des décalages, de part et d'autre de la diagonale.
  - . évasée : permet en outre de combler des retards plus importants à la fin de la phrase, ce qui peut être utile lors de comparaisons interlocuteurs.
  - . losange : permet d'ajuster le début et la fin de la phrase tout en laissant au milieu assez de liberté au chemin (comparaison interlocuteurs).
- de la largeur de la fenêtre,
- du type de progression.

La comparaison dans une cellule n'est en effet pas vraiment une progression dynamique dans la mesure où, si le point initial est bien fixé, le point final est libre de se déplacer sur les bords de la cellule (nous avons en effet pu constater qu'imposer le point (N,M) comme état final revenait à imposer la diagonale sur tout le parcours). Nous avons défini (schéma 4) :

- une progression symétrique : pour une cellule, le point final est le point le plus proche du point initial appartenant aux bords de la cellule.
- une progression dissymétrique : le point final est le point le plus proche du point initial appartenant à la dernière colonne. Ceci se justifie par le fait qu'effectuant une normalisation, il nous paraissait judicieux de privilégier la référence.

La courbe normalisée est alors ainsi construite :

Etant donné le treillis et le chemin optimum précédemment définis, on donne pour valeur à l'échantillon N(i) de la courbe normalisée (c.a.d. résultant de la normalisation), la valeur de la courbe test T(j) correspondant à l'échantillon (schéma 5).

III - MOYENS ET RESULTATS

L'Institut de Phonétique de Grenoble dispose d'un miniordinateur LSI 11/2 (Digital) relié à un Vocodeur (C.I.T. - Alcatel) pour lequel a été élaboré* un système logiciel permettant (COTTET & REILLER, 1979) :

- la lecture de données,
- la synthèse de phrases préalablement enregistrées.

*. Grâce à l'aide de D. DEGRYSE, P. GOYER, B. GUERIN (Laboratoire de la Communication Parlée de l'E.N.S.E.R.G.).

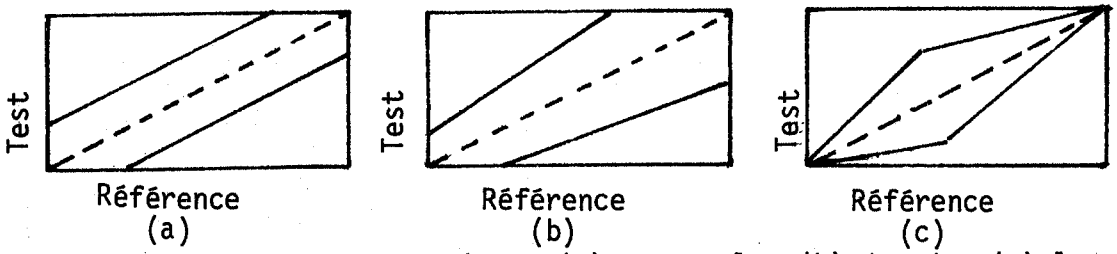
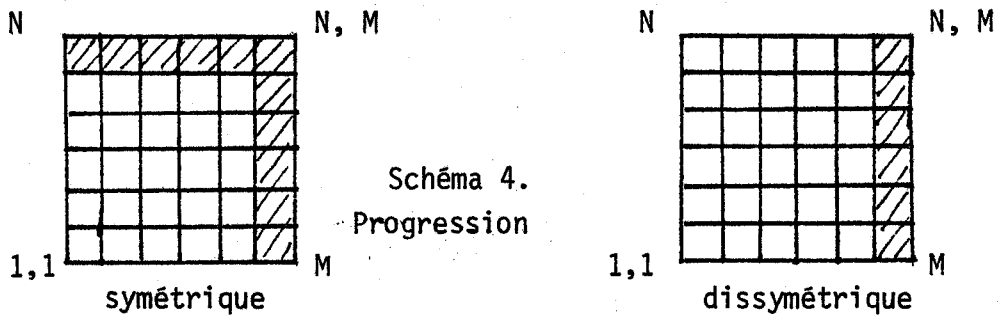


Schéma 3. Différents types de fenêtre: (a) rectangle, (b) évasée, (c) losange.



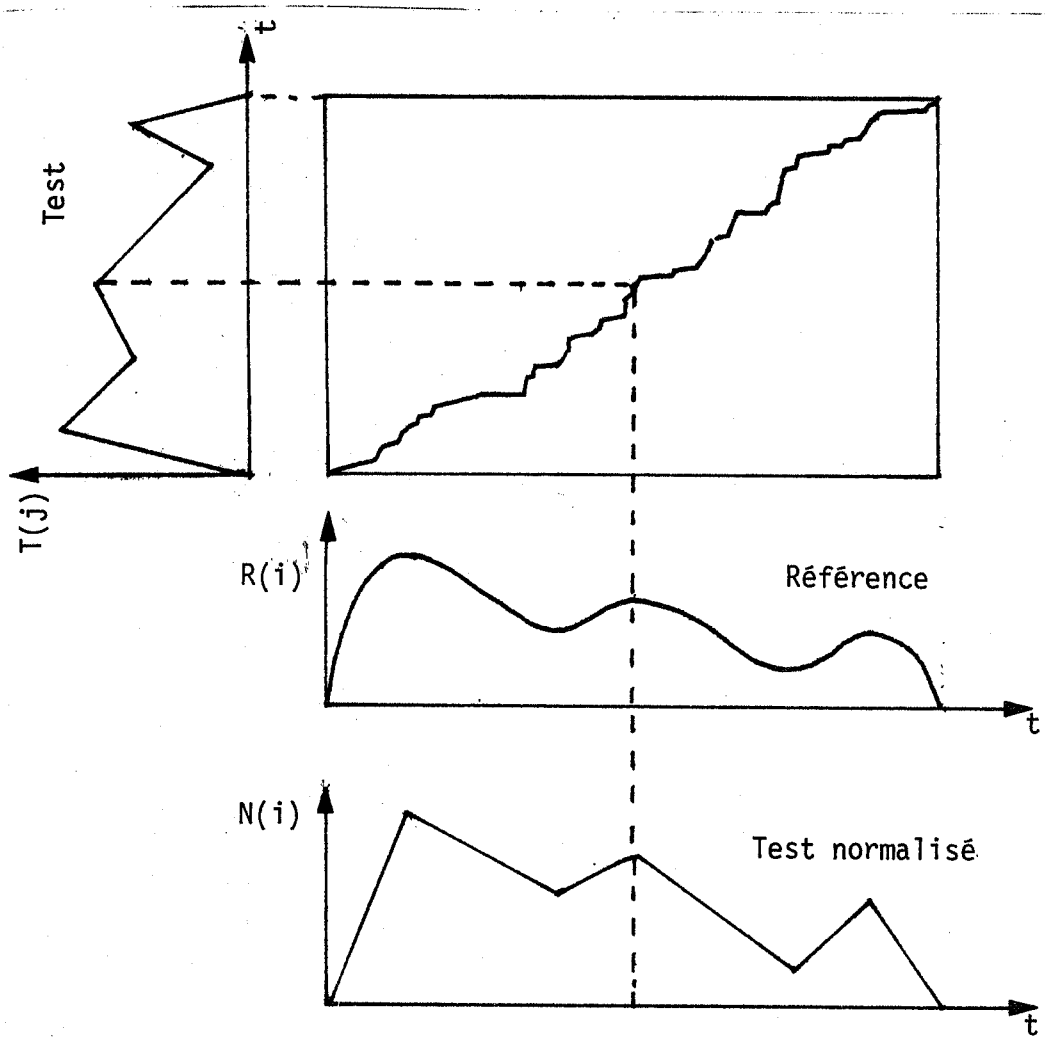


Schéma 5. Etapes de principe de la normalisation.

C'est cet appareillage que nous avons utilisé. En effet, il opère une analyse fine sur la fréquence fondamentale laryngienne.

La première étape de notre travail a consisté en la création d'un corpus de données par un programme qui opère en deux temps :

- il procède à l'acquisition du fondamental grâce au vocodeur et détermine début et fin de phrase,
- il stocke en mémoire les valeurs de F_0 correspondantes.

Nous avons choisi une phrase : "Amalia nous lira le roman" du type énonciatif qui présente un double avantage :

- elle est réalisée sans pause : ce qui dans un premier temps simplifie le problème,
- tous les sons sont voisés. Les erreurs possibles de détection de F_0 dues à la mesure opérée par le vocodeur sont ainsi limitées.

Nous avons enregistré trois hommes, dix fois chacun et quatre femmes, une fois chacune.

La normalisation intra-locuteur a porté sur les fichiers masculins, et la normalisation inter-locuteur sur l'ensemble des fichiers.

III - 1. Normalisation intra-locuteur

Seule une normalisation temporelle est effectuée.
Le procédé choisi correspond à une comparaison dynamique dans laquelle :

. la progression est symétrique,
. la taille de la cellule est de 5 échantillons,
. la largeur de la fenêtre est de 5 échantillons,
. la fenêtre est rectangulaire et droite.

Dans de telles conditions, on obtient une normalisation très satisfaisante (figure 0 et 1). Les choix précédents s'expliquent ainsi :

- une progression symétrique permet de combler les écarts temporels dans les deux sens (avance ou retard) tandis qu'un progression assymétrique ne permet d'annuler que l'avance éventuelle qu'aurait l'échantillon test sur l'échantillon de référence (figure 2). Cette deuxième solution est donc rejetée.
- une cellule de taille plus petite correspond à une procédure qui examine autour d'un point un nombre d'échantillons représentant une durée inférieure au décalage qu'il peut y avoir d'une réalisation à l'autre et ne peut donc pas la rattraper.

La taille choisie, 5 échantillons, correspond à peu près à la durée d'un demi-son (50 ms - l'échantillonnage de F_0 se faisant à 80 Hz) et à l'écart maximum qu'il y a généralement, pour un même locuteur, d'une forme à une autre.

Une cellule de taille plus grande n'améliore sensiblement pas les résultats (figure 3) mais alourdit inutilement le temps de calcul car un plus grand nombre de chemins sont examinés.

. Pour une cellule de taille 5 : temps moyen de calcul 3 s.
Pour une cellule de taille 20 : temps moyen de calcul 6 s.

. Pour une cellule de taille 5 : distance moyenne test/réf. : 466
Pour une cellule de taille 20 : distance moyenne test/réf. : 432

- Une fenêtre de type évasée ou en forme de losange est inutile. En effet, pour un même locuteur, les écarts temporels ne sont pas plus importants à la fin de la phrase qu'au milieu et ne risquent pas de s'accumuler.
- Une largeur de fenêtre plus importante ne convient pas non plus. En effet, le chemin est alors plus libre de diverger et certains écarts fréquentiels sont comblés. L'échantillon test est donc déformé et aucune comparaison n'est alors valable (figure 4).

La taille de la fenêtre choisie correspond aussi à la durée d'une demi-syllabe de part et d'autre de la diagonale.

III - 2. Normalisation inter-locuteur

D'un locuteur à l'autre, les réalisations de F_0 changent d'allure. Cependant une procédure de comparaison dynamique (quelles que soient les contraintes et le type de progression) ne permet pas d'affirmer que le locuteur test est ou n'est pas le locuteur de référence (Vérification du locuteur). En effet, les scores (coûts totaux associés à la comparaison dynamique), significatifs en reconnaissance de parole (VESIN & JOUANNELLE) sont ici du même ordre de grandeur.

La programmation dynamique équivaut à une reconnaissance de forme, et les formes en présence sont ici trop proches pour pouvoir être différenciées selon cette procédure.

III - 2.1. Normalisation fréquentielle

On définit un registre pour chaque locuteur

$$F_0 \in [\ \overline{F}_0 - 2\sigma,\ \overline{F}_0 + 2\sigma\ ]$$

Pour effectuer cette normalisation, un algorithme transforme une réalisation de F_0 en une courbe de même forme mais :

. de moyenne 118 Hz,

. d'écart type σ = 20 Hz

III - 2.2. Normalisation temporelle

La procédure choisie pour effectuer cette normalisation est la même que dans le cas d'une comparaison inter-locuteur. C'est-à-dire une programmation dynamique dans laquelle :

- la progression est symétrique
- la fenêtre est rectangle droite,
- la largeur de la fenêtre est de 5 échantillons,
- la taille de la cellule est de 5 échantillons.

(figures 5 et 6).

IV - CONCLUSIONS

L'ensemble de l'étude a montré que, pour une normalisation temporelle utilisant le principe de la programmation dynamique, le choix de la fenêtre est primordial.

Les fenêtres larges (quelles que soient leurs formes) ne permettent pas d'obtenir des résultats satisfaisants parce que, de même qu'une programmation dynamique pure, elles ajustent au mieux la courbe test à la courbe référence et par conséquent ne se contentent pas de la déformer dans l'échelle temps, ce qui est souhaité, mais aussi dans l'échelle fréquentielle.

Il est donc nécessaire de limiter le chemin de manière assez stricte autour de la diagonale (fenêtre droite de largeur 5 échantillons, ici 50 ms.)

La fenêtre finalement choisie étant étroite, la taille de la cellule n'a alors que peu d'influence sur le résultat final puisque de toute façon, le chemin est fortement limité.

La taille choisie (5 échantillons - un demi-son) est un compromis entre :

- la méthode pas à pas,
- la programmation dynamique.

Elle présente l'avantage d'être pour la normalisation intra-locuteur, plus performante que la première, et dans tous les cas, plus rapide et plus simple que la seconde procédure.

L'algorithme choisi compare donc "son-à-son" les deux réalisations de F_0 ce qui semble indiquer que du point de vue de ce paramètre deux sons successifs sont décorrélés.

La difficulté majeure qu'il reste à résoudre est de définir, pour une comparaison inter-locuteur, un critère de qualité de la normalisation étudiée.

En effet, aucune grandeur mesurable classique (distance, corrélation...) ne peut directement en rendre compte car l'objectif est de rapprocher la courbe test de la courbe référence selon un critère perceptif que seule une série de tests (synthèse des phrases normalisées) peut nous amener à préciser.

Ces résultats sont destinés à être utilisés pour :
- l'étude des variations intra-locuteur (CONTINI & al., 1981) et inter-locuteur (BELLET & al., 1981);
- la détermination de "schémas intonatifs type " pour le français;
- la reconnaissance du locuteur (VESIN & JOUANNELLE, 1980);
- la reconnaissance de la parole tenant compte de paramètres prosodiques;
- la synthèse de la parole.

BIBLIOGRAPHIE

BELLET (G.) & al., 1981, *Normalisation temporelle de F_0 inter- et intra-locuteur en vue de la synthèse et de la reconnaissance.* - Symposium on Intonation - Toronto.

BELLMAN (R.), 1957, *Dynamic programming.* - Princeton University Press.

BOË (L.J.), CONTINI (M.) & RAKOTOFIRINGA (H.), 1975, *Etude statistique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques du français.* - Phonetica 32, p. 1-23.

BOË (L.J.) & RAKOTOFIRINGA (H.), 1972, *A statistical analysis of laryngeal frequency : its relationship to intensity level and duration.* - Language & Speech (1975), 18, p. 1-13.

CONTINI (M.) & al., 1981, *Variations intralocuteurs de F_0 dans la réalisation d'une phrase énonciative en français : normalisation temporelle. Méthode et résultats.* - 12e J.E.P., Montréal.

CHEVALIER (A.) 1977, *La programmation dynamique.* - Dunod, Paris.

COELHO (M.F.) & FLEURY (M.P.), 1980, *Système d'apprentissage et de reconnaissance de la parole.* - Rapport de stage - Institut de Programmation de Grenoble. ENSERG.

COTTET (O.) & REILLER (J.P.), 1979, *Acquisition, manipulation et traitement de données vocodeur.* - Projet de fin d'Etudes. ENSIEG - Institut de Phonétique de Grenoble.

HATON (J.P. 1974 , *Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole.* - Thèse Etat - Université Nancy I.

HUNT (M.J.), LENNIG (M.) & MERMELSTEIN (P.), 1979, *Use of dynamic programming in a syllable-based continuous speech recognition system.* - Bell Northern Research, Montreal.

ITAKURA (F.), 1975 , *Minimum prediction residual principle applied to speech recognition.* - IEEE Trans. Acoustics, Speech and Signal Processing ASSP-23, n°1, p. 67-71.

JOUANNELLE (G.), 1980, *Programmation dynamique et normalisation intra/inter locuteur.* - Rapport de stage ENSIEG - Institut de Phonétique de Grenoble.

LIENARD (J.S.), CASTELLENGO (M.), LEIPP (E.), MLOUKA (M.), RENARD (G.), SAPALY (J.), TEIL (D.), 1973, *Quelques idées directrices en reconnaissance automatique de la parole.* - Automatisme 3, p. 89-98.

RABINER (R.), ROSENBERG (A.), LEVINSON (S.), 1978, *Considerations in dynamic Time warping algorithms for discrete words recognition.* - IEEE Trans. Acoustics, Speech, Signal Processing ASSP-26 n°6, p. 575-582.

SAKOE (H.) & CHIBA (S.), 1971, *A dynamic programming approach to continuous speech recognition.* - 7th Int. Congr. Acoust. Paper 20C13.

SAKOE (H.) & CHIBA (S.), 1979, *Dynamic programming algorithm optimization for spoken word recognition.* - IEEE Trans. Acoustics, Speech, Signal Processing ASSP-27, n°6, p. 43-49.

SMITH (A.R. & SAMBUR (M.R.), 1977, *Hypothesizing and verifying words for speech recognition.* - in Trends in Speech Recognition LEA W.A. ed. Prentice Hall, London.

VESIN (J.M.) & JOUANNELLE (G.), 1980, *Vérification du locuteur.* - Rapport de stage de fin d'Etudes. ENSIEG - Institut de Phonétique de Grenoble.

VELICHKO (V.M.) & ZAGORUYKO (N.G.), 1970, *Automatic recognition of 200 words.* - Inst. J. Man-Machine Stud. 2, 223.

WHITE (G.) & NEELY (R.), 1976, *Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming.* - IEEE Trans Acoustic, Speech, Signal Processing ASSP-24, p. 183-188.

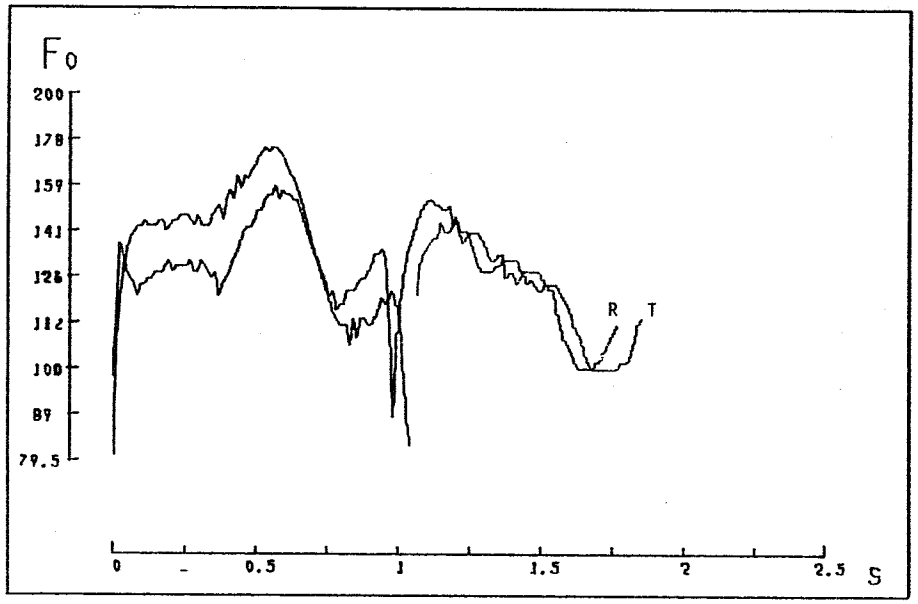


Figure 0. Normalisation intra locuteur.
Courbe de référence (R). Courbe test (T).

Intra speaker normalization.
Reference (R). Test (T).

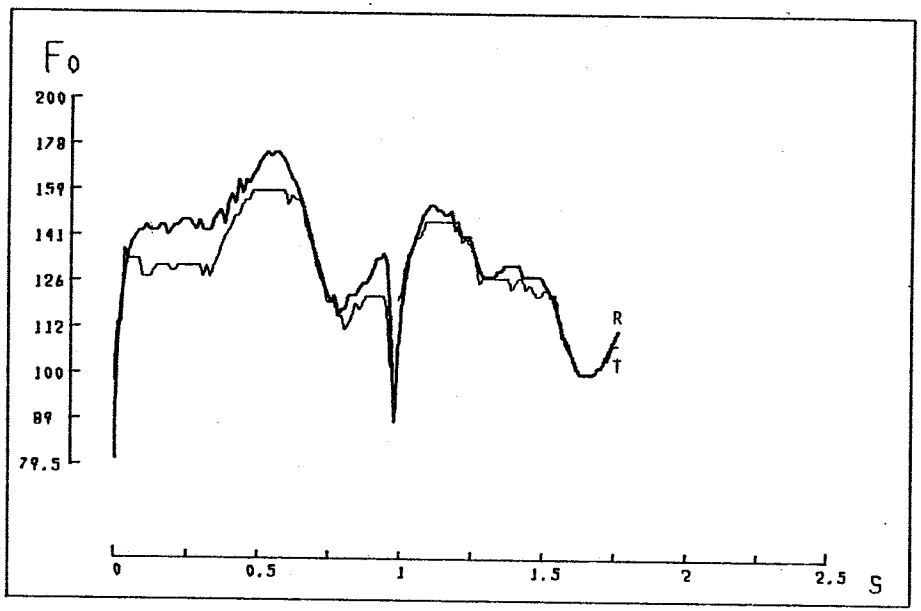


Figure 1. Normalisation intra locuteur: courbe de référence et courbe test normalisée.
Fenêtre : rectangle ; largeur de la fenêtre : 5 éch.
Largeur cellule : 5 éch. ; Progression : symétrique.

Intra speaker normalization: reference (R) and normalized test pattern (T).

Window : rectangular ; window size : 5 samples.
Step size : 5 samples; Form : symetric.

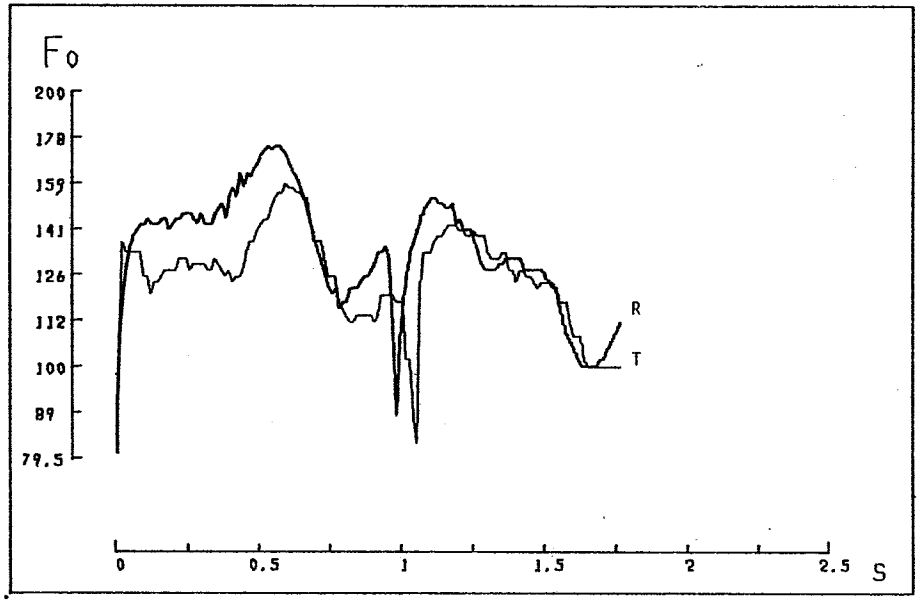


Figure 2. Normalisation intra locuteur.
Référence (R) et courbe test normalisée (T).
Fenêtre : rectangle; largeur fenêtre : 5 éch.
Largeur cellule : 5 éch. ; Progression :asymétrique.

Intra speaker normalization : reference (R) and normalized test pattern (T).
Window : rectangular; window size : 5 samples.
Step size : 5 samples; Form : asymetric.

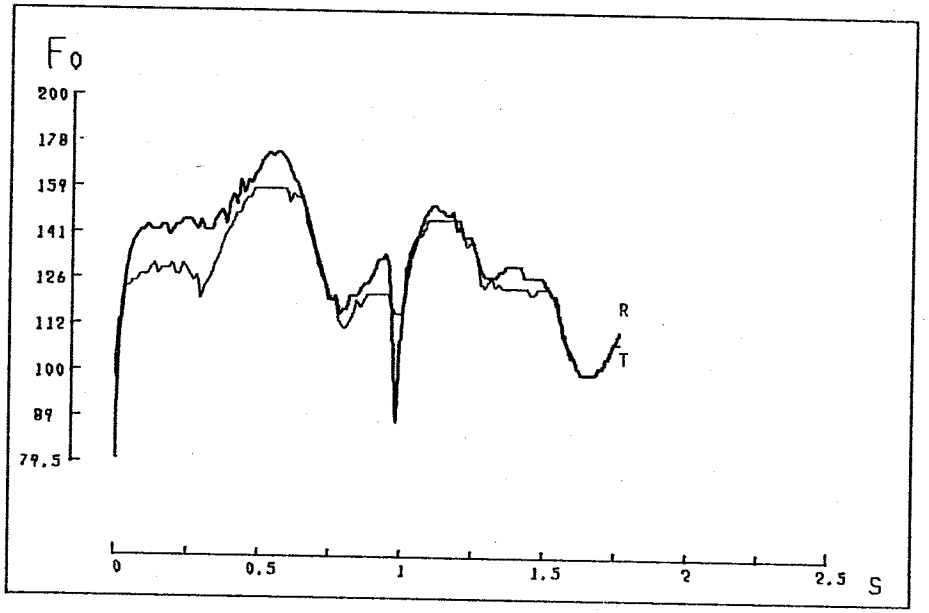


Figure 3. Normalisation intra locuteur.
Référence (R) et courbe test normalisée (T).
Fenêtre : rectangle; largeur fenêtre : 5 éch.
Largeur cellule : 20 éch.; Progression : symétrique

Intra speaker normalization : reference (R) and normalized test pattern (T).
Window : rectangular; window size : 5 samples.
Step size : 20 samples; Form : symetric.

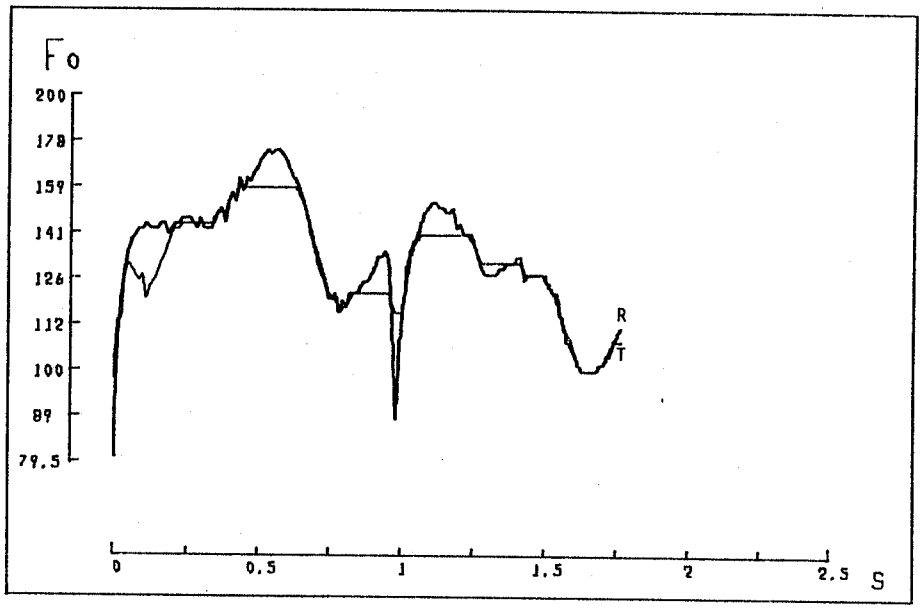


Figure 4. Normalisation intra locuteur.
Référence (R) et courbe normalisée (T);
Fenêtre : rectangle; largeur de la fenêtre : 20 éch.
Largeur cellule : 5 éch. ; Progression : symétrique.
Intra speaker normalization. Reference (R) and normalized pattern (T).
Window : rectangular ; window size : 20 samples
Form : symetric.

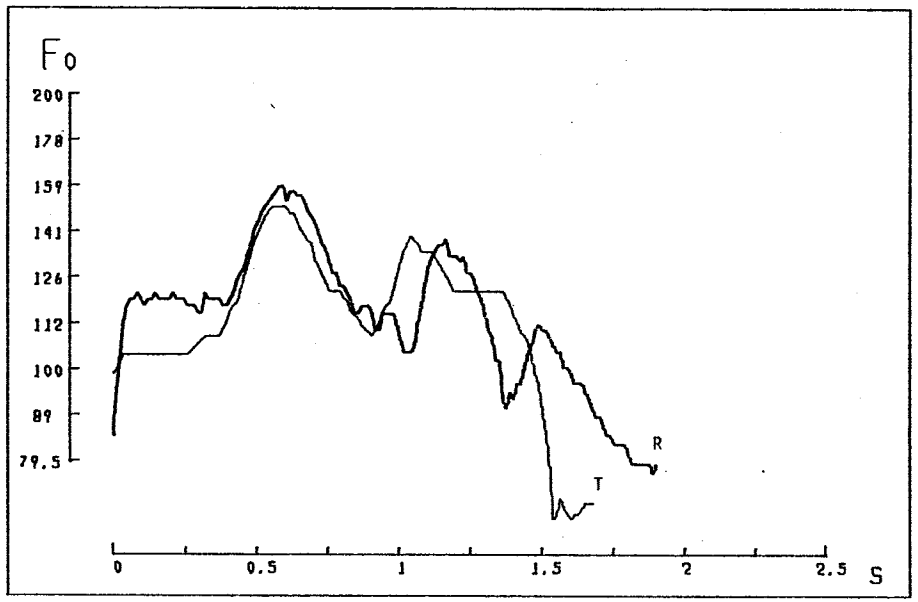


Figure 5. Normalisation inter locuteur.
Référence (R) et courbe test (T).

Inter speaker normalization.
Reference (R) and test pattern (T).

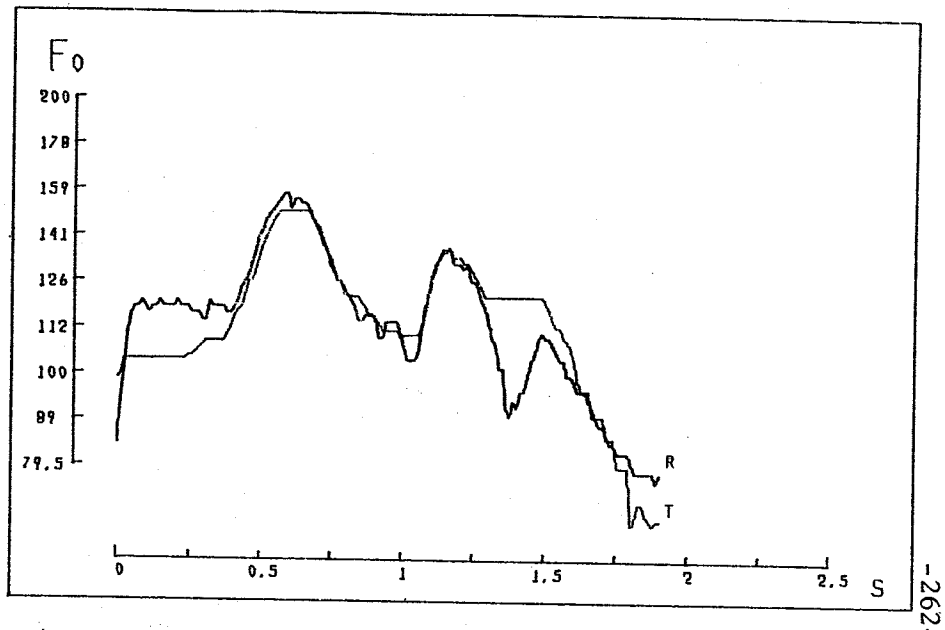


Figure 6. Normalisation inter locuteur.
Référence (R) et courbe test (T) normalisée.
Fenêtre : rectangle; largeur fenêtre : 5 éch.
Largeur cellule : 5 éch.; Progression : symétrique.

Inter speaker normalisation.: reference (R) and normalized test pattern (T).
Window : rectangular; window size : 5 samples.
Step size : 5 samples; Form : symetric.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ANALYSE IMPULSIONNELLE DE LA PAROLE : EXPERIENCES PRELIMINAIRES

J.S. LIENARD et F. MANCERON

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

(L.I.M.S.I. - C.N.R.S.) B.P. 30 - 91406 ORSAY Cedex FRANCE

.RESUME

Les méthodes usuelles d'analyse du signal de parole fournissent une évaluation du spectre sur un intervalle de temps de 20 à 40 ms, plus long que la plus longue période de voisement, qui élimine beaucoup d'information phonétiquement pertinente. Notre approche consiste à décomposer le signal en une séquence d'impulsions, définie dans le domaine spectro-temporel. Le travail décrit dans le présent article utilise un banc de filtres suivis d'intégrateurs à faible constante de temps. Une "fonction de cohérence impulsionnelle" est définie ; ses maximums sont utilisés pour marquer les débuts des impulsions. Les expériences de synthèse faites à partir de cette fonction montrent que les phénomènes brefs (explosions des occlusives), les bruits (des constrictions), le voisement et F_o sont correctement reproduits.

. ABSTRACT

The usual methods for speech signal analysis give a spectrum evaluation on a time interval (20 to 40 ms) larger than the largest pitch period, which eliminates most phonetically relevant information. Our approach consists in a decomposition of the signal into a string of "pulses" defined in the frequency-time domain. The work described in the present paper is done using a filterbank followed by short-time integrators. A "pulse coherence function" is defined ; its maxima are used to mark the beginnings of the pulses. The synthesis experiments made using this function show that short acoustical phenomena (of plosives), noises (of fricatives), voicing and pitch are correctly reproduced.

I. INTRODUCTION

Pour étudier, transmettre ou reconnaître le signal de parole on pratique habituellement une analyse spectrale, ou une transformation qui peut s'y ramener, dont les définitions en temps et en fréquence sont liées par une loi de réciprocité. Pour éviter que l'estimation instantanée du spectre ne soit perturbée par les variations rapides dues au fondamental on est amené à analyser à chaque instant une tranche de signal de durée supérieure à la plus longue période laryngienne, soit au moins 20 millisecondes, pendant lesquelles on considère le spectre comme stationnaire, et le fondamental est mesuré indépendamment du spectre. Que l'analyse spectrale soit faite par FFT, LPC ou par un banc de filtres, le réglage des paramètres d'analyse (largeur de fenêtre ou fréquence de coupure des intégrateurs) revient toujours à considérer que le spectre évolue lentement.

Cette approche traditionnelle, qui remonte au VOCODER de DUDLEY (1939), est satisfaisante en première approximation, mais présente certains inconvénients : les bruits d'explosion des consonnes constrictives sont éliminés ou fortement atténués, ainsi que les variations rapides du spectre que l'on peut observer dans certaines transitions liées aux consonnes ; de plus le spectre mesuré intègre indistinctement tout phénomène acoustique présent dans le signal, qu'il représente de la parole ou du bruit parasite ; enfin le degré de voisement et la fréquence F_0 sont mesurés indépendamment du spectre de parole, ce qui peut être rapproché des problèmes bien connus que pose l'évaluation de ces paramètres.

C'est pour tenter, à terme, d'améliorer cette situation que nous proposons de considérer le signal de parole comme une séquence d'impulsions ayant chacune leur spectre propre. Nous reprenons ainsi une idée émise par J.C. LAFON (1958). Lorsqu'on observe un sonagramme établi en bande large ($\Delta F = 300$ Hz), on constate que, dans les parties voisées, chaque impulsion glottique se manifeste par une trace verticale distincte, d'intensité plus forte dans les zones formantiques. Ceci tant que la fréquence de récurrence des impulsions (F_0) reste inférieure à une certaine valeur (environ 300 Hz), au-delà de laquelle elles fusionnent et donnent lieu à une structure de raies spectrales (et dans ce cas F_0 est repéré par la fréquence de la première de ces raies). Dans les parties bruitées, l'analyse en bande large fait apparaître des traces verticales, dont l'intensité, la fréquence centrale et la largeur temporelle ou fréquentielle sont distribuées de manière non périodique : ces impulsions représentent alors les "pointes de bruit". Enfin certaines impulsions apparaissent isolément, et correspondent par exemple aux explosions des consonnes occlusives.

Nous définirons donc une impulsion dans le domaine spectro-temporel : c'est la portion de signal donnant lieu à une énergie maximum dans plusieurs zones spectrales simultanément, ou dans une seule zone spectrale suffisamment large.

Les termes de cette définition sont volontairement peu explicites. L'énergie d'un signal dans une zone spectrale donnée dépend de la largeur de cette zone et de l'intervalle de temps sur lequel le spectre a été évalué. Si cet intervalle se trouve être plus petit que la période correspondant à une zone fréquentielle donnée, la notion d'énergie spectrale devient mal définie. La notion de simultanéité est également relative à un certain intervalle de temps. Enfin le nombre et la largeur des zones spectrales considérées dans la définition sont également à préciser.

Nous décrirons, dans le présent article, les modalités d'analyse qui nous ont permis de repérer les impulsions dans un signal de parole, et de reconstituer celui-ci. Notre démarche possède une certaine parenté avec celles qui considèrent l'oreille essentiellement comme un analyseur temporel (par exemple

LEIPP, 1971 et 1978), avec celles qui modélisent l'appareil auditif (par exemple DOLMAZON et coll., 1977, et J. CAELEN, 1979), et avec celles qui, pour mesurer F_o , utilisent les ondulations résiduelles dans diverses bandes spectrales (par exemple FRIEDMAN, 1979).

II. STRUCTURE D'ANALYSE ET DE SYNTHESE

Notre approche n'étant pas sans rapports avec le sonagraphe d'une part, et la modélisation de l'audition d'autre part, nous avons choisi une analyse par banc de filtres, de préférence à la FFT ou le codage prédictif. Dans une première expérimentation on s'est contenté d'extraire de l'analyse une fonction appelée "fonction de cohérence impulsionnelle" (FCI), dont les maximums indiquent les instants d'apparition des impulsions, et de comparer ces instants à ceux qui sont mis en évidence par un sonagramme en bande large. Dans d'autres expérimentations on a utilisé cette FCI pour exciter un synthétiseur, réalisant ainsi une structure analogue à celle d'un vocodeur à canaux.

II. 1. Première expérimentation (analyse et FCI)

a) Le banc de filtre

Le banc d'analyse comporte 32 filtres passe-bande, dont les fréquences centrales suivent l'échelle des Barks. Une formulation analytique de cette échelle, due à A.J. FOURCIN, est citée dans (J.S. SEVERWRIGHT, 1980)

$$f_{Hz} = 600.\sinh(Z/6)$$

(Z en Barks)

Cette échelle se situe entre l'échelle linéaire, qui donne trop d'importance à l'aigu, et l'échelle logarithmique, qui donne trop d'importance au grave. Nous utilisons 32 filtres couvrant la bande 50-5000 Hz, de largeur 1 Bark chacun - c'est-à-dire B_1 = 102 Hz à F_1 = 87 Hz, B_{15} = 203 Hz à F_{15} = 1058 Hz, B_{32} = 683 Hz à F_{32} = 4626 Hz - avec recouvrement de 0.5 Bark entre filtres successifs. Une telle précision est sans doute excédentaire, mais nous l'avons maintenue car nous ne sommes pas encore dans la phase d'optimisation du processus.

Dans la plupart de nos expériences les filtres, du 4^{e} ordre, sont constitués de deux cellules récursives du 2^{e} ordre (2 zéros, 4 pôles) à gain unité, avec pentes symétriques de 12 dB/octave. Nous pouvons les utiliser soit comme des filtres réels, en appliquant successivement les deux cellules du 2^{e} ordre au signal lu dans le même sens (temps croissant), soit comme des filtres non-causaux en appliquant les deux cellules successivement, mais en temps inversé ; dans ce dernier cas le filtrage est à déphasage nul.

Une préaccentuation de 6 dB/octave à partir de 1000 Hz est appliquée dans certaines expériences au signal avant analyse, au moyen d'un filtre non récursif.

b) Estimation de l'amplitude

Pour préserver les variations rapides du signal on est amené à estimer l'énergie du signal filtré, sur une durée de l'ordre du plus court intervalle de temps que l'on puisse rencontrer entre impulsions successives. Nous avons fixé ici cet intervalle à 1 ms, ce qui signifie qu'au-delà de F_o = 1000 Hz les impulsions successives fusionnent, et que l'approche impulsionnelle n'est plus valable.

La procédure d'estimation de l'énergie, qui consiste à redresser le signal (en double alternance) et à appliquer un filtrage passe-bas (de F_c comprise entre 500 et 1000 Hz), conduit à un résultat intéressant dans les voies élevées (c'est-à-dire un maximum par impulsion, en général), mais à une estimation erronée dans les voies dont la fréquence centrale est inférieure à 1000 Hz, chaque impulsion entraînant l'apparition de plusieurs maximums (figure 1). Des erreurs de ce type subsistent même si le redressement est effectué en simple alternance.

Pour qu'à chaque impulsion ne corresponde en général qu'un seul maximum nous avons dû recourir à un traitement plus complexe, dont le principe est donné dans la figure 2 : le signal filtré est remplacé par une interpolation linéaire entre maximums successifs, suivie d'un lissage (filtrage passe-bas non déphaseur). De cette manière on obtient, dans toutes les voies de fréquence supérieure à F_o, des signaux ne présentant en général qu'un maximum par impulsion. L'ensemble de ces signaux, qui étaient définis jusqu'à présent à chaque pas d'échantillonnage (c'est-à-dire toutes les 100 μs) peut maintenant être sous-échantillonné (défini par exemple toutes les 500 μs) et fournit une représentation analogue à un sonagramme très agrandi dans l'échelle de temps (figure 3).

On peut constater sur la figure 3 que les impulsions se détachent bien les unes des autres, mais que, avec l'agrandissement de l'échelle de temps, elles n'apparaissent pas purement verticales : on observe un retard considérable, qui peut atteindre plusieurs millisecondes, des portions d'impulsion correspondant aux formants. Un lissage vertical (selon la dimension de fréquence) permet d'atténuer ces sinuosités.

Fonction de cohérence impulsionnelle et repérage des impulsions

Pour repérer les impulsions nous avons expérimenté plusieurs méthodes. L'une consiste à sommer à chaque instant les valeurs obtenues à la sortie des filtres par le traitement précédent (ou leurs logarithmes). Cette fonction comporte bien un maximum à chaque impulsion, si le filtrage a été effectué à déphasage nul. Mais elle dépend du niveau du signal, ce qui rend difficile l'utilisation de seuils pour sélectionner les pointes correspondant aux impulsions principales. Une autre manière de faire, qui semble donner satisfaction, consiste à compter à chaque instant le nombre de voies dans lesquelles on observe une dérivée (première ou seconde) positive. Cette fonction de "cohérence impulsionnelle" est sensiblement indépendante du niveau, et présente des maximums et minimums francs. En ne considérant, dans une première approche, que les maximums supérieurs à un seuil donné (voisin de 10, dans nos exemples comportant 32 filtres), on obtient un repérage des impulsions comparable à celui que l'on peut effectuer visuellement sur un sonagramme (figure 3). On remarquera que les impulsions des plosives sont bien présentes, et que les parties bruitées du signal (fricatives sourdes, par exemple) fournissent des impulsions dont la structure temporelle est non-périodique.

Les essais précédents utilisaient des filtres d'analyse du 4^{e} ordre non déphaseurs. Un résultat presque analogue peut être obtenu avec les mêmes filtres en montage déphaseur (réalisables en temps réel), en compensant le temps de propagation dans chaque filtre (dans le cas où les bandes passantes diffèrent d'une voie à l'autre).

II. 2. Expérimentation d'analyse-synthèse

Nous avons ici mis en oeuvre un vocodeur un peu particulier (figure 4). Un premier banc de 19 filtres (inspiré du travail de HOLMES, 1978) sert à l'analyse spectrale ; un autre banc de filtres, placé en parallèle, sert à déterminer la FCI (en fait les voies d'analyse supérieures à 1300 Hz sont communes). Il est constitué de 30 à 35 filtres selon les expérimentations, dont la répartition

linéaire correspond en fait à une large prédominance de la bande aiguë. Le redressement simple alternance et le filtrage passe-bas (4^{e} ou 6^{e} ordre) à F_c = 500 Hz fournissent dans chaque voie un seul maximum par impulsion en général, puisque la plupart des filtres se situent au-dessus de 500 Hz. De la FCI obtenue on extrait une suite d'impulsions unitaires, utilisée d'une part pour déterminer les instants de mesure du spectre, d'autre part pour alimenter un banc de filtres de synthèse, analogue au banc d'analyse.La figure 5 montre les sonagrammes de la séquence originale et de la séquence reconstituée. On peut constater quelques défauts dans le bas du spectre et dans les largeurs de bande, mais ceci ne nous semble pas mettre en cause le principe même de la décomposition en impulsions, qui fournit par ailleurs une excellente restitution des phénomènes brefs, ainsi que des parties bruitées.

Dans cette expérimentation on n'a pas cherché à expliciter F_o et le degré de voisement, qui sont tous deux implicites dans la suite d'impulsions unitaires déduite de la FCI. Mais on observera que la FCI est sensiblement indépendante, non seulement du niveau sonore, mais aussi de la bande de fréquence couverte par le second banc de filtres. En particulier les expériences faites en filtrant le signal original par un passe-haut d'ordre élevé, coupant toute l'énergie au-dessous de 350 Hz, ont donné une FCI comparable et la même évolution perceptive de F_o.

III. CONCLUSION ET PERSPECTIVES

Les expérimentations décrites ci-dessus avaient pour but de montrer la validité d'une décomposition du signal de parole en impulsions. Nous avons explicité les problèmes rencontrés (estimation de l'énergie sur un court intervalle de temps, distorsion temporelle des impulsions, repérage au moyen de la fonction de cohérence impulsionnelle, synthèse à partir d'un signal d'excitation déduit de la FCI et ne différenciant pas le voisement du bruit), et montré que ceux-ci pouvaient être résolus de manière certainement non optimum. Mais l'essentiel de nos objectifs est atteint, puisque l'analyse proposée préserve les phénomènes brefs de la parole. Accessoirement nous avons mis en évidence une autre propriété de l'analyse impulsionnelle, qui réside dans la quasi-insensibilité de la FCI vis-à-vis de la bande considérée et du niveau ; cette propriété nous permet d'envisager une application à la détection de F_o dans la bande téléphonique.

Ce travail devrait se poursuivre dans une optimisation du processus, et dans la recherche explicite de F_o et du voisement à partir de la structure impulsionnelle, ce qui permettrait de réduire le débit d'information transmis.

REFERENCES

CAELEN (J.), 1979, "Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique" - Thèse d'Etat, Univ. de Toulouse.

DOLMAZON (J.M.), BASTET (L.), SHUPLJAKOV (V.S.), 1977, "A Functional Model of Peripheral Auditory System in Speech Processing" - IEEE-ICASSP, Hartford.

DUDLEY (H.), 1939, "Remaking Speech" - JASA 11, pp. 169-177.

FRIEDMAN (D.H.), 1979, "Multichannel Zero-crossing Intervals Pitch Estimation" - IEEE-ICASSP, Washington.

HOLMES (J.N.), 1980, "The JSRU Channel Vocoder" - IEE Proc. 127, PtF, n° 1.

LAFON (J.C.), 1958, "L'impulsion acoustique, unité de phonation et d'audition" - Congrès SOC. ORL, Paris.

LEIPP (E.), 1977, "La machine à écouter" - Masson, Paris.

LEIPP (E.), 1971, "Acoustique et Musique" - Masson, Paris.

SEVERWRIGHT (J.S.), HUGUES (A.R.B.), 1980, "Research in Acoustic Analysis for Speech Recognition" - Final report, contract n° F7A/252/77, THORN-EMI/JSRU.

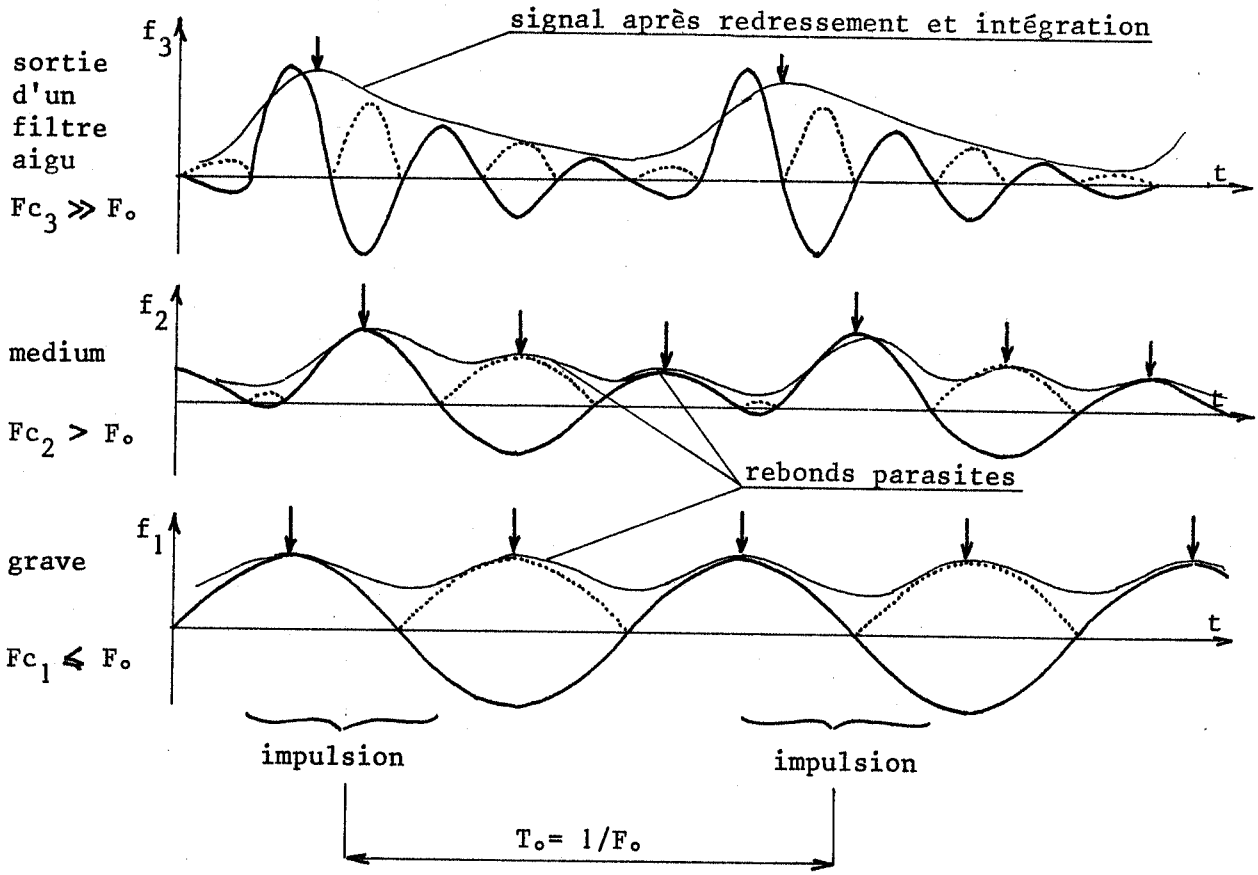


Figure 1 : Estimation de l'énergie instantanée par intégration à courte constante de temps ; problème du repérage des composantes impulsionnelles.
Estimation of instantaneous energy by short-time integration ; problem of the spotting of impulse constituents.

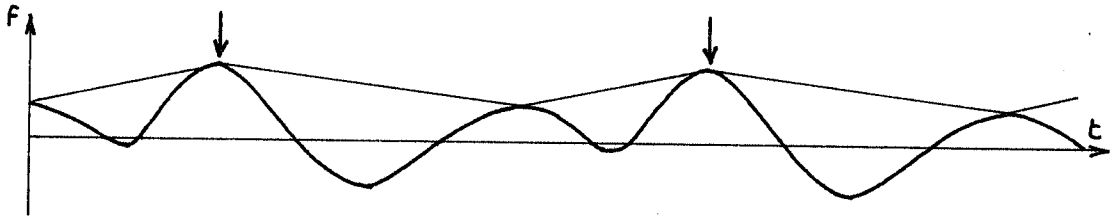


Figure 2 : Elimination des rebonds parasites par interpolation entre maximums successifs.
Rebond elimination by interpolation between successive maxima.

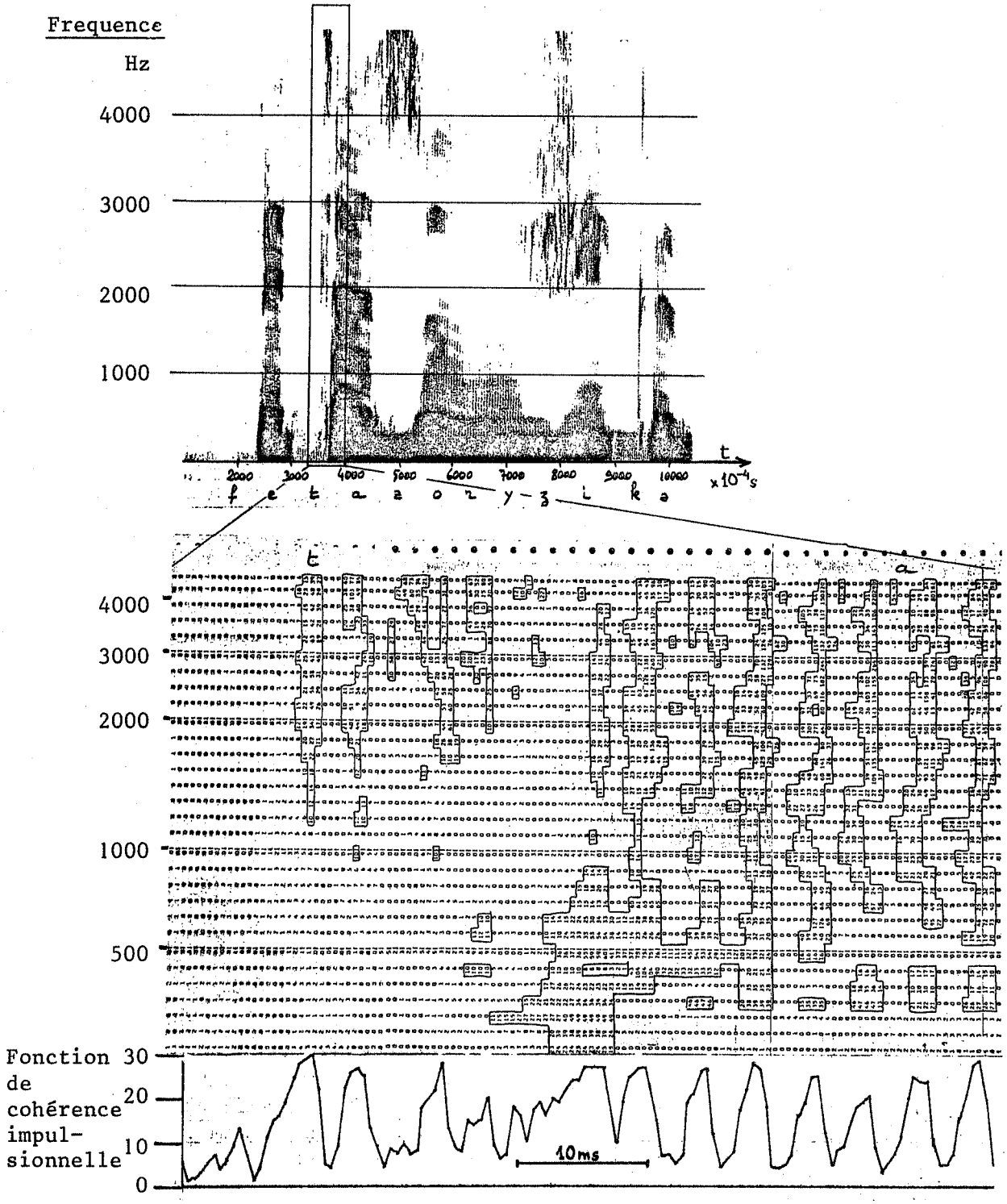


Figure 3 : Repérage des impulsions au moyen de la fonction de cohérence impulsionnelle.
en haut : sonagramme en bande large d'une séquence de parole (voix féminine).
en bas : analyse par filtres non-déphaseurs et F.C.I.

Spotting of the impulses by means of the impulse coherence function
top : wide-band spectrogram of a speech signal (female voice)
bottom : analysis by zero-phase filters and impulse coherence function

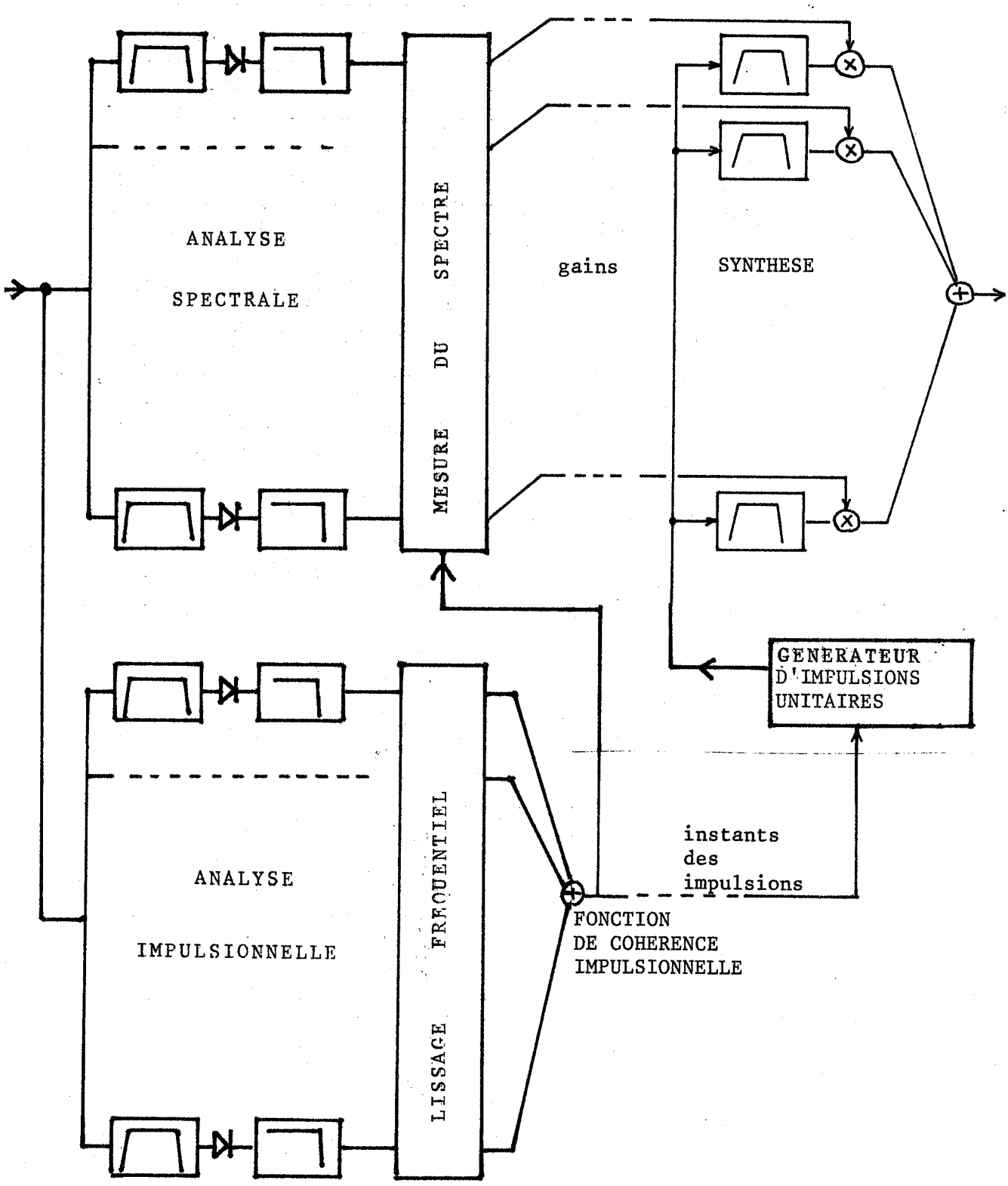


Figure 4: Structure d'analyse - synthèse à canaux , utilisant la fonction de cohérence impulsionnelle.
Analysis - synthesis structure , using the impulse - coherence function.

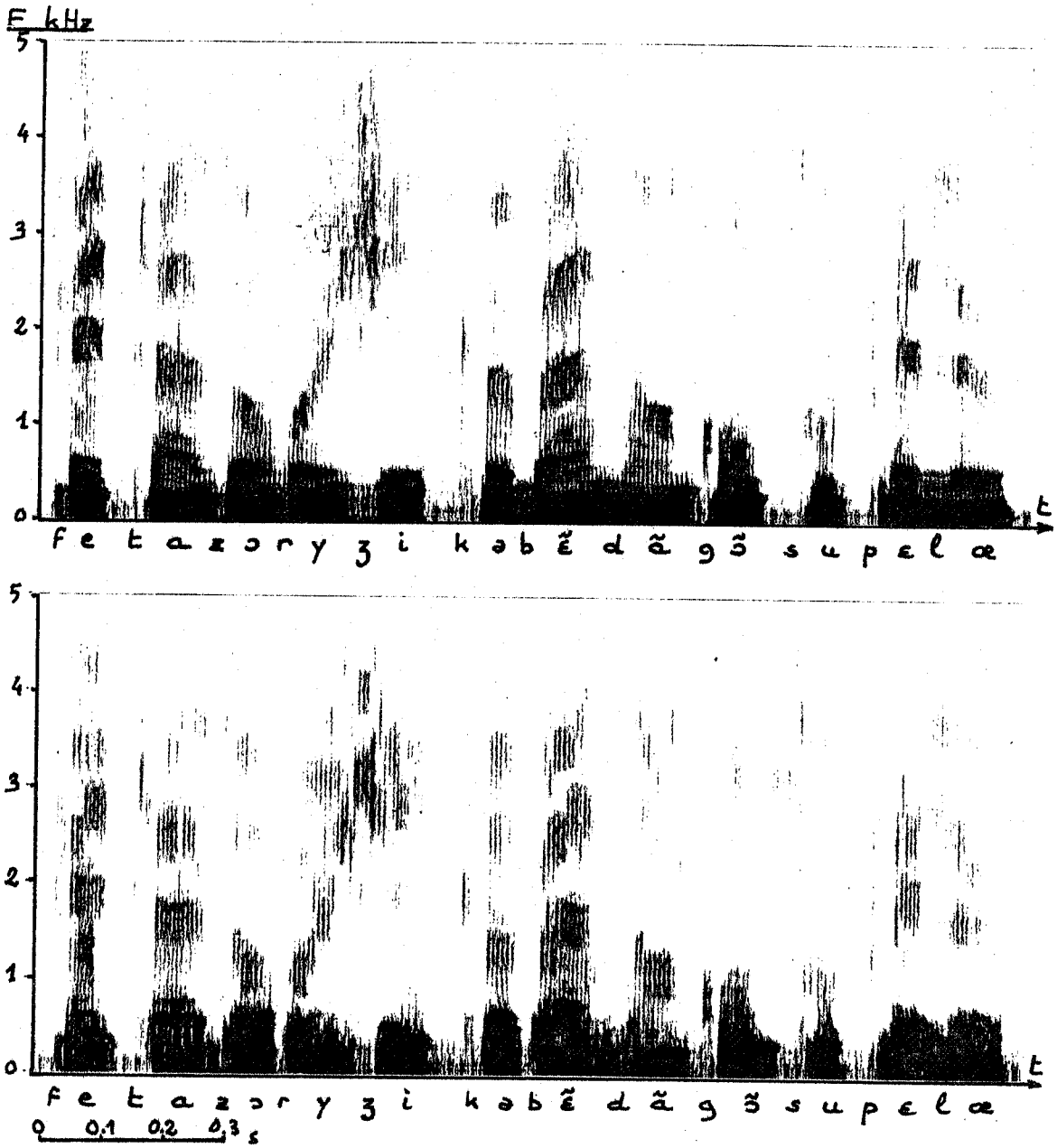


Figure 5 : Synthèse par banc de filtres après analyse impulsionnelle (voix masculine). En haut : original. En bas : synthétique. Les irrégularités apparaissant dans les fréquences graves sont dues à la synthèse et non à l'analyse.

Synthesis using a filterbank after impulse analysis (male voice). Top :original . Bottom : synthesis. The irregularities appearing in the low frequencies are due to the synthesis , not to the analysis.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

VARIATIONS INTRALOCUTEURS DE F_0 DANS LA REALISATION D'UNE PHRASE ENONCIATIVE EN FRANCAIS : NORMALISATION TEMPORELLE.

Michel CONTINI - Gaston BELLET - Louis-Jean BOË - Institut de Phonétique de Grenoble

. RESUME

Dans la perspective de l'étude de la variabilité intralocuteur de F_0 cette recherche propose une méthode permettant la neutralisation des différences de durée. Dans un corpus de 47 phrases énonciatives sont retenues 16 répétitions d'une même phrase prononcées par deux locuteurs féminins. Dans une première étape, une référence est choisie parmi ces dernières en utilisant la méthode des Nuées dynamiques suivie de tests perceptifs sur les phrases retenues; cette référence permet ensuite une normalisation temporelle par programmation dynamique.

Ainsi il est possible de comparer les seules différences de F_0 pour l'ensemble des répétitions d'une même phrase par un locuteur donné.

. ABSTRACT

In this paper the authors propose a method to describe intra-speaker variability in fundamental voice frequency (F_0) using a time normalization procedure to eliminate the influence of duration.

This study has been conducted on enunciative French sentences uttered by two speakers.

In a first phase a reference sentence is chosen from sixteen utterances using a dynamic clusters method followed by a neutrality perception test on the sentences retained.

This reference is then used to normalize duration of all the other sentences by dynamic programming : it is thus possible to compare the different repetitions of the same sentence for a given speaker.

VARIATIONS INTRALOCUTEURS DE F_0 DANS LA REALISATION D'UNE PHRASE ENONCIATIVE EN FRANCAIS : NORMALISATION TEMPORELLE. Méthode et résultats.

Michel CONTINI
Gaston BELLET
Louis-Jean BOË

Institut de Phonétique de Grenoble

I - INTRODUCTION

Les procédures de normalisation sur le plan segmental commencent à être bien étudiées. On pourrait se référer, entre autres, aux présentations de FERRARI-DISNER (1979, 1980) et de BOË & al. (1980). Des disciplines comme la socio-linguistique et la dialectologie, confrontées continuellement au problème de la variation intra- et interlocuteur, ont montré un intérêt particulier au développement des techniques de normalisation (LABOV, 1979; ABRY & BOË, 1979).

Au niveau suprasegmental, ces dernières n'ont pas fait l'objet, à notre connaissance, d'études approfondies. La plupart des travaux ont été consacrés au domaine fréquentiel (F_0 moyen et écart-type) (JASSEM, 1971; BOË & al., 1975).

Les différences temporelles, qu'elles soient intra- ou interlocuteur, sont pourtant loin d'être négligeables. Après les travaux de KAISER (1939), qui marquent le point de départ de ces études, la variabilité idiolectale a été chiffrée, à l'analyse, entre autres par FONAGY & MAGDICS (1960), GROSJEAN & DESCHAMPS (1973), LUCCI (1973 et 1974), SAINT-BONNET & BOË (1979). ATKINSON (1973, 1976) relève que les différences temporelles intra-locuteur sont de taille comparable aux différences inter-locuteur et ATAL (1972) propose ces dernières comme paramètres pertinents dans la vérification du locuteur.

Ces variations ne sont contournables qu'au prix de nombreuses précautions, notamment dans la constitution du corpus (phrases relativement courtes, si possible de durée égale, même structure syllabique, logatomes dans des phrases porteuses, phrases logatomes, etc.). Voir à ce propos les travaux de LINDBLOM & RAPP (1973), de LIBERMANN & STREETER (1976), NAKATANI & SCHAFFER (1976), LIBERMAN (1978).

Le but de ce travail est d'utiliser la programmation dynamique, bien connue des spécialistes de la reconnaissance de la parole, pour normaliser temporellement les variations de F_0.

II - PROCEDURE ET RESULTATS

II - 1. Les corpus et les locuteurs

Notre étude a été menée à partir de seize répétitions les plus neutres possibles de l'énoncé :

A MOULINS ON AIMAIT LES LILAS

insérées dans un corpus de 47 phrases affirmatives, de durée et de complexité variables. Deux répétitions successives de la phrase ci-dessus sont toujours séparées par deux phrases différentes.

Afin de réduire au maximum les variations dues aux caractéristiques intrinsèques et co-intrinsèques, le choix des consonnes a été limité aux liquides et aux nasales[1].

Le corpus a été lu par dix locuteurs féminins, âgés de 30 à 40 ans[2].

Les 16 réalisations de la phrase ci-dessus ont été analysées à l'aide d'un ensemble fréquencemètre logarithmique - mingographe (BOË & BELLET, 1979). Au vu des tracés obtenus ont été retenus les deux locuteurs présentant le maximum de réalisations sans pause, après le syntagme initial.

Les courbes ont été ensuite "lissées" afin d'éliminer au mieux les microvariations de période à période.

II - 2. Choix de la référence

Toute normalisation présuppose le choix d'une référence. Dans ce but, deux démarches sont possibles. L'une objective, par une méthode statistique donnée (par exemple, choix de la phrase se rapprochant le plus de la moyenne des phrases). L'autre perceptive, basée sur des critères linguistiques (phrase perçue comme la plus neutre, ou la plus naturelle).

La solution que nous avons adoptée tient compte des deux approches.

Dans un premier temps, par une démarche statistique, nous avons essayé de déterminer des classes parmi l'ensemble des phrases. Pour ce faire, chacune de ces dernières a été caractérisée par les points suivants : a. Attaque et fin, b. Maximums, c. Minimum(s), d. Tiers de pente entre l'attaque et le premier maximum, e. Moitié de pente entre le dernier maximum et la fin (ou le dernier minimum dans le cas d'une montée finale (voir schéma 1)).

Ce choix qui peut paraître quelque peu arbitraire a été dicté à la fois par : - des considérations objectives : décrire le contour des points caractéristiques (maximum et minimum);

- des travaux d'ordre perceptif qui ont mis en évidence l'importance du troisième tiers montant (ROSSI,1971) et de l'attaque des max et de fin (CONTINI& BOË,1979).

A partir de cette paramétrisation un programme dit des "nuées dynamiques" a été utilisé pour permettre un regroupement des énoncés en classes. Ce programme exploite la méthode d'agrégation autour de centres variables qui peut être résumée très schématiquement comme suit.

Soit un ensemble de N individus caractérisés par p paramètres. Ce sont donc n points dans un espace R^p, munis d'une distance euclidienne d. On désire constituer k classes au maximum.

1° On affecte au hasard chaque point-individu à une classe.
2° On calcule le centre de gravité de chaque classe.
3° On affecte de nouveau chaque individu à la classe dont le centre de gra-

---

[1]. D'un point de vue purement expérimental il aurait été plus judicieux de choisir les voyelles dans une même classe (ouverte ou fermée) : pour des raisons de naturel nous avons préféré utiliser des phrases plus acceptables.

[2]. Les locutrices sont des enseignantes ou des membres du personnel de l'Université de Grenoble III. Elles se caractérisent par une prononciation non marquée.

vité ou le noyau[1] sont le plus proches.

4° On recalcule le centre de gravité de chaque classe, puis on réaffecte ... On s'arrête soit parce que le nombre maximum d'itérations est atteint, soit parce qu'il y a stabilité (aucun élément n'a changé de classe entre deux itérations.

Il est évident que ces classes dépendent du tirage au sort initial. On pallie cet inconvénient en faisant plusieurs partitions et en ne retenant comme groupements stables (= formes fortes) que les ensembles d'individus qui ont toujours été affectés à une même classe dans chaque partition.

Ce procédé a permis de regrouper les énoncés en quatre classes pour le locuteur JO.EL et en trois classes pour le locuteur DO.MI, comme le montre le tableau 1.

Pour chaque classe a été retenue ensuite la phrase présentant la durée la plus proche de la durée moyenne.

Des auditeurs, au nombre de douze, ont écouté quatre passages des phrases retenues pour chaque locuteur.

A partir du deuxième (le premier étant un passage à blanc), ils devaient noter sur une grille de dépouillement la (les) phrase(s) leur paraissant le (les) plus neutre(s) ou la (les) moins marquée(s).

Précisons qu'à chaque passage l'ordre d'apparition des phrases était modifié.

La présentation du test peut être explicitée par le schéma n°2.

La phrase ayant été choisie le plus grand nombre de fois a été retenue comme référence pour le programme de normalisation temporelle.

II - 3. Normalisation temporelle de F_0

II - 3.1. Entrée des données

La courbe de référence ainsi que neuf autres courbes représentatives des classes dégagées ci-dessus, d'une durée de 1,5 s maximum, ont été acquises au moyen d'une tablette de digitalisation puis échantillonnées toutes les 7,5 ms et visualisées sur table traçante (voir la figure 1).

II - 3.2. Caractéristiques de la normalisation

Dans une étude préliminaire (JOUANNELLE & al., 1980), ont été déterminées les caractéristiques les plus satisfaisantes pour ce type de normalisation, à savoir :

- progression symétrique,
- taille de la cellule[2],
- largeur de la fenêtre[2],
- forme de la fenêtre.

---

[1]. Est appelé noyau d'une classe un ensemble de n individus les plus proches du centre de gravité de cette classe.

[2]. Pour notre étude la taille de la cellule et la largeur de la fenêtre ont été réduites à 37,5 ms pour que les courbes normalisées ne présentent de paliers manifestement peu conformes aux tracés originels.

II - 3.3. Résultats

L'analyse montre que les deux locutrices se différencient par :

a) la fréquence laryngienne moyenne (BOË & BELLET, 1979) qui se situe à 254 Hz pour JO.EL et à 232 Hz pour DO.MI. Notons aussi que chez la première le F_0 moyen, pour chaque phrase, présente une plus grande stabilité;
b) le registre, nettement plus étendu chez la locutrice JO.EL (8 tons contre 5 1/2 tons);
c) la forme de contours. Les tracés de JO.EL se caractérisent par deux maximums situés sur la syllabe finale des deux premiers syntagmes. Ceux de DO.MI ont toujours un troisième sommet sur l'avant dernière syllabe (-lilas) du syntagme final;
d) les variations temporelles. Elles sont plus importantes chez DO.MI, 60 ms au maximum, l'autre locutrice réalisant les 10 phrases avec un écart de 10 ms au plus.

Sur la figure 1 sont visualisés les 10 contours choisis pour les deux locutrices représentant la référence et les neuf phrases-test. La distance (D) entre ces dernières et la phrase de référence (R) est en moyenne de 3380 pour DO.MI et de 2340 pour JO.EL (Tableau II). C'est par rapport à ces valeurs qu'on pourra juger des performances de la procédure de normalisation pour la réduction des écarts.

La figure 2 montre la progression de la normalisation à l'intérieur de la fenêtre : elle est visualisée avec inscription du score correspondant à cette solution suboptimale. Nous remarquerons que ce dernier est plus élevé, en moyenne pour DO.MI (2540) que pour JO.EL (810) (Tableau II), ce qui reflète le comportement différent des deux locutrices dans la réalisation d'énoncés identiques.

Le résultat de ce traitement peut être apprécié au vu de la courbe normalisée et de la distance résiduelle entre phrases-test et référence. A la figure 3 est donné un exemple de normalisation pour les deux locutrices avec les courbes de référence, test et normalisée. Au-dessous sont tracées les variations de la pente d'accroissement de la distance entre :

- la phrase test et la référence,
- la phrase normalisée et la référence.

L'effet de la normalisation apparaît dans la diminution des distances.

Le résultat pour l'ensemble des neuf phrases test est présenté sur la figure 4, avec la moyenne normalisée encadrée par les courbes à plus ou moins l'écart type. On peut constater que la courbe de référence (R) est bien située à l'intérieur de cette enveloppe, ce qui confirme la validité de la double procédure de son choix et l'efficacité du programme de normalisation. Ce résultat est confirmé quantitativement au tableau II. La distance résiduelle (entre R et N) n'est plus que de 2064 pour DO.MI et de 664 pour JO.EL. Le pourcentage de réduction 100 (1-DR/D) est de 40 % pour DO.MI alors qu'il atteint 60 % pour JO.EL. Avec une fenêtre plus large, ce résultat aurait été sûrement plus élevé, mais, cela, au détriment de la forme des courbes normalisées.

Il semble intéressant d'étudier les courbes moyennes et les courbes à plus ou moins l'écart-type entre les neuf phrases-test et la référence (figure 5) et entre les neuf phrases normalisées et la référence (figure 6). Les écarts maximums se situent pour DO.MI sur le premier maximum (syllabe finale de syntagme initial), sur le deuxième (syllabe finale ou syntagme verbal) et sur la voyelle finale. Pour la deuxième locutrice les différences n'apparaissent pas localisées

de la même façon : elles sont maximales sur la voyelle initiale du deuxième syntagme qui correspond au premier minimum de F_0 et pour la voyelle finale. La normalisation réduit notablement ces différences jusqu'à les faire disparaître (deuxième maximum de JO.EL (fiure 6).

CONCLUSIONS

Cette étude nous a permis :

. D'élaborer une procédure de sélection d'une référence parmi n répétitions d'un même énoncé par un même locuteur. Cette démarche (d'un maniement relativement lourd) associe test perceptif et sélection statistique. Dans cette deuxième phase le programme des nuées dynamiques s'est révélé particulièrement souple et efficace; il pourrait être utilisé par la suite comme procédé unique pour la mise en évidence d'une phrase de référence.

. De mettre à l'épreuve un programme de normalisation temporelle. Au vu des tracés des courbes normalisées (et il est bien évident que seuls des tests perceptifs réalisés à partir de synthèse, permettront de vérifier que le naturel intonatif est conservé) et des résultats chiffrés ce procédé apparaît relativement efficace puisqu'avec les deux locutrices retenues il permet de réduire 40% à 60% des écarts, les différences résiduelles étant vraisemblablement dues à des décalages fréquentiels et non plus temporels.

. De préciser des différences intra-individuelles au niveau fréquentiel et temporel et de souligner la variabilité inter-locuteur. En particulier, on peut observer que les décalages entre les répétitions n'apparaissent pas, pour les deux locutrices sur les mêmes segments.

Ce travail va être poursuivi par une normalisation fréquentielle (qui va permettre de chiffrer et situer les résidus) et une normalisation temporelle inter-locuteur. Cet ensemble de traitement sera appliqué à un troisième locuteur à la suite de quoi sera dégagée une procédure simplifiée de choix de la référence pour les 7 autres.

REMERCIEMENTS

Les programmes d'acquisition graphique ont été implantés à l'Institut de Phonétique de Grenoble avec l'aide de G. JOUANNELLE. Ils ont été développés au CNET à Lannion; nous remercions R. DESCOUT, J. GENIN et S. MAEDA pour l'aide qu'ils nous ont apportée dans cette mise au point.

Le programme des Nuées dynamiques nous a été fourni par Y. GRENIER de l'ENST; il a été implanté par J. RAYMOND à l'Institut de Phonétique de Grenoble.

Nous remercions C. SORIN pour ses conseils relatifs à la préparation des tests.

ANNEXE

1) Matériel utilisé

- analogique
  - magnétophone NAGRA IV.
  - analyseur de F_0 de l'Institut de Phonétique de Grenoble.
  - inscripteur à jet d'encre oscillomink L. Siemens.
- informatique
  - LSI 11/02 (32 K mots d'U.C.).
  - 2 Floppy disques (2 x 256 K mots).
  - Visu graphique 4012 Tektronix.
  - Télétype ASR 33.
  - Tablette de digitalisation 4953 Tektronix.
  - Table traçante XY 7015 B Hewlett Packard

2) Logiciel

Système RT 11 V2.
Acquisition graphique (CNET).
Nuées dynamiques (ENST).
Normalisation (Institut de Phonétique).

BIBLIOGRAPHIE

ABRY (C.) & BOË (L.J.), 1979, Pour une idiolectologie : aspects phonétiques de l'identité. - Colloque Int. Production et affirmation de l'Identité. Toulouse.

ATAL (B.S.), 1972, Automatic speaker recognition based on pitch contours. - J.A.S.A. 52, p. 1687-1697.

ATKINSON (J.E.), 1973, Aspects of intonation in speech : implications from an experimental study of fundamental frequency. - Diss. Univ. of Connecticut.

ATKINSON (J.E.), 1976, Inter- and intraspeaker variability in fundamental voice frequency. - J.A.S.A. 60, pp. 440-445.

BOË (L.J.), ABRY (C.) & CORSI (P.), 1980, Les problèmes de normalisation interlocuteurs. Méthode d'ajustement aux limites. - XI° J.E.P., Groupe Communication Parlée, p. 143-162.

BOË (L.J.), BELLET (G.), 1979, Mesure et statistique de la fréquence laryngienne. Appareillage et programmes. - Bulletin de l'Institut de Phonétique de Grenoble VIII, pp. 23-34.

BOË (L.J.), CONTINI (M.) & RAKOTOFIRINGA (H.), 1975, Etude statistique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques du français. - Phonetica 32, pp. 1-23.

CONTINI (M.), BOË (L.J.), 1979, Etude quantitative de l'intonation en français. Premiers résultats. - In Recherches sur la Prosodie du Français, pp. 117-129. Institut de Phonétique de Grenoble.

DIDAY (E.), 1971, La méthode des nuées dynamiques. - Revue de Statistique Appliquée 19/2, pp. 19-34.

FERRARI-DISNER (S.), 1979, Cross-linguistic normalization. - 9th Int. Cong. Phonetic Sci. I, p. 262.

FERRARI-DISNER (S.), 1980, Evaluation of vowel normalization procedures. - J.A. S.A. 67, pp. 253-261.

FONAGY (I.) & MAGDICS (K.), 1960, Speed utterances in phrases of different lengths.- Language and Speech 3, pp. 179-192.

GROSJEAN (F.) & DESCHAMPS (A.), 1973, Analyse des variables temporelles du français spontané. - II. - Phonetica 28, pp. 191-226.

INSTITUT DE PHONETIQUE DE GRENOBLE, 1979, Recherches sur la prosodie du français.- Publications de l'Université des Langues et Lettres de Grenoble.

JASSEM (W.), 1971, Pitch and compass of speaking voice. - J. Int. Phonet. Ass. 1, pp. 59-68.

JOUANNELLE (G.), 1980, Programmation dynamique et normalisation intra / inter-locuteur. - Rapport de stage ENSIEG - Institut de Phonétique de Grenoble.

JOUANNELLE (G.), BOË (L.J.), BELLET (G.), 1980, Variation intra- et inter-locuteurs de F_0 : procédures de normalisation. Etude préliminaire. XIIe JEP, AGLF Comm. Parlée.

KAISER (L.), 1939, Biological and statistical research concerning the speech of 216 Dutch students. I. - Archives Néerlandaises de Phonétique Expérimentale 15, p. 1-76.

LABOV (W.), 1979, A sociolinguistic approach to the problem of normalization. - 9th Int. Congr. Phonetic Sci. I, p. 441.

LEBART (L.), MORINEAU (A.), FENELON (J.-P.), 1979, Traitement des données statistiques. Méthodes et programmes. - Dunod, Paris.

LIBERMAN (M.), 1978, Modeling of duration patterns in reiterant speech. - In Lingustic variation. Models and methods. Ed.by S. SANKOFF, pp. 127-138. Academic Press - New-York.

LINDBLOM (B.), & RAPP (K.), 1973, Some temporal regularities of spoken Swedish. - Pub. Inst. of Linguistics. University of Stockholm.

LUCCI (V.), 1973, Etude phonostylistique du rythme et de la variabilité de la longueur en français parlé et français lu. - Bull. Inst. Phon. Grenoble 2, pp. 139-161.

LUCCI (V.), 1974, Rythme et longueur du message parlé. La conversation. - Bull. Inst. Phon. Grenoble 3, pp. 139-152.

NAKATANI (L.H.) & SCHAFFER (J.A.), 1976, Hearing "words" without words. - Speech prosody and word perception. J.A.S.A. 60 (S1) S 28.

RAYMOND (J.), 1980, Adaptation d'un programme "des nuées dynamiques" à l'étude des caractéristiques individuelles. - Stage. Institut de Phonétique de Grenoble.

ROSSI (M.), 1971, Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. - Phonetica 23, pp. 1-33.

SAINT-BONNET (M.) & BOË (L.J.), 1979, Les pauses et les groupes rythmiques : leur durée et distribution en fonction de la vitesse d'élocution. - In Recherche sur la prosodie du français, pp. 199-208. Institut de Phonétique de Grenoble.

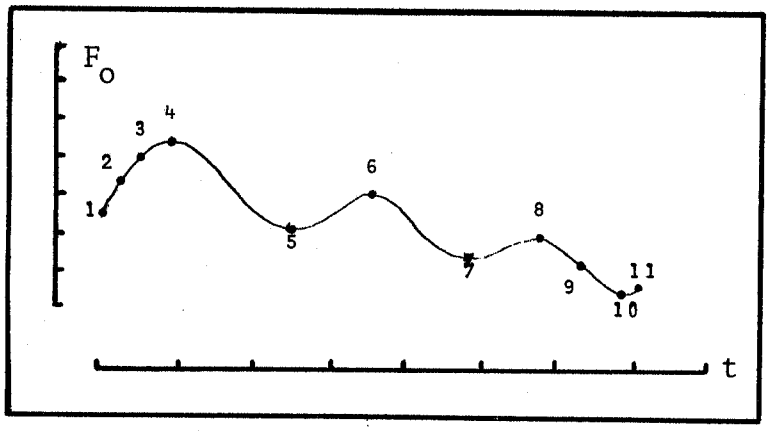


Schéma 1 - Choix des points "caractéristiques" du contour de F_o.

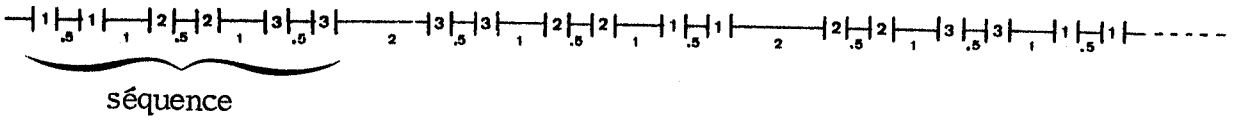


Schéma 2 - Présentation du test

1 - 2 - 3 : n° des répétitions.
0,5- 1 - 2 : temps entre chaque répétition et séquence en seconde.

LECTURE DU FICHIER RESULTAT

Test	1	0	1	3	2	2
"	7	0	1	3	2	2
Réf.	2	4	3	1	1	1
Test	3	0	3	1	1	1
"	4	0	3	1	1	1
"	6	0	3	1	1	1
"	12	0	3	1	1	1
"	8	4	2	2	3	3
"	9	0	2	2	3	3
"	10	0	2	2	3	3
"	11	0	2	2	3	3
"	13	0	2	2	3	3
"	14	0	2	2	3	3
"	15	0	2	2	3	3
"	16	0	2	2	3	3

Tableau I - Fichier résultat montrant les 3 classes retenues pour la locutrice DO.MI.

JO.EL	Phrases-test								
Référence 1	2	3	4	5	6	7	8	9	10
Distance	2164	1880	1556	2022	1930	3074	1738	2902	3822
Score	1482	634	664	700	832	1408	556	458	550
Distance résiduelle	1416	592	576	650	674	682	500	400	490
% Réduction	35	68	63	68	65	78	71	86	87

DO.MI	Phrases-test								
Référence 1	2	3	4	5	6	7	8	9	10
Distance	4364	1854	2102	1930	4536	3606	3266	4342	4388
Score	3084	1302	1398	1300	4664	2822	1692	3208	3308
Distance résiduelle	2436	844	1256	1132	3554	2420	1466	2648	2822
% Réduction	44	55	40	41	22	31	55	36	36

Tableau II - Score : évalué au cours de la programmation dynamique.
Distance (D) : entre référence et courbe-test.
Distance résiduelle (DR) : entre référence et courbe normalisée.
% de réduction : 100 (1-DR/D).

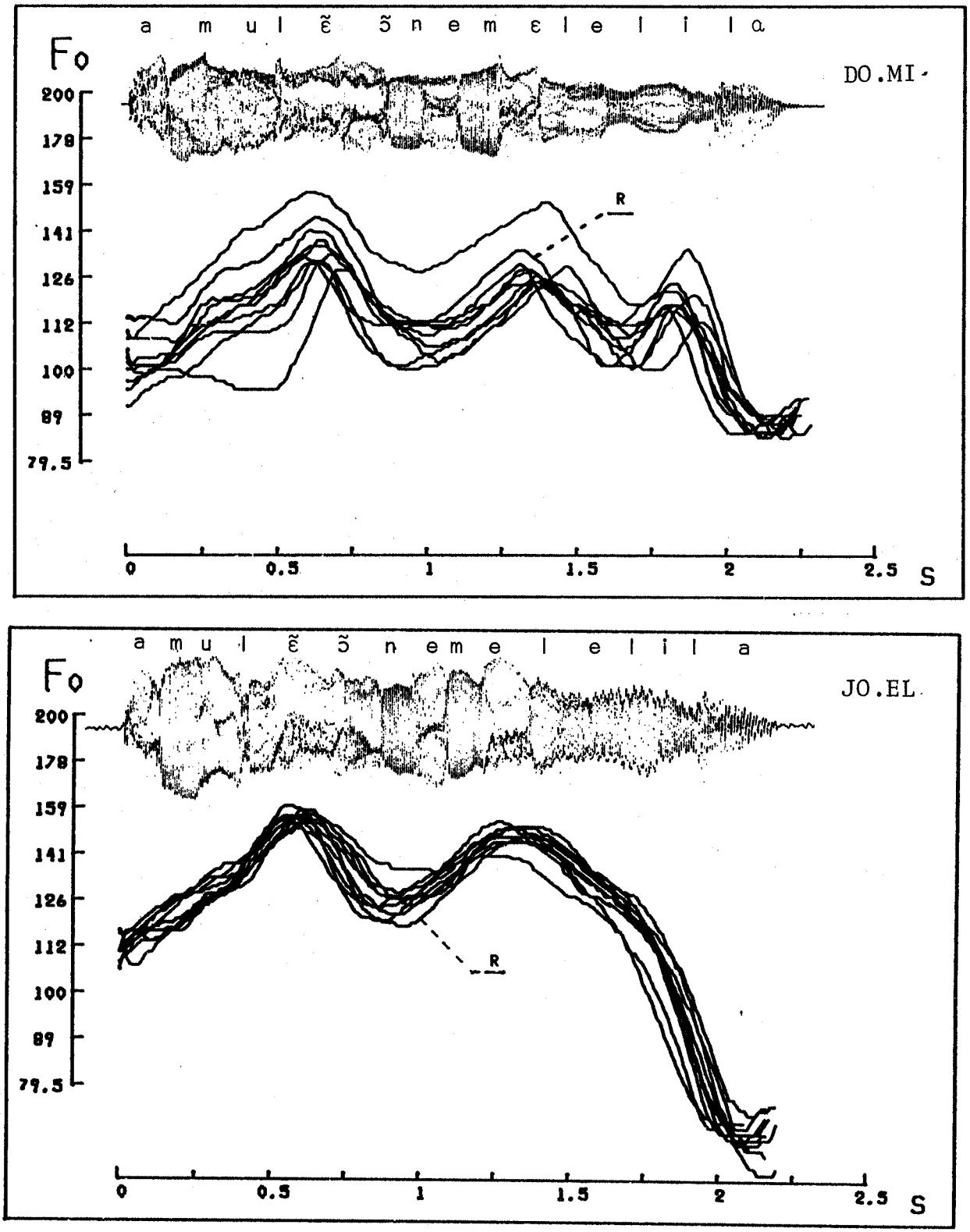


Figure 1 - Tracés des courbes de F_0 correspondant à 10 répétitions de la même phrase (A Moulins on aimait les lilas). Une d'entre elles sera choisie comme référence (R) pour chaque locutrice (DO.MI et JO.EL).
(Facteurs d'échelle : - fréquence à multiplier par 2
- temps à multiplier par 3/5)

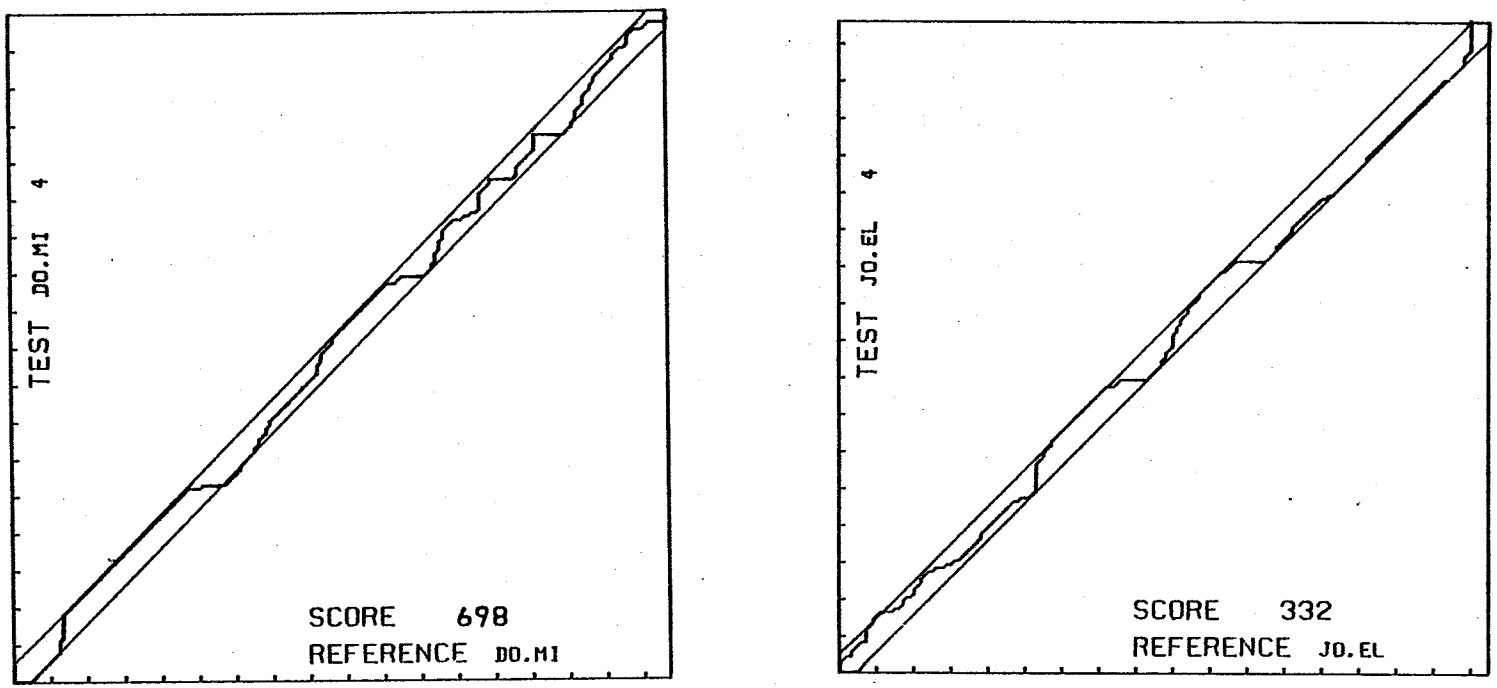


Figure 2 - Exemple de visualisation de la progression de la normalisation à l'intérieur de la fenêtre pour une des répétitions (phrase test) par rapport à la référence. Loc. JO.EL, répétition n° 4; loc. DO.MI, répétition n° 4).

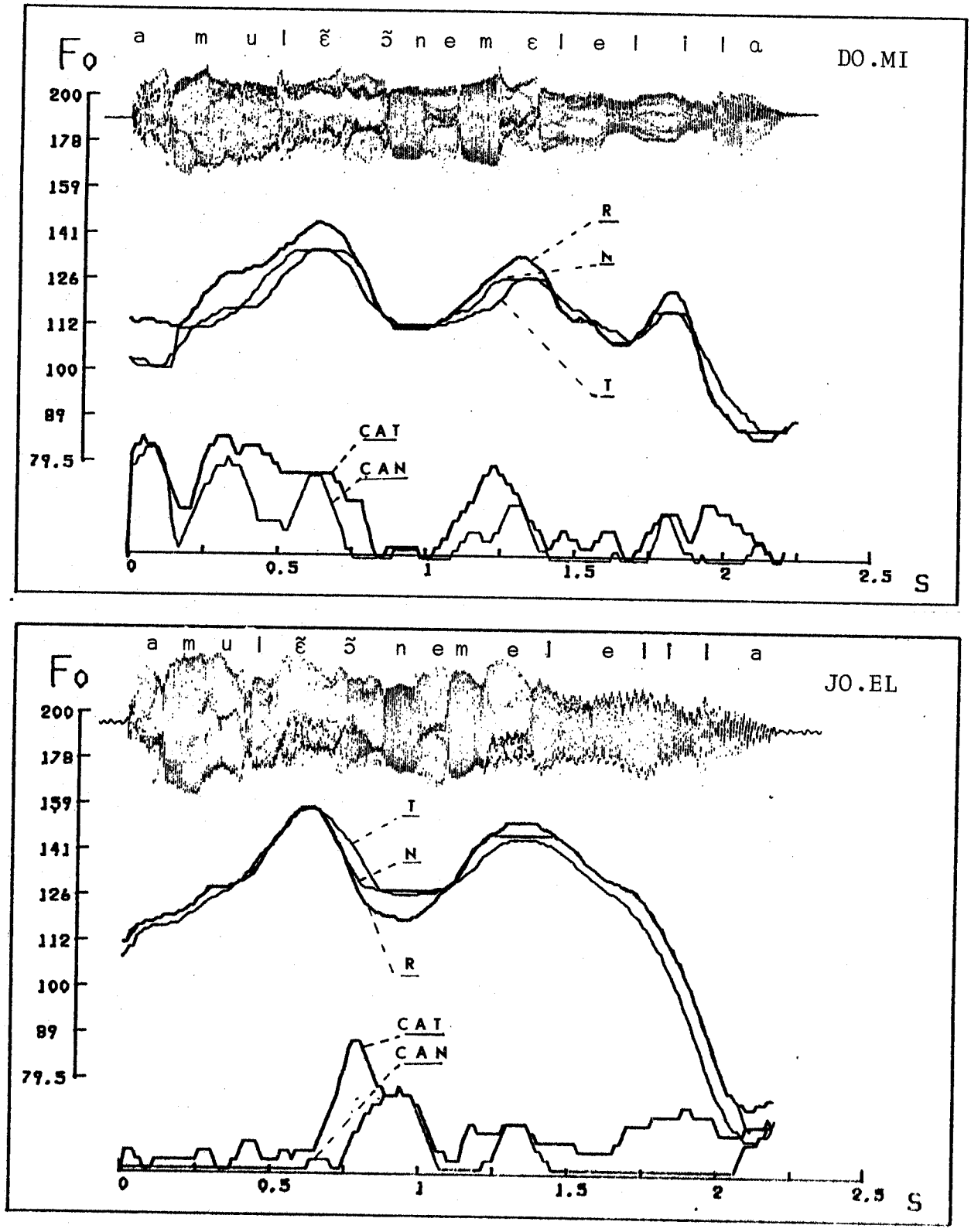


Figure 3 - Courbes de F_0 de la phrase de référence (R) d'une phrase-test (T) et de la phrase normalisée (N).
Courbes de la pente d'accroissement de distance entre :
- phrase-test et référence (CAT)
- phrase normalisée et référence (CAN).
(Facteurs d'échelle : - fréquence à multiplier par 2
- temps à multiplier par 3/5).

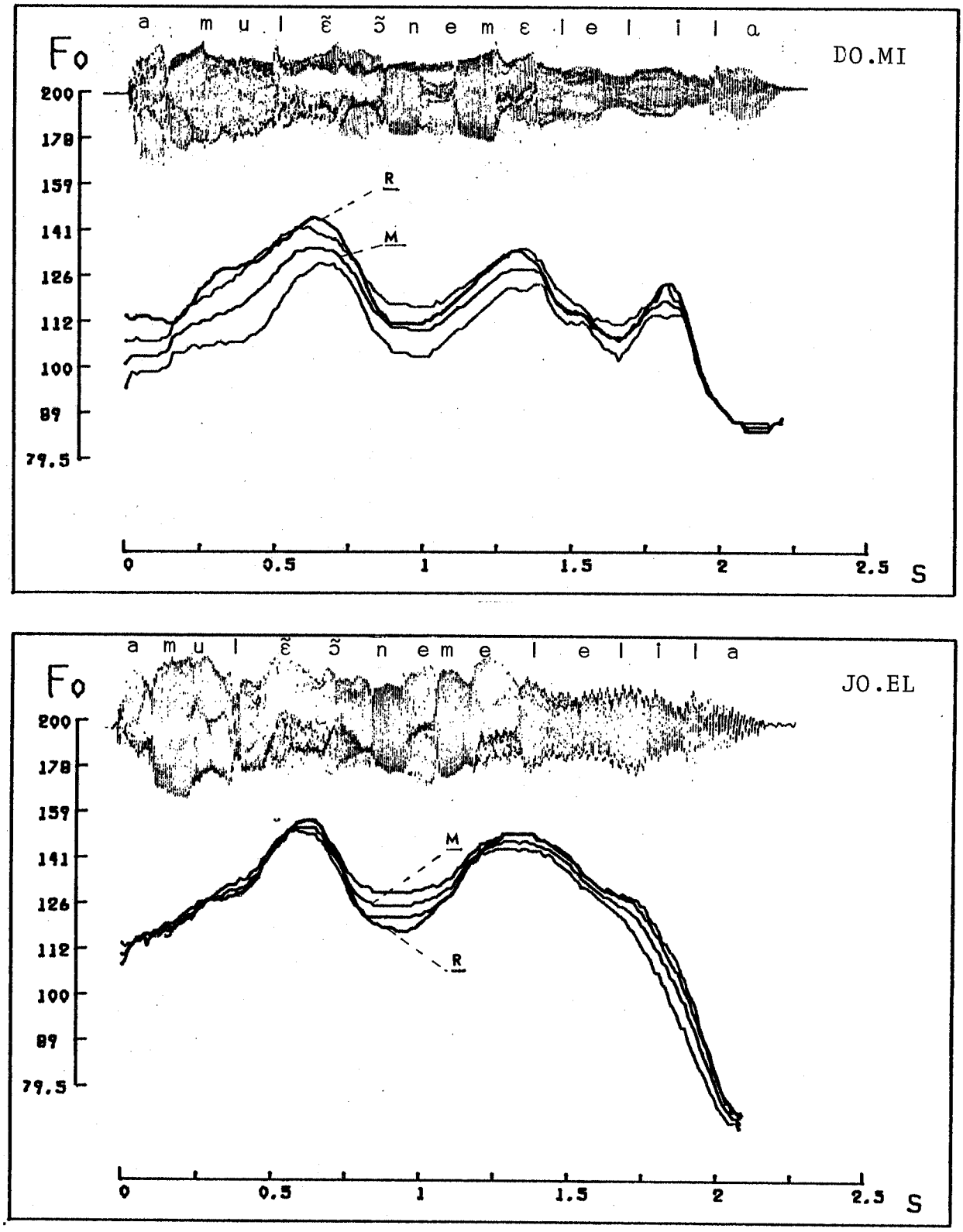


Figure 4 - Résultat statistique sur les 9 courbes normalisées : courbe moyenne (M), et courbes à ± 1'écart-type. A été ajoutéela courbe de référence (R). (Facteurs d'échelle : - fréquence à multiplier par 2 - temps à multiplier par 3/5).

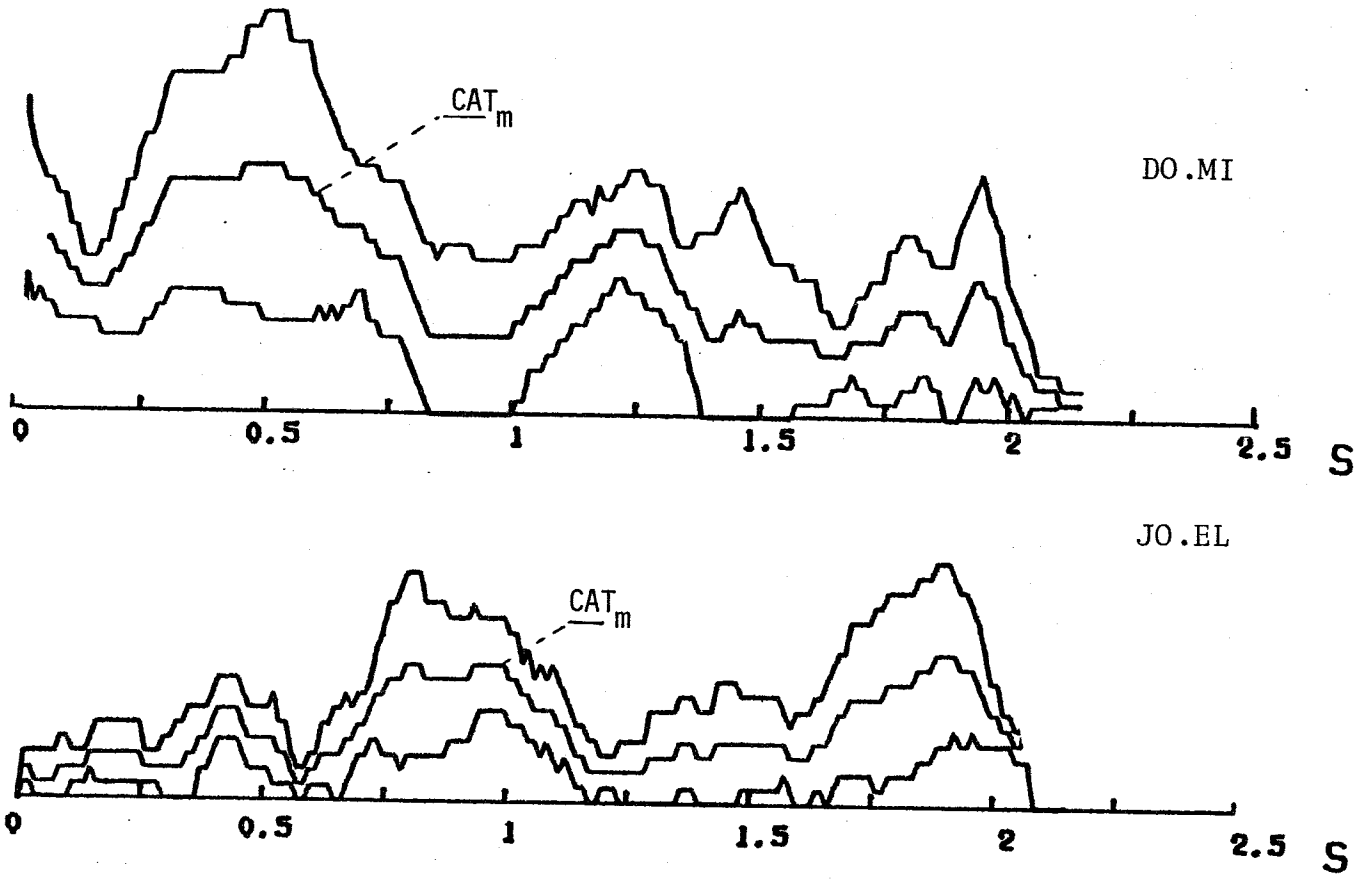


Figure 5 - Statistique sur les pentes d'accroissement entre la référence et les 9 phrases-test : - courbe moyenne CAT_m
- courbes a ± l'écart-type
pour les deux locutrices
(facteur d'échelle : temps à multiplier par 3/5).

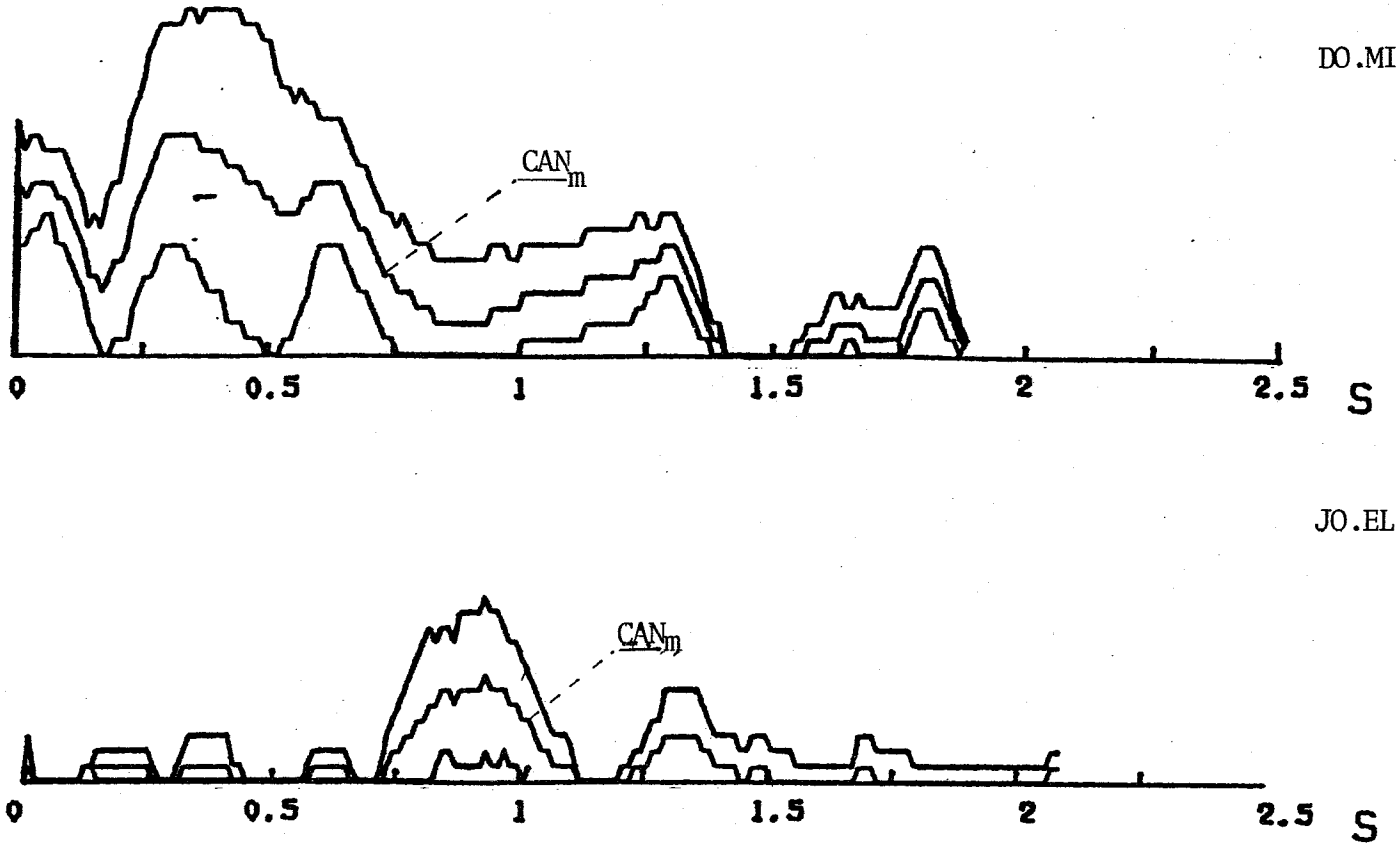


Figure 6 - Statistique sur les pentes d'accroissement entre la référence et les 9 phrases normalisées : - courbe moyenne CAN_m
- courbes à ± l'écart-type
pour les deux locutrices.
(Facteur d'échelle : temps à multiplier par 3/5).

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

UTILISATION DE LA THEORIE DE L'INFORMATION POUR UNE ETUDE QUANTITATIVE DE L'AMBIGUITE EN LECTURE LABIALE

Maria Domenica Di Benedetto
Francis Destombes
Jean-Pierre Tubach

IBM France, Centre Scientifique
36, Avenue Raymond Poincaré
75116 Paris

.RESUME

La méthode de réception de la parole privilégiée par les sourds profonds est la lecture labiale. Cependant, cette méthode est difficile et fatigante car l'image labiale est ambiguë.

Dans cet article, nous nous proposons:

- d'étudier, de façon quantitative, l'effet de cette ambiguïté sur le taux d'information du langage
- de tirer des conclusions utiles à l'assistance automatique à la lecture labiale

. ABSTRACT

For profoundly hearing-impaired people, lip-reading plays a key-role in reception of speech. However, it is a tiring and imprecise method since many sounds of speech look alike on the lips.

This article pursues two goals:

- to use statistical methods to evaluate the impact of this ambiguity on the message information content
- to draw conclusions in view of the development of automatic devices to assist lip-reading

1.0 CADRE DE L'ETUDE

1.1 LE PRINCIPE DE LA LECTURE LABIALE

Cette étude s'inscrit dans le cadre du projet "Enfants sourds" du Centre Scientifique IBM France visant à apporter une aide aux jeunes sourds profonds grâce à l'utilisation de techniques informatiques (principalement le traitement de la parole). Le but de la partie "aide à la lecture labiale" de ce projet est de faciliter la lecture labiale en réduisant son ambiguïté et de donner ainsi à l'enfant une compréhension claire des sons du langage parlé.

En effet, le principe de la lecture labiale est de percevoir le message d'après les mouvements des lèvres. Or, des phonèmes très différents peuvent avoir une apparence labiale identique. C'est le cas, par exemple, de /f/-/v/, de /p/-/b/-/m/, de /o/-/ɔ̃/ etc. Cette ambiguïté rend la lecture labiale non seulement imprécise mais aussi très fatigante. D'où l'idée d'apporter un complément d'information qui indique le choix à faire entre plusieurs "sosies labiaux".

Le "Cued Speech", méthode manuelle inventée par le Dr.CORNETT (1967), constitue une preuve de l'utilité de ces compléments labiaux. Son principe est le suivant: l'orateur accompagne sa parole de compléments indiqués par la forme et la position de sa main placée près de sa bouche afin que ses gestes soient perçus en même temps que le mouvement des lèvres. L'enfant non seulement apprend à comprendre le Cued Speech mais se rend compte ainsi de l'existence des différents sons du langage.

Le fait de lever l'ambiguïté de la lecture labiale facilite donc l'acquisition du langage de la part des jeunes sourds profonds et comporte de nombreuses et importantes conséquences positives: développement intellectuel, communication, facilité de lecture,etc... L'inconvénient d'une méthode manuelle telle que le Cued Speech est qu'elle nécessite un apprentissage de la part de l'interlocuteur de la personne handicapée. Cela explique l'intérêt de produire automatiquement des compléments labiaux. Le traitement de la parole, dans l'avenir, pourra sans doute permettre d'extraire du signal vocal une information phonétique complémentaire de l'information donnée par le mouvement des lèvres, afin de la communiquer au sourd profond, en temps réel, par des moyens appropriés (par exemple un dispositif lumineux intégré à des lunettes spéciales, ou un dispositif tactile) (CORNETT, BEADLES, WILSON, 1977 et DI BENEDETTO, DESTOMBES, TUBACH, 1980).

1.2 BUT ET REALISATION DE L'ETUDE

Nous nous proposons, dans cette étude sur corpus phonétique, de considérer l'influence des confusions labiales sur le taux d'information du langage, l'objectif étant de déterminer quels sosies labiaux il est le plus avantageux de discriminer pour fournir une aide efficace au sourd profond.

Les données utilisées sont la transcription phonétique d'un volumineux corpus du français oral constitué à l'Institut de Phonétique de Grenoble. Il comporte plusieurs sections, représentant plusieurs niveaux de langue, du français cultivé au français littéraire soigné: des conférences, des interviews, des discussions de France-Culture et de France-Inter, correspondant à environ 3 heures de parole. Il comprend plus de 150000 caractères: phonèmes, séparateurs de mots, de groupes rythmiques, de phrases, marques de liaison et diacritiques divers (accentuation, allongement, assourdissement, sonorisation, etc...). Lors des traitements informatiques effectués pour cette étude, seuls ont été pris en compte les 36 phonèmes

du français codifiés suivant l'alphabet phonétique international ainsi que la marque de fin de groupe rythmique ou de phrase (/):

a ɑ u e ɛ i o y ɔ œ ø ə ɑ̃ ɛ̃ ɔ̃ œ̃ j w ɥ p t k b d g s ʃ f z ʒ v m n ɲ l R /

Le corpus ainsi constitué comporte 95000 phonèmes. Rappelons que d' autres études statistiques sur corpus ont été effectuées: par exemple, concernant l'anglais écrit (SHANNON, 1951), l'anglais oral (DENES 1963), le français écrit (MOREAU, 1962), le français oral (TUBACH, 1969 et HATON, LAMOTTE, 1971).

L'exploitation du corpus est effectuée en APL sur un système IBM 370/145 fonctionnant avec le système d'exploitation VM/370. Cet article a été composé à l'aide du programme de composition de documents SCRIPT/VS et imprimé sur un système d'impression IBM 3800. L'utilisateur de ce matériel peut définir par programme des polices de caractères spécifiques. C'est ce qui a été fait ici pour les caractères phonétiques.

1.3 REMARQUES

Il faut, en lisant ce texte, tenir compte des remarques suivantes:

- il serait prétentieux de vouloir tirer des conclusions absolues de ce genre d' étude, car la lecture labiale est mal connue. Par exemple, son efficacité dépend, dans une large mesure, des conditions d'éclairage, de la distance de l'interlocuteur, de sa manière d'articuler. Par ailleurs, tous les spécialistes ne sont pas nécessairement d'accord sur une définition précise des sosies labiaux (GARRIC, 1974). La définition de 12 groupes de sosies adoptée ci-après présente donc une part d'arbitraire:

P	T	V	S	H	r	A	O	0	Y	I	E
p	t	f	s	ʃ	k	ɑ	o	ɔ	y	e	œ̃
b	d	v	z	ʒ	g	ɑ̃	ø	œ	u	ɛ̃	ɛ
m	n				R		ɔ̃			i	

  Chaque symbole de la première ligne identifie un groupe de sosies labiaux et suggère, dans la mesure du possible, le phonème le plus fréquent du groupe: par exemple, P correspond à /p/, /b/, /m/. On considère comme étant non ambigus les 7 phonèmes suivants: a ə j w ɥ ɲ l. Si, dans le corpus, chaque phonème est remplacé par l'identificateur du groupe de sosies auquel il appartient, on obtient un texte qui correspond, en quelque sorte, à une transcription "phonétique" du message labial (voir Annexes, FIG.1 et FIG.2). On peut observer que la lecture de ce texte n'est pas aisée bien que les blancs entre les mots n'aient pas été supprimés. Pour un sourd profond, la décodification de ce message est encore plus difficile en raison de l'absence de séparateurs de mots dans le langage oral.

- Par ailleurs, cette matrice définit une ambiguïté au niveau du phonème isolé, indépendamment de son contexte et néglige donc les effets de coarticulation. L'approche suivie est donc forcément partielle, mais c'est malheureusement la seule possible actuellement: en effet, il n'existe aucune étude quantitative précise mesurant les effets de la coarticulation sur le taux de reconnaissance des phonèmes ou syllabes du Français en lecture labiale. Or, les calculs détaillés par la suite font intervenir les fréquences de phonèmes isolés, de

diphones et de triphones afin de mesurer la quantité d'information du message phonétique perçu par un entendant (sans ambiguïté) ou par un sourd profond (avec ambiguïté). Les résultats sans ambiguïté sont donc précis, mais ceux qui font intervenir l'ambiguïté sont entachés d'une certaine inexactitude (assez limitée toutefois car les phonèmes "masquants" tels que /ʃ/ ou /u/ notamment sont relativement peu fréquents). Il faut donc en retirer seulement une indication de tendance et ces calculs devraient être effectués à nouveau lorsqu'une matrice de confusion en contexte sera disponible (Michèle Gentil poursuit actuellement dans notre projet une étude dans ce but).

- L'étude est menée essentiellement au niveau phonétique, les aspects morphologique, syntaxique, sémantique, contextuel n'entrant en ligne de compte qu'au niveau de leur répercussion sur les fréquences de groupements de phonèmes. Or, un sourd profond connaissant déjà le langage parlé ferait justement appel à des informations complémentaires liées à la syntaxe de la langue, à sa sémantique, au contexte. Un jeune enfant, au contraire, ne possède ni la structure ni le vocabulaire du langage parlé et ne peut donc pas s'appuyer sur ces connaissances pour percevoir les informations présentes sur les lèvres de ses interlocuteurs. Il faudra donc examiner les résultats sous deux angles différents, suivant qu'il s'agit d'un sourd profond ne connaissant pas encore le langage parlé (qu'on appellera par la suite "enfant sourd"), ou d'un sourd profond connaissant déjà le langage parlé (qu'on appellera "adulte sourd").

2.0 CALCULS D'ENTROPIE AVEC OU SANS CONFUSIONS LABIALES

On va chercher ici à déterminer les groupes de confusion labiale dont la discrimination améliorerait le plus la perception du message.

2.1 RAISONNEMENT INTUITIF

Intuitivement, on peut estimer que la résolution de l'ambiguïté pour un groupe de deux ou trois phonèmes très fréquents est globalement plus efficace que pour un groupe de phonèmes peu fréquents. Utilisons les fréquences de phonèmes obtenues par traitement du corpus (que l'on trouvera en annexe, FIG.3) et calculons la somme des fréquences des phonèmes de chaque groupe de confusion. Cette quantité, ou fréquence cumulée, représente le taux de couverture du message phonétique par les phonèmes du groupe. Suivant ce critère ,on obtient le classement des groupes de confusion ci-après:

GROUPE	FREQUENCE CUMULEE(%)
T tdn	12.86
r kgR	12.81
I eɛ̃i	12.75
P pbm	7.92
S sz	7.68
E œ̃c	5.96
A ɑɑ̃	4.25
Y yu	3.80
O oøɔ̃	3.52
V fv	3.44
0 ɔœ	2.79
H ʃʒ	1.67

On va maintenant affiner ces résultats au moyen de calculs plus complexes fondés sur la théorie de l'information, en particulier sur les notions d'entropie et de redondance. Cela permettra de prendre en compte non seulement les fréquences des phonèmes, mais aussi la structure linguistique du message, telle qu'elle transparaît dans les fréquences des diphones et triphones.

2.2 DEFINITIONS

Nous appellerons f(1),f(2),.....f(N-1) les 36 phonèmes du français et noterons par commodité f(37) les marques de fin de groupe rythmique ou de phrase (représentées respectivement par / et // dans la transcription du corpus). Les pauses dans le discours sont des indications importantes dans la perception de la parole, aussi bien pour les entendants qu'en lecture labiale, et il nous a paru souhaitable de les prendre en compte pour nos calculs de quantités d'information. Nous travaillerons donc avec un ensemble de N=37 symboles, soit F =(f(1),f(2),.....f(N)) que nous appellerons "phonèmes". Nous appellerons de même diphone tout élément de F×F et triphone tout élément de F×F×F.

Le comptage dans le corpus fournit les nombres d'occurrences:

n(i) des phonèmes f(i)
n(i,j) des diphones f(i)f(j)
n(i,j,k) des triphones f(i)f(j)f(k)
i,j,k =1,2,...,N

En assimilant les fréquences aux probabilités (voir Annexe 6.1), on obtient les probabilités:

p(i) des phonèmes (probabilités d'ordre 1)
p(i,j) des diphones (probabilités d'ordre 2)
p(i,j,k) des triphones (probabilités d'ordre 3)

puis les probabilités conditionnelles:

$p(j|i)$ du phonème f(j) après le phonème f(i)
(probabilité conditionnelle d'ordre 2)
$p(k|i,j)$ du phonème f(k) après le diphone f(i)f(j)
(probabilité conditionnelle d'ordre 3)

On a: $p(i,j) = p(i)p(j|i)$ et $p(i,j,k) = p(i,j)p(k|i,j)$.

Introduisons quelques grandeurs qui vont être utilisées dans la suite de l'article (tous les logarithmes figurant dans les formules sont des logarithmes de base 2).

La quantité H1, ou entropie d'ordre 1, mesure le degré d'incertitude d'un phonème parmi N, de probabilités p(i) i=1,2,...,N.

$$H1 = -\Sigma p(i) \log p(i)$$

C'est une quantité d'information exprimée en bits. Elle doit être comparée à la quantité H0 =-log N qui mesure ce que serait ce degré d'incertitude si les N phonèmes étaient équiprobables; la redondance au niveau des phonèmes isolés, ou redondance d'ordre 1, est donnée par:

$$R1 = 1 - H1/H0$$

La quantité H2, ou entropie conditionnelle d'ordre 2, mesure le degré d'incertitude du phonème f(j) lorsque le phonème précédent f(i) est connu et rend compte des contraintes sur les suites de deux phonèmes amenées par les niveaux morphologique, syntaxique, etc... du langage:

$$H2 = -\Sigma p(i,j) \log p(j|i) \qquad i,j=1,2,\ldots,N$$

$$= -\Sigma p(i,j) \log p(i,j) - (-\Sigma p(i) \log p(i))$$

La redondance d'ordre 2, sur les suites de deux phonèmes, est alors:

$$R2 = 1 - H2/H0$$

De la même façon, à l'ordre 3, on obtient l'entropie H3 et la redondance R3.

Les notions qui précèdent ont été introduites par SHANNON (1951) dans une étude sur un corpus de langage écrit. On en trouvera un exposé didactique dans YAGLOM et YAGLOM (1969) par exemple.

2.3 RESULTATS GLOBAUX

Les résultats obtenus dans le cas de notre corpus phonétique sont les suivants (les redondances sont exprimées en pourcentage):

	H0	H1	H2	H3
SANS AMBIGUITE	5.209	4.728	3.835	3.053
REDONDANCE		9.2	26.4	41.4

La diminution de l'entropie, c'est à dire de l'incertitude d'un phonème quand on prend en compte des contextes de plus en plus longs (ordre 2, ordre 3), illustre les contraintes apportées par la structure du langage: si l'on connaît deux phonèmes consécutifs, le suivant est choisi parmi 2 à la puissance H3 en moyenne, soit ici environ 8 et non parmi N=37.

Effectuons maintenant la confusion des différents sosies labiaux en remplaçant dans le corpus chaque phonème faisant partie d'un groupe de confusion labiale par l'identificateur de ce groupe, comme défini au chapitre précédent. On obtient de nouvelles probabilités d'ordre 1,2,3 et les quantités précédentes prennent les nouvelles valeurs suivantes :

	H0	H1	H2	H3
AVEC AMBIGUITE	4.322	3.917	3.311	2.938

Les diminutions d'entropie sont :

	D0	D1	D2	D3
DIFFERENCE	.888	.811	.525	.114
DIFFERENCE (PCT)	17.04	17.15	13.68	3.75

On constate qu'il y a diminution d'entropie à tous les ordres 0, 1, 2 et 3. Cette diminution correspond à l'appauvrissement du message provoqué par la confusion entre les phonèmes.

D'autre part, les variations d'entropie en valeur relative diminuent considérablement de l'ordre 1 à l'ordre 3 (de 17% à 3,7%). Ce fait illustre à la fois la possibilité et la difficulté de la lecture labiale. Un "adulte sourd" est capable d'utiliser les informations de contexte et se trouve pour le moins au niveau H3, où la perte d'entropie, c'est à dire d'information, est assez faible (3,7 %). La réalité confirme qu'il peut effectivement lever la plupart des ambiguïtés. Bien entendu, nous ne prétendons pas qu'il le fasse en s'appuyant exclusivement sur des probabilités qui ne sont qu'une des manifestations extérieures des structures linguistiques. Par contre un "enfant sourd" ne peut pas utiliser ces informations de contexte et se trouve au mieux au niveau H1, avec une nette perte d'information (17 %) due aux sosies labiaux, qu'il ne peut pas compenser.

2.4 RESULTATS DETAILLES: VARIATIONS D'ENTROPIE

Supposons que l'on puisse présenter au sourd profond une information qui lève l'ambiguïté de l'un des groupes de sosies labiaux, que nous appellerons le groupe X, les autres groupes restant ambigus. On obtient alors de nouvelles valeurs de H0, H1, H2 et H3 indiquées dans le tableau 1. Dans ce tableau, D1, D2, D3 sont les variations d'entropie par rapport aux valeurs de la première ligne (aucune ambiguïté éliminée).

Les groupes sont indiqués par ordre décroissant d'influence sur l'entropie H3. On obtient donc un classement des groupes suivant la dégradation qu'ils introduisent dans le message et donc selon l'importance pour la compréhension du langage qu'il faut attribuer à la résolution de leur ambiguïté. Par exemple,distinguer /t/,/d/,/n/ semble très utile, tandis que la distinction entre /ɔ/ et /œ/ n'apporte pratiquement rien.

En se souvenant du raisonnement intuitif du paragraphe 2.1, on peut remarquer que les cinq groupes de confusion classés ici en tête sont ceux dont les fréquences cumulées sont les plus importantes, mais on constate d'importantes variations de classement à l'intérieur de cet ensemble.

Levons maintenant l'ambiguïté de tous les groupes de sosies successivement, dans l'ordre indiqué par TAB.1. Les nouvelles valeurs de l'entropie sont reportées en TAB.2. Les différences D1, D2, D3 entre ces nouvelles valeurs et les entropies de même ordre du message non ambigu calculées au paragraphe précédent sont reportées également en TAB.2.

	H0	H1	D1	H2	D2	H3	D3
AUCUN	4.322	3.917	.000	3.311	.000	2.938	.000
T tdn	4.459	4.097	.180	3.440	.129	3.013	.075
I eɛ̃i	4.459	4.073	.156	3.436	.125	2.999	.060
P pbm	4.459	4.020	.103	3.382	.071	2.962	.024
r kgR	4.459	4.054	.137	3.386	.076	2.960	.021
S sz	4.392	3.968	.051	3.346	.035	2.955	.017
O oøɔ̃	4.459	3.959	.042	3.342	.031	2.951	.013
V fv	4.392	3.947	.030	3.336	.025	2.951	.012
Y yu	4.392	3.952	.035	3.336	.026	2.946	.008
A ɑɑ̃	4.392	3.942	.025	3.327	.016	2.944	.006
E œ̃ɛ	4.392	3.941	.023	3.327	.016	2.941	.002
H ʃʒ	4.392	3.929	.012	3.318	.008	2.940	.002
0 ɔœ	4.392	3.933	.016	3.321	.011	2.939	.001

TAB.1

	H0	H1	D1	H2	D2	H3	D3
T tdn	4.459	4.097	.630	3.440	.396	3.013	.039
I eɛ̃i	4.585	4.254	.474	3.564	.271	3.059	.006
P pbm	4.700	4.356	.371	3.633	.202	3.071	.019
r kgR	4.807	4.493	.234	3.701	.134	3.074	.022
S sz	4.858	4.544	.183	3.731	.104	3.079	.026
O oøɔ̃	4.954	4.586	.141	3.756	.080	3.076	.024
V fv	5.000	4.617	.111	3.780	.055	3.078	.026
Y yu	5.044	4.651	.077	3.799	.036	3.070	.017
A ɑɑ̃	5.087	4.676	.052	3.807	.029	3.066	.014
E œ̃ɛ	5.129	4.699	.028	3.819	.016	3.060	.008
H ʃʒ	5.170	4.712	.016	3.826	.009	3.058	.005
0 ɔœ	5.209	4.728	.000	3.835	.000	3.053	.000
TOUS	5.209	4.728	.000	3.835	.000	3.053	.000

TAB.2

Les différences d'entropie, tant au niveau H1 qu'au niveau H2, deviennent faibles dès que l'ambiguïté des cinq premiers groupes de sosies est levée (2 à 3%). Si l'on veut fournir une aide efficace aux "enfants sourds", il faut donc porter notre attention au moins sur les cinq premiers groupes de la liste ci-dessus et être capable de lever, par traitement automatique, l'ambiguïté des phonèmes leur appartenant.

Au niveau H3, par contre, cette différence est toujours négligeable. Cela confirme le fait qu'un "adulte sourd" a beaucoup moins besoin qu'un "enfant sourd" d'une aide à la lecture labiale puisqu'il a la possibilité de suppléer aux ambiguïtés par sa connaissance de la langue. Cependant, cette tâche nécessite un effort d'attention considérable et pourrait être grandement facilitée par des informations complémentaires.

3.0 PREDICTION DE TAUX D'IDENTIFICATION DE DIPHONES CV OU VC

La définition adoptée pour C (consonnes) et V (voyelles) est la suivante:

C : j w ɥ p t k b d g s ʒ f z n v m n ɲ l R
V : a ɛ u e y i o ø ə ɑ̃ ɛ̃ ɔ̃ œ̃

La distribution des semi-voyelles j, w, ɥ étant beaucoup plus proche de celle des consonnes que de celle des voyelles, nous les avons assimilées à des consonnes pour avoir une partition des phonèmes en deux classes: consonnes et voyelles.

Dans ce chapitre, on ne prend pas en compte tous les diphones mais uniquement ceux ayant la forme CV ou VC qui, en effet, constituent respectivement 43.29 et 40.83% des diphones du corpus et sont donc particulièrement intéressants. De plus, l'unité syllabique CV (ou VC) a une importance fondamentale au niveau de l'apprentissage du langage. Pour s'en rendre compte, il suffit de prêter attention aux termes utilisés par les parents lorsque leurs enfants sont très jeunes: ils sont simples et essentiellement constitués de ce type de syllabes. D'autre part, l'unité de base du Cued Speech étant la syllabe CV, il semble utile d'étudier ce cas particulier de diphone. Le problème crucial de l'aide aux enfants malentendants est donc de les aider à percevoir les diphones de type CV ou VC.

Supposons maintenant effectuer un test de lecture labiale où l'on présente à un malentendant des diphones CV ou VC sans signification (aucun contexte linguistique), avec une fréquence proportionnelle à leur fréquence dans le langage. Admettons également que le malentendant, n'ayant pas à sa disposition de contexte sur lequel s'appuyer pour faire un choix opportun, réponde au hasard dans un même groupe de confusion. Nous allons calculer quel serait, dans ces conditions d'expérience, les taux d'identification de diphones CV ou VC.

Considérons un diphone de type CV (le même raisonnement est valable pour le type VC) .Soit fcfv ce diphone, formé des phonèmes fc, qui est une des 20 consonnes possibles, et fv, qui est une des 16 voyelles. Désignons par Mc et Mv les nombres de phonèmes des groupes de confusion auxquels appartiennent fc et fv (si fc ou fv ne sont pas ambigus, ces nombres sont égaux à 1). Le produit Mc.Mv est le nombre d'identifications possibles pour le diphone fcfv et 1 / Mc.Mv est, avec nos hypothèses, la probabilité d'obtenir ce diphone comme réponse.

On connaît, pour chaque diphone CV, sa fréquence dans le corpus: p(fcfv) .Calculons la quantité:

$$Q(fcfv) = \frac{p(fcfv)}{\Sigma p(fCfV)}$$

la somme du dénominateur étant prise pour toutes les voyelles fV et toutes les consonnes fC.

Q(fcfv) représente la fréquence relative du diphone fcfv par rapport aux autres diphones de même type (CV). Au cours de l'expérience décrite précédemment, le taux d'identification du diphone fcfv serait donc:

$$Q(fcfv)/(Mc.Mv)$$

On effectue ce calcul dans les deux cas suivants:

- tous les groupes de sosies sont ambigus sauf un (colonnes ELIMINATION ISOLEE en TAB.3), pour voir quelle est l'influence de l'élimination de l'ambiguïté de ce seul groupe sur le taux d'identification. Les colonnes ELIMINATION PROGRESSIVE correspondent au cas où l'on élimine successivement l'ambiguïté des différents groupes.

- un seul groupe est ambigu (colonnes UN SEUL GROUPE AMBIGU en TAB.3) pour voir quelle est l'influence de la confusion des phonèmes de ce groupe sur le taux d'identification

Les résultats de ces calculs (en %) sont reportés dans le tableau 3.

GROUPE	ELIMINATION ISOLEE CV	ELIMINATION ISOLEE VC	ELIMINATION PROGRESSIVE CV	ELIMINATION PROGRESSIVE VC	UN SEUL GROUPE AMBIGU PCT CV	UN SEUL GROUPE AMBIGU PCT VC
AUCUN	28.78		27.19			
Ieĉi	38.78	35.94	38.78	35.94	81.07	82.23
Ttdn	37.94	35.49	51.50	47.05	83.78	84.22
rkgR	36.47	35.59	61.45	57.57	86.87	85.62
Ppbm	33.75	33.62	67.35	65.53	91.03	89.51
Ssz	32.15	31.17	71.87	71.07	93.73	92.64
Eɛ̃ɛ	31.90	30.46	78.12	78.16	93.29	92.77
Ooøɔ̃	31.65	29.10	83.02	82.69	94.79	95.27
Vfv	30.39	29.67	85.93	86.12	96.67	96.12
Yyu	30.71	29.18	90.21	90.66	95.64	95.39
Aɑã	30.79	29.00	94.93	94.68	95.22	95.95
Oɔœ	30.50	28.76	98.36	98.41	96.55	96.22
Hʃʒ	29.75	28.23	100.00	100.00	98.36	98.41

TAB.3

Les groupes sont classés par ordre décroissant d'influence sur le taux d'identification global des diphones (l'ordre est le même pour les diphones CV que pour les diphones VC). On voit, d'après ce tableau, que si, par exemple, tous les groupes de sosies demeurent ambigus, les taux d'identification de diphones CV et VC sont respectivement de 28,8% et de 27,2%. Des tests de lecture labiale de syllabes effectués par NICHOLLS (1979) pour l'anglais ont abouti à des résultats du même ordre de grandeur (environ 30%).

Le classement obtenu en TAB.3 illustre les conséquences de l'ambiguïté des différents groupes de sosies sur le taux d'identification des diphones CV et VC. Par exemple, si l'ambiguïté du groupe T est levée (c.à.d. que le malentendant reçoit une information telle qu'il ne confond plus le /t/ avec le /d/ ou avec le /n/) le taux d'identification de diphones CV est de 37,9%. Si, par contre, l'ambiguïté du groupe V est levée, le taux d'identification de diphones CV est de 30,4%. Par ailleurs, si le groupe T est le seul dont l'ambiguïté n'ait pas été levée, le taux d'identification de diphones VC est de 84,2%.

D'autre part, on obtient ainsi les groupes pour lesquels il est le plus intéressant de lever l'ambiguïté si l'on veut obtenir un taux d'identification acceptable: si, par exemple, l'objectif fixé est d'obtenir un taux de 70%, il est

facile de voir qu'il sera atteint en levant l'ambiguïté des cinq premiers groupes de confusion labiale. Observons, d'ailleurs, que ce sont les mêmes groupes classés en tête au paragraphe précédent (ils correspondent aussi à ceux qui ont les fréquences cumulées les plus élevées).

4.0 ETUDE AU VU DES CONNAISSANCES ACTUELLES EN TRAITEMENT DE LA PAROLE

Il faut tenir compte maintenant des difficultés inhérentes au traitement automatique de la parole, et examiner l' "état de l'art" en matière de systèmes de reconnaissance automatique de phonèmes, afin de déterminer quels sont les groupes de confusion dont il sera le moins difficile de lever l'ambiguïté. Intuitivement, on pourrait penser, par exemple, que les groupes faisant intervenir une opposition de voisement (du type /f/ - /v/) seront plus aisés à traiter que ceux qui reposent sur une différence de mode d'articulation (nasal - non nasal par exemple). Pour dépasser ce type d'impressions et être plus précis, nous avons voulu utiliser des statistiques représentatives des erreurs commises par les reconnaisseurs de phonèmes fonctionnant réellement aujourd'hui. Nous avons trouvé des données détaillées dans la Thèse de J.CAELEN (1979) (pages 149 à 151), pour un corpus d'un millier de phonèmes dit par deux locuteurs masculins. Les résultats globaux de son système sont: un taux de reconnaissance de 71% en moyenne pour tous les phonèmes et 86% pour les voyelles. Nous n'avons pas trouvé d'autres informations de ce type pour des systèmes récents reconnaissant le Français.

Nous disposons donc des probabilités P(x->y) que le phonème x soit reconnu comme y. Dans un groupe de confusion labiale X ,contenant les phonèmes x et y, x sera reconnu comme y dans P(x->y)% des cas et y comme x dans P(y->x)%. Ces taux d'erreur, pondérés par les fréquences, p(x) et p(y), des phonèmes x et y dans la langue, nous fourniront l'Erreur Moyenne Pondérée (EMP) dans le groupe X. Nous prendrons cette quantité comme critère de la difficulté à éliminer l'ambiguïté dans le groupe par traitement automatique.

$$EMP = \frac{p(x)\,P(x->y) + p(y)\,P(y->x)}{p(x)+p(y)}$$

(et une expression équivalente pour les les groupes de confusion de trois phonèmes).

Le groupe k,g,R présente une particularité: le /k/ et le /g/ sont assez facilement confondus (P(k->g)=4,5% et P(g->k)=7%), mais le /R/ est de loin le phonème le plus fréquent des trois et n'est jamais confondu avec /k/ ni /g/. Ceci confère à ce groupe une EMP faible (1,8%).

Les données ne sont pas disponibles pour classer le groupe /ɑ/,/ɑ̃/ car le /a/ et le /ɑ/ ne sont pas distingués dans les données utilisées. Nous classerons ce groupe dans la même zone que ceux qui font intervenir l'opposition nasale/ non nasale pour des voyelles.

La classification s'établit comme suit :

	Groupe	EMP (%)
Y	y,u	0
E	œ̃,ɛ	1,2
O	o,ø,õ	1,3
A	ɑ,ɑ̃	?
r	k,g,R	1,8
S	s,z	1,9
O	ɔ,œ	2,7
I	e,ɛ̃,i	3,7
H	ʃ,ʒ	4,9
T	t,d,n	6,3
V	f,v	6,8
P	p,b,m	13

Les confusions dans les groupes de fricatives interviennent dans le sens sonore->sourde et jamais dans le sens contraire : les fricatives sonores sont souvent plus brèves et le voisement peut n'avoir pas le temps de se manifester.

Ce classement est assez semblable, pour les consonnes, à celui fourni par la matrice de confusion publiée par H.SILVERMAN et R.DIXON (1976) pour un système de reconnaissance de l'anglais américain. Aucune comparaison n'est bien entendu possible en ce qui concerne les voyelles.

	H0	H1	D1	H2	D2	H3	D3
Y yu	4.392	3.952	.776	3.336	.499	2.946	.106
E œ̃ɛ	4.459	3.975	.753	3.352	.483	2.949	.104
O ɔœ	4.524	3.991	.737	3.363	.472	2.949	.104
A ɑɑ̃	4.585	4.016	.712	3.379	.456	2.954	.098
r kgR	4.700	4.153	.575	3.453	.382	2.969	.084
S sz	4.755	4.204	.524	3.484	.351	2.980	.073
O oøõ	4.858	4.246	.482	3.512	.323	2.987	.065
I eɛ̃i	4.954	4.402	.325	3.631	.205	3.031	.022
H ʃʒ	5.000	4.415	.313	3.638	.198	3.030	.022
T tdn	5.087	4.595	.133	3.754	.082	3.059	.007
V fv	5.129	4.625	.103	3.777	.058	3.060	.007
P pbm	5.209	4.728	.000	3.835	.000	3.053	.000

TAB.4

Nous allons maintenant, comme au paragraphe 2.4, lever successivement l'ambiguïté de tous les groupes de sosies labiaux, mais cette fois-ci dans l'ordre de difficulté croissante qui vient d'être défini. Nous obtenons un tableau TAB.4, qui est à comparer au tableau TAB.2. D1, D2, et D3 désignent toujours les différences par rapport aux entropies H1, H2, et H3 sans aucune ambiguïté.

On constate que les variations d'entropie les plus fortes interviennent encore lorsque l'on prend en compte les groupes qui se sont déjà révélés importants aux paragraphes précédents. L'efficacité de ces groupes demeure, même si d'autres groupes plus mineurs ont été traîtés auparavant.

Par exemple, sur H1 ou H2, il fallait dans l'ordre optimal, prendre les 3 premiers groupes pour rattraper 50% de la perte d'entropie, et les 5 premiers pour 70%. Ici, ces chiffres deviennent:
4 (pour H2) ou 5 (pour H1) pour 50%
9 (pour H2) ou 10 (pour H1) pour 70%.

Ce fait montre bien que c'est sur l'amélioration de la reconnaissance des phonèmes "difficiles" que doit porter l'effort de recherche principal.

5.0 CONCLUSION

Les différentes approches utilisées dans cette étude (approche intuitive, taux d'identification de diphones, entropie), se recoupent dans une assez large mesure: cinq groupes de confusion semblent très importants: T, I, P, r et S. Leur importance relative varie, cependant, en fonction de l'approche employée et nous retiendrons l'ordre résultant du calcul le plus précis (calcul d'entropie).

Cependant, il faut noter que les phonèmes correspondant à ces groupes (en particulier T et P) sont aussi les plus difficiles à reconnaître par un traitement automatique. L'état actuel de la technique doit déjà permettre de fournir à un "adulte sourd" des compléments d'information qui lui faciliteront la lecture labiale, compte tenu de ses connaissances linguistiques. Par contre, pour favoriser l'acquisition du langage chez de jeunes enfants sourds, des informations plus précises et plus fiables sont nécessaires.

Les recherches visant à obtenir une meilleure reconnaissance phonémique trouveront donc ici un champ d'application nouveau et important. Dans l'attente de leurs résultats, il apparaît souhaitable dans une première étape de réduire la complexité du problème en se limitant à des réalisations destinées à un seul locuteur, avec apprentissage. C'est en particulier l'approche qu'a décicé de suivre notre Projet "Enfants Sourds".

6.0 ANNEXE

6.1 PRECISION DES RESULTATS

L'erreur commise en prenant pour probabilité p d'un groupement sa fréquence f dans un corpus de longueur L est:

$|p-f| < 2\sqrt{(p(1-p))/L}$ dans 95% des cas

Pour notre corpus de 95000 phonèmes, l'application numérique indique que les fréquences des phonèmes sont connues avec précision (erreur relative 2 à 7%), celles des diphones et triphones les plus fréquents de façon assez précise (erreur relative de l'ordre de 10 à 20%). L'imprécision sur les fréquences des groupements très peu fréquents est par contre assez forte.

6.2 EXEMPLES DE TEXTE TRAITE

sɑ̃ k ɔ̃ pɥis ale / osi lwɛ̃ / kə sø ki vwa / katRəvɛ̃sɛ̃ puR sɑ̃ / də la pɔpylasjɔ̃ / gRupe / dɑ̃ le sɑ̃tRəz yRbɛ̃ / tɛl kə nu le z amenaʒɔ̃ / mɛ̃tnɑ̃ / mwɛ̃ de kɔ̃sɑ̃tRasjɔ̃ / kə dez ɛR / dɑ̃ lekɛl / lez eʃɑ̃ʒ / pɛRmetRɔ̃ / yn gRɑ̃d / kɔ̃sɑ̃tRasjɔ̃ ymɛn / e ɑ̃tRə lekɛl / l ɛspɑs RyRal / dəvjɛ̃dRa / vRɛsɑ̃blablәmɑ̃ / də plyz ɑ̃ ply elemɑ̃tɛR / o pwɛ̃ d vy / dez aktivite / ekɔnɔmik //ɛ s kə

FIG.1: Exemple de texte phonétique original

SA r O PЧIS alI / OSI lwI / rə SO rI Vwa / raTrəVISI PYr SA
/ Tə la POPYlaSjO / rrYPI / TA lI SATrəS YrPI / TEl rə TY lI
S aPITaHO / PITTA / PwI TI rOSATraSjO / rə TIS Er / TA lIrEl
/ lIS IHAH / PErPITrO / YT rrAT / rOSATraSjO YPET / I ATrə
lIrEl / l ESPAS rYral / TəVjITra / VrESAPlaPləPA / Tə PlYS A
PlY IlIPATEr / O PwI T VY / TIS arTIVITI / IrOTOPIr//E S rə

FIG.2: Après introduction des confusions labiales

6.3 FREQUENCES DES PHONEMES (%)

R	7.67	a	6.84	l	6.23	s	6.14	i	5.94	e	5.66	t	5.44
ɛ	5.43	k	4.55	d	4.51	ə	4.12	p	3.62	ɑ̃	3.56	m	3.28
n	2.91	ɔ	2.36	v	2.12	õ	2.09	j	2.05	y	1.95	u	1.85
z	1.55	f	1.31	ɜ	1.25	ɛ̃	1.15	o	1.04	b	1.01	w	.79
ɑ	.69	g	.58	œ̃	.53	Ч	.48	œ	.43	ʃ	.43	ø	.40
ɲ	.05												

FIG.3

7.0 REFERENCES

CORNETT (R.O.), 1967, Cued Speech. -American Annals of the Deaf, 1967, Volume 112, 3-13.

CORNETT (R.O.), BEADLES (R.), WILSON (B.), 1977, Automatic Cued Speech. -Conference on Speech-Analyzing Aids for the Deaf, Gallaudet College, Washington D.C., May 1977.

CAELEN (J), 1979, Un modèle de l'Oreille, Analyse de la Parole, Reconnaissance Phonémique. -Thèse d'Etat, Toulouse, 1979.

DENES (P.B), 1963, On the Statistics of Spoken English- Journal of the Acoustical Society of America, Volume 35, p.892, 1963.

DI BENEDETTO (M.D), DESTOMBES (F), TUBACH (J.P), 1980, Assistance Automatique à la Lecture Labiale. -Séminaire International sur la Labialité (GALF), CNET Lannion, Fév. 1980, 12 pages.

GARRIC (J), 1974, Lecture sur les Lèvres et Conservation de la Parole. -Editions Musicales et Universitaires, 1974, 109 pages.

GENTIL (M), 1980, Labialité en Français Etude Phonétique et Aspects physiologiques: Géométrie, Cinématique, Electromyographie et Idiolectologie. -Thèse de 3ème Cycle, 1980.

HATON (J.P), LAMOTTE (M), 1971, Etudes Statistiques des phonèmes et diphonèmes dans le Français parlé. -Revue d'Acoustique, 16, Déc. 1971.

MOREAU (R), 1962, Linguistique et Télécommunications. -L'Onde Electrique, Sept. 1962, pp 731-737.

NICHOLLS (G.), 1979, Cued Speech and the Reception of Spoken Language. -School of Human Communication Disorders, Mc Gill University, Montreal, 163p.

SILVERMAN (H), DIXON (R), 1976, A Comparison of several Speech Spectra Classification Methods. -IEEE Trans. ASSP 24 n.4, Aug 1976, pp 289-295.

SHANNON (C.E), 1951, Prediction and Entropy of Spoken English. -Bell System Techn. Journal., 30, 1951, 50-64.

TUBACH (J.P), 1969, Etude des contraintes Statistiques des Groupements Phonématiques. -Congrès Informatique au service de l'Homme, Grenoble, 1969, pp 31-48.

YAGLOM (A.M), YAGLOM (I.M), 1969, Probabilité et Information. -Dunod,1969, 320 pages.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

APPLICATION DE LA SYNTHESE DE LA PAROLE A LA COMMUNICATION DES HANDICAPÉS

Françoise EMERARD
Patrick GRAILLOT
Georges CYNE
Jean - Jacques LUCAS

CNET – LANNION

.RESUME

Cette étude s'intègre dans un programme de recherche concernant l'aide à la communication des handicapés.

L'idée d'utiliser pour les infirmes moteurs cérébraux privés de la parole les résultats des travaux sur la synthèse de la parole est séduisante. Le CNET dispose d'un système de synthèse opérationnel mais l'utilisation d'un clavier standard pour composer un message est inadapté pour des personnes atteintes gravement dans leurs fonctions gestuelles.

C'est pourquoi, pour deux «prothèses de parole» actuellement à l'étude, l'entrée du système de synthèse se compose de tableaux de lettres et/ou de mots : à l'aide d'un détecteur approprié, l'utilisateur opère un balayage du tableau et sélectionne les cases qui représentent les éléments de son message.

Quand le message est rédigé sous une forme orthographique, la sortie sonore est directement réalisée à partir de la transcription orthographique - phonétique ; mais quand le message initial ne représente qu'une suite de mots - clés, la chaîne phonétique ne peut être reconstituée qu'après intervention d'un module syntaxique.

Ces prothèses sont en cours d'expérimentation dans des Centres de Réadaptation. C'est maintenant aux utilisateurs et au personnel de rééducation d'évaluer leur efficacité comme outil pédagogique et comme outil de communication.

SPEECH PROSTHESIS TO AID THE COMMUNICATION OF HANDICAPPED PERSONS WITH VOCAL DIFFICULTIES

. ABSTRACT

During the last few years, the production of non - vocal communication systems for severely handicapped persons, with essentially normal hearing but unable to speak, has considerably furthered. The development of studies about speech synthesis has drawn the attention of many researchers on an entirely new solution : the functional speech prosthesis.

We present a work being conducted using a talking computer and an orthographic (Figure 1) and a Blissymbol scanning board.

Blissymbolics is a global communication system using a set of visual ideograms which represent the meaning of words and ideas in a simple way readily associated with a familiar referent.

The Blissymbol scanning board we use is arranged with 500 lexical and syntactic cells (Figure 2).

The user indicates the desired sequence by pressing a button to start the scanner moving and by pressing it again to illuminate the desired symbols. Then with the selected symbols, the computer produces - with the aid of a grammar prepared especially for the purpose (Graillot, 1981) - both orthographic and phonetic, syntactically reconstructured strings. Then, the phonetic string is translated into a vocal output with an automatic prosodic processor and the orthographic string is printed on a small screen and on a type-writer.

Experiments are conducted in Readaptation Centers with disabled children. They will determine whether the goals of this study have been reached :

- to further the range of communication in an oral and/or graphic form,
- to encourage different ways of learning to read and spell, and of grasping the syntactical structures in French.

APPLICATION DE LA SYNTHESE DE LA PAROLE A LA COMMUNICATION DES HANDICAPÉS

Françoise EMERARD
Patrick GRAILLOT
Georges CYNE
Jean - Jacques LUCAS

CNET – LANNION

INTRODUCTION

Aujourd'hui, le CNET dispose pour la langue française d'un système de synthèse opérationnel (EMERARD et GRAILLOT, 1978). A partir d'une entrée orthographique sur clavier, il est possible d'obtenir la sortie vocale de n'importe quel message de la langue avec une parole intelligible et relativement naturelle.

Le champ d'application de la synthèse est vaste, du secteur du jouet à celui des banques de données en passant par les télécommunications. Marginales sans doute quant aux débouchés, les applications concernant les handicapés figurent pourtant parmi les plus intéressantes : elles impliquent des dispositifs de synthèse relativement élaborés. Si dès maintenant on peut envisager, grâce à la synthèse du texte, des équipements réalistes pour une population qui ne présente que des troubles de la parole (muets et laryngectomisés par exemple), il n'en va pas de même pour des personnes gravement atteintes sur le plan moteur et ayant des déficits d'expression orale.

C'est le cas d'une étiologie comme l'Infirmité Motrice Cérébrale où les enfants sont porteurs de séquelles de lésions cérébrales, anté et péri - natales, non évolutives et sans déficits intellectuels prépondérants. C'est également le cas de certains traumatismes crânio - cérébraux.

Pour ces populations, un grand nombre d'outils de communication existe déjà. Mais ils n'autorisent, pour la plupart, que des échanges **non - verbaux** pour des individus en contact direct : l'utilisateur parvient à désigner les cases d'un tableau de caractères alphabétiques, de mots ou d'images, et une personne de l'entourage intervient pour assurer la diffusion orale du message ainsi composé (VANDERHEIDEN, 1976 et 1978).

Avec le développement des techniques de synthèse de la parole, d'autres voies de recherche sont explorées, principalement pour l'instant en Amérique du Nord : des outils réalisent la sortie vocale de mots sélectionnés sur un tableau de désignation. En fait, il ne s'agit ici que de donner une correspondance orale en synthèse par mots au contenu des cases sélectionnées ; les messages autorisés sont donc obligatoirement limités au vocabulaire stocké ; et, si le procédé est suffisant pour assurer la compréhension, il présente l'inconvénient de laisser l'enfant handicapé dans le ghetto d'un langage non structuré (du type «moi - vouloir - manger»). Dans ces conditions, le personnel de rééducation ne peut pas se dispenser d'intervenir pour restructurer syntaxiquement le message à partir de ces composantes de base.

Les deux limites exposées ci-dessus :

— limites d'une communication purement écrite qui nécessite la présence et un rôle actif de l'entourage,

— limites de la technique de synthèse par mots qui ne donne accès qu'à des possibilités d'expression orale très rudimentaires, nous ont conduit à concevoir deux «prothèses de parole» - organisées autour de tableaux de désignation spécialement conçus et du système de synthèse par diphones - destinées à fournir la sortie sonore de tout message construit à partir de caractères alphabétiques, de mots ou de symboles.

DESCRIPTION ET FONCTIONNEMENT DE LA PROTHESE SAHARA I

(Synthétiseur Assurant aux Handicapés une Aide dans leurs Relations avec les Autres)

Des pressions successives sur un détecteur d'entrée permettent d'effectuer un balayage sur un tableau de désignation avec déplacement d'un spot lumineux et de sélectionner des suites de cases correspondant à des unités linguistiques différentes.

Une fois le message achevé et inscrit sur un écran, l'organe de synthèse - utilisant le synthétiseur d'un vocodeur à 14 canaux durant la phase expérimentale - en réalise la sortie sonore.

Le tableau de désignation :

Le tableau (Figure 1) comporte 81 cases (9 x 9) de 2,5 cm de côté dont chacune peut être éclairée par un spot lumineux.

GRILLES / Fonctions	1	2	3	4	5	6	7	8
SYNTHESE DIPHONES	ESPACE	T	R	ESPACE	U	W	K	accent aigu
Répétition phrase	E	A	N	D	M	V	Q	accent grave
RETOUR	O	S	C	Y	G	J	?	accent. circonflexe
changement grille	I	H	F	Z	1	4	7	trema
	L	P	X	;	2	5	8	cedille
	⇇	B	, virgule	–	3	6	9	E collé
	.	' apostrophe	←	*	ESPACE	0	@	ESPACE
	.	.	/	#	$	&	=	ESPACE

Fig. 1 : Tableau de désignation «SAHARA I»

Le déplacement du spot lumineux s'effectue en X, Y de façon automatique ou manuelle selon les préférences ou les capacités de l'utilisateur.

Au repos, le spot se trouve en haut à gauche (0, 0). La première colonne de gauche est réservée à des fonctions :

(0, 1) : sortie vocale du message avant traitement prosodique,

(0, 2) : sortie vocale du message après traitement prosodique,

(0, 3) : après synthèse, retour en début de programme pour une nouvelle composition de message,

(0, 4) : changement de grille.

La première rangée correspond à des numéros de grille. Une grille contient les 64 cases qui ne sont ni dans la première rangée ni dans la première colonne. Si aucune commande (0, 4) n'est effectuée, c'est la grille n° 1 (Figure 1) qui est prise en compte (option par défaut) : cette grille permet de composer caractère par caractère tout message sous forme orthographique. Si une commande (0, 4) est associée à un numéro de grille (8 grilles sont prévues), c'est au vocabulaire proposé par la grille choisie que l'utilisateur va avoir accès (substantifs, verbes conjugués, adjectifs, expressions courantes, généralement représentés par des images) après sa mise en place par une personne valide de l'entourage.

Deux cases sont prévues à des fins de correction :

- caractère ⟵ pour effacer la dernière case sélectionnée,
- caractère ⟵⇇ pour effacer l'ensemble du message.

Les marqueurs prosodiques les plus fréquents sont également inscrits pour les utilisateurs qui ont déjà une bonne connaissance de la syntaxe du français.

Les détecteurs d'entrée :

Chaque utilisateur présente des possibilités physiques très spécifiques. Les détecteurs ont été conçus de manière à pouvoir réagir à l'action de la tête, d'un pied, d'une ou deux mains, etc. . .

Plusieurs types de détecteurs (infra-rouge, ultra-son, radar, acoustique ou d'humidité de la langue), adaptés au handicap, pourront être reliés aisément à un connecteur unique du tableau de désignation.

Dans sa version prototype, la prothèse n'utilise que des détecteurs de type pneumatique et par contact électrique :

— la pression à exercer est réglable électroniquement de façon à être adaptée à différentes forces des membres,

— une temporisation réglable à l'aide d'un potentiomètre situé sur la face latérale du tableau agit sur la validité de la commande et permet d'éviter les pressions parasites dûes à un geste incontrôlé.

Chaque type de détecteur est réalisé selon trois modes différents de sélection :

— Détecteur à un seul bouton et à trois pressions :

1e pression : déplacement sur l'axe des X,
2e pression : déplacement sur l'axe des Y,
3e pression : sélection et validation de la case.

Le déplacement du spot de case en case se fait automatiquement et la vitesse de déplacement est définie au départ par chaque utilisateur.

— Détecteur à deux boutons :

1e bouton : progression sur l'axe des X,
2e bouton : déplacement sur l'axe des Y,
1e bouton à nouveau : sélection et validation de la case sur laquelle est arrivé le spot.

Le déplacement du spot est réalisé manuellement sur décision de l'utilisateur.

— Détecteur à trois boutons avec déplacement automatique de case en case :

1e bouton : déplacement sur l'axe des X,
2e bouton : déplacement sur l'axe des Y,
3e bouton : sélection et validation de la case.

L'écran de visualisation :

Le prototype de prothèse est équipé d'une console classique qui permet la visualisation du contenu des cases choisies.

La présence de cet écran est impérative :

— pour contrôler visuellement que ce qui est affiché correspond bien à ce qui a été sélectionné,

— pour vérifier à tous moments et mémoriser la progression de la composition du message : le temps qui s'écoule entre la sélection de deux caractères ou de deux mots est parfois si long (entre 8 et 10 secondes pour un handicapé «habile») que la mémorisation du contenu de la case précédente représente une tâche réellement complexe et justifie la visibilité de tous les caractères déjà validés.

A la réception du caractère «point» qui doit nécessairement terminer chaque message, la transcription orthographique - phonétique est effectuée, puis le traitement prosodique, enfin la synthèse du message.

Depuis octobre 1980, cet outil est opérationnel et est expérimenté dans un Centre de Réadaptation pour Infirmes Moteurs Cérébraux de la banlieue parisienne.

Pour des raisons de facilité, l'organe de synthèse (synthétiseur du vocodeur à canaux et calculateur) demeure au CNET ; au Centre de Réadaptation, seuls sont visibles le clavier de désignation et la console de visualisation. Les échanges se font par l'intermédiaire de deux circuits téléphoniques.

Il est évidemment trop tôt pour tirer des conclusions, mais on remarque déjà combien les enfants - habitués à manipuler des outils identiques de désignation non-verbale - se sont adaptés rapidement et sans difficultés au maniement de ce tableau.

Comme l'outil en lui-même présentait un aspect familier et un mode de fonctionnement assez peu complexe, les enfants n'ont pas été perturbés et se sont montrés immédiatement très fascinés et très excités par les possibilités nouvelles offertes par le synthétiseur de parole. Leur motivation s'est accrue dans deux directions : au niveau de l'apprentissage de l'écriture et au niveau de la communication. Quand le son obtenu ne correspond pas au désir d'expression, on observe que les enfants cherchent l'erreur et interviennent spontanément - sans pression de l'entourage - pour modifier l'orthographe de leur message. Le fait qu'il n'y ait plus d'intermédiaire humain entre le graphisme et le son crée une sensation d'autonomie (la possibilité d'interprétation, de censure de la part de l'entourage est supprimée), de «responsabilité» vis-à-vis des mots qui semble stimulante au niveau de l'envie de communiquer.

Ces remarques rejoignent les propos - rapportés lors d'une expérience similaire menée en Suède (CARLSON et al., 1980) - d'un enfant IMC de 15 ans : «Quand on communique toujours à travers les autres, ça entraîne la passivité».

DESCRIPTION ET FONCTIONNEMENT DE LA PROTHESE

SAHARA II

Destinée également aux infirmes moteurs cérébraux, cette prothèse vocale a une vocation plus universelle : elle veut pouvoir s'adresser indifféremment :

— à une population scolarisée pouvant rédiger des messages sous une forme orthographique,

— à une population en cours de scolarisation et ne maîtrisant ni l'écriture ni la lecture,

— à une population pour qui la composition d'un message caractère par caractère représente un effort et un temps trop considérables.

Les approches traditionnelles d'apprentissage du langage et de la communication pour les enfants IMC ont été pendant longtemps concentrées sur des tableaux à base de lettres, de mots, d'images ou d'expressions courantes de la langue ; ces tableaux faisaient référence à un niveau sémantique et syntaxique mais de façon très partielle. Les limites de ces approches, liées soit au caractère restrictif du vocabulaire accessible, soit aux difficultés de compréhension par l'entourage ont conduit un certain nombre de cliniciens à adapter d'autres systèmes de communication existants. C'est ainsi que le système Bliss a été utilisé.

Le Blissymbolisme est l'adaptation d'un langage universel à base de symboles visuels conçu par Charles K. Bliss pour «développer une sorte d'Esperanto idéographique ayant une logique et une éthique qui amènerait une meilleure compréhension entre les peuples de la terre et surmonterait ainsi les barrières culturelles», (Bliss, 1965).

Ce système comprend :

— un nombre restreint de symboles pictographiques arbitraires qui représentent de façon simple le sens des mots et qui peuvent se combiner entre eux pour former de nouvelles significations,

— un ensemble de signes diachritiques pour indiquer à quelle catégorie grammaticale réfère le mot (objet, action) et pour prendre en compte quelques notions syntaxiques (passé, présent, futur, pluriel. . .).

Dans l'esprit des responsables du Centre de handicapés physiques de Toronto, promoteurs de ce système symbolique de communication, «les symboles en tant que substituts à la parole, peuvent fournir un pré-requis à l'actualisation des modes de communication plus avancés tels que la lecture et l'écriture» (SYLVESTRE, 1979).

C'est un tableau de type Bliss qui a été utilisé ici comme entrée du système de synthèse ; il comporte 486 cases qui donnent pour chaque unité linguistique une double représentation (Figure 2) :

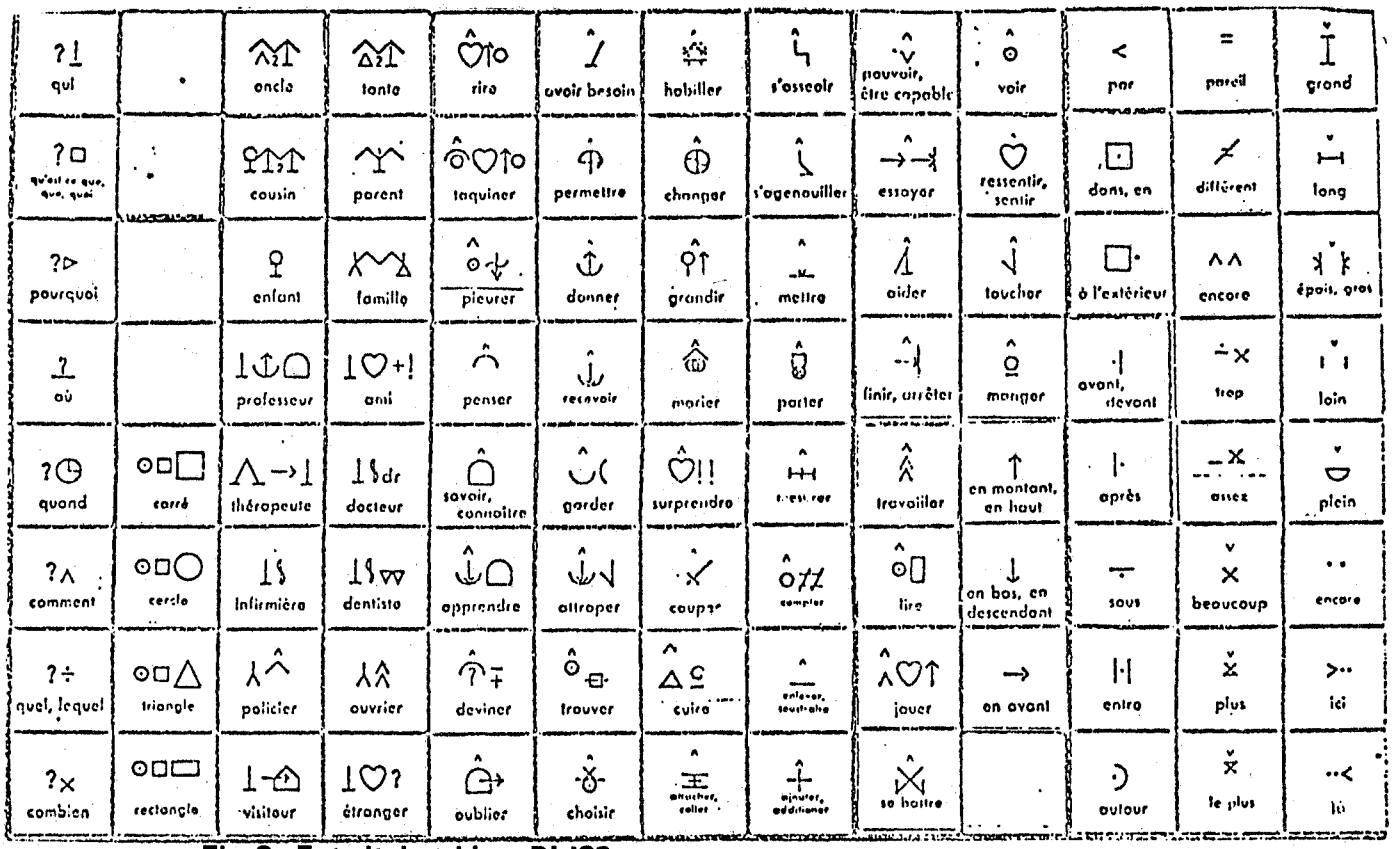


Fig. 2 : Extrait du tableau BLISS

— sous sa forme orthographique,
— sous une forme «idéographique» dont l'utilisateur connait la signification.

Quand ce tableau est utilisé comme moyen de communication non verbale, l'utilisateur pointe successivement les cases qui contiennent les mots qu'il veut énoncer et une personne intervient pour reconstituer à haute voix le message à partir de ses éléments de base.

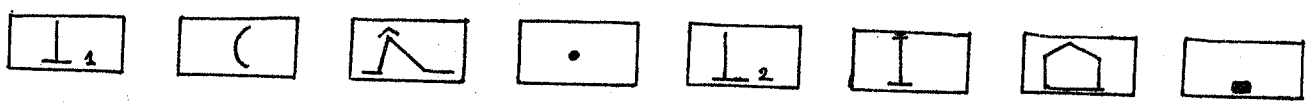
Ici, pour réaliser une prothèse vocale, le clavier de désignation de type Bliss a été connecté au système de synthèse par diphones utilisant un synthétiseur à codage prédictif et à un programme permettant de transformer la suite des cases sélectionnées (les mots - clés) en une phrase syntaxiquement correcte (GRAILLOT, 1981).

Comme pour la prothèse précédemment décrite, un détecteur permet à l'utilisateur de désigner les cases qui contiennent les maillons de son message : un spot lumineux se déplace de case en case ; une pression sur le détecteur réalise la sélection et la validation de la case sur laquelle est arrivée le spot.

Au fur et à mesure de la sélection des cases, leur contenu sous forme orthographique va s'inscrire sur un petit afficheur solidaire du clavier ; son rôle est double :

— permettre le contrôle visuel de la case choisie,
— faciliter la mémorisation de l'orthographe d'un mot et donc habituer l'utilisateur à faire le lien entre l'idéogramme et l'écriture.

Si par exemple, les cases suivantes ont été retenues



apparaîtra sur un premier écran la suite



Ensuite, le passage de la suite de ces mots - clés à une suite syntaxique et grammaticale est effectuée : un lexique de tous les concepts du tableau Bliss a été stocké. Il comporte les radicaux de tous les mots (inscrits à la fois sous le code phonétique défini pour la synthèse et sous le code orthographique), leur catégorie lexicale, et pour chaque catégorie un renvoi aux règles correspondantes de conjugaison.

Deux traitements distincts sont effectués sur la suite des mots désignés :

— les conjugaisons et élisions en code orthographique et phonétique, et les liaisons en code phonétique,

— le positionnement automatique des marqueurs prosodiques définis pour la synthèse (EMERARD, 1977).

De cette façon, on peut réaliser une seconde sortie visuelle du message après sa construction - «j'irai dans ta grande maison» - et une sortie sonore avec prosodie et sans les possibles erreurs provoquées par le programme de transcription orthographique - phonétique dont on s'affranchit ici.

REMARQUES GÉNÉRALES SUR LA MÉTHODE

* Le vocabulaire proposé, bien qu'il ait été étudié pour satisfaire les besoins de communication élémentaires, est très limité (moins de 500 termes) et souvent l'utilisateur devra avoir recours à l'écriture orthographique pour compléter ses messages. Dans le cas d'une extension du vocabulaire de base :

— la connaissance de l'orthographe est nécessaire,

— la vitesse de composition est réduite,

— le programme de transcription orthographique - phonétique est utilisé et peut provoquer certaines erreurs.

* La sélection d'une case par balayage d'un tableau de grande dimension demeure longue. Il faudra envisager sans doute d'autres procédés pour réduire la taille du clavier sans appauvrir le lexique ; par exemple, un outil comme Handivoice (HC 110 - Model II) ne comporte que 128 cases qui peuvent donner accès à quatre niveaux différents de vocabulaire.

* La méthode utilisée présente un caractère restrictif dans la mesure où elle s'adresse à des enfants qui, par mémorisation visuelle et auditive, sont supposés avoir acquis une maîtrise de la syntaxe suffisante pour connaître l'ordre correct d'apparition des mots dans une phrase.

* Les programmes sont indépendants des mots eux-mêmes puisqu'ils travaillent sur les catégories lexicales : on peut ainsi très facilement changer des mots dans le tableau Bliss pour personnaliser le vocabulaire aux besoins de chaque utilisateur ou même introduire d'autres données d'entrée.

* La grammaire retenue est suffisamment riche pour autoriser des formes relativement complexes. En ce sens, l'outil doit se révéler être un bon instrument d'apprentissage de la syntaxe du français.

* Quand les symboles proposés sur le tableau sont suffisants pour satisfaire les besoins d'expression verbale, l'utilisateur est dispensé, pour communiquer, de la connaissance de l'orthographe.

CONCLUSION

Il est aujourd'hui prématuré de porter des jugements sur les effets thérapeutiques et éducatifs de telles prothèses.

Des outils très voisins sont en cours de développement aux USA (EULENBERG RAHIMI, 1978 ; COLBY et al., 1978) et les témoignages recueillis durant les phases d'expérimentation font état d'impressions très diverses : alors que certains adultes IMC regrettent de n'avoir pû bénéficier dans leur enfance de telles technologies, d'autres considèrent que ces substituts électroniques présentent un caractère dégradant, susceptible d'accentuer le handicap, et peuvent être vécus comme une rupture de liens exclusifs de dépendance avec l'entourage (SYLVESTRE, 1979).

Ces remarques montrent que la plus grande prudence est nécessaire, tant dans la conception de l'outil que dans l'information dispensée auprès des utilisateurs et du personnel de rééducation.

En ce sens, l'expérimentation de quatre prototypes de prothèse qui va être effectuée dans différents Centres de Réadaptation durant 1981 sera décisive avant d'envisager le développement de prothèses individuelles et portables.

BIBLIOGRAPHIE

BLISS, C.K, 1965, «Semantography» ; Semantography Blissymbolics Publications ; 2e édition ; SYDNEY, AUSTRALIA.

CARLSON, R., GALYAS, K., GRANDSTROM, B., PETERSON, M., ZACHRISSON, G., 1980, «Speech synthesis for the non - vocal in training and communication», KTH, Speech Transmission Laboratory, Quarterly Report 1, 13 - 27.

COLBY, K.M., CHRISTINAZ, D., GRAHAM, S., 1978, «A computer - driven, personal, portable and intelligent speech prosthesis» ; Computer and Biomedical Research, 11, 4, 337 - 343.

EMERARD, F., 1977, «Synthèse par diphones et traitement de la prosodie» ; Thèse, Grenoble III.

EMERARD, F., et GRAILLOT, P., 1978, «Test d'intelligibilité de la parole synthétique transmise par téléphone» ; Institut de Phonétique de Grenoble, volume VII.

EULENBERG, J.B., et RAHIMI, M.A., 1978, «Toward a semantically accessible communication aid» ; Proceedings of the National Electronics Conference, Vol. XXXII.

GRAILLOT, P., 1981, «Commandes pour la synthèse à partir de mots - clés», XIIe JEP, GALF, MONTREAL.

SYLVESTRE, A., 1979, «Modes de communication comme substituts ou compléments à la parole chez les Infirmes Moteurs Cérébraux ayant des déficits d'expression orale» ; XVIIe Journées de l'Association de Psychologie Scientifique de Langue Française, Barcelone 20 - 22 sept.

VANDERHEIDEN, G.C., et GRILLEY, K., 1976, Non - vocal communication techniques and aids for the severely physically handicapped», University Park Press, BALTIMORE.

VANDERHEIDEN, G.C., 1978, Non - vocal communication resource book, University Park Press, BALTIMORE.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

SIMULATION D'UN LECTEUR AUTOMATIQUE DU FRANCAIS

O'SHAUGHNESSY, Douglas
LENNIG, Matthew
MERMELSTEIN, Paul

INRS-Télécommunications
3, Place du Commerce
Ile des Soeurs, Québec

DIVAY, Michel

Institut Universitaire de
Technologie de Lannion-France

.RESUME

Nous exposons les résultats concernant la simulation sur ordinateur d'un système qui permet de convertir un texte français dactylographié en discours sonore. La version pratique d'un tel lecteur devrait pouvoir accepter une page d'un livre ou d'un journal et lire le texte de cette page à haute voix, à une vitesse pouvant atteindre 200 mots par minute et avec un naturel raisonnable. Ces lecteurs existent actuellement pour l'anglais, mais pas pour le français. Pour effectuer ce project il a fallu développer un logiciel pour la reconnaissance optique des caractères (ROC) afin de reconnaître la sortie d'un dispositif d'exploration, écrire un nouveau logiciel et modifier des algorithmes existants pour le traitement linguistique et la génération de la parole. La recherche a été axée principalement sur quatre domaines:

a) Le développement d'algorithmes pour convertir la sortie d'un dispositif ROC en texte français.
b) Le développement d'un ensemble adéquat de règles pour la conversion de graphèmes en phonèmes pour le français.
c) L'élaboration d'un modèle d'intonation pour le français.
d) La modification d'un synthétiseur actuel de parole de l'anglais pour l'appliquer au français.

Le système permet actuellement la génération en temps non réel, mais d'un manière complètement automatique, d'un discours continu hautement intelligible.

. ABSTRACT

SIMULATION D'UN LECTEUR AUTOMATIQUE DU FRANCAIS

O'SHAUGHNESSY, Douglas
LENNIG, Matthew
MERMELSTEIN, Paul

INRS-Télécommunications
3, Place du Commerce
Ile des Soeurs, Québec

DIVAY, Michel

Institut Universitaire de
Technologie de Lannion-France

SUMMARY:

A reading machine for the blind generates speech output from typed text input. It consists of: an optical character recognizer, a linguistic processor, and a sound generation algorithm. In the first component, a hardware scanner produces a binary image from the printed page, and software interprets the image as a string of letters and punctuation, using template matching techniques. Assuming a single-font as input, the program gives virtually error-free performance in non-real time. The second component translates letters into phonemes, and provides intonation appropriate to the syntax of the message. In French, this requires deciding when to pronounce usually-silent word-final letters, and interpreting letters in general by rules different from English, such as with nasalized vowels. Similarly, the pitch and duration of individual sounds is different for French. Syllables have more regular durations in French, and pitch has a greater tendency to rise at the end of a French word than an English word. Without such rules, the French speech would have an English accent. The final component interprets the phonemes as variations in vocal tract resonances and amplitudes, and drives a speech synthesizer to yield sound output. It is based on a formant synthesis algorithm developed by Klatt of M.I.T., but modified for French by the addition of nasalized and front rounded vowels, as well as other changes related to different pronunciations in French and English. Without making such acoustic rule modifications, the speech would have a heavy English accent. The final stage of actual synthesis is currently in software, but will soon be replaced by an LSI chip.

SIMULATION D'UN LECTEUR AUTOMATIQUE DU FRANCAIS

O'SHAUGHNESSY, Douglas
LENNIG, Matthew
MERMELSTEIN, Paul

INRS-Télécommunications
3, Place du Commerce
Ile des Soeurs, Québec

DIVAY, Michel

Institut Universitaire de
Technologie de Lannion-France

1) INTRODUCTION

Un lecteur automatique acceptant à l'entrée un texte imprimé (sous forme de feuilles détachées, de journaux, de livres, ou même d'images de télévision) et produisant à la sortie un discours sonore (la version parlée du texte), est une entreprise socialement désirable et représente un défi technique. Ce lecteur serait d'un grande utilité pour les personnes dont la vue est handicapée, puisque de nos jours une grande partie de l'information est imprimée. Un petit nombre d'aveugles seulement peuvent lire le Braille, mais virtuellement tous comprennent le langage parlé et un lecteur automatique permettrait l'accès à l'information qui doit pour le moment être enregistrée sur bandes par des personnes dont la vue n'est pas handicapée et qui lisent le texte à haute voix.

Il y a environ trois ans, un lecteur a été développé pour la langue anglaise (KURZWEIL, 1976) et un deuxième dispositif similaire est sur le point d'apparaître (CALDWELL, 1978). Cependant, il n'existe actuellement aucun lecteur automatique parlant français, et à notre connnaissance personne n'en a commencé le développement. Nous nous sommes efforcés de simuler un lecteur automatique afin de démontrer la qualité de la parole synthétique quand l'entrée consiste en un texte français.

2) LES COMPOSANTS D'UN LECTEUR AUTOMATIQUE

Un lecteur automatique comprend trois composants: reconnaissance optique des caractères, traitement linguistique et synthèse de la parole.

2.1 Reconnaissance optique des caractères (ROC)

Ce composant effectue la conversion image-graphème, c'est à dire qu'il convertit l'image d'un texte imprimé en une suite de caractères (lettres, chiffres et symboles) contenus dans l'image. Dans notre système le premier étage du ROC utilise un explorateur automatique standard de page, qui produit une représentation numérique de la page, en interprétant l'image comme un ensemble constitué par des millions de points noirs et blancs. Sur une page blanche, évidemment, des ensembles avoisinants de points noirs sont interprétées par l'oeil humain comme des lettres. Notre dispositif d'exploration peut résoudre une image avec une précision de 82 points par centimètre; si l'on considère un exemple standard de 2.5 millimètres par caractère, chaque symbole serait échantillonné 20 fois dans sa largeur. Chaque échantillon est quantifié au moyen de 8 bits sur une échelle noir-blanc. Un seuil permet de décider quels sont les points gris que l'on devrait considérer comme étant noirs (faisant partie du symbole) ou blancs (faisant partie de l'arrière fond). Deux cents points par pouce sont suffisants pour résoudre les variations fines de presque tous les caractères, bien qu'un texte écrit fin comme par exemple un annuaire téléphonique et les avis publicitaires d'un journal pourrait exiger une resolution plus fine.

On a développé un logiciel faisant la conversion de l'image numérique en une suite de graphème. Nos programmes acceptent une fonte standard IBM, et produisent un taux d'erreur d'environ 0.2%, avec une méthode de reconnaissance de forme utilisant un gabarit de comparaison standard.

Les programmes segmentent l'image en régions correspondant approximativement aux lettres, en respectant les constraintes provenant de la dimension des lettres et en tenant compte du fait que des pages dactylographiées consistent généralement en lignes horizontales espacées d'une manière uniforme le long des dimensions verticales et horizontales. Dans un texte français, les accents se trouvent souvent dans une région séparée du reste de la lettre, mais un autre programme permet de les joindre ensemble, en se servant du contexte. Chaque région est comparée à une série de gabarits mis en mémoire, mais certains symboles spéciaux (comme "$"), dont la représentation numérique peut être très différente à cause de la résolution du dispositif d'exploration, possèdent deux gabarits mémorisés. Il est possible d'utiliser un ensemble simple de gabarits uniquement parce qu'on utilise à l'entrée du système une seule fonte.

Les gabarits sont ordonnés selon leur probabilité d'apparition, déduite d'une étude statistique grossière de la langue française. Ainsi, les voyelles qui surviennent plus fréquemment comme "e" et "a" sont les premières à être considerées, et les consonnes rares comme "x" et "k" sont les dernières. Puisque la comparaison se termine quand une correspondance adéquate a été trouvée, en considérant d'abord les candidats les plus probables, il est possible de réduire le temps moyen de calcul.

Au cours de l'exploration, certaines versions du même symbole peuvent être plus grandes que d'autres. Aussi la comparaison commence par centrer toutes les régions selon leurs limites maximum et ensuite elle les décale horizontalement et verticalement pour obtenir une

meilleure correspondance si l'entrée et les gabarits mémorisés sont suffisamment identiques.

2.2 Traitement linguistique

Ce composant du système effectue la conversion graphème-phonème, c'est à dire qu'il convertit la suite des lettres et des signes de ponctuation en une suite de phonèmes du français. Il produit aussi un ensemble de marqueurs d'intonation suffisamment détaillés pour pouvoir spécifier la hauteur et la durée de chaque phonème. La première tâche est effectuée au moyen de règles de transformation lettre-phonème, qui permettent la conversion de graphèmes en phonèmes en tenant compte des propriétés inhérentes du langage. On peut décrire un langage par un ensemble de lois qui lui sont propres. Certains langages peuvent être décrits d'un manière plus compacte que d'autres. Par exemple, l'Espagnol possède un ensemble de règles lettre-phonème relativement petit, alors qu'un ensemble de règles complet pour l'Anglais serait considérablement plus complexe. Le Français se trouve en quelque sorte entre les deux.

2.2.1 Conversion lettre à phonème

Une méthode lettre-phonème avec un petit lexique constitue probablement le meilleur moyen pour la conversion. Un ensemble compact de règles lettre-phonème peut tenir compte de la majorité des mots d'un langue, tandis que le lexique consiste principalement en mots que les règles lettre-phonème n'arrivent pas à décrire facilement (ce qui inclus souvent des mots très répandus, ainsi que des mots étrangers). Pour augmenter l'efficacité du lexique, on pourrait mémoriser les préfixes et les suffixes répandus, et les mots pourraient être d'abord décomposés en affixes et en racines avant de consulter le dictionnaire. Il n'est donc pas nécessaire d'enregistrer toutes les dérivations et conjugaisons d'un mot du lexique, mais de conserver uniquement ses composants de base. Cette méthode s'applique particulièrement bien au français dont les verbes possèdent de nombreuses terminaisons.

Le système de conversion lettre-phonème développé à l'Université de Rennes en collaboration avec le CNET-Lannion (Centre National d'Études des Télécommunications) (DIVAY et GUYOMARD, 1977) consiste en un ensemble de règles qui s'appliquent dans une situation orthographie donnée. Les règles sont ordonnées pour qu'on puisse examiner d'abord les liaisons et les enchaînements, puis le /s/ final du mot (qui peut être silencieux, par exemple "chats", "chiens", ou que l'on doit prononcer, par exemple "hélas"), ensuite les mots répandus, les mots réguliers et finalement l'élision de certaines lettres à la fin d'un mot. Cette manière de procéder spéciale pour les mots répandus permet au programme de fonctionner plus rapidement, puisqu'un grand pourcentage des mots n'auront pas à passer à travers l'ensemble des règles lettre-phonème. Les règles sont fonctions du contexte et se présentent sous la forme:

$$a \longrightarrow b \;/\; c + d$$

(la suite de symboles "a" est transformée dans la suite "b" si on trouve "a" dans le contexte "c-d", c'est à dire que "c" est à gauche et "d" est à droite). Par exemple, une voyelle suivie par "m" ou "n" serait considérée comme une voyelle nasalisée si la lettre suivante consiste en une consonne non-nasalisée, mais ne serait pas considérée

comme une voyelle nasalisée si elle était suivie par une voyelle (par exemple "lent" par apposition à "ennemi").

2.2.2 Intonation

Les paramètres fondamentaux de l'intonation, le hauteur et la durée, varient avec plusieurs paramètres du texte entrant (BEAUCHEMIN, 1972; BENGUEREL, 1971). La variation phonétique (par exemple les voyelles basses présentent une hauteur faible et de plus longues durées) peut être traitée directement à partir de la liste d'entrée de phonèmes. Mais puisque la hauteur et les durées des composants d'un mot dépendent aussi de la structure syntaxique et de la signification sémantique (CHOPPY et LIENARD, 1975), il est nécessaire d'effectuer une autre analyse linguistique du texte. L'absence de modèle adéquat de l'intonation pour de nombreux synthétiseurs est une cause principale de l'accent "étranger" de la parole synthétisée au moyen de règles.

On a développé un algorithme pour accepter à l'entrée la ponctuation et une analyse simple des parties du discours, ainsi que les phonèmes d'une phrase, et cet algorithme produit à sa sortie la fonction d'intonation. Le lexique comprend une liste de mots syntaxiques (les articles, les pronoms, et les prépositions) qui permet de distinguer les mots sémantiquement moins importants des mots de contenu (substantifs, verbes, adjectifs); ceux-ci ont des durées plus longues et des variations de hauteur plus importantes.

Nous avons modelé la fréquence fondamentale et les durées des phonèmes d'après le modèle décrit dans une analyse de la parole naturelle (O'SHAUGHNESSY, 1981; SMITH, 1977; WAJSKOP, 1979; METTAS, 1969). On a trouvé que la durée d'une voyelle varie énormément en fonction de la nasalité et des caractéristiques des consonnes suivantes (les voyelles nasalisées et les voyelles qui se trouvent devant les consonnes voisées sont allongées). Aussi, les consonnes pouvaient former des agrégats plus ou moins longs, dépendant de la difficulté d'articulation et de la proximité des points d'articulation (par exemple les durées de /s/ et de /t/ dans "reste" sont plus courtes que dans "messe" ou "mette", à cause de la similarité de l'articulation). La fréquence fondamentale tend à augmenter pour les mots importants, et demeure constante pour les mots moins importants, mais la rapidité des variations semble plus petite qu'en anglais.

L'algorithme de durée attribue une durée à chaque phonème d'une phrase donnée, et la modifie ensuite selon son contexte phonétique. Ainsi, par exemple, on attribue aux voyelles en général et aux voyelles nasalisées en particulier de longues durées (voir Tableau I), tandis que des durées plus courtes sont attribuées aux consonnes non voisées. Les durées de base sont alors ajustées par le contexte immédiat:

a) si une consonne apparaît dans un agrégat, sa durée est écourtée si les consonnes ont des caractéristiques phonétiques similaires, ou est allongée si l'agrégat de consonnes est difficile à prononcer

b) les voyelles sont écourtées si elles sont suivies de consonnes non voisées et allongées si elles sont suivies par des consonnes voisées (surtout les sons fricatifs). Pour le contexte de plus haut niveau, celui du mot, les durées, sont écourtées pour les mots

de plusieurs syllabes et pour les mots que l'on présume être moins importants et par conséquent moins appuyés.

La ponctuation affecte la durée, car les durées sont allongées pour le mot qui vient juste devant un signe de ponctuation (on suppose que ce signe nécessite une courte pause dans le discours). Puisque de longues phrases ne comportent pas nécessairement des signes de ponctuation, une analyse syntaxique simple choisit les candidats les plus probables pour les frontières phonologiques, et allonge les mots qui les précèdent.

L'algorithme de la fréquence fondamentale (F_0) attribue une direction générale, appelée ligne de déclinaison, aux membres principaux de la phrase. Dans ce cas F_0 est décalé dans le sens positif et diminue graduellement. Les modifications principales apportées à cette méthode générale sont dues à la ponctuation: un point indiquant la fin d'une phrase entraîne une diminution importante de F_0 tandis que pour un point d'interrogation F_0 demeure élevé tout au long, avec une augmentation majeur vers la fin de la phrase. (Si la question contient un mot d'interrogation, par exemple "qui", "pourquoi", "comment", le F_0 diminue). D'autres signes de ponctuation, comme les virgules, entraînent une plus petite augmentation de F_0 juste avant la pause qui marque l'emplacement de la ponctuation. A l'intérieur d'une phrase (unités phonologiques entre les pauses présumées), F_0 varie selon l'importance présumée du mot et le nombre de syllabes qui composent le mot. L'importance d'un mot dépend du niveau d'emphase de la phrase: soit appuyé, soit non appuyé. Les mots non appuyés possèdent un F_0 qui diminue (mais qui n'atteint pas une valeur assez basse pour être au-dessous de la ligne de base de déclinaison, que la fin d'une phrase pourrait indiquer). Les mots appuyés nécessitent une augmentation de F_0 dans leur syllabe finale et parfois dans la première syllabe aussi (pour les cas contenant plusieurs syllabes, les syllabes du milieu nécessitent une légère diminution de F_0).

2.3 Synthèse de la parole

La synthèse de la parole (QUINIO et TEIL, 1970; VASSIERE, 1971; FLANAGAN et RABINER, 1973; FLANAGAN et al, 1970) convertit une suite de phonèmes et de marqueurs d'intonation en un discours, et consiste en deux parties: un convertisseur phonème-paramètre et un modèle du conduit vocal. Le modèle du conduit vocal est actuellement simulé au moyen d'un logiciel FORTRAN, mais sera remplacé bientôt par une pastille LSI pouvant effectuer la synthèse en temps réel. Le convertisseur phonème-paramètre constitue un choix de méthode fondamental. Il serait possible d'enregistrer les élements de la voix humaine et de les traiter au moyen d'une analyse prédictive linéaire (PL) pour réduire la mémoire utilisée. Alors, lorsque différents sons devraient apparaître en séquence dans une phrase donnée, les composants mémorisés pourraient être transformés et enchaînés pour produire le

discours de sortie. Alternativement, on peut effectuer la synthèse phonétiquement en utilisant une simulation directe des paramètres établis de la parole (c'est à dire, manipulation des résonnances des formants, des bandes passantes et des amplitudes).

La méthode d'enchaînement par PL possède un accent machine moins prononcé mais elle est limitée à une simulation de la voix du locuteur original. Elle est sujette aux dégradations du codage PL et elle nécessite une mémoire plus grande. La méthode phonétique peut potentiellement produire un discours parfait mais pratiquement, elle possède en général un accent machine. Nous avons choisi la deuxième méthode et avons modifié un logiciel existant de synthétiseur développé par KLATT (1980) au M.I.T. (et utilisé actuellement pour le lecteur automatique de Telesensory Systems) pour s'appliquer aux phonèmes de français. Par exemple, nous avons ajouté les voyelles nasalisées et les voyelles arrondies antérieures que l'on trouve dans la langue française mais pas dans la langue anglaise, et nous avons éliminé le phénomène anglais de diphtongaison. Un autre changement majeur a été apporté aux consonnes "liquides" /l/ et /r/. Le /l/ français est articulé en avançant plus la langue vers l'avant et avec un contact moindre avec le palais. Ceci tend à augmenter le deuxième formant comparé avec le /l/ anglais. Le /r/ anglais est un son rétroflèchi avec un troisième formant très bas. Le /r/ français, d'autre part, est prononcé sans rétroflexion du bout de la langue et généralement avec un point de constriction vélaire. L'allophone non voisé du /r/ est souvent accompagné d'un son fricatif. Cet allophone apparait quand le /r/ est adjacent à une consonne non voisée de la même syllabe (par exemple "parc", "trac" - non voisés: "large", "vrai" - voisés). Le Tableau II indique les valeurs couramment utilisées pour les trois premiers formants de toutes les sonorités françaises, et le Tableau III indique les valeurs pour les consonnes ordonnées selon les points d'articulation (puisque les consonnes ayant la même articulation, comme /t/, /d/ et /n/, possèdent les mêmes cibles).

3) CONCLUSIONS

Au cours de la première année consacrée au développement du lecteur automatique parlant français, la recherche et le développement ont été orientés vers la possibilité d'un démonstration en temps non réel. Nous pouvons maintenant démontrer une simulation de lecteur automatique pour lequel:

a) Le ROC fournit une suite de symboles à partir d'une page dactylographiée.

b) Cette suite de symboles est ensuite traitée par des logiciels successifs qui effectuent une reconnaissance des formes, un traitement linguistique et l'élaboration des paramètres de la parole.

c) La suite de paramètres sert à commander un synthétiseur logiciel qui calcule le signal numérique de parole. La parole synthétique, produite en temps non réel, a été jugée intelligible et de qualité acceptable au cours de plusiers tests d'audition. D'autres essais perceptuels objectifs seront effectués très prochainement.

Au cours de l'année qui vient, nous construirons les dispositifs nécessaires pour obtenir une lecture en temps réel et pour remplacer certaines fonctions du logiciel. L'utilisation d'un petit microprocesseur devrait aboutir à une économie dans les temps de calcul. Lorsque l'interface de tous les composants sera réalisé, il sera possible d'obtenir une démonstration en temps réel, produisant un discours alors que la page du texte est en train d'être explorée. Cette deuxième démonstration devrait mettre en valeur le fonctionnement du système dans les conditions d'utilisation désirées, c'est à dire, avec une personne manoeuvrant le ROC et écoutant le discours de sortie simultanément.

REFERENCES

Beauchemin, N. (1972) "Corrélation des durées sous l'accent en français", 7th Intern. Cong. on Phonetic Sciences (1971), A. Rigault et R. Charbonneau (Mouton: the Hague), 860-865.

Benguerel, A., (1971) "Duration of French vowels in unemphatic stress", Language & Speech 14, #4, 383-391.

Caldwell, J., (1978) "Flexible, high performance speech synthesizer", J. Acoust. Soc. Am. 64, S1, S72.

Choppy, C. et Lienard, J.S., (1975) "Un algorithme de prosodie automatique sans analyse syntaxique", J.E. Toulouse.

Divay, M. et Guyomard, R., (1977) "Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique", thèse de 3ème cycle, Université de Rennes.

Flanagan, J., Coker, C., Rabiner, L., Shafer, R., et Umeda, N., (1970) "Synthetic voices for computers", IEEE Spectrum 7, 22-45.

Flanagan, J. et Rabiner, L., (1973) "Speech Synthesis", (Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania).

Klatt, D., (1980) "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. 67, 3, 971-995.

Kurzweil, R., (1976) "The Kurzweil reading machine: a technical overview", Science, Technology, and the Handicapped, AAAS Meeting, 3-7.

Mettas, O., (1969) "Etude sur la durée des consonnes dans l'un des parlers parisiens", Studia Linguistica 22, #2, 91-103.

O'Shaughnessy, D., (1981), "A Study of French vowel and consonant durations", J. of Phonetics, à apparaître.

Quinio, J. et Teil, D., (1970) "La synthèse de la parole à partir des diagrammes phonétiques", Revue Acoustique 3, no. 9.

Smith, A., (1977) "The timing of French, with reflections on syllable timing", Works in Progress, U. of Edinburgh 10, 97-108.

Vassière J., (1971) "Contribution à synthèse par règles du français", thèse de 3ème cycle, Université de Grenoble.

Wajskop, M., (1979) "Segmental durations of French intervocalic plosives", Frontiers of Speech Communication Research, B. Lindblom & S. Ohman (Academic Press: N.Y.), 109-123.

TABLEAU I - Durées de base des phonèmes (en millisecondes)

PHONEMES	DUREE
i,y,u	90
ø,ɛ,ɔ,a,oe	90
e,o,ɑ	140
ə	70
ɛ̃,ɑ̃,ɔ̃,œ̃	200
p,t,k	116
b,d,g	109
f,s,ʃ	176
v	115
z	128
ʒ	150
m	108
n	123
ɲ	142
l	101
r	113
w,j,ɥ	143

TABLEAU II - Valeurs cibles des 3 premiers formants pour les sonorités (en Hz)

PHONEME	EXEMPLE	F1	F2	F3
i	lit	250	2500	3000
e	bébé	360	2010	2600
ɛ	lait	480	1820	2470
ɑ	pâte	660	1160	2410
a	patte	720	1320	2400
ɔ	bonne	630	920	2430
o	beau	420	690	2400
u	bout	290	810	2130
ə	debout	470	1320	2070
y	sur	300	1800	2200
ø	feu	420	1510	2120
oɛ	boeuf	560	1300	2120
ɛ̃	fin	500	1600	2500
œ̃ɛ	brun	500	1500	2500
ɑ̃	sans	600	1000	2250
ɔ̃	bon	400	750	2200
w	bois	285	610	2150
j	biais	240	2070	3020
ɥ	huile	330	1400	2200
l	lit	330	1400	2400
r	rouge	450	900	2350

TABLEAU III - Valeurs cibles des 3 premiers formants pour les consonnes (en Hz)

PHONEMES	F1	F2	F3
p,b,f,v,m	200	900	2100
t,d,s,z,n	200	1400	2700
ʃ, ʒ	200	1650	2400
k,g, ɲ	250	1600	1900

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

PARTICULARITES ARTICULATOIRES ET ACOUSTIQUES DE L'OCCLUSIVE [g]

ZERLING Jean-Pierre

INSTITUT DE PHONETIQUE DE LILLE

.RESUME

Cette étude a pour but de présenter trois phénomènes de coarticulation qui affectent la réalisation de l'occlusive [g].

Nous rappelons tout d'abord quelques particularités articulatoires essentielles des voyelles, mises en évidence dans un travail antérieur (ZERLING, 1979.b), notamment celles qui concernent la forme adoptée par l'apex.

Deux phénomènes de déplacement du lieu occlusif sont présentés: l'antériorisation liée au contexte vocalique et le glissement systématique indépendant du contexte.

Les variations de la forme apicale pendant l'occlusion et ses transitions sont décrites: elles dépendent à la fois de la voyelle suivante et du sens de la transition.

L'analyse spectrale confirme l'importance de la position apicale pendant l'occlusion. Cette position est responsable de la convergence des seconds formants vocaliques vers une zone commune (locus) en contexte I-A (voyelles dites antérieures)et de leur divergence à partir d'une zone virtuelle ("anti-locus") avec les voyelles U-A (dites postérieures). Elle permet enfin de rendre compte de la dissymétrie de certaines transitions du troisième formant dans les groupes [gVg].

PARTICULARITES ARTICULATOIRES ET ACOUSTIQUES DE L'OCCLUSIVE [g].

(SOME ARTICULATORY AND ACOUSTICAL PECULIARITIES OF VELAR STOP CONSONANTS)

ZERLING Jean-Pierre — Institut de Phonétique de Lille

. ABSTRACT

X-ray lateral views of the vocal tract and corresponding acoustic data were analysed for [əgVg] utterances pronounced by two speakers, for vowels [i,ɛ,a,ɔ,u].

First, some essential vowel articulatory features are described, which concern mainly apical shape. According to this feature, vowels may be divided into two categories: retracted(U-A axis) and non-retracted (I-A axis). For a more complete description, see ZERLING (1979,b).

Then, muscular activity for stop consonant [g] is presented: main muscles involved are genioglossus, palatoglossus and styloglossus. For our speakers, the place of articulation is medio or post-palatal.

Movements and shape of the tongue depend on several parameters: the following vocalic context, closure muscular energy and transition direction ([Vg] or [gV]). The nature and value of these parameters are responsible for several variations of closure and transitions articulation:

- The place of closure obviously moves forward when following vowel is alveolar (eg. [i,y]). It slightly moves backwards when vowel is velo-pharyngeal (eg. [u,o,ɔ,a]).
- A forward systematic glide of the place of closure is observed, which may depend on genioglossus activity. This phenomenon seems to increase with muscular tenseness. It is not vowel dependant.
- Finally, apical shape variability during closure and transitions is described. It seems to depend both on the following vowel and on transition direction: Tongue tip goes flat and rather high during plosive transitions ([gV]) if following vowel belongs to I-A axis (non-retracted). In all other cases, it takes a low or retracted position,either as a passive movement or under motor control.

Spectral analysis of the sound wave confirms the important role of apical shape during velar closure. It is responsible for second formant converging towards a locus with I-A type vowels, and diverging from a virtual "anti-locus" with U-A vowels. It also accounts for third formant dissymetrical transitions in [gVg] utterances: Onset frequency T3 is superior to vowel formant frequency F3 only in [gV] transitions when following vowel is of I-A type. If not, or if transition is [Vg], T3 is inferior to F3.

PARTICULARITES ARTICULATOIRES ET ACOUSTIQUES DE L'OCCLUSIVE [g].

ZERLING Jean-Pierre Institut de Phonétique de Lille

I— INTRODUCTION

Cette étude rassemble des observations articulatoires et acoustiques émises à propos de l'occlusive [g] prononcée dans des contextes vocaliques variés et dans des positions différentes. Les remarques concernent à la fois l'occlusive et ses transitions.

Nous présentons tout d'abord un rappel sur l'activité musculaire vocalique et en particulier sur le rôle important joué par la musculature intrinsèque. Trois phénomènes de coarticulation sont ensuite décrits. Deux d'entre eux ne semblent pas avoir été mentionnés auparavant. Des remarques acoustiques sont enfin proposées qui confirment les observations articulatoires.

II - PROCEDURE EXPERIMENTALE

L'étude s'appuie sur une série de profils articulatoires du conduit vocal vu latéralement des lèvres jusqu'au larynx. Les clichés ont été obtenus à partir de films radiologiques accompagnés d'une bande-son synchrone. La cadence de prise de vues est de 60 images par seconde. Les logatomes de type [əCVC] ont été prononcés pour les voyelles [i,ɛ,a,ɔ,u] et les occlusives [b,d,g] par deux sujets masculins (FL et JPZ). L'étude spectrographique de la bande sonore a permis d'associer à chaque cliché ses caractéristiques acoustiques propres.

III - RAPPEL SUR L'ACTIVITE MUSCULAIRE VOCALIQUE

Nous avons présenté récemment plusieurs études comparées de l'activité musculaire et des positions linguales pour les voyelles(ZERLING,1979a,b). Le système musculaire envisagé correspond à ceux décrits par HARDCASTLE (1976) et ZEMLIN(1978). Il comporte deux catégories de muscles: extrinsèques et intrinsèques (fig. 1).

L'activité des muscles extrinsèques de la langue est responsable de deux mouvements distincts que nous baptisons: axes naturels I-A et U-A(fig.2). Le premier est parcouru en modifiant l'équilibre entre deux tensions musculaires antagonistes: celle du génioglossus qui tire la langue vers l'avant, et celle de l'hyoglossus qui la tire vers le bas. L'action conjuguée des deux muscles a pour effet de maintenir constamment la langue éloignée de la paroi vélaire. L'axe U-A est parcouru en modifiant la résultante de deux tensions orthogonales: celle du styloglossus qui tire la langue vers le voile et celle de l'hyoglossus qui la tire vers le bas. La combinaison des deux tensions maintient en permanence la langue vers l'arrière, à proximité de la paroi postérieure. Le mouvement lingual le long de l'axe I-A permet de réaliser les voyelles alvéolaires [i,e,y,ø] , neutres [ɛ,œ] et pharyngales [æ,a] . Le lieu articulatoire des voyelles de l'axe U-A se déplace du voile du palais [u] au bas pharynx [ɑ,ɒ] , en passant par la région vélo-pharyngale [o,ɔ,ʌ] (fig. 2).

En ce qui concerne le rôle des muscles intrinsèques, WILLIAM(1965) et PERKELL(1969) croient à l'existence d'un "tonus "vocalique unique responsable d'une forme linguale identique pour toutes les voyelles. Les observations que nous avons réalisées d'après nos données et celles d'autres auteurs nous mènent à penser qu'il existe deux formes vocaliques de la langue qui se distinguent en particulier par une forme différente de l'apex: celui-ci est tantôt rétracté sous la tension du muscle longitudinal inférieur, tantôt plate ou même creusée sous la tension du longitudinal supérieur. Ces deux

degrés de la tension apicale permettent de répartir les voyelles en deux catégories que nous nommons: rétractées et non-rétractées.

Il apparait que les voyelles de l'axe I-A sont non-rétractées et celles de l'axe U-A rétractées. Selon nous, ce couplage entre les activités musculaires extrinsèque et intrinsèque serait un phénomène moteur puisqu'on peut trouver des articulations où il n'apparait pas.

La nature musculaire très différente des deux axes rend instable toute position vocalique intermédiaire. En effet, les positions situées entre les axes ne semblent correspondre qu'à des articulations transitoires(diphtongues, glides, etc.), ou à des articulations temporaires(évolution diachronique d'une voyelle vers une autre). L'instabilité musculaire des articulations vocaliques situées entre ces deux axes explique sans doute l'existence de la "zone formantique interdite" mise en évidence par F. LONCHAMP(1978, p.217) et contribue à son explication articulo-perceptive de la mutation [u > y] qui s'est produite dans de nombreux dialectes (1978, pp.220-260).

Après avoir rappelé ces quelques caractéristiques articulatoires des voyelles, nous allons présenter trois phénomènes de coarticulation consonantique. Les deux premiers sont liés à l'activité musculaire extrinsèque, le troisième à l'activité musculaire intrinsèque.

IV - ANTERIORISATION DU LIEU OCCLUSIF DE [g] DANS CERTAINS CONTEXTES VOCALIQUES.

Le lieu occlusif de [g] tel qu'on peut l'observer pour le français sur les radiographies de CHLUMSKY(1938) ou sur les vues dynamiques de ROCHETTE(1973) est situé dans les régions "palato-vélaire ou vélo-palatale". D'après les observations cinéradiographiques de P. SIMON(1967, p.258), "le lieu d'articulation est franchement palatal", c'est-à-dire plus en avant que pour les auteurs précédents.

Chez nos sujets, l'occlusion est réalisée le plus souvent dans la région post-palatale(fig.3). Cependant, on observe un avancement du lieu occlusif lorsque la consonne précède une voyelle alvéolaire([i,y], etc.). C'est sans doute l'action anticipée du génioglossus avant ce type de voyelles qui entraine un déplacement de la masse linguale vers l'avant et donc une antériorisation du contact occlusif. Notons que cette palatalisation (et même alvéolarisation) ne se produit pas avec [a] , contrairement à ce que laisse présager la terminologie habituelle "voyelle antérieure". Face à ce phénomène, [a] ne se comporte pas comme une voyelle d'avant mais comme une voyelle pharyngale. C'est une des raisons qui nous a amené à définir le lieu articulatoire des voyelles comme emplacement du rétrécissement maximal du conduit vocal(ZERLING, 1979,a).

V - GLISSEMENT SYSTEMATIQUE DU CONTACT OCCLUSIF DE [g].

Ce phénomène est distinct de celui décrit précédemment: il ne semble pas être contextuel. Il se produit toujours vers l'avant, quelle que soit la nature de la voyelle subséquente.

Nous en avons cherché confirmation chez divers auteurs pour le français et pour l'anglais. Nous l'avons trouvé décrit plusieurs fois, mais aucune explication satisfaisante ne semble être proposée. Notons au passage que le "glissement" mentionné par OUELLON(1979) se rapporte à l'effet d'antériorisation décrit au paragraphe précédent.

P. SIMON(1967,p.212) observe un déplacement du lieu occlusif vers l'avant "devant une articulation aussi bien antérieure que postérieure". L'auteur attribue ce mouvement de la langue à un "roulement" de la langue contre la paroi palatale.

J. PERKELL(1969,p.25-27) décrit les mouvements d'un repère métallique fixé sur le dos de la langue pour les logatomes /hə'kV/. Le déplacement du repère indique que le mouvement réalisé pendant l'occlusion est un glissement plutôt qu'un roulement de la langue(fig. 4a). L'auteur attribue ce phénomène à l'articulation des segments phoniques qui précèdent et qui suivent

l'occlusive. Cette explication nous parait incorrecte car en fait le glissement éloigne la langue de sa position vocalique subséquente. Il semble donc que ce mouvement ne soit pas contextuel.

T. GAY(1977) décrit également ce phénomène. Lorsque la voyelle suivante est un [i] ou un [a], la masse linguale continue à s'élever tout en glissant vers l'avant pendant toute la durée de l'occlusion. En revanche, si la voyelle subséquente est un [u] , le glissement vers l'avant est neutralisé(fig.4.b). Selon l'auteur, les données de HOUDE(1967) et de KENT(1970), qui n'ont pas été publiées, permettent d'observer le même type de mouvement. GAY décrit donc le phénomène, mais n'en donne aucune explication.

Devant l'inexistence ou l'inadéquation des solutions proposées, nous avons cherché une explication à ce phénomène. Selon nous, le mouvement est purement mécanique: lors de l'occlusion, l'élévation linguale vers la région palato-vélaire est essentiellement due à la contraction du génioglossus. Cette activité musculaire est non seulement responsable de la montée de la langue, mais aussi de son avancement (cf. fig. 1). Il est donc permis de penser que lorsque la tension articulatoire augmente pour renforcer l'occlusion, la langue est du même coup tirée vers l'avant et qu'elle peut glisser contre la voûte palatine.

Ce mouvement vers l'avant peut se produire quel que soit le contexte vocalique. Néanmoins, pendant l'occlusion, la langue est soumise à d'autres forces, notamment à celles qui anticipent la voyelle subséquente.L'amplitude du glissement vers l'avant dépend donc de la résultante de ces forces.

Le caractère systématique de ce mouvement permet de comprendre certains phénomènes curieux comme celui observé par P. SIMON(1967,p.212): "parfois le glissement vers l'arrière est aussitôt suivi d'un glissement vers l'avant bien que la voyelle subséquente ait un lieu d'articulation plus reculé".

Indépendemment des déplacements contextuels ou systématiques du lieu occlusif décrits jusqu'à présent, nous avons relevé un autre phénomène qui ne semble pas avoir été remarqué auparavant et dont les conséquences acoustiques ne sont pas négligeables.

VI - VARIABILITE DE LA FORME APICALE PENDANT L'OCCLUSION ET LES TRANSITIONS DE [g].

Ce phénomène apparait lié à la fois au sens de la transition([Vg] ou [gV]) et au contexte vocalique.

Pendant les transitions implosives [Vg] , le mouvement lingual est identique quelle que soit la voyelle: dès la fin de celle-ci, le dos de la langue s'élève vers la région palato-vélaire. La partie antérieure de la langue accompagne le mouvement avec un certain retard. En quelque sorte, l'apex semble se laisser entrainer passivement par le dos de la langue. Il en résulte qu'à l'instant de l'occlusion l'aspect de la cavité buccale est quasiment constant, indépendamment de la voyelle qui précède(fig. 5a). La rétraction apicale confère au résonateur alvéo-palatal un volume important.

Au contraire, dans les transitions explosives [gV] , la forme de la langue n'est pas unique. On sait quel'articulation est souvent anticipante en français. D'autre part, nous avons montré que les voyelles se distinguent par des formes très variées du conduit vocal, en particulier au niveau de l'apex. Il n'est donc pas étonnant d'observer que les profils articulatoires de l'occlusive et de sa transition diffèrent selon la voyelle suivante. Lorsque celle-ci est non-rétractée, l'apex s'élève pendant l'occlusion et adopte une forme plate relativement parallèle à la paroi alvéo-palatale. Au contraire,avant une voyelle rétractée, le volume de la cavité alvéolaire est plus grand(fig. 5b, 6). Contrairement au mouvement passif de l'apex observé pendant les transitions [Vg], l'élévation et la rétraction apicale sont ici des mouvements moteurs.

Nous avons cherché la confirmation de l'existence de ce phénomène chez d'autres auteurs. Or, pour la plupart, ils ne décrivent les mouvements de la langue qu'au point d'occlusion. Nous n'avons pu trouver une description

du mouvement apical pour [k] et [g] que chez SIMON(1967) et PERKELL(1969). Nous regrettons que ces auteurs ne présentent les transitions qu'en contexte vocalique non-rétracté. Chez SIMON, on peut observer nettement l'élévation de la langue antérieure pendant la tenue de [k] et [g] suivies de [a] (1967, fig. 112, 116-118 124, 125). Chez PERKELL(1969,fig.3.14), les variations de l'espace alvéolaire font apparaitre une forte montée de l'apex pendant l'occlusion dans le logatome /hə'kɛ/ (fig. 7 et 8).

Pour nos sujets, les variations comparées de l'espace alvéolaire en fonction du temps pour plusieurs voyelles montre d'une part qu'il y a bien catégorisation de la position apicale selon le contexte, d'autre part que le mouvement apical est amorcé pendant la tenue occlusive (fig. 9).
Sur le tracé, il est clair que la voyelle [a] s'apparente aux voyelles de l'axe I-A quoique son articulation linguale soit en fait très voisine de celle de [ɔ].

De ces différentes observations, nous retiendrons que l'articulation de l'occlusive [g] se distingue tout particulièrement lorsqu'elle est suivie d'une voyelle non-rétractée (axe I-A): la langue antérieure adopte alors une position plate et élevée. Dans tous les autres cas, l'apex est relativement rétracté, que le mouvement soit moteur (anticipation d'une voyelle de l'axe U-A) ou passif (transition [Vg]).

Après avoir décrit ces phénomènes de variabilité articulatoire des transitions de [g], il est intéressant d'en chercher la trace au plan acoustique.

VII - PARTICULARITES SPECTRALES DES TRANSITIONS DE [g].

L'étude spectrographique des transitions de [g] a permis de rassembler des observations dont certaines nous semblent importantes. La figure 12 fournit une représentation schématique simplifiée de l'allure des transitions des 4 premiers formants pour nos deux sujets.

Pour les premier et quatrième formants, aucune remarque nouvelle n'a pu être dégagée: pour F1, la valeur formantique de la transition (on-glide ou off-glide) est toujours plus faible que celle de la voyelle; au contraire. pour F4, elle est supérieure quel que soit le sens de la transition et la voyelle associée. En revanche, pour les deux autres formants quelques observations importantes ont pu être réalisées.

VII-1 - Un "anti-locus" pour F2.

Il serait trop long de résumer ici l'évolution des théories sur les "hubs" et les "loci". Rappelons seulement que POTTER et al.(1947,p.81-103) furent parmi les premiers à remarquer la convergence des formants vocaliques vers des zones qui dépendent de l'articulation de l'occlusive. DELATTRE et al.(1955) ont ensuite développé une théorie plus élaborée, assez largement adoptée. L'étude d'EMERIT(1974-76) est la plus récente et la plus complète en la matière. L'auteur montre en particulier que DELATTRE a tord d'attribuer deux loci distincts à l'occlusive [g], l'un pour les voyelles antérieures, l'autre pour les postérieures. En revanche, il propose un locus avec les premières, et un "anti-locus" avec les autres, mais il ne fournit pas d'explication rationnelle de la discontinuité des transitions de [g].

Nous avons expliqé plus haut en quoi les transitions articulatoires de cette occlusive diffèrent selon la nature de la voyelle suivante:

- En présence des voyelles de l'axe I-A, la position linguale est quasiment constante au moment de l'explosion(fig.5a). Il n'est donc pas étonnant d'observer une zone réelle de convergence des formants vocaliques (locus), située pour nos sujets aux environs de 1900 à 2250 Hz (fig. 12a).
- En présence des voyelles de l'axe U-A, la transition du 2° formant pour [gV] n'est pas toujours descendante comme l'affirment DELATTRE et al.(1955) (fig.10). Les mêmes auteurs montrent d'ailleurs en utilisant la synthèse acoustique qu'une forme plate ou légèrement montante de F2 permet de percevoir les transitions de [gɔ] et [gu] sans ambiguité (LIBERMAN et al.,1954)(fig.11). En fait, la convergence des formants ne se fait pas vers la consonne, mais

vers la voyelle. EMERIT(1975) baptise cette zone virtuelle de convergence "anti-locus", mais ne précise pas son origine. Nous acceptons l'existence de cet anti-locus et en proposons l'explication suivante: On sait que [u] est la position vocalique la plus voisine de celle de [g]. Il est donc peu surprenant d'observer que les transitions de F2 pour [u] sont presque plates et ont une valeur proche de 900 Hz. D'autre part, FANT(1960) a montré que lorsque le conduit vocal présente une forte constriction vélo-pharyngale, la moindre variation articulatoire peut avoir des conséquences acoustiques importantes. Or, nous avons montré que l'occlusive [g] est fortement sujette à la coarticulation. On observe notamment de grandes différences de longueur du conduit selon la voyelle qui suit(ZERLING, 1979b). Au moment du relâchement d'occlusion, la valeur du second formant sera donc d'autant plus éloignée de celle de [u] que l'articulation vocalique subséquente sera éloignée de [u] sur l'axe U-A. Par exemple, les valeurs d'attaque du second formant pour [u] et [ɔ] sont séparées d'environ 1000 Hz. En revanche, plus on relâche l'occlusion, plus le conduit vocal s'élargit dans la région vélaire, et moins les conséquences acoustiques de la coarticulation se font sentir: l'écart entre les F2 pour les deux mêmes voyelles n'est plus que de 400 Hz. Lors du passage de l'occlusive à la voyelle, il se produit donc une convergence apparente des F2 vers un point situé dans la voyelle aux environs de 900 Hz. Ce point virtuel de convergence peut-être baptisé "anti-locus" (fig.12b).

VII - 2 - Des transitions dissymétriques de F3.

Pour les transitions du troisième formant, on remarque comme pour celles de F2, une catégorisation selon la nature du contexte vocalique. Avec les voyelles de l'axe U-A, la valeur du formant vocalique est toujours supérieure à celle de son attaque ou de sa disparition. Si on nomme T3 cette dernière valeur, on a donc F3 > T3. Chez nos sujets, T3 reste comprise entre 2000 et 2250 Hz. En revanche, pour les voyelles de l'axe I-A, F3 est tantôt supérieur, tantôt inférieur à T3, selon le sens de la transition: F3 > T3 pour les transitions [Vg] , F3 < T3 pour les transitions [gV](fig. 12).

Cette dissymétrie spectrale de part et d'autre de la consonne reflète la dissymétrie articulatoire évoquée plus haut: l'apex est toujours en position basse lors de l'implosion, alors que pour l'explosion sa position peut-être plate ou rétractée, selon la nature de la voyelle suivante. Ce comportement particulier de F3 est la seule trace acoustique que nous ayons relevée du phénomène. A notre connaissance, ces particularités n'ont pas été décrites auparavant. Elles correspondent pourtant à une réalité acoustique prévisible: FANT(1960) montre en effet que F3 est particulièrement sensible au volume de la cavité qui précède le point de constriction maximale du conduit. Le choix d'une position apicale tantôt plate ou tantôt rétractée est donc déterminant pour la valeur de F3.

VIII - CONCLUSIONS.

Il ressort de cette étude que le comportement articulatoire de [g] est relativement complexe. L'occlusion nécessite tout d'abord une montée verticale de la langue réalisée essentiellement par la tension du génioglossus du styloglossus et du palatoglossus. Le contact est médio ou post-palatal. Les mouvements et la forme globale de la langue dépendent de plusieurs paramètres: le contexte vocalique, l'énergie musculaire dépensée pour réaliser l'occlusion, et le sens de la transition ([Vg] ou [gV]). Selon la nature et la valeur de ces paramètres, l'occlusion et ses transitions articulatoires peuvent subir des modifications importantes:

- un déplacement de l'occlusion vers l'avant si la voyelle subséquente est alvéolaire, ou un léger recul si elle est vélo-pharyngale;
- un glissement systématique du contact occlusif vers l'avant, d'autant plus net que la force occlusive est grande. Ce phénomène n'est pas contextuel.
- une modification radicale de la forme de l'apex: élévation pour les transitions explosives [gV] si la voyelle subséquente appartient à l'axe I-A,

ou rétraction (active ou passive) dans tous les autres cas.

Ce dernier phénomène, la catégorisation de la forme apicale en fonction du contexte vocalique subséquent, a des conséquences non négligeables au plan acoustique: le second formant converge vers un locus situé aux environs de 2000 Hz avec les voyelles de l'axe I-A, et vers un "anti-locus" situé vers 950 Hz avec celles de l'axe U-A. Pour F3, la fréquence d'attaque T3 est supérieure à celle du formant vocalique si la voyelle suivante est du type I-A; dans tous les autres cas, T3 < F3.

REFERENCES BIBLIOGRAPHIQUES

CHLUMSKY J. (1938) Radiographie des voyelles et des semi-voyelles françaises, Prague.

DELATTRE P., LIBERMAN A.M. et COOPER F.C. (1955) "Acoustical loci and transitional cues for consonants", JASA, 27. 4, 769-773, in Readings in Acoustic Phonetics, LEHISTE ed., 1967.

EMERIT E. (1974,75,76) "Nouvelle contribution à la théorie des locus", Phonetica, 30, p.1-30; 31, p.6-37; 33, p.425-466.

FANT G. (1960) Acoustic theory of speech production, Mouton, The Hague, 2° ed., 1970.

GAY T. (1977) "Articulatory movements in VCV sequences", JASA, 62,1,p.183-193.

HARDCASTLE W.J. (1976) Physiology of speech production

HOUDE R.A. (1967) A study of tongue body movement during selected speech sounds, Ph. D. Thesis, University of Michigan.

KENT R.D. (1970) A cinefluorographic-spectrographic investigation of the component gestures in lingual articulation, Ph.D. Thesis, University of Iowa.

LIBERMAN A.M.,DELATTRE P.,COOPER F.S. et GERSTMAN L.J.(1954) "The role of consonant-vowel transitions in the perception of stop and nasal consonants" Psychological Monographs,68,8, in FRY, Acoustic Phonetics, 1976.

LONCHAMP F. (1978) Recherches sur les indices perceptifs des voyelles orales et nasales, Thèse de 3° Cycle, Université de Nancy II.

OUELLON C. (1979) "Phénomènes liés à l'enchainement des consonnes françaises dans la chaîne parlée", 9° ICPS, Copenhague, p. 206.

PERKELL J.S. (1969) Physiology of speech production, M.I.T. Press, Cambridge, Mass., USA.

POTTER R.,KOPP G.A. et GREEN H.L.(1947) Visible Speech,New York Van NorstrendCy

ROCHETTE C.(1973) Les groupes de consonnes en français, Klincksieck, Paris.

SIMON P. (1967) Les consonnes du français. Mouvements et positions articulatoires à la lumière de la cinéradiographie, Klincksieck, Paris.

WILLIAMS J.T. (1965) A static model of the tongue during vowel articulation, M.S. Thesis, M.I.T.

ZEMLIN W. (1978) Speech and hearing science: anatomy and physiology, Englewood Cliffs, Prentice Hall Inc., New Jersey.

ZERLING J.P. (1979,a) "Description de cinq voyelles orales du français en contexte, et nouvelle classification articulatoire", Travaux de l'Institut de Phonétique de Nancy, 2, p.55-87.

ZERLING J.P. (1979,b) Articulation et coarticulation dans des groupes occlusive-voyelle en français, Thèse de 3° Cycle, Université de Nancy II.

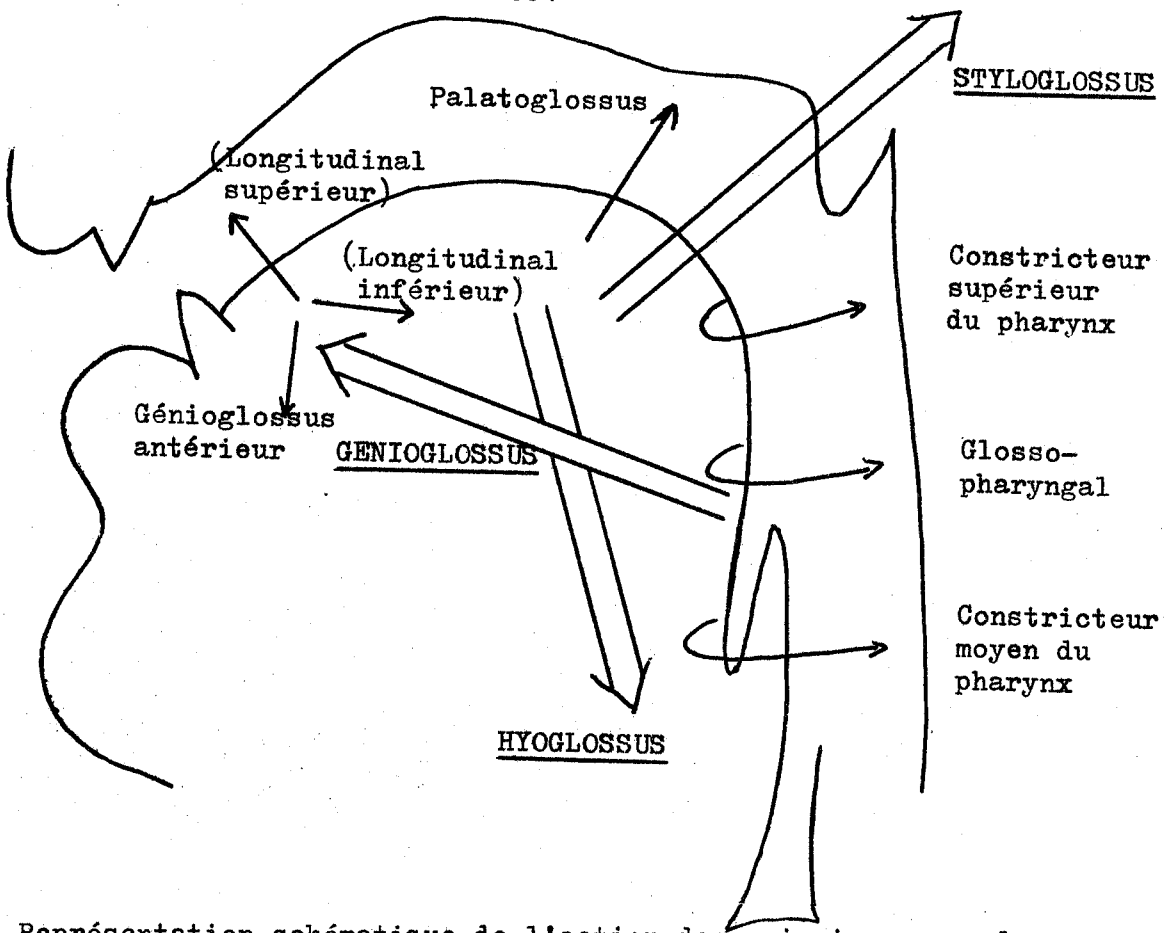


fig. 1 Représentation schématique de l'action des principaux muscles qui agissent sur la langue. Les muscles intrinsèques sont indiqués entre parenthèses.

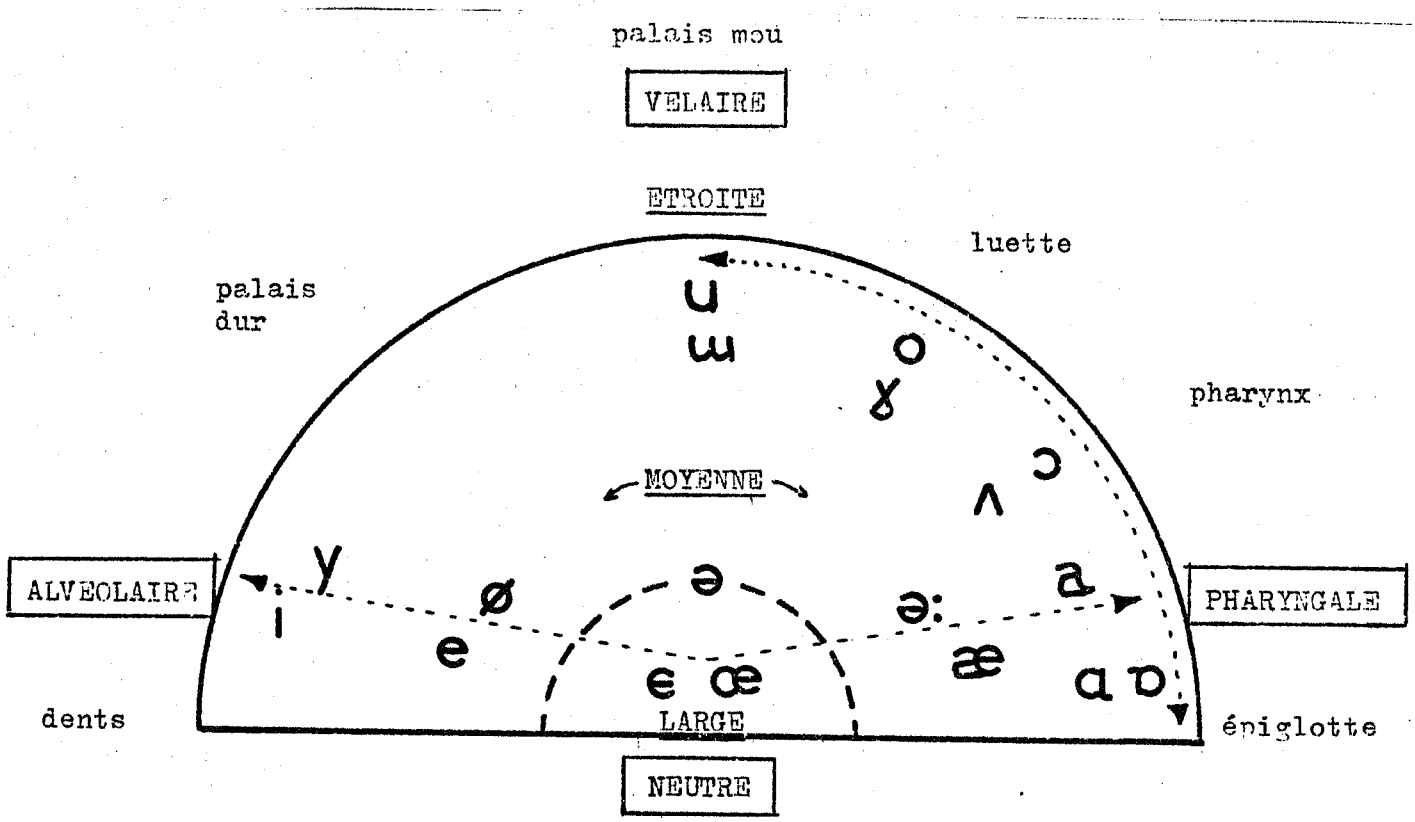


fig. 2 Classement articulatoire des voyelles du français et des voyelles cardinales en fonction du lieu et du degré de rétrécissement du conduit vocal. Les axes U-A et I-A(Y-A)sont indiqués en pointillés.

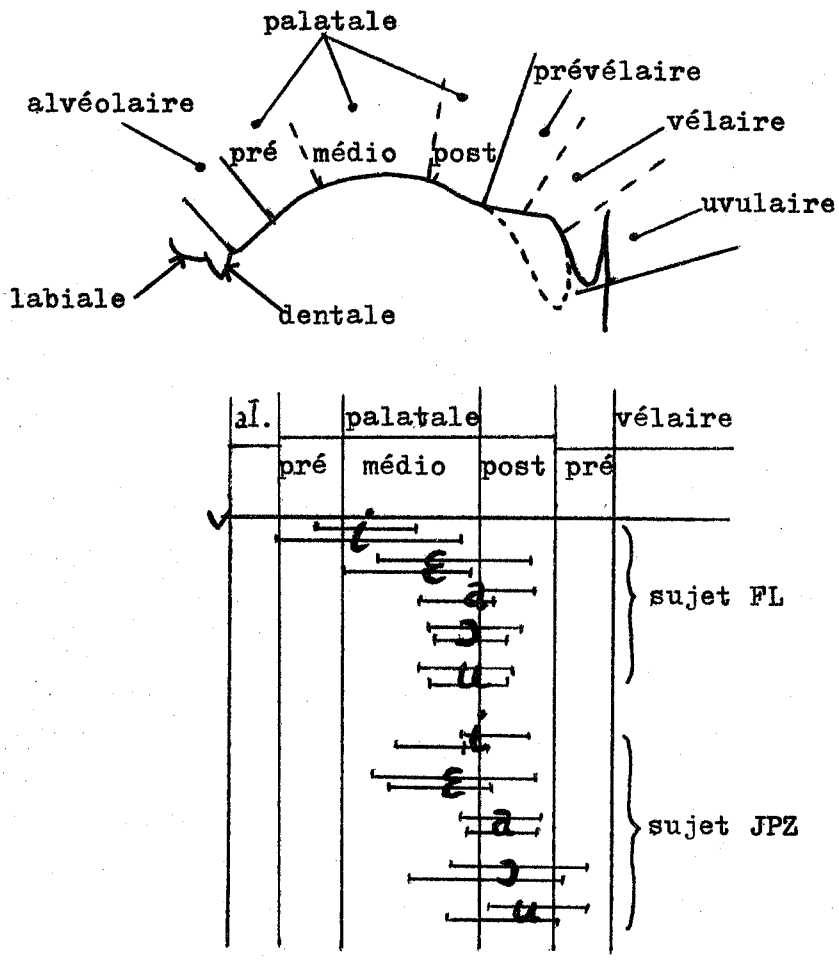


fig. 3 Emplacement des zones de contact occlusif de [g] en fonction du contexte vocalique chez les deux sujets.

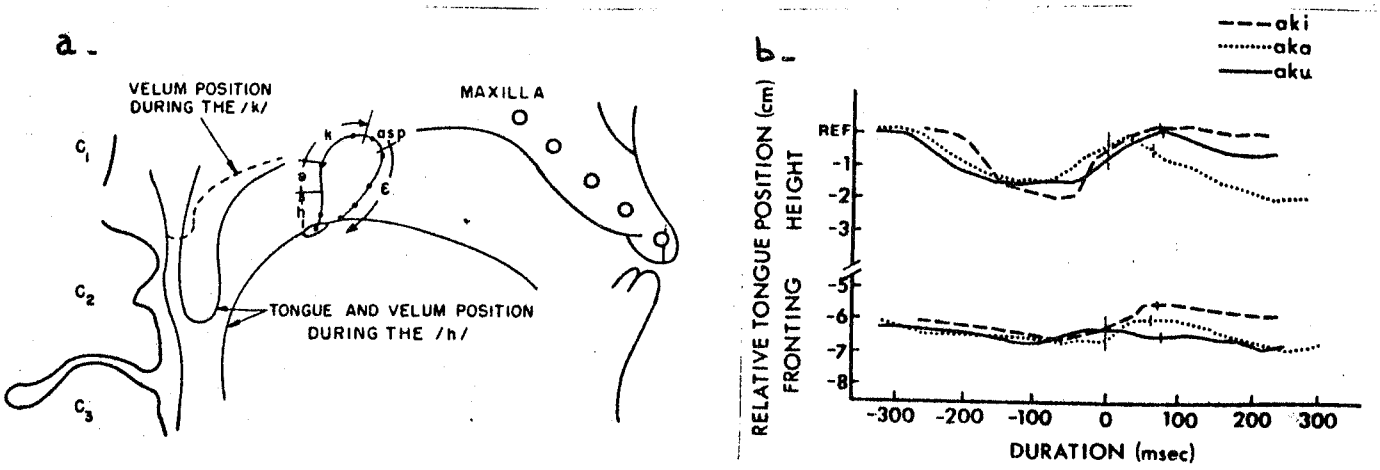


fig. 4 Mouvements d'un repère métallique placé sur le dos de la langue pendant l'occlusion de [k]:
a - d'après PERKELL (1969, fig.3.15), logatome /hə'kɛ/
b - d'après GAY (1977, fig. 7), logatomes /aki, aka, aku/
Les courbes supérieures représentent les déplacements verticaux, les courbes inférieures, les déplacements horizontaux.

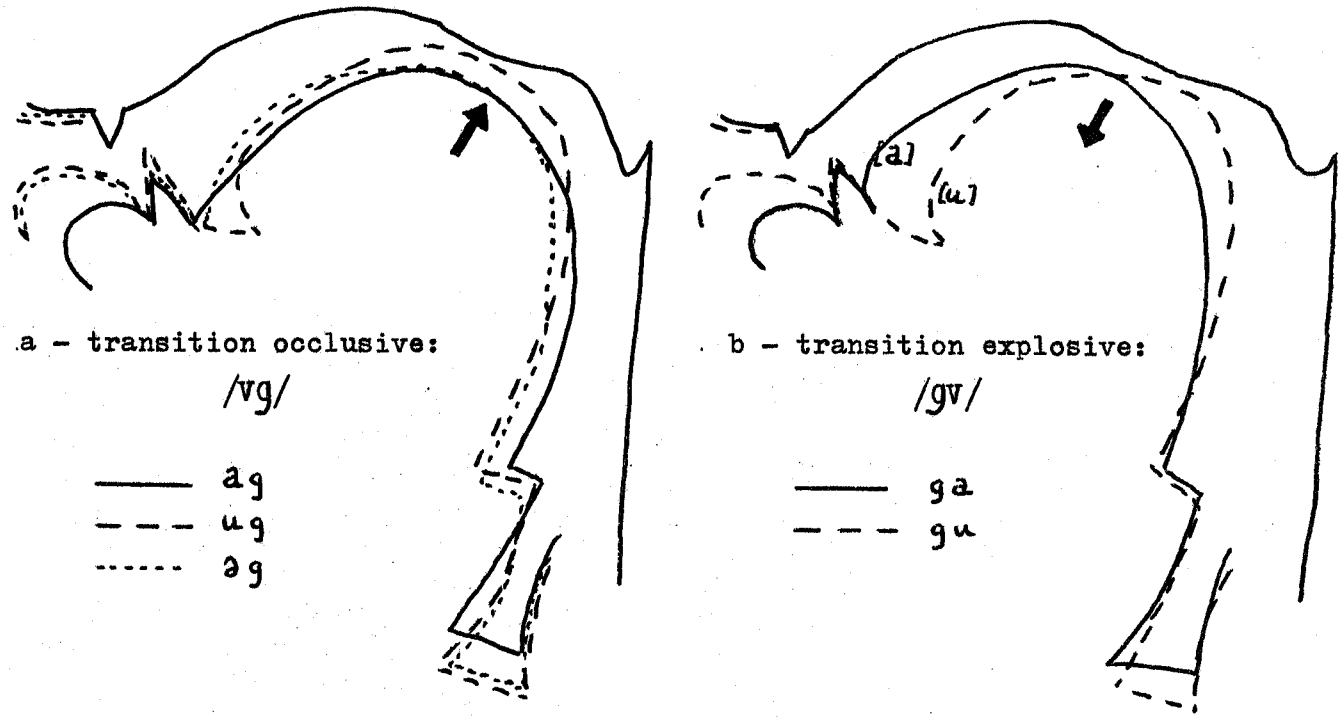


fig. 5 Comparaison des profils transitoires du conduit vocal du sujet JPZ pour la consonne [g],
a - directement avant l'occlusion
b - directement après l'occlusion.

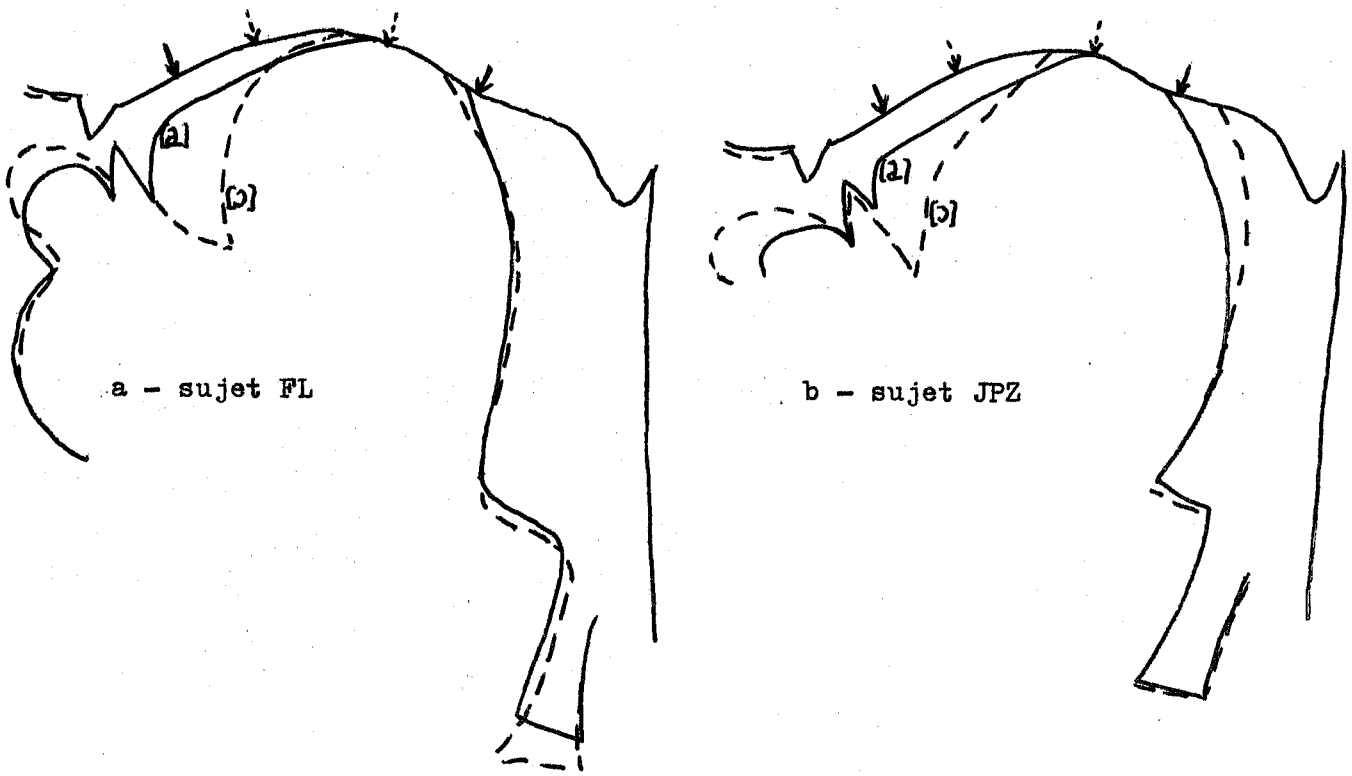


fig. 6 Comparaison des profils occlusifs de [g] dans les contextes vocaliques [a] et [ɔ].

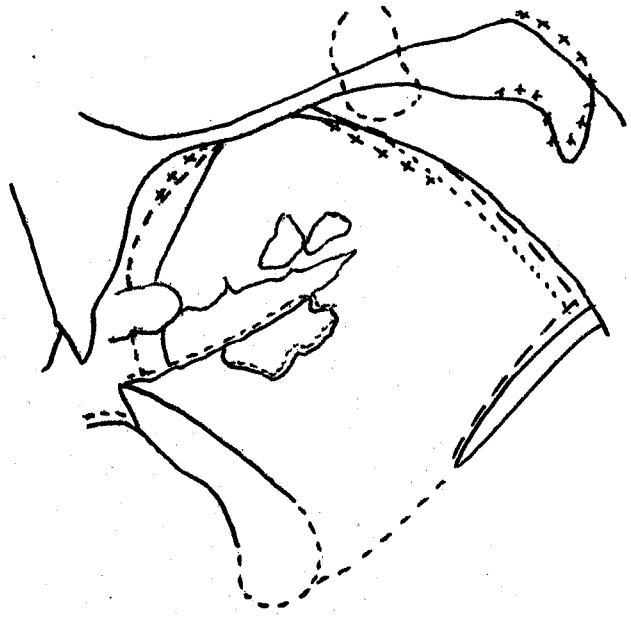


fig. 7

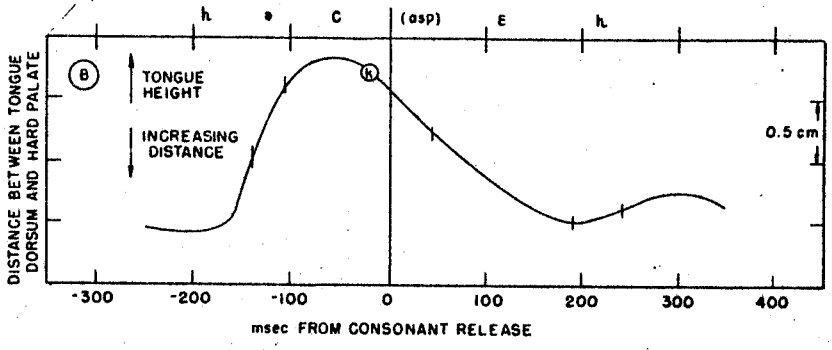
Elévation de la langue antérieure pendant l'occlusion de [k] suivie de [a], dans le mot "placards". Quatre images successives sont superposées:
1——— 2– – – – 3····· 4·-·-·-

D'après SIMON (1967, fig. 112)

fig. 8

Variations de l'espace dorso-palatal pendant la réalisation de /haˈkɛ/

D'après PERKELL (1969, fig. 3.14)



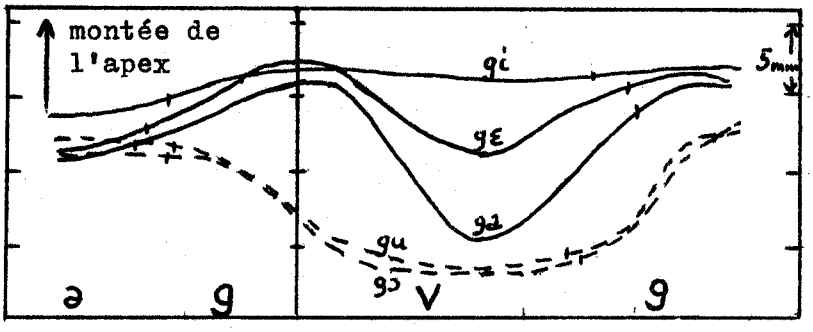


fig. 9 Comparaison des mouvements de l'apex pendant l'occlusive [g] selon le contexte vocalique, pour le sujet JPZ.

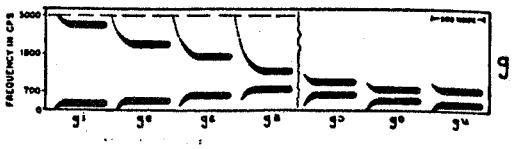


fig. 10 Représentation schématique des transitions des deux premiers formants de plusieurs voyelles après l'occlusive [g]. La ligne pointillée indique la fréquence d'un locus. D'après LIBERMAN et al.(1959, fig. 2 et 7).

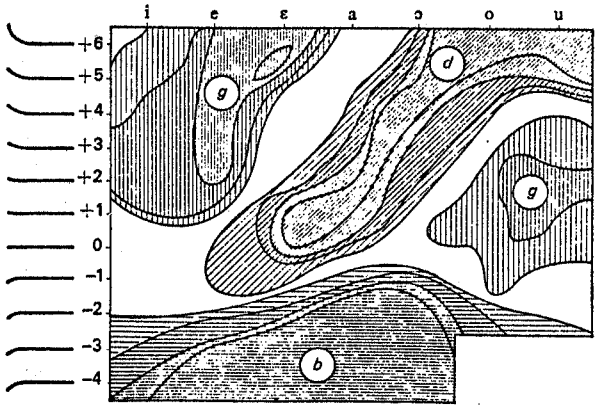
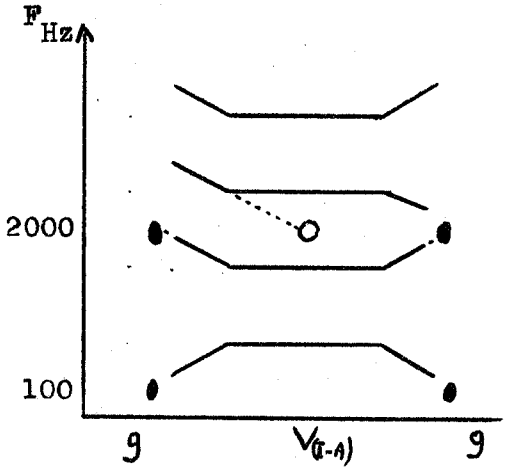
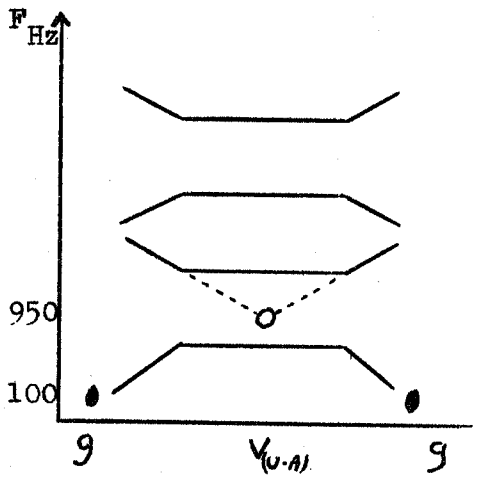


fig. 11 Perception des occlusives en fonction de l'allure croissante ou décroissante de la transition du second formant: carte des régions pour lesquelles la reconnaissance de [b,d,g] est dominante. D'après LIBERMAN et al.(1954, fig. 4).



a - Occlusive [g] et voyelles de l'axe I-A (antérieures)



b - Occlusive [g] et voyelles de l'axe U-A (postérieures)

fig. 12 Représentation schématique des transitions occlusives des quatre premiers formants vocaliques en présence de [g] chez nos deux sujets.

Les zones de convergence réelles sont indiquées par un point noir (●) (locus); les zones de convergence virtuelles (anti-locus), par un cercle (O). Les fréquences approximatives de ces zones sont indiquées.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ETUDE EMG DE QUATRE MUSCLES LABIAUX DANS DES SYLLABES FINALES DE MOTS EN FRANCAIS : L'EFFET DE LA COARTICULATION

Michèle GENTIL

Institut de Phonétique de Grenoble
Actuellement : Centre Scientifique IBM
36, Av. Raymond Poincaré, 75116 Paris

.RESUME

Cette étude, utilisant les techniques électromyographique et cinématographique, analyse en la reliant au mouvement l'activité coordonnée de 4 muscles labiaux : OO INF, MENT, DLI (concernant la lèvre inférieure) et LLSA (pour la lèvre supérieure). Les enregistrements EMG sont réalisés avec des électrodes de surface, type "paint-on", pour un locuteur féminin et un masculin.

Cette recherche s'inscrit dans le cadre d'une étude générale (GENTIL, 1980) sur les sibilantes qui nous paraissent des plus représentatives de la labialité en français et fait suite à une première observation EMG (GENTIL & al., 1980) relative à des monosyllabes du français composées principalement de sibilante + voyelle, qui avaient été produites par le locuteur féminin.

L'étude présente traite de syllabes finales de mot lequel est intégré dans une phrase porteuse. Celles-ci sont composées d'une sibilante [s, z, ʃ, ʒ] combinées avec les voyelles [i, y, e, ø]. Les effets coarticulatoires ont particulièrement été mis en évidence du fait de l'interaction consonne - voyelles existant dans des syllabes où des sibilantes [± lab] sont suivies et précédées d'une voyelle [± lab]. A partir des phénomènes observés, quelques règles de coarticulation sont proposées.

EMG STUDY OF FOUR LABIAL MUSCLES IN LAST SYLLABLES OF SOME FRENCH WORDS : EMPHASIS ON COARTICULATORY EFFECTS

Michèle GENTIL — Institut de Phonétique de Grenoble
Presently : Centre Scientifique IBM France
36, Av. Raymond Poincaré, 75116 Paris

. ABSTRACT

This study analyses the activity from four labial muscles : orbicularis oris inferior, mentalis, depressor labii inferioris, levator labii superioris alaeque nasi, connected with lip movement. The techniques of electromyography ("paint-on" electrodes) and cinematography (labiographic film 16 mm, front and profile, 66 frames/sec) were used.

This research belongs to a general study (GENTIL, 1980) related to sibilants which are consonantal productions most representative of labiality in french and which have not yet been analysed. It follows a preliminary EMG study (GENTIL & al., 1980) related to french monosyllables which are composed especially of sibilant + vowel and produced by a frenchwoman.

The present study concerning two speakers, deals with the last syllables of french words, composed of sibilant [s, z, ʃ, ʒ] + vowel [i, e, y, ø]. These words are embedded in sentence frames. It is an attempt to define effects of coarticulation on EMG activity recorded from lips and to correlate this activity with lip movement.

Not very protruded syllables [ʃi, ʒi, ʃe, ʒe] are characterized by :
- a slight anticipatory labial coarticulation with protrusion and mentalis-activity which precede the sibilant
- a carryover coarticulation effect with activity of mentalis and orbicularis oris inferior, only when the vowel before the consonant is [- lab].

Protruded syllables [sy, zy, sø, zø]
The following vowel [+ lab] has a great influence on the sibilant [s, z].
In a [- lab] preconsonantal context, this influence begins during the production of the sibilant. We notice from the sibilant a protrusion and the activity of mentalis and orbicularis oris inferior.
In a [+ lab] preconsonantal context, we observe that the protrusion and mentalis-activity decrease slightly during the production of sibilant.

Very protruded syllables [ʃo, ʒo, ʃy, ʒy]
Two labial productions explain an anticipatory labial coarticulation more important than for not very protruded syllables.

ETUDE EMG DE QUATRE MUSCLES LABIAUX DANS DES SYLLABES FINALES DE MOTS EN FRANCAIS : L'EFFET DE LA COARTICULATION

Michèle GENTIL — Institut de Phonétique de Grenoble
Actuellement : Centre Scientifique IBM-France
36, Av. Raymond Poincaré, 75116 Paris

INTRODUCTION

L'EMG, technique consistant à enregistrer les courants d'action musculaire, est exploitée à partir des années cinquante dans l'étude de la parole. En ce qui concerne les lèvres, LYSAUGHT & al. (1961) fournissent les premières données qui vont être complétées par beaucoup d'autres études portant sur l'anglais-américain (BELL BERTI & al., 1978, DANILOFF & al., 1979, GAY, 1979) et le suédois (MacALLISTER, 1978, LUBKER & al., 1980) pour n'en citer que quelques unes, parmi les plus récentes, mais pour le français, l'activité musculaire des lèvres a été très peu étudiée et ces rares travaux (RIORDAN, 1976, BONNOT & al., 1980) ne considèrent qu'un seul muscle, l'orbiculaire. Or quand bien même ce muscle est essentiel, la grande mobilité des lèvres est le fait du jeu de plus d'une dizaine de muscles.

D'autre part, d'une façon générale, les sibilantes n'ont jamais fait l'objet d'études EMG. Nous nous devions, particulièrement en français où la participation des lèvres est très évidente pour ce type de production, de combler cette lacune. C'est la raison de la présente étude traitant de l'activité de quatre muscles : l'orbiculaire inférieur (OO INF), le mentalis (MENT), le carré du menton ou depressor labii inferioris (DLI) et le releveur superficiel ou levator labii superioris alaeque nasi (LLSA), pour deux locuteurs dans des syllabes finales de mot, lequel est intégré dans une phrase porteuse*. Cette recherche vient compléter les résultats d'une observation EMG préliminaire (GENTIL & al., 1980) concernant des monosyllabes du français réalisées par un locuteur.

Ici l'accent est mis sur les effets coarticulatoires. En effet, nous avons les quatre sibilantes [s, z, ʃ, ʒ] combinées avec les voyelles subséquentes [i, e, y, ø] et précédées d'une voyelle [a lab]. L'intéraction consonne - voyelles est examinée (pour un locuteur du moins) eu égard à l'activité musculaire associée au mouvement. Celui-ci n'est pas quantifié mais apprécié qualitativement à partir du labiofilm.

PROCEDURE EXPERIMENTALE

1 - Electrodes

Les enregistrements électromyographiques ont été effectués à l'Institut de Linguistique de Stockolm* au moyen de double

---

* Nous devons à Louis-Jean BOE de l'Institut de Phonétique de Grenoble et Raymond DESCOUT du CNET Lannion, l'initiative de cette expérimentation qui a été réalisée à l'Institut de Linguistique de Stockolm, grâce à Björn LINDBLOM et ses collaborateurs. Qu'ils soient ici tous remerciés !

* par Louis-Jean BOE (Institut de Phonetique de Grenoble) et Raymond DESCOUT (CNET Lannion)

"paint-on electrodes", électrodes de surface faites d'un mélange de poudre d'argent, acétone et adhérant à l'aide d'un peu de colle scotch. La peau est soigneusement dégraissée à l'alcool avant leur placement.
Pour leur positionnement, les indications données par KENNEDY III & al. (1979) ont été suivies. Soit :
- pour l'OO INF, à la limite inférieure du vermillon de la lèvre inférieure, à une distance de moitié entre l'aile du nez et la ligne du plan sagittal médian ;
- pour le DLI, à 15 mm de la commissure et à une distance de moitié entre le bord inférieur du menton et la limite supérieure du vermillon de la lèvre inférieure ;
- pour le MENT, à 10 mm de la ligne du plan sagittal médian et 15 mm du bord inférieur du menton ;
- pour le LLSA, à 3 mm de la partie latérale de l'aile du nez et à 10 mm de son bord inférieur.

2 - Appareillage

Il a été utilisé les amplificateurs différentiels fabriqués au laboratoire de l'Institut de Linguistique de Stockolm.

Les mouvements des lèvres (pour le locuteur féminin) ont été filmés en couleur à 66 images/sec., au moyen d'une caméra Beaulieu 16 mm. Les images étaient prises de face et de profil avec miroir à 45° et numérotées par compteur digital branché sur prise synchro de la caméra.

L'ensemble des informations suivantes a été collecté sur enregistreur FM, Racal (8 pistes, 30 ips) :
- les signaux EMG des muscles (OO INF, MENT, DLI & LLSA)
- le son
- la synchro caméra
- une référence à 100 Hz.

Ces enregistrements ont été tracés* à l'aide d'un Oscillomink Siemebs, vitesse de défilement : 10 cm/sec.

A chaque signal EMG a été ajouté le logarithme de sa valeur efficace.

3 - Sujets

Deux sujets âgés d'une trentaine d'années ont été choisis : l'un féminin (CS) et l'autre masculin (RD). Ils appartiennent au milieu chercheur (parole). Le premier est né à Nantes et le second dans la région parisienne. Tous deux résident actuellement en Bretagne et s'expriment sans accent régional marqué.

4 - Corpus

Il comporte 32 syllabes, finales de mots, composées de sibilantes [s, z, ʃ, ʒ] combinées avec les voyelles [i, y, e, ø] soit 16 types de réalisations syllabiques précédées d'une voyelle [– lab] ou [+ lab]. Afin de donner à ces productions le plus de naturel possible, les mots ont été placés dans des phrases porteuses du genre : C'est bien saisi/concis. Celles-ci ont été lues et accentuées normalement. Il convient de préciser toutefois que la vitesse de lecture du locuteur masculin était, dans l'ensemble, légèrement supérieure à celle du locuteur féminin.

---

* par Louis-Jean BOE et Raymond DESCOUT au CNET Lannion

5 - Dépouillement et mesures

5 - 1. Mouvement

Les commentaires suivants concernent uniquement le locuteur féminin pour lequel nous disposions d'un film.

Pour toute réalisation, nous avons remarqué à la visionneuse et noté chacune des images correspondant à une nouvelle séquence de mouvement. Elles ont été numérotées dans l'ordre et les numéros* sont reportés sur une ligne de déroulement temporel des gestes labiaux et du maxillaire qui est repérée par rapport au signal acoustique.

Nous avons consigné les mouvements à partir du début de la voyelle précédant la syllabe, ceci afin de donner la position des lèvres voire du maxillaire au moment où débute le signal acoustique pour la consonne.

Par ailleurs, nous avons relevé et décrit les images correspondant au :

Début de la voyelle préconsonantique, début de la consonne, milieu de celle-ci, début de la voyelle subséquente, milieu et fin de cette dernière.

Nous avons ainsi une suite de croquis commentés qui jalonnent la réalisation et permettent de mieux suivre l'évolution des gestes labiaux.

Nous avons mentionné les mouvements de la mâchoire même si l'activité d'aucun muscle la concernant n'a été enregistrée car nous estimons importante la liaison lèvres-maxillaire. En effet, il a été prouvé que les lèvres supérieure et inférieure et la mâchoire jouent réciproquement dans l'ouverture de la bouche par exemple. Quand la mâchoire se déplace beaucoup, les lèvres bougent peu et inversement (ABBS & al. 1973, HUGHES & al., 1976).

5 - 2. Activité musculaire

Celle-ci fait l'objet d'un dépouillement manuel.

Pour la consonne et la voyelle de chaque syllabe, nous avons mesuré sur le signal brut :

La durée de l'activité musculaire, l'amplitude du pic max en μV, le temps entre ce pic max et le début du signal acoustique, enfin le temps entre l'entrée en activité musculaire et le début du signal acoustique.

Afin de rendre cette série de mesures facilement interprétable, nous avons adopté pour les divers types de syllabes, une schématisation très simple. Nous avons représenté par une ligne le déroulement temporel de chacun des signaux EMG repérés par rapport au signal acoustique en reprenant toujours la

---

* D'une syllabe à l'autre, le même numéro ne correspond pas à une séquence identique ; cf pour chacune des syllabes le type de mouvement en relation avec le numéro.

même distribution : MENT, OO INF, DLI et/ou LLSA, quel que soit le moment d'entrée en jeu de ces muscles. Ces lignes figurent au-dessous de celle du déroulement temporel des gestes labiaux. A l'emplacement des pics max, nous avons noté des flèches. L'amplitude de ces mêmes pics max est reportée sur une échelle graduée en 10^2 µV. Il ne semble pas toutefois qu'il faille relier très précisément ces pics max au mouvement.

L'activité musculaire concerne la syllabe CV uniquement, bien entendu avec l'arbitraire que cela implique compte tenu des effets coarticulatoires. Nous avons essayé de procéder au mieux quant au découpage de l'activité musculaire et son affectation.

Pour le locuteur masculin, nous n'avons pas représenté les mesures de l'action musculaire. En ce qui le concerne, notre étude sera limitée et ce d'autant plus que de nombreux artéfacts ne nous ont pas permis l'exploitation du mentalis. Il nous servira d'élément de comparaison pour apprécier les résultats du locuteur féminin.

RESULTATS ET DISCUSSIONS

L'objectif est d'observer le travail coordonné de 4 muscles dont 3 pour la lèvre inférieure, importante dans la production des sibilantes et d'examiner principalement les effets coarticulatoires au niveau de l'activité musculaire laquelle est toujours mise en relation avec le mouvement labial. Pour cette analyse, nous nous intéresserons à 3 groupes* de syllabes mis en évidence d'après le mouvement :

- Syllabes faiblement protruses avec sibilantes [+ lab] + voyelles [– lab], soit [si, zi, se, ze]
- Syllabes protruses avec sibilantes [– lab] + voyelles [+ lab], soit [sy, zy, sø, zø]
- Syllabes fortement protruses avec sibilantes [+ lab] + voyelles [+ lab], soit [ʃy, ʒy, ʃø, ʒø].

Pour chacun de ces groupes de syllabes où les effets de coarticulation labiale sont évidents nous avons retenu pour examen le type de syllabe le plus représentatif de l'ensemble* avec contexte préconsonantique [+ lab] et [– lab].

A la suite de nos observations, quelques règles de coarticulation sont proposées.

1 - SYLLABES FAIBLEMENT PROTRUSES

1 - 1. Syllabe [ʃe], contexte préconsonantique [+ lab] (fig.1)
Elle entre dans la phrase : "C'est bien fauché".
Trois muscles jouent un rôle : MENT, OO INF & LLSA.
Cette syllabe se caractérise par le passage d'une protrusion importante, en son début, à un étirement des lèvres final. En fait, à la voyelle labiale [o] succède la sibilante labiale [ʃ]. Ceci justifie la forte action du MENT (cf amplitude du pic max). La voyelle subséquen-

* Il ne sera pas question ici du groupe des syllabes étirées [si, zi, se, ze] pour lequel le contexte préconsonantique [± lab] n'a pas grand pouvoir sur l'activité musculaire.

* cf l'analyse de chaque syllabe (GENTIL, 1980, pp 181 - 196)

te n'est pas concernée par l'activité de ce muscle. A son niveau, la protrusion arrive à son déclin et il semble qu'il n'y ait pas du point de vue labial l'influence progressive du [ʃ] à laquelle nous pouvions nous attendre.
L'OO INF vient renforcer l'activité du MENT. Sa fonction est synergique.
Le LLSA intervient pour relever la lèvre supérieure en découvrant les dents dans la protrusion.

1 - 2. Syllabe [ʒe], contexte préconsonantique [- lab]
Elle figure dans la phrase "Il faut rager".
Trois muscles travaillent : MENT, OO INF & DLI.
Nous constatons un effet discret de coarticulation labiale d'anticipation se traduisant par une protrusion légèrement antérieure à la sibilante et l'action du MENT correspondant. Celle-ci dure toute la protrusion et cesse avec elle, c'est-à-dire approximativement au milieu de la voyelle. Dans ce cas, nous remarquons un effet coarticulatoire rétentif, le [ʒ] influençant le [e].
L'OO INF est actif lors de la protrusion mais nous expliquons mal la fin d'activité de ce muscle alors qu'il n'y a plus protrusion et que la lèvre inférieure s'abaisse. Compte tenu de l'importance de l'abaissement, nous ne pouvons relier ce dernier à une réduction de l'activité de l'OO INF (cf les propriétés passives du muscle, CHEVRIE-MULLER, 1980).
En raison d'artéfacts, nous ne savons pas comment agit le DLI.

1 - 3. Comparaison interlocuteurs
Dans ce groupe de syllabes, nous confronterons les observations précédentes recueillies pour le locuteur féminin avec celles concernant le locuteur masculin (RD), étant entendu qu'il s'agit d'un parallèle fragmentaire du fait de l'absence de données relatives au MENT. Pour l'OO INF, nous notons une action plus longue et de plus grande amplitude par rapport à RD, pas d'activité du DLI qui existe systématiquement chez le locuteur masculin, une similitude dans l'action éventuelle du LLSA. Dans l'ensemble, quelques divergences peu importantes et tout à fait plausibles compte tenu du jeu articulatoire assez spécifique à chaque individu.

2 - SYLLABES PROTRUSES

Ces syllabes sont caractérisées par une sibilante [- lab] très influencée par la voyelle [+ lab] subséquente.

Nous examinerons la syllabe [zy] à la fois en contexte préconsonantique [– lab] et [+ lab], soit dans les phrases "C'est deux bizuths" et "t'es décousu".

2 - 1. Dans la syllabe avec contexte préconsonantique [– lab] (fig. 2) nous constatons l'activité des MENT, OO INF & DLI.
La protrusion démarre avec la consonne, autrement dit l'influence anticipative de la voyelle [+ lab] remonte jusque là. Le MENT responsable en partie de cette protrusion commence à agir, comme il se doit, un peu plus tôt et travaille tout au cours de la poussée des lèvres en avant et avec une amplitude élevée.
L'OO INF a une action synergique au MENT dans la protrusion. Il joue également dans le rapprochement des lèvres.
Le DLI fonctionne au début de la réalisation dans l'abaissement de la lèvre inférieure. A partir du début de la voyelle, nous ne pouvons plus expliquer ainsi son activité qui est vraisemblablement antagoniste à celle de l'OO INF.

2 - 2. En contexte préconsonantique [+ lab], la syllabe [zy] se caractérise par une protrusion qui a des degrés différents tout au long de la réalisation. En fait, la sibilante [– lab] est encadrée par deux productions [+ lab]. La protrusion décroît au niveau de la consonne pour croître à nouveau avec la voyelle [y]. L'action du MENT correspondant suit ces fluctuations (amplitude inférieure du pic max pour la consonne) et accompagne toute la protrusion (cf fig. 3).
L'OO INF travaille en synergie du MENT et le DLI comme antagoniste de l'OO INF.

2 - 3. Parallèle entre les deux locuteurs
Dans ce groupe, la comparaison des activités musculaires entre les deux locuteurs fait ressortir la même fonction de l'OO INF synergique au MENT encore qu'il soit difficile de juger le travail de l'orbiculaire chez le locuteur masculin, des mesures manquant en raison d'artéfacts.
Le DLI paraît plus efficace chez RD, c'est d'ailleurs l'impression générale que donne ce muscle plus présent et actif pour ce locuteur.

3 - SYLLABES FORTEMENT PROTRUSES

3 - 1. [ʒø], contexte préconsonantique [+ lab] (fig. 4)
Cette syllabe figure en finale de la phrase "C'est bien au jeu".
La protrusion est considérablement longue et s'explique par la suite de réalisations labiales. Elle a déjà lieu pendant la voyelle précédant la syllabe et, à ce niveau, est même supérieure à celle constatée durant la syllabe.
Le MENT assure ce mouvement aidé par l'OO INF.

3 - 2. [ʒø], contexte préconsonantique [– lab]
Nous trouvons cette syllabe dans la phrase "C'est bien

grégeux".
Nous constatons l'influence de la consonne [+ lab] sur la voyelle précédente, à savoir la protrusion antérieure à la consonne qui est provoquée par l'activité du MENT et de l'OO INF à titre synergique. Ce dernier muscle intervient encore dans le rapprochement des lèvres alors qu'il est contrôlé par le DLI.

3 - 3. Différence interlocuteurs

Nous observons, dans ce groupe de syllabes, pour le locuteur masculin, une plus grande activité de l'OO INF (auxiliaire du MENT) et ceci aussi bien pour la voyelle que pour la consonne. De même pour le DLI nous avons une activité de plus forte amplitude chez RD. Nous pensons que cette différence d'activité musculaire labiale peut provenir d'une différence de débit de parole entre les 2 locuteurs, les productions de RD étant sensiblement plus rapides. En effet, il a été remarqué (GAY & al., 1973) pour les consonnes [p, w] un phénomène semblable : activité musculaire labiale plus forte à débit plud rapide allant de pair avec une vitesse de déplacement plus grande des articulateurs.

CONCLUSION

Au terme de cette étude, il convient de résumer les constatations faites et nous sommes amenée à proposer quelques règles de coarticulation qui se dégagent des phénomènes observés.

Les syllabes faiblement protruses sont marquées par :

- Un effet regressif discret de coarticulation labiale se traduisant par une protrusion antérieure à la sibilante (sans considération de frontière syllabique) et l'action du MENT correspondante. Cette remarque tient compte bien entendu des effets progressifs de coarticulation labiale qui auraient pu éventuellement influencer la voyelle [– lab] précédant la sibilante [+ lab] et provenant de phonèmes antérieurs à celle-là.
- Lorsque le contexte préconsonantique est [– lab], un effet progressif de coarticulation labiale (imputable à la sibilante) sur la voyelle subséquente, avec activité des MENT & OO INF.

Les syllabes protruses sont caractérisées par la perméabilité de la sibilante [– lab] à l'influence de la voyelle [+ lab] subséquente.
En contexte préconsonantique [– lab] cette influence se concrétise par la protrusion constatée dès le début de la consonne et qui sur le plan de l'activité musculaire correspond au travail du MENT (protagoniste) et OO INF (synergique).
En contexte préconsonantique [+ lab] la protrusion est antérieure au début de la consonne et s'infléchit quelque peu durant cette dernière. Au niveau de l'activité musculaire, nous avons une action du MENT un peu décroissante pendant la consonne. Pour ce muscle, l'amplitude du pic max EMG relatif à cette dernière est inférieure à celle du même pic dans les syllabes en contexte [– lab].

En ce qui concerne les syllabes fortement protruses, la conjugaison de deux réalisations labiales entraîne une protrusion importante qui, en contexte préconsonantique [–]ab] se manifeste avant la consonne, sans respect de frontières syllabiques. L'action du MENT précède un peu le mouvement d'avancée des lèvres. Cet effet d'anticipation est plus remarqué dans ce groupe que dans celui des syllabes faiblement protruses. Néanmoins l'influence de la labialité consonantique dans le sens droite - gauche est nettement moins considérable que celle de la labialité vocalique dans ce même sens (cf [sy], etc).

Nous proposons donc les règles suivantes :

(1) Voy. [- lab] // [ʃ, ʒ] + [i, e] ⟶ Voy. labialisée // [ʃ, ʒ] + [i, e]

(2) Voy. [+ lab] // [ʃ, ʒ] + [i, e] ⟶ Voyelle [+ lab] // [ʃ, ʒ] + [i, e]

(3) Voy. [- lab] // [s, z] + [y, ø] ⟶ Voyelle [- lab] // [s, z] + [y, ø]

(4) Voy. [+ lab] // [s, z] + [y, ø] ⟶ Voyelle [+ lab] // [s, z] + [y, ø]

(5) Voy. [- lab] // [ʃ, ʒ] + [y, ø] ⟶ Voy. labialisée // [ʃ, ʒ] + [y, ø]

Ces règles sont établies sur la base de l'activité musculaire en liaison avec le mouvement. Il peut nous être reproché que le nombre restreint des réalisations et des sujets ne permettent pas de tirer des conclusions statistiquement valables. Par ailleurs, une analyse quantitative et non qualitative du mouvement serait souhaitable. Nous sommes tout à fait consciente de ces insuffisances et ces règles sont à prendre comme des motions. Cette étude, tributaire des moyens manuels, est forcément limitée et doit être considérée comme préliminaire à des études ultérieures bénéficiant du dépouillement automatique et qui offriront d'autres possibilités.

REFERENCES

ABBS (J.) & NETSELL (R.), 1973, "Coordination of the jaw and lower lip during speech production", ASHA, Détroit

BELL-BERTI (F.) & HARRIS (K.S.), 1978, "Anticipatory coarticulation Some implications from a study of lip rounding", SR on SR 54, pp 121 - 125

BONNOT (J.F.), GREINER (G.F.), MATON (B.) & CHEVRIE-MULLER (C.), 1980, "Etude électromyographique de l'organisation temporelle (timing) des consonnes simples, géminées et des groupes de consonnes labiales en français", Préimprimés Séminaire sur la labialité, GALF

CHEVRIE-MULLER (C.), 1980, "Organisation neuromotrice et physiologie de la motricité de l'effecteur labial", Séminaire sur la labialité, GALF

DANILOFF (R.G.) & TATHAM (M.A.), 1979, "Segmental invariance reconsidered", 9 th Int. Cong. Phonetic Sci, 2, pp 332 - 336

GAY (T.), 1979, "Coarticulation in some consonant - vowel and consonant cluster - vowel syllables" in Frontiers of speech communication, London, Academic Press, Lindblom & Ohman (eds), pp 69 - 77

GAY (T.) & HIROSE (H.), 1973, "Effect of speaking rate on labial consonant production. A combined electromyographic/High

Speed Motion picture study, Phonetica 27, pp 44 - 56

GENTIL (M.), 1980, "Labialité en français, étude phonétique et aspects physiologiques - Des lèvres : géométrie, cinématique, électromyographie et idiolectologie", thèse de doctorat de 3e cycle, Grenoble III, vol. 1 & 2

GENTIL (M.), BOE (L.J.) & DESCOUT (R.), 1980, "Etude EMG des lèvres (OO INF, MENT, DLI & LLSA) dans la réalisation de monosyllabes du français, préimprimés Séminaire sur la labialité, GALF

HUGHES (O.) & ABBS (J.), 1976, "Labial-mandibular coordination in the production of speech", Phonetica 33, pp 199 - 221

KENNEDY III (J.G.) & ABBS (J.), 1979, "Anatomical studies of the perioral motor system : Foundations for studies in speech physiology" in Speech and Language : Advances in basic research and practice, 1, Academic Press Inc. pp 211 - 270

LUBKER (J.), MacALLISTER (R.) & LINKER (W.), 1980, "Anticipatory labial coarticulation: Extent and influencing variables" préimprimes Séminaire international sur la labialité,GALF

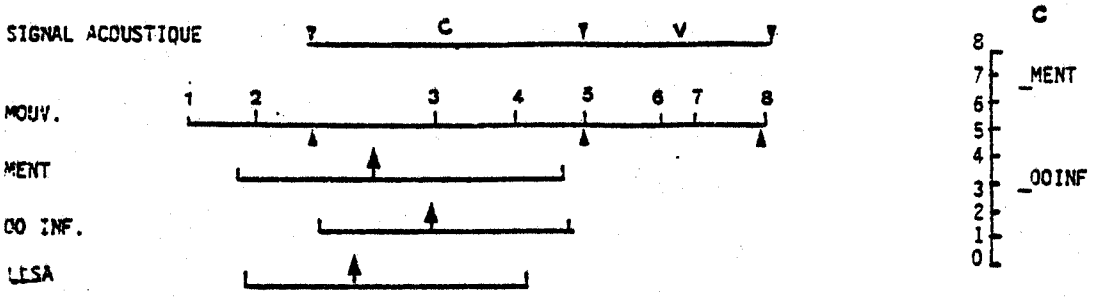
LYSAUGHT (G.), ROSOV (R.J.) & HARRIS (K.S.), 1961, "Electromyography as a speech research technique with an application to labial stops", JASA 33, p 842 (A)

MacALLISTER (R.), 1978, "Temporal asymmetry in labial coarticulation", Stockolm, pilus 35, pp 1 - 16

RIORDAN (C.J.), 1976, "Electromyographic correlates of the phonological /y/ - /u/ distinction in french, J. Phonetics 4, pp 1 - 16

En ce qui concerne les échelles, sur les figures 1 à 4 suivantes, il y a lieu de tenir compte d'une réduction de 30 % intervenue lors de la reproduction de ces documents

[ʃe] dans : "C'est bien fauché"



Déroulement temporel des gestes labiaux et des signaux EMG, repérés par rapport au signal acoustique pour la syllabe CV [ʃe]
(2 mm = 1/100 seconde). Les amplitudes des pics max du signal EMG brutsont notées en 102µV.

1 : Lèvres protruses
2 : Abaissement de la lèvre inférieure
3 : Max protrusion
4 : Abaissement de la lèvre et du maxillaire inférieurs. Nette diminution de la protrusion
5 : Fin de protrusion
6 : Etirement des commissures
7 : Remontée de la lèvre inférieure et abaissement de la lèvre supérieure
8 : Fin de mouvement

[oʃe] dans : "C'est un fauché"

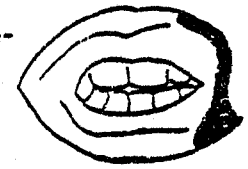
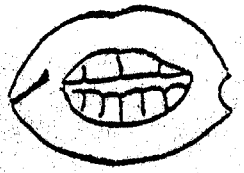
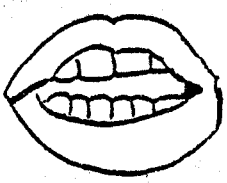
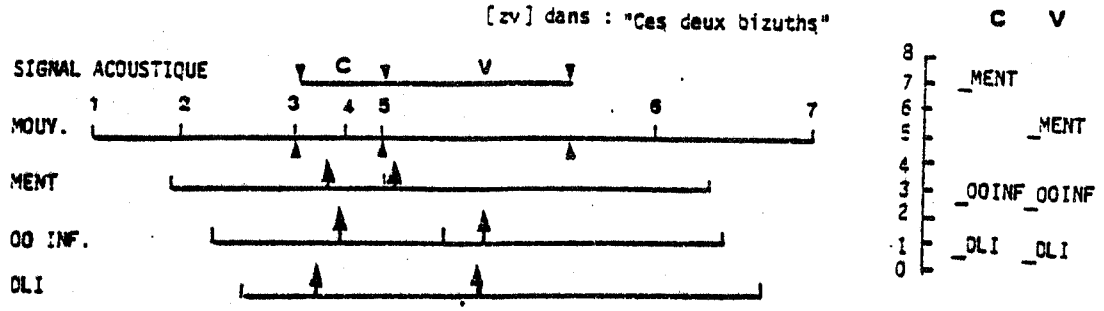
Image N°	Correspondance signal acoust.	Description de l'image	
93	Début de la voyelle [o]	Lèvres protruses notamment la lèvre supérieure découvrant deux incisives	
102	Début de la consonne [ʃ]	La protrusion est plus importante et l'abaissement de la lèvre inférieure laisse apparente les 4 incisives du bas	
105	Milieu " "	Lèvres très protruses découvrant les dents légèrement déserrées	
111	Début de la voyelle [e]	Les lèvres sont encore un peu protruses, les dents sont déserrées et la langue est apparente	
115	Milieu " "	La lèvre inférieure est abaissée au maximum en même temps que les commissures commencent à s'étirer	
122	Fin " "	La bouche est presque fermée, laissant apparaître seulement le bas des incisives supérieures	

Fig. 1



Déroulement temporel des gestes labiaux et des signaux EMG, repérés par rapport au signal acoustique pour la syllabe CV [zv]
(2 mm = 1/100 seconde). Les amplitudes des pics max du signal EMG brut sont notées en 10^2uV.

1 : Ouverture de la bouche
2 : Abaissement de la lèvre inférieure et montée de la lèvre supérieure
3 : Début de protrusion
4 : Remontée de la lèvre inférieure
5 : Max protrusion
6 : Fin de protrusion
7 : Fin de mouvement

[izy] dans : "Ces deux bizuths"

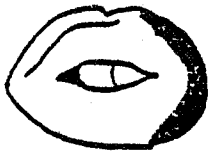
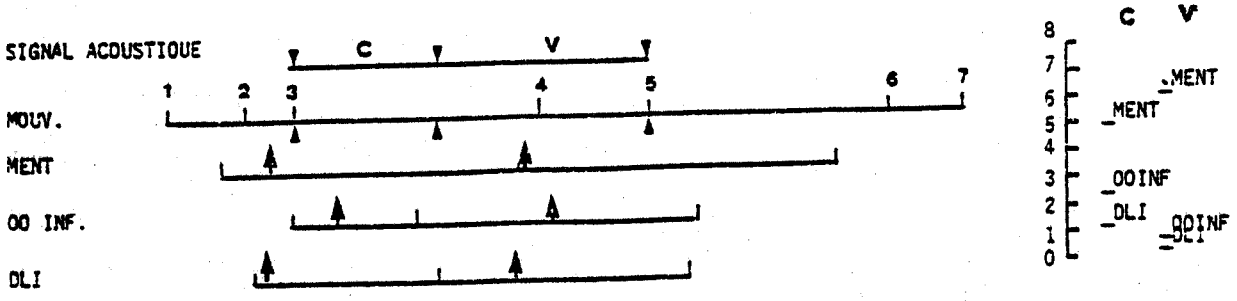
Image N°	Correspondance signal acoust	Description de l'image	
3332	Début de la voyelle [i]	La bouche s'ouvre à peine sans découvrir encore les dents	
3344	Début de la consonne [z]	Les lèvres commencent à être protruses et laissent apparaître les dents	
3346	Milieu " "	La protrusion s'est accrue en même temps que la bouche s'est fermée	
3349	Début de la voyelle [y]	La protrusion des lèvres est au maximum	
3355	Milieu " "	La bouche est pratiquement obturée et les lèvres encore protruses	
3361	Fin " "	La protrusion des lèvres diminue nettement et la bouche est presque fermée	

Fig. 2

[zy] dans : " T'es décousu"



Déroulement temporel des gestes labiaux et des signaux EMG, repérés par rapport au signal acoustique pour la syllabe CV [zy]
(2 mm = 1/100 seconde). Les amplitudes des pics max du signal EMG brut sont notées en $10^{2}\mu V$.

1 : Lèvres protruses
2 : Abaissement de la lèvre inférieure
3 : Décroissance de la protrusion
4 : Max protrusion
5 : Remontée de la lèvre inférieure
6 : Fin de protrusion
7 : Fin de mouvement

[uzy] dans : "tes décousu"

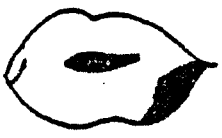
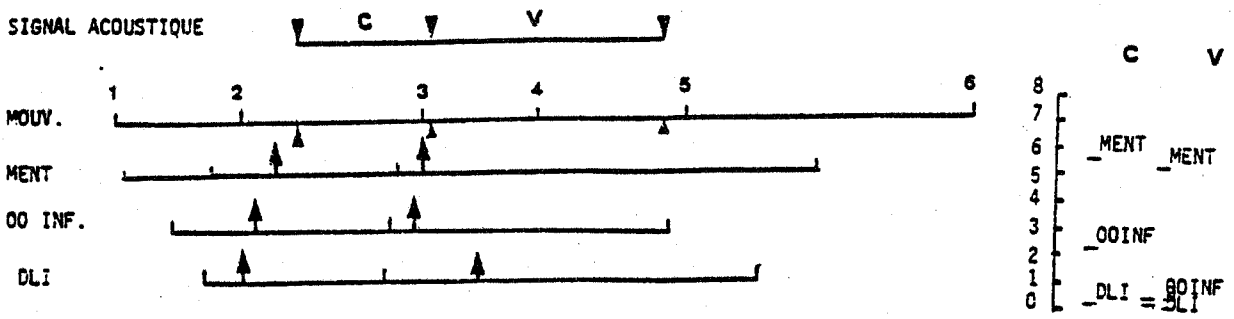
Image N°	Correspondance signal acoust	Description de l'image	
1431	Début de la voyelle [u]	Les lèvres sont protruses et la bouche presque fermée laisse apparaitre un peu des incisives médianes supérieures	
1438	Début de la consonne [z]	La protrusion a décru, les deux incisives médianes supérieures sont découvertes	
1442	Milieu " "	Les lèvres sont protruses et l'on voit les deux incisives médianes supérieures	
1446	Début de la voyelle [y]	Pas de différence notable avec l'image 1442	
1451	Milieu "	Les lèvres encore protruses se resserrent	
1458	Fin "	Les lèvres sont pratiquement fermées	

Fig. 3

[ʒø] dans : "C'est bien au jeu"



LLSA (artéfact)

Déroulement temporel des gestes labiaux et des signaux EMG, repérés par rapport au signal acoustique pour la syllabe CV [ʒø]

(2 mm = 1/100 seconde). Les amplitudes des pics max du signal EMG brut sont notées en 10^{-5}µV.

1 : Lèvres très protruses
2 : Remontée de la lèvre supérieure
3 : Abaissement de la lèvre et du maxillaire inférieurs
4 : Remontée de la lèvre inférieure
5 : Fin de protrusion
6 : Fin de mouvement

[oʒø] dans : "C'est bien au jeu"

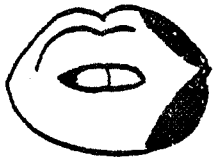
Image N	Correspondance signal acoust'		Description de l'image	
1650	Début de la voyelle	[o]	Lèvres très protruses, bouche à peine entr'ouverte on devine une à deux incisives	
1660	Début de la consonne	[ʒ]	Lèvres encore protruses, on voit les deux incisives supérieures	
1663	Milieu	" "	Même disposition mais incisives un peu plus découvertes	
1667	Début de la voyelle	[ø]	Léger abaissement du maxillaire inférieur, les lèvres sont protruses	
1673	Milieu	" "	Remontée de la lèvre inférieure, on aperçoit une partie des deux incisives supérieures	
1679	Fin	" "	Les lèvres se sont rapprochées l'une de l'autre mais sont encore un petit peu protruses	

Fig. 4

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ESPACE ARTICULATOIRE MANDIBULAIRE : METHODE ET PREMIERS RESULTATS

Louis-Jean BOË - Eric BOGNÁR — Institut de Phonétique de Grenoble
Thomas GAY — Department of Oral Biology, Farmington

RESUME

Le but de cette étude est de présenter l'espace articulatoire mandibulaire d'un locuteur français, à travers l'organisation des positions du point incisif inférieur pendant la réalisation de logatomes de type CVCV. Les mouvements de ce point de référence ont été enregistrés à l'aide d'un appareillage magnétométrique. Les données recueillies ont été soumises à un traitement statistique permettant d'obtenir des ellipses de dispersion représentant les consonnes et les voyelles analysées. Les résultats montrent, pour ce locuteur, la validité de la notion de cible consonantique et la faible résistance des voyelles à la coarticulation.

ABSTRACT

The purpose of this study is to determine the articulatory space of the jaw through the organisation of the positions of the incisal point in speech production. Speech material consists of 24 CVCV non sense utterances (where C = [p, t, k, f, s, ʃ] and V = [a, i, y, u]) each repeated twenty times by a French male speaker. Horizontal and vertical jaw movements have been recorded with a Mandibular Kinesiograph. The data have been submitted to statistical analysis and 48 ellipses of dispersion have been plotted for C_2 and V_2. The dispersion of the consonants [p, t, k] and [f s ʃ] in the same vocalic environment and their respective concentration in different environments support the validity of the notion of consonantal target for the jaw. Inversely, the separation of the same vowels in different consonantal environments and their concentration in the same consonantal environment indicate the weakness of their resistance to coarticulation.

ESPACE ARTICULATOIRE MANDIBULAIRE : METHODE ET PREMIERS RESULTATS

Louis-Jean BOË - Eric BOGNÁR — Institut de Phonétique de Grenoble
Thomas GAY — Department of Oral Biology, Farmington

Par rapport aux structures peu opaques aux radiations telles que les lèvres ou le voile, la mandibule, qui impose sa densité sur les documents radiographiques, semble d'un abord facile : parmi les distances traditionnellement mesurées, aucune n'est plus évidente que celle qui sépare les incisives et qui fournit ainsi une image de l'ouverture maxillaire. Cette facilité d'accès a cependant contribué à masquer la complexité des mouvements mandibulaires (régis par une articulation reconnue comme l'une des plus complexes du corps humain) (LEJOYEUX, 1967; POSSELT, 1969), puisque la distance interincisive n'apporte aucune information sur les déplacements horizontaux de l'articulateur. Aussi de nouvelles techniques, permettant de suivre l'évolution de la mandibule dans les deux plans (HIXON, 1971; ABBS & GILBERT, 1973; MAC CUTCHEON & al., 1977; MULLER & ABBS, 1979), ont vu le jour, tandis que parallèlement la mâchoire a pu accéder au devant de la scène grâce à une controverse apparue au cours du dernier Congrès International des Sciences Phonétiques, par laquelle s'affrontaient notamment P. LADEFOGED, qui considère que le rôle de la mandibule dans la détermination de la forme du conduit vocal est négligeable et B. LINDBLOM, pour qui ce rôle est "primordial" (MAC NEILAGE & al., 1979). Quel que soit le point de vue que l'on adopte, une étude systématique de cet articulateur paraît indispensable car la suppression ou l'admission d'un paramètre relatif à la mâchoire dans un modèle articulatoire doit se faire à partir de données précises sur les mécanismes cinétiques de la mandibule, sur la constance ou l'inexistence de positions articulatoires cibles, sur le taux de résistance à la coarticulation, sur le rôle de l'accent, celui de la vitesse d'élocution, l'existence éventuelle de constantes d'un locuteur à l'autre, etc.

L'étude que nous présentons ici, qui s'inscrit dans la perspective d'une telle monographie, traite de l'espace articulatoire de la mandibule : notre but est de décrire, d'une manière statique, les positions du point incisif inférieur pendant la réalisation de la syllabe accentuée de plusieurs logatomes, afin de dégager des tendances préliminaires à une analyse plus approfondie dans le domaine de la coarticuation consonantique et vocalique.

ACQUISITION ET TRAITEMENT DES DONNEES

Le locuteur (LJB) qui s'est prêté à l'expérimentation servant de base à cette étude est âgé de quarante ans, originaire du sud-ouest de la France, connaissant le but de l'expérience. Des radiographies sagittales montrent qu'il se caractérise par un recouvrement incisif; celui-ci n'entraîne pas de défaut perceptible.

Le corpus a été réalisé de telle sorte qu'il permette l'observation d'écarts maxima de la position du point incisif, ainsi que l'examen de la réaction mandibulaire à l'opposition antérieure de labialité. C'est pourquoi ont été retenues les occlusives [p], [t], [k] et les constrictives correspondantes [f], [s], [ʃ], combinées avec les voyelles [a], [i], [y], [u]. Ces réalisations apparaissent dans 24 logatomes formés de deux syllabes CV identiques. Ces logatomes,

qui sont insérés dans des phrases porteuses ("Dis CVCV six fois"), apparaissent chacuns 20 fois, selon un ordre aléatoire, dans le corpus lu par le sujet.

L'enregistrement des mouvements mandibulaires a été opéré grâce à un appareil récemment mis au point pour la recherche et la pratique odontologiques, le MANDIBULAR KINESIOGRAPH (modèle K-5, réalisé par Myotronics Research Inc.; JANKELSON & al., 1975). Cet appareil permet notamment l'enregistrement continu des déplacements d'un point mandibulaire dans les plans frontal et transversal (avec une précision dépassant le dixième de millimètre). Il se compose de plusieurs magnétomètres fixés sur une structure d'aluminium assujettie à des lunettes spéciales que chausse le locuteur et qu'un opérateur ajuste de telle sorte que les capteurs se trouvent dans un plan perpendiculaire au plan de Francfort. Les magnétomètres principaux, au nombre de trois, enregistrent les modifications du champ magnétique dues au mouvement vertical et antéro-postérieur d'une plaquette aimantée fixée contre les incisives inférieures centrales du sujet. Un magnétomètre auxiliaire, qui capte uniquement le champ magnétique terrestre, permet de le déduire des informations recueillies par les autres. Les signaux captés par les magnétomètres sont amplifiés puis enregistrés sur un magnétophone MF à 11 pistes, à partir duquel une sortie graphique permet de les visualiser, sur un écran oscilloscopique ou sur un visicorder.

Outre la sensibilité et la précision extrême de ses capteurs, cet appareil présente deux avantages pour l'étude phonétique. D'une part, la pastille magnétique est de taille suffisamment réduite (1,28 x 0,64 x 0,32 cm), pour ne pas altérer la prononciation. D'autre part, le dispositif soutenant les magnétomètres est si léger qu'il peut être supporté sans gêne durant l'enregistrement d'un corpus extrêmement long. En conséquence, cet appareillage permet l'acquisition de très nombreuses données, basées sur des répétitions systématiques qui assurent une plus grande fiabilité à l'analyse des résultats.

Le traitement des données s'est opéré en deux étapes, et en deux lieux : grâce à un programme graphique et statistique mis au point par Ian COVILLE, implanté sur un calculateur du Laboratoire d'Oral Biology, la moyenne et l'écart-type des déplacements horizontaux et verticaux ainsi que leur coefficient de corrélation pour quinze (la première et les quatre dernières apparitions n'étant pas entrées en ligne de compte) des vingt répétitions de chacun des logatomes ont été calculés, tracés et listés tous les 5 ms durant un laps de temps (1"22) dépassant la durée maximum des items. Naturellement, du fait de l'uniformité de l'étalonnage temporel les calculs ne tiennent pas compte des inévitables différences de durée des segments CVCV enregistrés. Il s'ensuit une certaine erreur que seule une normalisation un peu lourde aurait permis d'éviter. Son importance a cependant été minimisée en la répartissant de part et d'autre de la frontière entre la consonne et la voyelle de la seconde syllabe, puisque l'instant zéro de l'échantillonnage temporel, correspond à la première apparition des vibrations sonores après le consonne sourde sur le signal acoustique. De cette manière, l'erreur, qui est nulle à l'instant zéro, s'accroît à mesure que l'on s'en éloigne : la conséquence en est que la seconde syllabe de nos logatomes, que nous étudions, est moins soumise aux irrégularités temporelles que la première.

La seconde étape du traitement des données s'est effectué à Grenoble, grâce à un programme permettant le calcul et le tracé d'ellipses de dispersion à partir des données obtenues auparavant. Ces ellipses, qui représentent un pourcentage donné de la dispersion théorique des valeurs, correspondent au centre des réalisations consonantiques et vocaliques de la seconde syllabe des items du corpus. Afin de garder à l'esprit le déplacement réel du point incisif et bien que les différences de valeurs sur l'axe horizontal soient deux fois plus

faibles que sur l'axe vertical, nous n'avons pas dilaté l'axe horizontal. Après le calcul de leurs coordonnées (longueurs des axes, aire et pente), les 48 ellipses (une pour chaque voyelle et pour chaque consonne de la seconde syllabe des 24 logatomes du corpus) représentant 90% de la dispersion ont été dessinées par groupe de trois ou quatre, selon la nature des réalisations à étudier.

HIERARCHIES ET ESPACES MANDIBULAIRES

A partir des données dont l'obtention vient d'être exposée, notre analyse a présenté deux aspects complémentaires : d'une part l'observation des valeurs moyennes de l'ensemble des réalisations en vue de l'étude de leur hiérarchie en fonction de l'ouverture et de la protraction, d'autre part, l'examen des espaces formés par les ellipses de dispersion de ces réalisations en relation avec l'opposition consonne/voyelle.

La première étape, préalable à chacune de ces deux tâches, a été la sélection, parmi la masse de valeurs dont nous disposons, des instants considérés respectivement comme centre de la consonne et comme centre de la voyelle étudiées. Ce choix qui a été réalisé par repérage à la visu sur les tracés simultanés des déviations en X et Y et le signal de parole, a été opéré selon le critère suivant : est considéré comme centre, le lieu de la déviation mandibulaire maximale (horizontale et/ou verticale), même dans les cas où ce lieu ne correspond pas exactement au centre du signal acoustique (Figure 1). Nous devons toutefois préciser que dans le cas d'un corpus de logatomes de type CVCV, l'effet coarticulatoire par lequel la mandibule "semble parfois ignorer certaines consonnes ou voyelles au cours de la chaîne parlée", observé notamment par WIOLAND (1971) est heureusement rare.

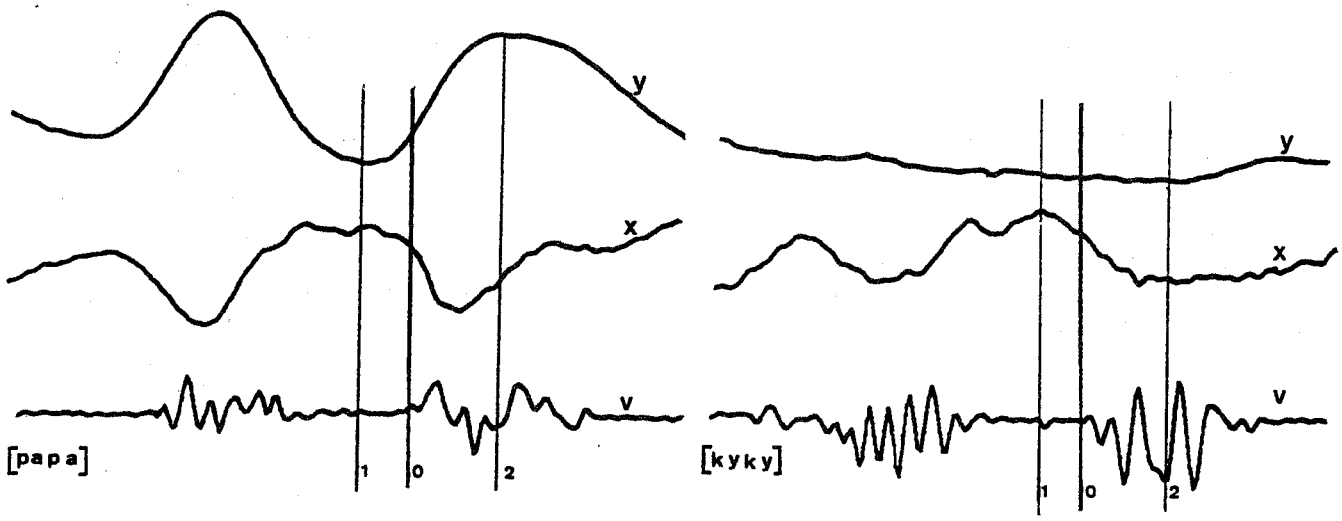


FIGURE 1 Tracés illustrant le choix des instants considérés respectivement comme centre de la consonne et comme centre de la voyelle de la seconde syllabe des logatomes de notre corpus en fonction du mouvement mandibulaire vertical (comme dans dans le cas de [papa]) ou horizontal (comme dans celui de [kyky]).

X = déplacement horizontal
Y = déplacement vertical
V = signal de la parole

0 = instant 0 de l'échantillonnage temporel
1 = centre de la consonne
2 = centre de la voyelle

Nous avons ensuite listé les coordonnées du point incisf inférieur (Ii) aux instants selectionnés. Ces coordonnées, dont il faut garder à l'esprit que chacune d'elles représente une moyenne sur 15 réalisations du même item, indiquent la distance parcourue par Ii à partir de sa position de référence (celle-ci s'obtient lorsque le sujet est en position d'intercuspidation). Les coordonnées des réalisations consonantiques, dans les 4 contextes vocaliques possibles, figurent dans le tableau 1, celles des réalisations vocaliques, dans différents entourages consonantiques, se trouvent dans le tableau 2.

		CONTEXTES VOCALIQUES			
		[a]	[ɪ]	[y]	[u]
CONSONNES	[p]	(-9,-76)	(-4,-56)	(0,-46)	(0,-58)
	[t]	(0,-20)	(0,-26)	(2, -8)	(-2,-16)
	[k]	(-11,-56)	(-19,-46)	(-8,-38)	(-7,-52)
	[f]	(-7,-60)	(-1,-60)	(6,-60)	(0,-61)
	[s]	(4,-16)	(4,-10)	(7,-16)	(8,-12)
	[ʃ]	(-15,-38)	(-20,-34)	(-21,-28)	(-10,-14)

TABLEAU 1 Ce tableau présente les coordonnées du point incisif inférieur lors des réalisations consonnantiques [p] [t] [k], [f], [s] et [ʃ] dans les contextes vocaliques [a] [ɪ], [y] et [u].
La première coordonnée, correspondant à l'axe horizontal, est positive lorsque le point Ii est antérieur à sa position de référence, et négative lorsqu'il est postérieur à cette même position ; la seconde coordonnée, correspondant à l'axe vertical, croît avec la fermeture mandibulaire.
L'unité utilisée est le 1/10^{e} de millimètre.

		CONTEXTES CONSONANTIQUES					
		[p]	[t]	[k]	[f]	[s]	[ʃ]
VOYELLES	[a]	(-4,-148)	(-1,-102)	(-10,-138)	(-4,-132)	(2,-88)	(-13,-88)
	[ɪ]	(-1,-64)	(1,-36)	(-19,-54)	(0,-62)	(5,-20)	(-19,-40)
	[y]	(4,-52)	(4,-20)	(-5,-42)	(5,-58)	(8,-18)	(-18,-32)
	[u]	(3,-58)	(-4,-28)	(-7,-56)	(-1,-28)	(8,-18)	(-9,-22)

TABLEAU 2 Ce tableau présente les coordonnées du point incisif inférieur lors des réalisations vocaliques [a], [ɪ], [y] et [u] dans les entourages consonantiques [p], [t], [k], [f], [s] et [ʃ].
Les valeurs des coordonnées et l'unité utilisée correspondent à celles du premier tableau.

Afin de faciliter l'interprétation de ces résultats, nous avons classé les valeurs des réalisations consonantiques et vocaliques, d'une part en fonction de l'accroissement de l'ouverture mandibulaire et d'autre part en fonction de la rétraction du point incisif inférieur. Cette réorganisation des résultats a été l'objet de 4 représentations schématiques dans lesquelles les symboles géométriques représentent les réalisations phonétiques, les colonnes, leurs contextes d'apparition et les rangées, leur classement respectif (Figure 2). L'intérêt de cette schématisation est de permettre de suivre de manière plus frappante la stabilité hiérarchique des réalisations observées en fonction des variations contextuelles. Les valeurs indexées aux symboles phonétiques indiquent le score de chaque réalisation (une réalisation 4 fois au sommet de la hiérarchie obtient 24 points, 4 fois à la seconde place, 20 points ...). Le classement des scores nous donne une image de la hiérarchie moyenne. Pour trivial que soit ce procédé, il permet néanmoins de visualiser instantanément d'une part la hiérarchie des éléments de notre corpus en fonction de la direction du mouvement, et d'autre part, la limite de validité de cette hiérarchie par l'indication des fluctuations de classement.

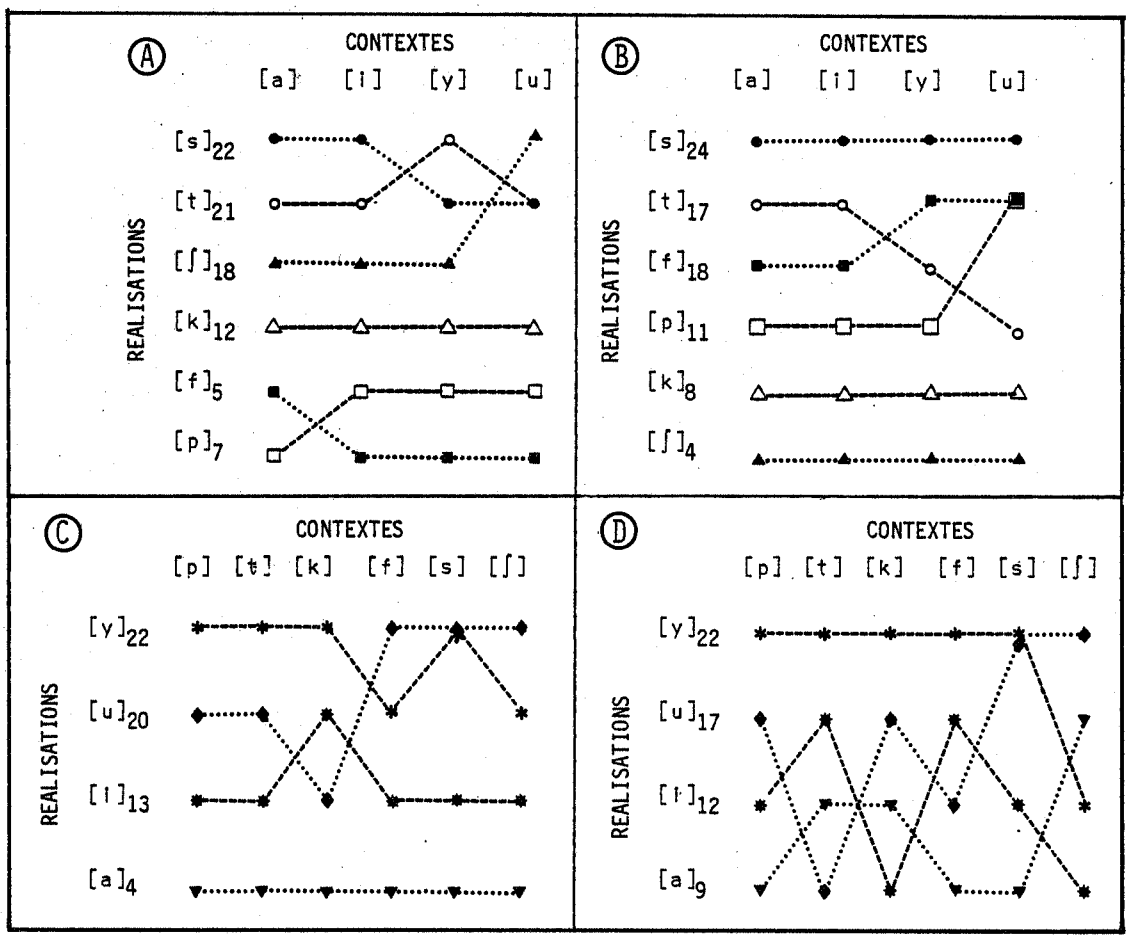


FIGURE 2 Fluctuations du classement de chaque réalisation phonétique selon le contexte. La valeur indexée aux réalisations indique leur score (voir le texte) : pour les figures 3a et 3c, le score le plus élevé indique la réalisation la plus fermée; pour les figures 3b et 3d, le score le plus élevé indique la réalisation la plus rétractée

a : classement des consonnes selon l'axe vertical
b : classement des consonnes selon l'axe horizontal
c : classement des voyelles selon l'axe vertical
d : classement des voyelles selon l'axe horizontal

En ce qui concerne les consonnes, les tendances qui se dégagent de ces schémas sont les suivantes : l'ouverture consonantique croît selon l'ordre [s], [t], [ʃ], [k], [f], [p], (figure 2a). Notons que ceci est conforme à la hiérarchie établie par ABBS & al. (1972) (cité par MAC NEILAGE & LADEFOGED, 1976) pour les occlusives. Par ailleurs le déplacement antéro-postérieur de la mâchoire suit l'agencement [s], [f], [t], [p], [k], [ʃ], cette dernière consonne se caractérisant par la protraction mandibulaire maximale. (figure 2b). Nous pouvons estimer ces hiérarchies comme assez stables chez notre sujet, dans la mesure où les fluctuations d'origine contextuelle dans ce classement sont peu importantes (en effet, sur les 12 consonnes des schémas 2-a et 2-b, 4 sont parfaitement stables, 5 ont des fluctuations hiérarchiques d'un rang et 3 de deux rangs). Au vu de ces données, les consonnes se caractérisent donc par leur relative stabilité hiérarchique.

Les figures 2-c et 2-d montrent une organisation vocalique totalement différente. En ce qui concerne les scores obtenus par les différentes voyelles, on peut observer le même classement pour l'ouverture et pour la rétraction mandibulaire : [y], [u], [i], [a]; [y] étant la voyelle la plus fermée et la plus rétractée. On note que pour l'ouverture, cette hiérarchie semble maintenant bien établie puisqu'on la retrouve chez de nombreux auteurs traitant de diverses langues; citons entre autres KANEKO (1957) pour le japonais* et l'anglais-américain, KIM & FUJISAKI (1974) pour le coréen et le japonais* , et WIOLAND (1971) dont les résultats concernent toutes les voyelles du français.

Cependant, l'examen des fluctuations des hiérarchies, tant horizontales que verticales, révèle leur instabilité contextuelle chez notre locuteur. La position du point incisif n'est hiérarchiquement stable que pour la réalisation de la voyelle [a] sur l'axe vertical. Toutes les autres voyelles connaissent des modifications de rang, notamment en ce qui concerne la localisation horizontale.

Cette première étape nous permet donc de conclure à l'existence d'importantes différences dans l'organisation vocalique et consonantique : en effet, la stabilité des hiérarchies consonantiques par rapport aux hiérarchies vocaliques permet de faire l'hypothèse qu'il existe un espace mandibulaire, défini par les mouvements verticaux et horizontaux de l'articulateur, dans lequel les consonnes occuperaient des lieux spécifiques, assimilables à des cibles, tandis que les voyelles seraient libres de tels lieux.

Notre deuxième tâche a consisté à confronter cette hypothèse avec la représentation de la variation des données par des ellipses de dispersion. Celles-ci se répartissent sur nos documents en deux groupes consonantiques (les espaces de chaque consonne suivie de toutes les voyelles et les espaces de toutes les consonnes suivies de chaque voyelle) et en deux groupes vocaliques (les espaces de chaque voyelle précédée de toutes les consonnes et les espaces de toutes les voyelles précédées de chaque consonne).

Une première confirmation de la rigueur de l'organisation hiérarchique des consonnes du point de vue des déplacements mandibulaires nous est fournie par les tableaux 3a et 3b. Ceux-ci, qui indiquent les limites horizontales et verticales (X MAX : rétraction maximale, X MIN : protraction maximale, Y MAX : fermeture maximale, Y MAX : ouverture maximale) des sous-ensembles d'ellipses consonantiques [p t k] et [f s ʃ] dans le même contexte vocalique, montrent la régu-

*. Le /u/ japonais se réalise en fait [ɯ].

valeurs extrêmes	contextes vocaliques [a]	[i]	[y]	[u]	valeurs extrêmes	contextes vocaliques [a]	[i]	[y]	[u]
X MAX	[k]	[k]	[k]	[k]	X MAX	[ʃ]	[ʃ]	[ʃ]	[ʃ]
X MIN	[t]	[t]	[p]	[t]	X MIN	[s]	[f]	[f]	[s]
Y MAX	[t]	[t]	[t]	[t]	Y MAX	[s]	[s]	[s]	[ʃ]
Y MIN	[p]	[p]	[p]	[p]	Y MIN	[f]	[f]	[f]	[f]

TABLEAU 3 Réalisations consonantiques occupant les zones extrêmes de l'espace mandibulaire dans les contextes vocaliques [a], [i], [y], [u], pour le sous-ensemble [p], [t], [k] (à droite), et pour le sous-ensemble [f], [s], [ʃ] (à gauche).

X MAX : rétraction maximale
X MIN : protraction maximale
Y MAX : fermeture maximale
Y MIN : ouverture maximale

larité de cette organisation*. Cette stabilité de l'espace consonantique se confirme par ailleurs grâce à deux séries d'indices. L'examen des groupes d'ellipses représentant différentes consonnes dans le même contexte vocalique révèle une disjonction nette de la plupart des réalisations. Dans l'espace [p -t -k], les surfaces de recoupement interconsonantiques sont mimimes, voire nulles entre [t] et [k]. Elles sont généralement plus importantes entre [k] et [p], la confusion étant la plus importante dans le contexte [a] (figure 3b). Dans l'espace [f -s -ʃ], mis à part un faible recoupement, toujours dans l'entourage de [a], les ellipses représentent des cibles totalement séparées les unes des autres. Inversement, les séries d'ellipses montrant la même consonne dans différents entourages vocaliques présentent, dans tous les cas, une forte cohésion, se traduisant par une zone commune à toutes les ellipses. L'exemple que nous présentons, concernant les réalisations de [t] n'a, à ce titre, rien d'exceptionnel (figure 3a).

De même que la hiérarchie des voyelles, qui varie selon l'entourage consonantique, les ellipses de dispersion vocaliques présentent des caractéristiques inverses des ellipses consonantiques.

Tout d'abord notons que la dispersion des réalisations d'une même voyelle en contextes consonantiques différents est deux fois plus importante que la dispersion consonantique analogue, et la dispersion vocalique la plus faible (celle de [y]) est du même ordre de grandeur que la dispersion consonantique la plus forte (celle de [k]). (N.B. Cette surface ne correspond pas à l'union des ellipses, mais au produit des différences entre les valeurs minimales et maximales). La conséquence de l'ampleur de ces dispersions est naturellement la difficulté à dégager des zones vocaliques cibles. Au contraire, nous pouvons observer que la tendance générale des ellipses vocaliques, notamment celles qui sont précédées de constrictives,est à l'absence de recoupement (figure 3c).

D'autre part, la comparaison des ellipses représentant les quatre voyelles précédées de la même consonne rend l'absence de cible vocalique particulièrement évidente. Les espaces vocaliques, qui s'organisent tous de manière identique, présentent d'une part l'ellipse isolée du [a] et d'autre part un agglomérat de trois ellipses de [i], [y] et [u] caractérisé par une zone de recouvrement très importante (figure 3d).

* Le "hasard" a voulu qu'après avoir dissocié la représentation des occlusives et des constrictives pour éviter la confusion qui naîtrait de la présence de 6 ellipses sur le même document, nous puissions adopter, sinon justifier, une étude séparée de deux espaces consonantiques parallèles.

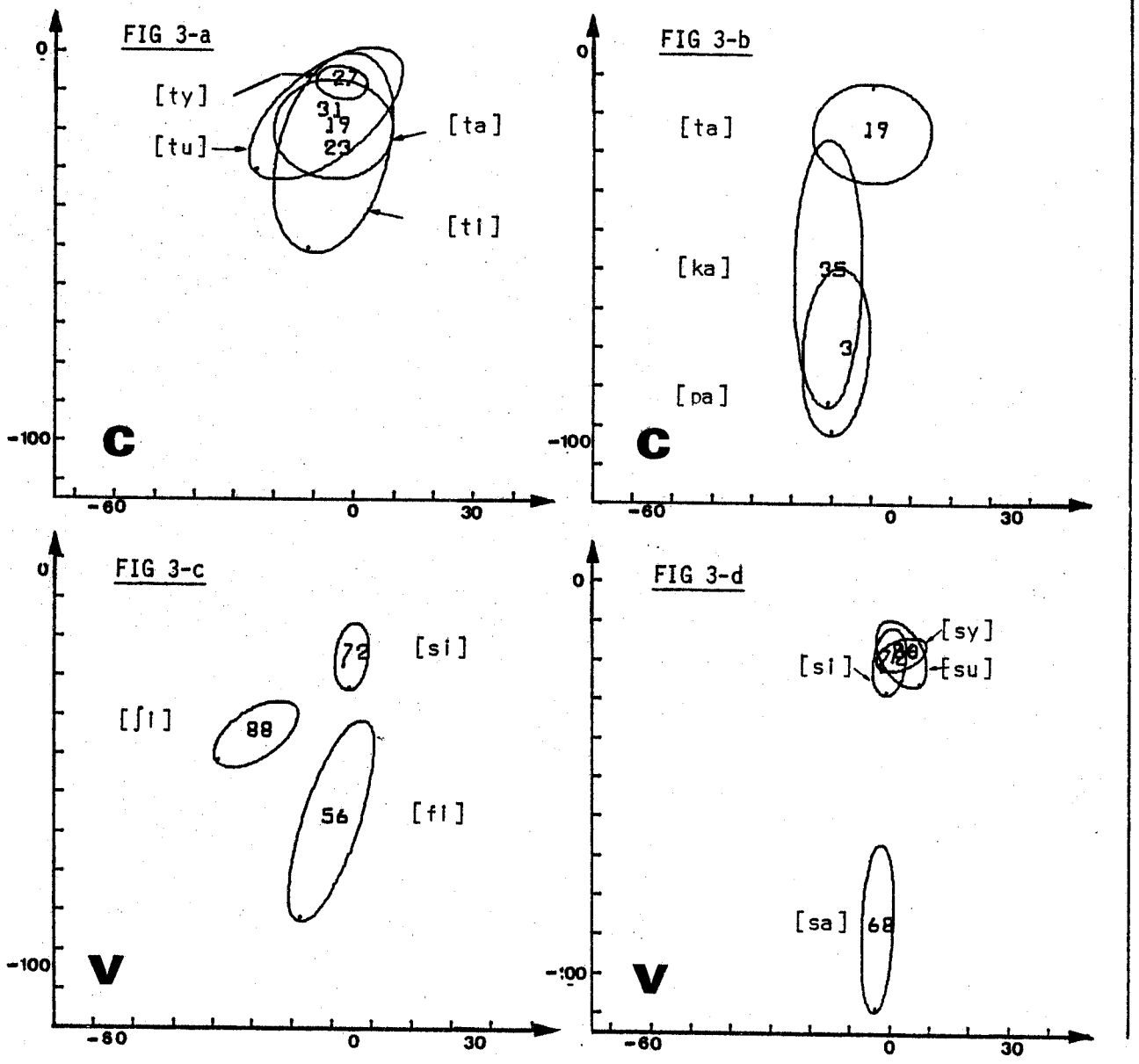


FIGURE 3 Tracés de quelques ellipses caractéristiques de l'opposition consonne/ voyelle représentant 90% de la dispersion théorique des valeurs mesurées à partir de 15 réalisations de chaque item de notre corpus (au total, 96 groupes d'ellipses de ce type ont été effectués). Les valeurs sont indiquées en dizième de millimètre.

a : dispersion de la consonne [t] dans les contextes vocaliques [a], [i], [y], [u].
b : dispersion des consonnes [p], [t] et [k] dans le contexte vocalique [a]
c : dispersion de la voyelle [i] dans les entourages consonnantiques [f], [s], [ʃ]
d : dispersion des voyelles [a], [i], [y], [u] dans le contexte [s]

o o
o

Cette étude, qui n'est qu'un premier pas dans l'utilisation de méthodes statistiques pour l'interprétation de données relatives au mouvement de la mâchoire, permet de souligner la différence totale des comportements vocalique et consonantique de cet organe articulatoire chez notre locuteur. La disjonction des consonnes dans le même contexte vocalique et leur cohésion respective dans des entourages vocaliques différents prouve l'importance de la notion de cible articulatoire consonantique, tandis qu'inversement, la disjonction des voyelles dans des entourages consonantiques différents et leur cohésion respective dans des contextes consonantiques identiques démontrent leur faible résistance à la coarticulation.

BIBLIOGRAPHIE

ABBS (J.) & GILBERT (B.), 1973, A strain gage transduction system for lip and jaw motion in two dimensions : design criteria and calibration data. - Journ. Speech Hearing Res. 16, pp. 248-256.

ABBS (J.), NETSELL (R.) & HIXON (T.J.), 1972, Variations in mandibular displacement velocity, and acceleration as a function of phonetic context. - Paper presented to the A.S.A., Denver., J.A.S.A. 51, p. 89 (A).

HIXON (T.J.), 1971, An electromagnetic method for transducing jaw movements during speech. - J.A.S.A. 49, pp. 603-605.

KANEKO (N.), 1957, Specific X-ray observations on the movements on the lower jaw, lips and tongue in pronounciation. - Study of Sounds 8, pp. 1-18.

KIM (B.) & FUJISAKI (H.), 1974, Measurements of mandibular control in vowels and its relevance to the articulatory description of vowel systems of Korean and Japanese. - Speech Communication Seminar, Stockholm. Preprint, Session 2 / Ann. Bull. Res. Inst. Logopedics Phoniatrics, Tokyo 8, pp.59-68.

JANKELSON (B.), SWAIN (C.W.), CRANE (P.F.) & RADKE (J.C.), 1975, Kinesiometric instrumentation : a new technology. - J.A.D.A. 90, pp. 834-840.

LEJOYEUX (J.), 1967, Prothèse complète : diagnostic, traitement. - Maloine, Paris.

MAC CUTCHEON (H.J.), FLETCHER (S.G.) & HASEGAWA (A.), 1977, Video-scanning system for measurement of lip and jaw motion. - J.A.S.A. 61(4), pp. 1051-1055.

MAC NEILAGE (P.) & LADEFOGED (P.), 1976, The production of speech and language. - In : CARTERETTE & FRIEDMAN, Handbook of Perception, V. 7, pp. 75-120.

MAC NEILAGE (P.), LADEFOGED (P.) & SAWASHIMA (M.), 1979, Report : Speech Production. - Proc. 9 ICPS 3, pp. 29-40.

MULLER (E.M.) & ABBS (J.H.), 1979, Strain gauge transduction of lip and jaw motion in the midsagittal plane : refinement of a prototype system. - J.A.S.A. 65 (2), pp. 481-486.

POSSELT (U.), 1969, Physiologie de l'occlusion et réhabilitation (traduction de FISSORE, Y.J.), Prélat, Paris.

WIOLAND (F.), 1971, Les mouvements du maxillaire dans la chaine parlée. - Trav. Inst. Phon. Strasbourg 3, pp. 57-119.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

VERS UN ESPACE DE LA LABIALITE VOCALIQUE : APPORT DES NUEES DYNAMIQUES ET DE L'ANALYSE DES CORRESPONDANCES.

Christian ABRY & Louis-Jean BOË Institut de Phonétique. de Grenoble

.**RESUME**

L'arrondissement est-il une dimension à part entière des systèmes vocaliques ? ...

L'essai présent ne se veut qu'une modeste tentative pour répondre à quelques questions concernant la puissance de représentation et de structuration des paramètres des lèvres pour les voyelles du français (i, e, y, ø, a; avec ou sans coarticulation de protrusion, soit avec les consonnes s, z et ʃ, ʒ).

Nous avons utilisé deux techniques d'analyse des données : les Nuées dynamiques de DIDAY et l'Analyse des correspondances de BENZECRI.

Nos résultats montrent que :

- l'espace de la labialité vocalique du français est essentiellement bipolaire (y, ø) vs. (i, e, a); ce qui revient à dire que la voyelle /a/ n'est pas à elle seule un pôle de classification différent des voyelles non arrondies (i,e);
- l'écartement aux commissures, la protrusion et l'aire étant tous trois honorablement significatifs pour représenter nos données, nous adopterons décidément une approche de cet espace de la labialité vocalique qui ne soit pas unidimensionnelle;ceci dans le but de préserver son interprétabilité pour d'ultérieures validations acoustique et perceptive.

. ABSTRACT

Is rounding a true dimension of vocalic systems ? ...

Our present attempt - a rather modest one - is to answer only some questions concerning representation and structuration powers of lip-parameters for French vowels (i, e, y, ø, a), with and without protrusion coarticulation effects (namely s, z and ʃ, ʒ).

Two data analysis techniques have been used, i.e BENZECRI's Analyse des correspondances factor analysis, and DIDAY's Nuées dynamiques clustering algorithm.

Our results show that :

- French vocalic lip-space is essentially 2-partitionnable (y, ø) vs. (i, e, a) : i.e. /a/ cannot form a third pole of classification different from unrounded vowels (i, e);
- lip-width, lip-protrusion and area being fairly well significant for representing our data, we have finally opted for a non unidimensioning of our vowel lip-space, thus procuring interpretability for subsequent acoustic and perceptual validations.

VERS UN ESPACE DE LA LABIALITE VOCALIQUE : APPORT DES NUEES DYNAMIQUES ET DE L'ANALYSE DES CORRESPONDANCES.

Christian ABRY & Louis-Jean BOË - Institut de Phonétique de Grenoble

0 - INTRODUCTION

Cet article ne voudrait être qu'une mise au point sur une petite formule plus ou moins imprudente que nous avions donnée dans la préface de Labialité et Phonétique. En effet, proposer "un triangle vocalique labial" (ABRY, 1980, p. 5) sur le modèle de la classique présentation des voyelles, soit en quelque sorte un triangle sur un triangle, peut paraître à la fois raisonnable et déraisonnable. Evident, lorsqu'on examine les données déjà anciennes de Vicky FROMKIN (1964) (figure 1); discutable, si l'on prend en compte une dénégation quasi complète du rôle de la labialité comme dimension vocalique que vient nous apporter la lecture d'un récent article de NEAREY (1980, p. 232-239)*.

Pour notre part nous ne ferons ici qu'avancer quelques propositions de représentation et de structuration de l'espace vocalique labial, tout en gardant à l'esprit deux ou trois a priori sur ce type d'entreprise.

I - RAPPEL D'A PRIORI

a) Sur les classes :

Les conditions d'une recherche de classes phonémiques ont été exposées ailleurs (ABRY & BOË, 1980). Rappelons simplement que, dans un premier temps, nous n'envisageons la possibilité de classes discriminantes que pour les gestes vocaliques extrêmes (type i, a, u), pourvus ou non d'une distinction phonologique (dans notre cas i/y), et dans certains contextes syntagmatiques bien définis, dits

*. Selon lui, "... 'rounding' must be regarded as an unsatisfactorily interpreted feature at the phonetic level" (p. 239). Mais ni les arguments "empiriques et théoriques", ni les inférences "historiques" utilisés ne nous semblent probants; si ce n'est pour constater une fois de plus la variabilité interindividuelle du trait d'arrondissement, sa non-orthogonalité avec les autres et l'équivocité manoeuvre articulatoire - résultat acoustique; d'où peut-être une plus forte dépendance diachronique et synchronique de ce trait dans les processus phonologiques.

optimaux (dans notre cas avec [s]). C'est à partir de ces contextes que nous testons la résistivité des classes pour des conditions que nous appellerons pessimales (avec [ʃ]).

b) Sur les caractères :

Nous adopterons une approche pluri-indicielle . En effet, la réflexion que nous avons pu mener sur la simultanéité des propriétés et des indices du trait de voisement (ABRY & BOË, 1979), notamment à partir des travaux de SERNICLAES (1973), SAWUSH & PISONI (1974), SUMMERFIELD & HAGGARD (1974), MASSARO & COHEN (1976)*, nous fait penser que, quelles que soient les techniques statistiques d'analyse des données utilisées, il est illusoire et même préjudiciable de réduire la stratégie d'étude des corrélats d'une opposition linguistique à la recherche d'un discriminant unique. Dans notre cas, nous avons pu ainsi proposer d'intégrer les 3 degrés de liberté, classiquement reconnus pour les lèvres, dans une fonction discriminante C/S, soit le rapport de la protrusion à l'aire, c'est-à-dire C/A.B au facteur K1 près (K1 = S/A.B; A : écartement aux commissures, B : aperture intérolabiale; ABRY, 1980; ABRY & BOË, 1980). Cette fonction discriminante, valable pour les voyelles, peut même être parfois moins efficace que ne l'est le paramètre de l'étirement (A) tout seul. Un tel choix ne se justifie par conséquent que si l'on considère la recherche des corrélats comme n'étant que l'amont des différentes validations perceptives (soit ici les propriétés visuelles et indices phonétiques pour la labiolecture).

c) Sur les locuteurs :

Sans reprendre ici la discussion sur la systématisation des différences entre nos locuteurs (BOË, ABRY & CORSI, 1980), nous n'utiliserons dans ce travail les résultats de notre typologie que pour justifier le choix du meilleur et du pire. Rappelons que DL est ainsi considérée comme "hyperdifférenciatrice", puisque pratiquant l'extension du contraste de protrusion, phonologique pour les voyelles, aux consonnes sibilantes; tandis que PC est lui un "hypodifférenciateur", qui maintient à grand peine ses oppositions phonémiques d'arrondissement.

---

*. Ceci pour ne rester qu'au niveau segmental; pour l'intonation on peut se référer à la démarche "pluriparamétrique" de l'équipe de l'Institut de Phonétique d'Aix-en-Provence.

II - METHODES

Disposant, dans le cas de notre corpus (DESCOUT & al., 1979) de classes a priori, nous en avons sélectionné trois : /i, y, a/ représentant pour le français les extrêmes de l'opposition de labialité (i/y) avec leur pôle de neutralisation /a/ (l'opposition a/ɒ n'étant pas systématique chez nos locuteurs). Ces classes sont décrites par un nombre respectable de paramètres, 12 en l'occurrence (Labialité et Phonétique, pp. 108-109).

En bonne méthode, un classement automatique permettrait déjà de vérifier, dans cet espace d'origine, la validité des classes a priori. Puis, une analyse de type factoriel pourrait proposer, après interprétation, une réduction de l'espace d'origine. C'est sur ce dernier espace que serait testée enfin la conservation ou l'amélioration des propriétés classificatoires de l'espace d'origine.

Dans notre cas, l'analyse des correspondances que nous avons menée antérieurement (GRAILLOT & al., 1980), comme l'analyse PARAFAC de Wendy LINKER (1979), nous ramènent en fait aux trois degrés de liberté des lèvres (paramètres articulatoires de LADEFOGED, 1979); les seules différences portant sur la composition de ces paramètres (A et C/B au lieu de S et A/B ou C dans notre cas). Nous avons donc repliqué sous une forme réduite notre analyse de l'ensemble consonne-voyelle sur les voyelles seules. Ce nouveau passage s'imposait,d'une part pour tester sur les voyelles l'interprétabilité de certains paramètres que nous avait sélectionné l'analyse factorielle; d'autre part, il était de première nécessité de tester un espace vocalique sans oublier une voyelle aussi "cardinale" que le /a/.

III - L'ANALYSE DES CORRESPONDANCES

Des cinq paramètres retenus pour cette analyse, A, B et C se veulent un retour, ainsi que nous venons de le dire, aux trois degrés de liberté articulatoire . S et C ont d'autre part des mérites acoustiques. Quant à A - B, c'est là une nouvelle tentative. En effet dans l'espace A,B où nous avions vu apparaître le triangle vocalique tel que le laisse supposer FROMKIN (1964), le fait que la géométrie articulatoire de la fente labiale ne présente jamais une aperture supérieure à l'étirement peut laisser supposer qu'une translation sur A de la valeur de B doit récupérer une partie de l'espace non utilisé à cause de cette contrainte.

. Les axes

Nous donnons (tableau I) pour les deux locuteurs (DL et PC) les deux premières valeurs propres (λ_i) et le pourcentage de la variance qu'elles expliquent ($\tau_i = \lambda_i/\Sigma\lambda$). Le premier facteur rend compte à lui tout seul de 94% à 95% de cette variance. Par rapport à l'analyse (GRAILLOT & al., 1980) que nous avions fait tourner sur l'ensemble consonne - voyelle (sans le a), un seul axe monopolise la représentation des données avec les 5 paramètres retenus. De la même façon les valeurs propres, comme nous l'avions déjà laissé entendre, sont plus élevées pour les voyelles seules, comme c'est le cas habituellement dans une dichotomie.

. Les paramètres

Nous retrouvons (figure 2) dans la projection de ceux-ci, la même disposition triangulaire dont les sommets sont S, A (ou A - B) et C. En examinant de plus près les contributions absolues et relatives (tableaux II et III), nous pouvons avancer les interprétations suivantes :

- pour le premier axe : le seul paramètre que nous ayons conservé et qui n'a vraiment pas de "chance" est B. Les autres ont tous des valeurs honorables : leur variance étant relativement bien expliquée par l'axe 1 et C étant celui qui de tous contribue le plus à la construction de cet axe.
- pour le deuxième axe : C et A - B arrivent ex aequo en contributions absolues, l'axe n'expliquant la variance que d'un seul paramètre, le "délaissé" de l'axe 1, B.

. Les classes

Les ellipses de dispersion (à 90 %) nous montrent le même pouvoir discriminateur de l'axe 1. Cet axe conserve pratiquement à lui tout seul les propriétés classificatoires de l'espace des paramètres retenus. Il reflète totalement :

a) l'opposition [$\pm$ rond] (i, e /y, ø) quel que soit le contexte sibilant;
b) l'impossibilité de distinguer labialement le /a/ du /i/, c'est-à-dire de le caractériser autrement qu'une voyelle non arrondie (figure 3).

Cette analyse :

- met en évidence le caractère bipolaire de la labialité vocalique du français;
- elle confirme d'autre part que la quasi équi-interprétabilité de paramètres aussi différents que A, A-B, C, S ne peut trouver de prolongements que dans le cadre d'une approche pluri-indicielle.

IV - LES NUEES DYNAMIQUES

Le retour à un espace, même réduit, mais avec des paramètres d'origine, s'impose pour qui a le souci premier de l'interprétabilité de ses dimensions descriptives (interprétabilité physiologique, acoustique et perceptive pour le phonéticien). Ce retour aux sources peut être validé par la vérification du maintien dans cet espace des classes mises en évidence à l'origine et conservées dans l'analyse factorielle. La méthode des "nuées dynamiques" fait partie de ces algorithmes de classification qui fonctionnent par agrégation autour de centre mobiles (DIDAY, 1972). Elle permet de dégager des groupements stables (formes fortes) et de limiter ainsi l'influence du tirage au sort des premiers centres. Très souple d'utilisation, cette technique peut être un préalable à une analyse factorielle ou une vérification de celle-ci.

C'est sous cette deuxième forme que nous l'avons menée avec les quatre paramètres les plus représentatifs de l'analyse factorielle précédente soit : A, A-B, C et S.

Les résultats pour nos locuteurs extrêmes sont les suivants :

- DL aboutit à deux formes fortes quel que soit le nombre de classes demandé a priori (jusqu'à 5). Ces deux classes (y, ø) et (i, e, a) ne peuvent être phonologiquement homogènes que du point de vue du trait d'arrondissement.
- PC nous donne, comme nous pouvions le prévoir, plus de fil à retordre dans nos essais classificatoires. Nos demandes de classes (2, 3, 4,etc.) sont toujours acceptées et nous ne rencontrons de classes phonologiquement homogènes qu'à partir de 3. Ce n'est en effet qu'à ce stade que la classe (y, ø) se trouve "épurée" de quatre voyelles (i, e); elle se maintient telle quelle dans la suite des partitions. Les deux autres classes sont à dominante (i, e) ou (a). Ce n'est donc, pour PC, qu'en retirant ces quatre voyelles litigieuses que nous retombons sur les deux classes homogènes de DL (y, ø) / (i, e, a) et ce dès la tripartition.

V - CONCLUSIONS

Nos résultats montrent sans aucun doute qu'au niveau des classes "naturelles" de l'espace vocalique - et ceci au moins pour les voyelles du français - la cardinale /a/ ne saurait être considérée comme un pôle à part entière. En fait

l'image reçue que nous avons de cette voyelle au niveau des lèvres ne doit pas nous faire oublier une autre image familière, elle, au praticien de la "labio-lecture" : que serait en effet, en face de /a/, un /i/ sans les dents*. D'autre part, le pouvoir de représentation et de structuration de cet espace revient, sans aucun doute aussi, à plusieurs paramètres. Certes, il n'en demeure pas moins qu'une représentation à plus de deux dimensions, quels que soient les mérites des paramètres retenus - mérites que nous avons su vanter par ailleurs (ABRY & BOË, 1980) - est toujours ressentie comme plus ou moins commode par le phonéticien qui veut les utiliser dans ses taxinomies; et c'est sans doute là le revers de la médaille de toute approche pluri-indicielle. Disons que ce n'est pas par un simple hasard de notre esprit - qu'il soit ou non phonéticien - si les débats autour de l'extension des espaces vocaliques, mêmes s'ils ne sont pas toujours aussi réducteurs que peut l'être une approche à laquelle nous avons fait d'emblée allusion (NEAREY, 1980), évoluent plus ou moins sur la réserve et les positions d'un LINDBLOM (9e Congrès Int. des Sci. Phon., Vol. III, p. 190) : elles ont au moins le mérite d'être d'une prudence qui sait s'avouer seulement provisoire.

REMERCIEMENTS

Un grand merci à Patrick GRAILLOT pour nous avoir initié à l'analyse des correspondances et à Yves GRENIER qui nous a transmis un programme des Nuées Dynamiques parfaitement transportable.

Ces deux types d'analyse ont été adaptés sur LSI 11 par Leila LTEIF et Jacques RAYMOND au cours d'un stage de l'ENSIMAG à l'Institut de Phonétique de Grenoble.

REFERENCES

ABRY, C., 1980, Préface et profils. - In Labialité et Phonétique, pp. 1-8.

ABRY, C & BOË, L.J., 1979, Le trait de voisement, ses indices et leurs propriétés. Implications pour les détections automatique et perceptive d'une analyse phonologique au-delà du trait. - In Recherches sur la Prosodie du Français, pp. 57-93.

ABRY, C. & BOË, L.J., 1980, A la recherche de corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français. - In Labialité et Phonétique, pp. 217-237.

BENZECRI, J.P., 1973, L'analyse des données, tome 1 : La taxinomie, tome 2 : L'analyse des correspondances. - Dunod, Paris (2e Ed., 1976).

*. C'est ici que reparaît le tendon d'Achille plusieurs fois évoqué de nos données, le paramètre M (Labialité et Phonétique, p. 108), qui ne nous a pas permis de relier quantitativement la structure péri-orale au maxillaire.

DESCOUT, R., BOË, L.J. & ABRY, C., 1980, Labialité vocalique et labialité consonantique en français. Un jeu des lèvres au féminin : l'idiolecte D.L. - In Labialité et Phonétique, pp. 111-126.

DIDAY, E., 1971, La méthode des nuées dynamiques. Revue de Statistique Appliquée, 19/2, pp. 19-34.

FROMKIN, V., 1964, Lips positions in American English vowels. - Language & Speech 7, pp. 215-225.

GENTIL, M., 1980, Labialité en français. Etude phonétique et aspects physiologiques. Géométrie, cinématique, électromyographie et idiolectologie. - Thèse de 3e Cycle Phonétique - Université des Langues et Lettres de Grenoble.

GRAILLOT, P., BOË, L.J., GENTIL, M. & ABRY, C., 1980, Analyse des correspondances de paramètres descriptifs du jeu des lèvres en français. - In Labialité et Phonétique, pp. 127-143.

INSTITUT DE PHONETIQUE DE GRENOBLE, 1979, Recherches sur la prosodie du français.- Publications de l'Université des Langues et Lettres de Grenoble.

INSTITUT DE PHONETIQUE DE GRENOBLE, 1980, Labialité et Phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales. - Publications de l'Université des Langues et Lettres de Grenoble.

LADEFOGED, P., 1979, Articulatory parameters. - 9th Int. Congr. of Phonetic. Sci. 1, pp. 41-47.

LEBART, L., MORINEAU, A, FENELON, J.-P., 1979, Traitement des données statistiques. Méthodes et programmes. - Dunod, Paris.

LINKER, W., 1979, A cross linguistic study of lip position in vowels. - 9th Int. Congr. of Phonetic Sci. 1, p. 201.

LTEIF, L. & RAYMOND, J., 1980, Adaptation d'un programme d'analyse des correspondances à l'étude des caractéristiques individuelles. - Projet de stage, ENSIMAG, Grenoble.

MASSARO, D.W. & COHEN, M.M., 1976, Fundamental frequency and voice onset time cues to the /zi/ - /si/ distinction. - J.A.S.A. 60, pp. 704-717.

NEAREY, T., 1980, On the physical interpretation of vowel quality : cine-fluorographic and acoustic evidence. - Journal of Phonetics 8, pp. 213-241.

SAWUSH, J.R. & PISONI, D.B., 1974, On the identification of place and voicing features in synthetic stops consonants. - Journal of Phonetics 2, pp. 181-194.

SERNICLAES, W., 1973, La simultanéité des indices dans la perception du voisement des occlusives. - 4° J.E.P. - Groupe Communication Parlée du GALF, pp. 359-369.

SUMMERFIELD, A.Q. & HAGGARD, M.P., 1974, Perceptual processing of multiple cues and context : effects of following vowel upon stop consonant voicing. - Journal of Phonetics 2, pp. 279-285.

	DL	PC
λ_1	0,13	$0,79.10^{-1}$
τ_1	95,1 %	94,0 %
λ_2	$0,67.10^{-2}$	$0,48.10^{-2}$
τ_2	4,7 %	5,7 %

Tableau I - Les deux premières valeurs propres λ_i et le pourcentage de la variance expliqué τ_i pour les deux locuteurs DL et PC.

	axe 1	axe 2	axe 1	axe 2
A	3548	201	853	124
B	16	96	1	28
C	15135	562	4938	214
S	4624	15	2843	28
A-B	5552	561	954	190
	Loc. DL		Loc. PC	

Tableau II - Contributions absolues des paramètres aux deux premiers axes factoriels pour les deux locuteurs DL et PC.

	axe 1	axe 2	axe 1	axe 2
A	94	5	87	13
B	11	61	1	58
C	96	4	96	4
S	99	0	99	1
A-B	91	9	83	17
	Loc. DL		Loc. PC	

Tableau III - Contributions relatives, en %, des paramètres aux deux premiers axes factoriels pour les deux locuteurs DL et PC.

Victoria Fromkin

Lip Positions in American English Vowels

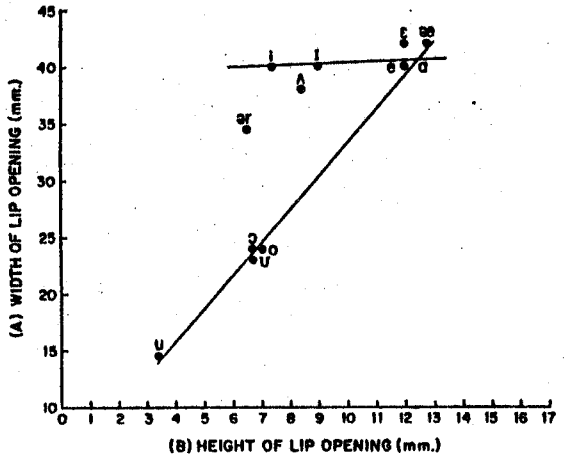


Fig. 1. Relation between mean height and width for the principal subject.

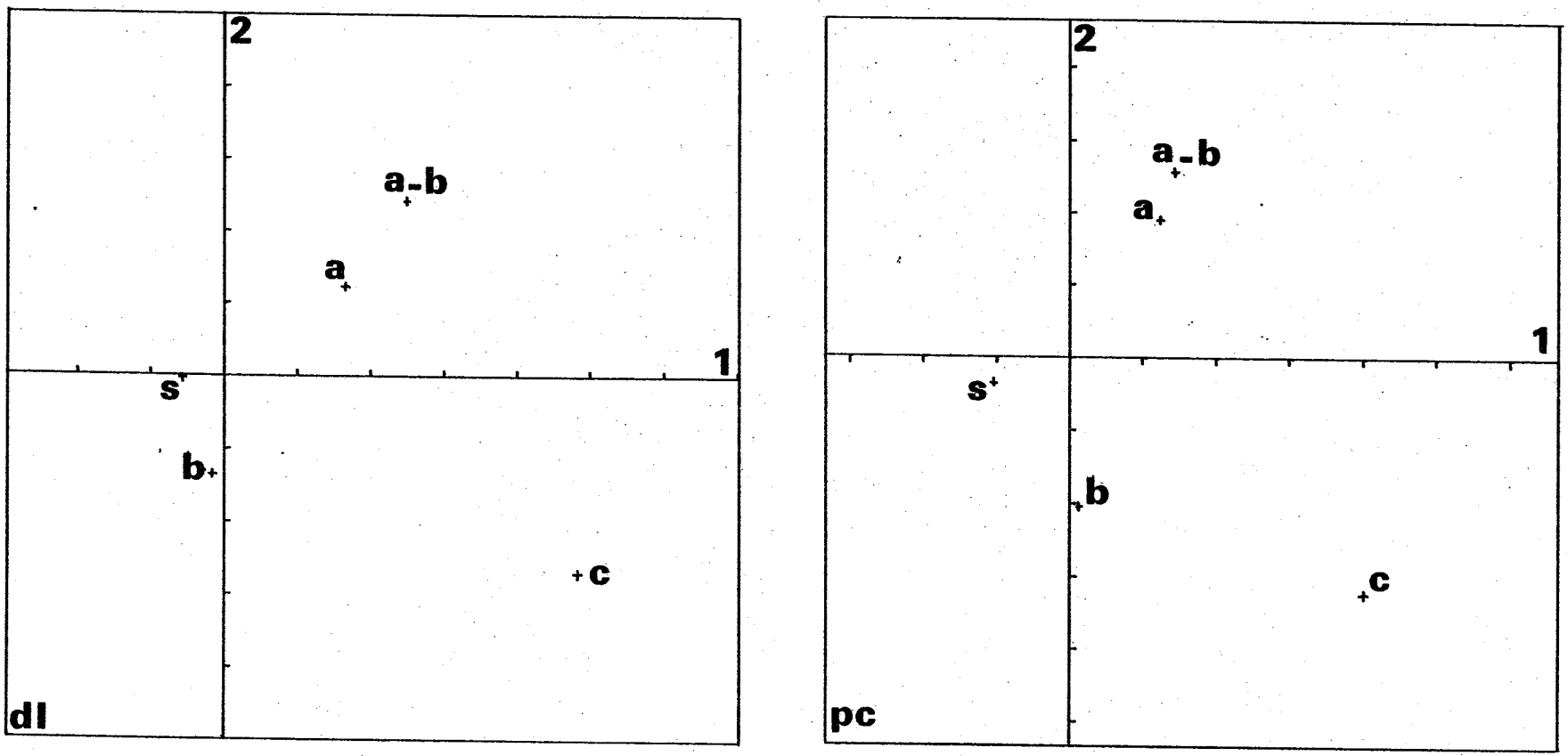


Figure 2 - Projection des paramètres dans le plan des deux premiers facteurs pour les deux locuteurs (DL et PC).

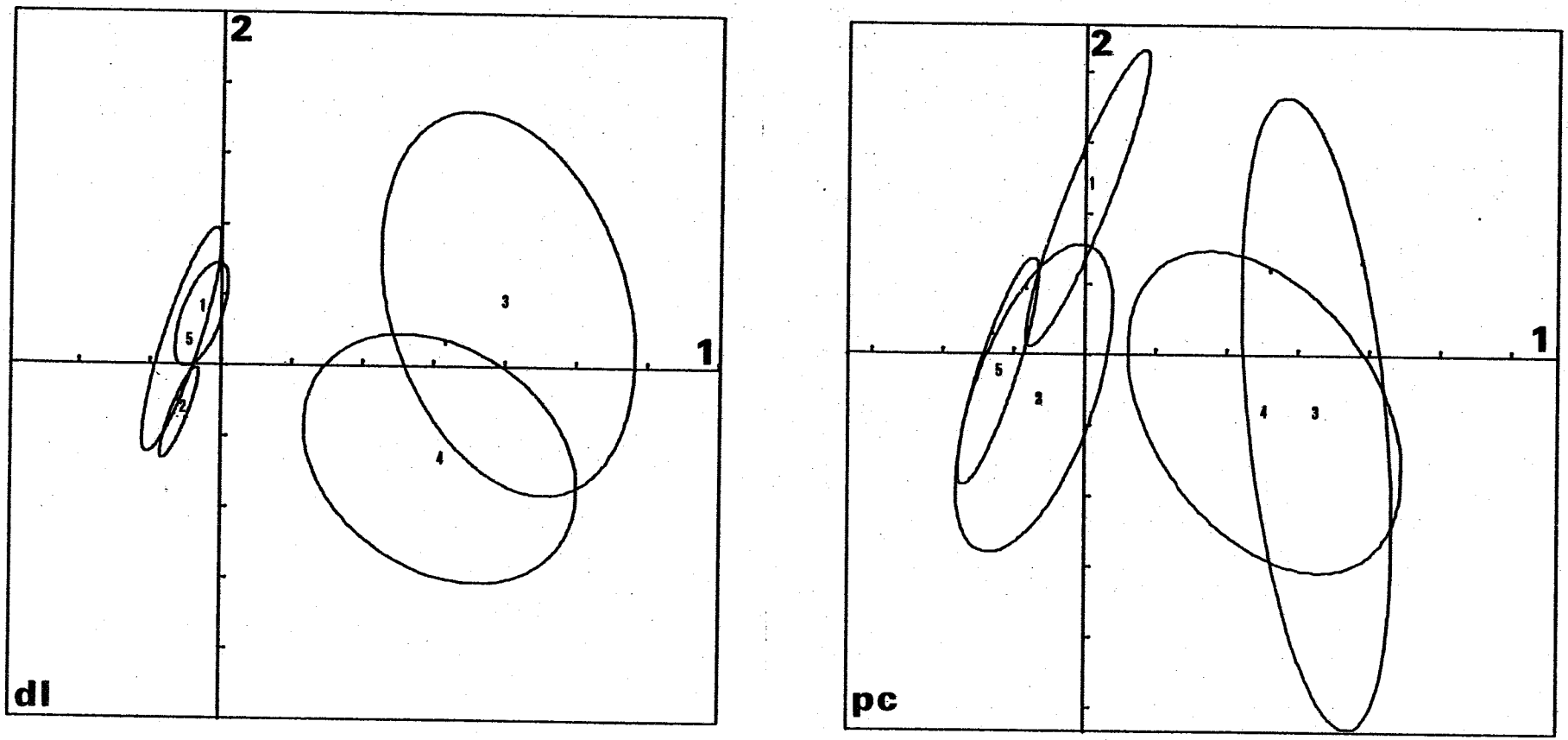


Figure 3 - Projection des réalisations vocaliques et consonantiques dans le plan des deux premiers facteurs (ellipses de dispersion à 90% de la variance) pour les deux locuteurs (DL et PC).

1. Voyelles i, e dans le contexte s, z
2. " i, e " ʃ, ʒ
3. " y, ø " s, z
4. voyelles y, ø dans le contexte ʃ, ʒ
5. " a " s, z, ʃ, ʒ

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

EXEMPLE DE TRAITEMENT DE L'OPPOSITION DE "SONORITE" PAR DES ENFANTS DE COURS PREPARATOIRE - UTILISATION DE LA METHODE ELECTROGLOTTOGRAPHIQUE[xx]

Jean-Pierre GOUDAILLIER

Laboratoire de Phonétique
U.E.R. de Linguistique
Générale et Appliquée
UNIVERSITE RENE DESCARTES
PARIS

.RESUME

Il s'agit de l'étude électroglottographique (électroglottographe FRØKJAER-JENSEN type 830) des articulations occlusives phonologiquement sourdes / p t k / *et sonores* / b d g / *prononcées en position initiale et intervocalique par 11 enfants de 6-7 ans d'une école du Nord de la France (Vervins, département de l'Aisne). Les occlusives* / b d g / *peuvent être prononcées sonores ou sourdes. Lorsque les vibrations laryngiennes ne parviennent plus à opposer* / p t k / *et* / b d g / (= [b̥ d̥ g̥]), *ce qui est relativement fréquent dans cette région, non seulement le VOT permet encore de distinguer les deux séries d'occlusives (valeurs de VOT plus élevées pour* / p t k /), *mais la durée de la phase d'occlusion et la durée totale des articulations établissent aussi une distinction. Ceci permet de conclure que les enfants analysés disposent de deux types de réalisations foncièrement différents de l'opposition* / p t k / - / b d g / : *une distinction due aux vibrations des cordes vocales (la seule mise en oeuvre dans des régions méridionales (Aix-en-Provence, etc...)); une autre faisant intervenir des différences de FORCE ARTICULATOIRE, dont le VOT est un des indices. Lorsque l'on connaît l'importance du VOT pour la perception, on apprécie au plus juste l'apport d'une telle étude en ce qui concerne les productions enfantines.*

. ABSTRACT

Thanks to an electroglottograph (FRØKJAER-JENSEN type 830), the stops, which in French are generally considered from a phonological point of view as voiceless (= / p t k /) and voiced (= / b d g /), are analysed, as they have been pronounced by eleven 6-7 year old children of a school of the North of France (Vervins, Aisne). The study of the different pronounciations noted reveals that they can be produced with or without laryngeal vibrations. When these vibrations do not oppose / p t k / and / b d g / (= [b̥ d̥ g̥]) (such allophones are relatively frequent in this region) any more, not only does the VOT establish the distinction between the two series (VOT values are more important for / p t k /), but also the duration of the closure period and the total duration of the articulations contribute to this distinction. Consequently it may be concluded that the analysed children have at their disposal two fundamentally different types of realizations of the opposition / p t k / - / b d g / : on the one hand a contrast attributable to the vibrations of the vocal cords (this is the only type of contrast noted in the Southern area of France (Aix-en-Provence, etc...)), on the other hand a contrast based upon differences of FORCE OF ARTICULATION, of which the VOT is one of the cues. Considering how important this cue is for perception, one can appreciate the contribution of such a study to the knowledge of children speech.

INTRODUCTION - CADRE DE L'ETUDE

Pour ce qui est des occlusives françaises, on analyse traditionnellement les faits en termes d'opposition de "sonorité". Encore s'agit-il de déterminer au plus juste ce que l'on recouvre par ce terme : s'agit-il d'une opposition entre présence et absence de vibrations des cordes vocales pour opposer / p t k / et / b d g / ? Cette opposition se matérialise-t-elle de la même manière en toutes positions de la chaîne parlée ? Des variations d'ordre régional ou imputables à l'appartenance sociale, aux activités socio-professionnelles des locuteurs peuvent-elles être relevées en matière de voisement des articulations occlusives ? Du point de vue génétique (dynamique), qu'en est-il au niveau des enfants, aux divers stades de l'acquisition du langage ? Autant de questions, auxquelles il conviendrait de répondre, si l'on désire apprécier au plus juste le fonctionnement de l'opposition de "sonorité" (au sens large du terme) en français et conférer de ce fait à une telle appréciation, tout comme à celle de tout phénomène linguistique, sa véritable dimension socio-linguistique (en l'occurrence, ici, socio-phonétique), car il s'agit de ne pas exclure la possibilité pour chaque locuteur d'avoir "un système de réalisation individué différent de celui des autres locuteurs" (*ABRY*,

BOË, 1980, p. 133).

Associé à un groupe de recherche de l'*I.N.R.P.* depuis plusieurs années[1], j'ai récemment procédé à l'examen expérimental/instrumental, entre autres, du voisement chez des enfants de Cours Préparatoire (*GOUDAILLIER, 1981b*). Les études de ce groupe de recherche ont pour finalité la détermination à court et moyen terme des systèmes phonologiques majoritairement employés par les enfants à l'entrée de l'Ecole primaire (systèmes vocaliques et consonantiques en place ou non ?), des variations régionales et sociales (systèmes sous-différenciés et sur-différenciés), ce en vue de l'élaboration, à plus long terme, d'outils pédagogiques.

Si l'on considère des enfants scolarisés dans le Midi de la France et dans le Nord (*GOUDAILLIER, 1981a* et *a'*)[2] au niveau du Cours Préparatoire, on est amené à constater deux traitements foncièrement différents de l'opposition de l'opposition de "sonorité", ce que résume parfaitement un seul chiffre : le pourcentage global d'articulations de / b d g v z ʒ / désonorisés partiellement ou complètement, toutes positions dans la chaîne parlée confondues, pour *10* enfants de chacune des deux régions; à Aix (*10* enfants : *633* articulations au total), seulement *4,4%* des articulations ne sont pas totalement voisées; à Vervins (*10* enfants : *600* articulations au total), ce chiffre est de *26,2%*[3], ce qui est beaucoup pour une région non véritablement répertoriée comme étant à désonorisation importante (telle l'Alsace, etc...).

Etant donné les résultats constatés pour Vervins lors de l'analyse d'une enquête effectuée en Octobre pour l'*I.N.R.P.*, j'ai effectué en Mai 1980 une autre enquête dans la même école avec les mêmes enfants, afin de déterminer, entre autres[4] :

1) dans quels pourcentages les articulations phonologiquement sonores sont dévoisées en fonction des positions dans la chaîne parlée prises en considération (évolution entre Octobre 1979 et Mai 1980 ?);
2) quelles sont les valeurs de VOT pour / b d g /, lorsque les articulations sont voisées, dévoisées partiellement, complètement sourdes (comment s'établit alors l'opposition avec la série / p t k /?);
3) quels sont les paramètres permettant de différencier les occlusives des deux séries.

Seuls les résultats concernant les positions initiale absolue (mot non précédé d'un article), initiale (présence d'un article devant le mot) et intervocalique pour les occlusives / p t k / et / b d g / sont présentés, car il importait dans un premier temps de centrer l'étude sur l'indice VOT; or, cet indice ne peut pas être mesuré pour les occlusives en position finale.

METHODE EXPERIMENTALE

- sur le terrain : enregistrement magnétique (REVOX A77, bande BASF LPR35, vitesse défilement : 19cm./sec.) du signal délivré par un microphone (voie 1) et de celui fourni par un électroglottographe FRØKJAER-JENSEN type 830) (cf., entre autres, *FRØKJAER-JENSEN, 1969*) (voie 2);
- en laboratoire (Laboratoire de Phonétique, U.E.R. de Linguistique Générale et Appliquée / Université René Descartes / Paris) : visualisation sur écran cathodique après stockage dans une mémoire numérique[5] des deux voies enregistrées (échelles choisies : *5ms*, *10ms* ou *20ms* par division d'1cm sur l'écran). Les mesures ont été faites sur l'écran de l'oscilloscope[6]. Après visualisation, le contenu de la mémoire numérique peut donner lieu à un enregistrement graphique (unité de sortie GOULD-ALLCO 4002 vers Oscillomink Type E de SIEMENS)[7].

RESULTATS (*11* enfants - *7* garçons et *4* filles - même milieu socio-professionnel[8] - dates de naissance comprises entre Janvier 1973 et Décembre 1973 - Enquête de Mai 1980)

A partir de l'analyse effectuée, on peut relever trois grands types de réalisations des phonèmes phonologiquement sonores / b d g / :

a) variantes sonores [b d g] : type [bɔs], [dexidǫ] (enfant O.O., planche 1), [buʒi] [bɛnᵂaʁ] (enfant D.H., planche 1), [bik$^{(h)}$ɛt$^{(h)}$] (enfant O.S., planche 2) (*l'ensemble de l'articulation est voisé : VOT négatif*);

b) variantes désonorisées [b̥ d̥ g̥] (*le début de l'articulation est voisé, absence de voisement au moment de l'explosion : présence d'un VOT négatif (= prévoisement)*[9] *et d'un VOT positif (décalage dans la mise en place des vibrations laryngiennes*[9]);

c) variantes sourdes [p̥ t̥ k̥] : type [t̥ʷɑ] (enfant D.H., planche 1), [gaᵉᵘsɔ̃], [gat$^{(h)}$ǫ], [p̥iʃ] (enfant O.S., planche 2) (*l'ensemble de l'articulation ne comporte pas de voisement : VOT positif*).

Le Tableau 1 récapitule l'ensemble des occurrences des trois types de variantes relevées pour les phonèmes / b /, / d / et / g / en positions initiale absolue, initiale et intervocalique. Quel que soit la position prise en considération, les pourcentages de dévoisement sont plus importants, lorsque l'on passe de / b / à / d / et à / g / : dans cet ordre, les articulations voisées représentent *79%*, *68%* et *31%* à l'initiale absolue, *90%*, *71%* et *47%* à l'initiale, *95%*, *82%* et *66%* à l'intervocalique. Quel que soit le phonème considéré, c'est en position initiale absolue que l'on trouve le plus de cas de désonorisation (ceci est confirmé par les chiffres du Tableau 2, qui comprend les données tous points d'articulation confondus).

Le Tableau 3 indique que la situation n'a guère évolué entre Octobre 1979 et Mai 1980 : ceci est particulièrement probant à l'initiale absolue et à l'initiale, puisque les pourcentages sont du même ordre de grandeur (positions initiale absolue et initiale confondues, les pourcentages sont pratiquement les mêmes : *26,8%* (*70/261*) en Octobre 1979 vs. *25,3%* (*116/459*) en Mai 1980. C'est une des constatations qui permettent d'établir que l'on est en présence d'un trait régional et non pas d'un stade d'acquisition de l'opposition (ceci est d'ailleurs confirmé par la langue des adultes de la région; pour ne prendre qu'un exemple, la maîtresse du C.P., duquel proviennent les enfants soumis à l'enquête, désonorise elle aussi dans des proportions importantes les consonnes "phonologiquement" sonores).

Depuis le début des années 60, le VOT est considéré comme un indice important pour l'analyse des consonnes occlusives (cf., entre autres, *LISKER/ABRAMSON, 1964*). Une utilisation nous en a été proposée l'an dernier aux XI° J.E.P. (cf. *BARTH, BEN FADHEL* et *MAJO, 1980*) (on se reportera aussi à la note[9] pour ce qui est des études menées dans ce domaine par les équipes de Grenoble et de Bruxelles).

Dans les cas où / p t k / et / b d g / sont différenciés par une opposition de VOT négatif (/ b d g /) à VOT positif (/ p t k /), c'est cet indice, qui peut être à lui seul à la base de l'opposition (importance sur le plan perceptuel de la distinction prévoisement / délai d'établissement du voisement). Les histogrammes (1), (3) et (4) représentent trois cas particulièrement nets de distinction due à une différence VOT positif / VOT négatif ((1) = enfant J.T. : [b] et [p] à l'initiale absolue; (3) = enfant V.M. : [b] et [p] en position initiale; (4) = enfant O.O. : [b d g] et [p t k] à l'initiale).

Tableau 1 : *Répartition des réalisations de / b,d,g / (Mai 1980)*

	/ b /			/ d /			/ g /			
	[b]	[b (A)]	[b ^]	[d]	[d (A)]	[d ^]	[g]	[g (A)]	[g ^]	
initiale absolue	*79% (63/80)*	*5% (4/80)*	*16% (13/80)*	*68% (112/165)*	*11% (18/165)*	*21% (35/165)*	*31% (4/13)*	*15% (2/13)*	*54% (7/13)*	*258 occurrences*
	80 occurrences			*165 occurrences*			*13 occurrences*			
initiale	*90% (122/136)*	*6% (8/136)*	*4% (6/136)*	*71% (34/48)*	*14,5% (7/48)*	*14,5% (7/48)*	*47% (8/17)*	*35% (6/17)*	*18% (3/17)*	*201 occurrences*
	136 occurrences			*48 occurrences*			*17 occurrences*			
intervocalique	*95% (37/39)*	*2,5% (1/39)*	*2,5% (1/39)*	*82% (18/22)*	*9% (2/22)*	*9% (2/22)*	*66% (12/18)*	*17% (3/18)*	*17% (3/18)*	*79 occurrences*
	39 occurrences			*22 occurrences*			*18 occurrences*			

(JPG 1/81)

Tableau 2 : Synthèse du Tableau 1

	[b,d,g] (sonores)	[b,d,g] (désoncrisées)	[b,d,g] (sourdes)
initiale absolue	*69,4% (179/258)*	*9,3% (24/258)*	*21,3% (55/258)*
initiale	*81,6% (164/201)*	*10,5% (21/201)*	*7,9% (16/201)*
intervocalique	*84,8% (67/79)*	*7,6% (6/79)*	*7,6% (6/79)*

Tableau 3 : % de désonorisations partielles et complètes ([b,d,g] + [b,d,g]) en Octobre 1979 et Mai 1980

	Octobre 1979 (318 occurrences)		*Mai 1980 (538 occurrences)*	
initiale absolue	*39,5% (15/38)*	*26,8% (70/261)*	*30,6% (79/258)*	*25,3% (116/459)*
initiale	*24,7% (55/223)*		*18,4% (37/201)*	
intervocalique	*5,3% (3/57)*		*15,2% (12/79)*	

Toutefois, si l'on considère les réalisations de / d / et / t / par l'enfant J.T. à l'inititiale absolue, on constate, du fait de la présence de variantes [d] (désonorisation partielle), l'existence de VOT positifs pour / d /. Dans un tel cas, il n'y a pas de chevauchements entre les intervalles (réels) de variations (valeur maximale de VOT pour / d / : *15ms* vs. valeur minimale pour / t / : *20ms*). La différenciation est encore assurée. Mais, pour 5 enfants, les faits ne sont pas directement appréhendables et nécessitent un traitement plus élaboré (test t de différenciation de moyennes), car certaines valeurs de VOT sont positives pour / b d g /, aucun prévoisement (VOT négatif) n'étant alors relevé (cf. Tableau 4).

Tableau 4 : Valeurs de V.O.T. positives pour / p,t,k / et / b,d,g / chez 5 enfants (Y.B., O.S., E.W., M.M. et S.F.)

	fortes /ptk/ V.O.T.			*P (test t)*	*faibles /bdg/ V.O.T.*		
	n	$\bar{x}$	*s*		*n*	$\bar{x}$	*s*
Y.B.							
initiale absolue	11	+ 20,00ms	5,92ms	≃ 0,008	6	+ 10,83ms	5,85ms
O.S. (10)							
initiale absolue	12	+ 44,17ms	12,40ms	≃ 0,01	3	+ 21,67ms	5,17ms
E.W.							
initiale absolue	16	+ 31,88ms	13,77ms	≃ 0,001	7	+ 11,43ms	2,44ms
initiale	30	+ 37,17ms	17,84ms	≃ 0,001	9	+ 16,11ms	4,17ms
M.M.							
initiale absolue	12	+ 28,75ms	8,56ms	≃ 0,10	8	+ 21,88ms	9,23ms
intervocalique	10	+ 37,50ms	19,61ms	≃ 0,10	3	+ 23,33ms	2,89ms
S.F.							
initiale absolue	25	+ 39,80ms	16,99ms	≃ 0,06	5	+ 24,00ms	9,62ms
	116				41		

En initiale absolue, les valeurs moyennes de VOT positif varient entre *10,83ms* et *24,00ms* pour / b d g /, entre *20,00ms* et *44,17ms* pour / p t k / (elles sont de *23,33ms* et *37,50ms* à l'intervocalique). La différenciation est établie, même si les articulations des deux séries présentent des VOT positifs, de que confirme les probabilités P lors de l'application du test t de différenciation des moyennes (le seuil 0,10 n'est jamais dépassé; on a même des valeurs de l'ordre de 0,001).

La différence de VOT (voicing lag / voicing lead, pour reprendre la terminologie *LISKER/ABRAMSON, 1964*, entre autres) ou celle de durée d'établissement du voisement (D.E.V.), lorsque le VOT est positif pour les deux types d'articulation, n'est pas le seul indice permettant d'établir que les enfants analysés font véritablement une différence entre les deux types d'occlusives. La durée de la "tenue" (= phase d'occlusion; cf., entre autres, *BEJSTER, 1975/1976*) et la durée totale de l'articulation sont aussi deux indices sûrs permettant, quel que soit le type de VOT relevé, d'établir une distinction (les données ne sont quantifiables pour ces deux indices que pour les positions initiale (non absolue) et intervocalique, puisque le début de la phase d'occlusion ne peut pas être déterminé à l'initiale absolue, étant donné la méthode d'investigation employée), ce qu'indiquent les Tableaux 5 et 6 (*11* enfants = *895* articulations au total : *630* / p t k / et *265* / b d g /).

Tableau 5 : Durée de la "tenue" pour / p,t,k / et / b,d,g / (11 enfants)

	fortes /ptk/			*P (test t)*	*faibles /bdg/*		
	n	$\bar{x}$	*s*		*n*	$\bar{x}$	*s*
N.L. (81 art.)							
initiale	30	100,83ms	24,04ms	< 0,001	16	72,19ms	24,90ms
intervocalique	24	115,42ms	26,16ms	< 0,001	11	82,27ms	19,41ms
J.T. (72 art.)							
initiale	30	131,17ms	46,81ms	≃ 0,03	10	96,50ms	26,78ms
intervocalique	26	116,15ms	20,85ms	≃ 0,02	6	92,50ms	22,08ms
Y.B. (82 art.)							
initiale	42	134,29ms	40,66ms	< 0,001	15	79,67ms	13,82ms
intervocalique	22	132,50ms	25,53ms	≃ 0,003	3	81,67ms	7,64ms
V.M. (144 art.)							
initiale	62	117,74ms	28,58ms	< 0,001	32	83,13ms	21,32ms
intervocalique	40	113,50ms	25,58ms	< 0,001	10	68,00ms	18,44ms
O.O. (82 art.)							
initiale	37	110,95ms	30,16ms	≃ 0,004	21	87,14ms	25,82ms
intervocalique	18	111,39ms	24,54ms	≃ 0,01	6	83,33ms	11,26ms
O.S. (66 art.)							
initiale	35	115,00ms	34,87ms	≃ 0,02	6	79,17ms	11,58ms
intervocalique	17	128,82ms	23,75ms	≃ 0,001	8	91,88ms	22,67ms
E.W. (95 art.)							
initiale	42	121,31ms	35,51ms	≃ 0,002	26	95,19ms	26,02ms
intervocalique	22	120,46ms	26,00ms	≃ 0,07	5	96,00ms	23,02ms
D.H. (33 art.)	21	114,05ms	20,35ms	< 0,001	12	51,67ms	18,13ms
M.M. (66 art.)							
initiale	34	132,50ms	36,69ms	≃ 0,001	6	78,33ms	25,63ms
intervocalique	18	115,56ms	21,48ms	< 0,001	8	79,38ms	15,68ms
V.G. (95 art.)							
initiale	37	122,84ms	32,01ms	< 0,001	27	84,44ms	19,23ms
intervocalique	21	137,38ms	26,15ms	< 0,001	10	91,50ms	19,44ms
S.F. (79 art.)							
initiale	28	114,11ms	35,07ms	< 0,001	20	76,25ms	17,54ms
intervocalique	24	118,54ms	22,24ms	< 0,001	7	75,00ms	13,54ms
	630				265		

Tableau 6 : Durée totale de / p,t,k / et / b,d,g / (11 enfants)

	fortes /ptk/			*P (test t)*	*faibles* /bdg/		
	n	$\bar{x}$	*s*		*n*	$\bar{x}$	*s*
N.L. (81 art.)							
initiale	30	127,00ms	27,72ms	< 0,001	16	76,88ms	23,73ms
intervocalique	24	134,38ms	24,24ms	< 0,001	11	83,64ms	19,51ms
J.T. (72 art.)							
initiale	30	163,00ms	47,14ms	< 0,001	10	98,50ms	25,39ms
intervocalique	26	139,42ms	20,41ms	< 0,001	6	92,50ms	22,08ms
Y.B. (82 art.)							
initiale	42	161,90ms	34,30ms	< 0,001	15	83,00ms	12,93ms
intervocalique	22	160,00ms	26,10ms	< 0,001	3	81,67ms	7,64ms
V.M. (114 art.)							
initiale	62	158,55ms	30,43ms	< 0,001	32	88,59ms	12,89ms
intervocalique	40	158,88ms	30,29ms	< 0,001	10	73,00ms	17,51ms
O.O. (82 art.)							
initiale	37	136,35ms	28,52ms	< 0,001	21	89,76ms	26,43ms
intervocalique	18	128,89ms	21,93ms	< 0,001	6	86,67ms	8,17ms
O.S. (66 art.)							
initiale	35	161,29ms	23,59ms	< 0,001	6	82,50ms	10,37ms
intervocalique	17	154,12ms	20,86ms	< 0,001	8	93,13ms	22,67ms
E.W. (95 art.)							
initiale	42	160,83ms	39,85ms	< 0,001	26	103,46ms	29,22ms
intervocalique	22	165,23ms	34,35ms	≃ 0,001	5	101,00ms	30,08ms
D.H. (33 art.) (11)							
initiale	21	141,67ms	27,63ms	< 0,001	12	71,25ms	22,68ms
M.M. (66 art.)							
initiale	34	173,24ms	35,99ms	< 0,001	6	85,83ms	22,89ms
intervocalique	18	154,17ms	31,77ms	< 0,001	8	93,13ms	21,20ms
V.G. (95 art.)							
initiale	37	153,11ms	33,57ms	< 0,001	27	85,93ms	18,71ms
intervocalique	21	161,14ms	28,09ms	< 0,001	10	94,50ms	19,92ms
S.F. (79 art.)							
initiale	28	150,36ms	36,46ms	< 0,001	20	79,75 ms	15,68ms
intervocalique	24	152,50ms	24,54ms	< 0,001	7	84,29ms	13,97ms
	630				265		

CONCLUSION

Les 11 enfants de l'Ecole Ceccaldi de Vervins (Aisne) soumis à l'analyse expérimentale (données électroglottographiques) présentent dans de nombreux cas d'importantes désonorisations des occlusives phonologiquement sonores / b d g / en positions initiale et intervocalique (env. 25% des cas). Ceci indique que l'opposition phonologique / p t k / - / b d g / n'est pas matérialisée dans tous les cas par les enfants de cette région - et par les adultes d'ailleurs, puisqu'il s'agit de l'une des caractéristiques régionales relevées - par une différence de voisement (présence vs. absence de vibrations des cordes vocales). La différence de *FORCE ARTICULATOIRE*(12) des occlusives / p t k / et / b d g /, qui est le *trait phonologique pertinent* (puisque permanent) (pour permanent = *pertinent*, cf. *GOUDAILLIER, 1981d*, p. 22-23) distinguant ces deux types d'occlusives, peut, dans les cas où les vibrations laryngiennes ne permettent plus la distinction, prendre le relai.

Cette étude doit être complétée par l'examen (en cours) des faits en position finale et être étendue aux articulations de type continu (/ f / vs. / v /, / s / vs. / z /, / ʃ / vs. / ʒ /) (en cours) avec estimation des valeurs de Fo des différentes consonnes "sonores" en fonction des contextes. D'ores et déjà, elle permet à nouveau d'établir que le français - comme toute autre langue ou tout autre parler - n'est pas *UNE* entité immuable(13), mais bien un ensemble de variétés, dont il s'agit de tenir compte lors de nos préoccupations de linguistes/phonéticiens (description) ou d'ingénieurs (reconnaissance, etc.).

REFERENCES

ABRY Christian & BOË Louis-Jean, 1979, Le trait de voisement, ses indices et leurs propriétés. Implications pour les détections automatique et perceptive d'une analyse phonologique au-delà du trait, Recherches sur la prosodie du français, p. 57-93.

ABRY Christian & BOË Louis-Jean, 1980, Système phonétique, idiolecte et différences individuelles, XI° J.E.P., Groupe Communication parlée du G.A.L.F., p. 133-142.

BARTH S., BEN FADHEL R. & MAJO G., 1980, Le D.E.V. dans la parole des déficients auditifs - comparaison avec les entendants, XI° J.E.P., p. 65-75.

BEJSTER Pierre, Le rôle de la durée de l'occlusion dans la perception du voisement chez l'enfant, Rapport d'activités de l'Institut de Phonétique de l'Université Libre de Bruxelles, 10/1, 1976, p. 23-38.

FRØKJAER-JENSEN, 1969, Construction and comparative tests of two different types of glottographs : the photo-electric glottograph and the Fabre glottograph, XVII° Northern Congress of Otolaryngology.

GOUDAILLIER Jean-Pierre, 1981a, Aix-en-Provence : antériorisation des réalisations de / A /; rotation vers l'avant des voyelles nasales; grandes variétés du / r / (à paraître dans un ouvrage collectif I.N.R.P.).

1981a', Vervins : réalisations marquées régionalement d'un système vocalique à trois degrés d'aperture; assourdissement des consonnes phonologiquement sonores (à paraître dans le même ouvrage).

1981b, Dynamique du trait de voisement, P.U.F., Collection "Le Linguiste" (ouvrage collectif sous la direction d'Anne-Marie HOUDEBINE).

1981c, Etude instrumentale du voisement des occlusives d'enfants de Cours Préparatoire dans le Nord de la France (Vervins; dépt. de l'Aisne) (à paraître dans un prochain volume des Miszellen des Hamburger Phonetische Beiträge).

1981d, Phonologie fonctionnelle et phonétique expérimentale; exemples empruntés au luxembourgeois, XII + 476 pages (Bd. N° 36 des Hamburger Phonetische Beiträge), Hamburg, Buske Verlag.

LISKER Leigh & ABRAMSON Arthur S., 1964, A cross-language study of voicing in initial stops : acoustical measurements, Word, 20, p. 384-422.

MACKEN Marlys A. & BARTON David, 1979, The acquisition of the voicing contrast in English : a study of voice onset time in word-initial stop consonants, Journal of child language, 7, p. 41-74.

SERNICLAES Willy & BEJSTER Pierre, 1978, Différences interlinguistiques dans le traitement perceptif des indices de voisement, Rapport d'activités de l'Institut de Phonétique de l'Université libre de Bruxelles, 10/1, p. 83-94.

NOTES

x Une version plus étendue de ce texte doit faire l'objet d'une publication prochaine (GOUDAILLIER, 1981c).

(1) Groupe Langue Orale, Français 1er Degré (coordination scientifique : Anne-Marie HOUDEBINE / Hélène ROMIAN). Cette recherche a été en partie financée par l'I.N.R.P..

(2) Ecole Les Milles I - Les Milles II, Les Milles (10-15 kms d'Aix-en-Provence) (Bouches-du-Rhône); Ecole Ceccaldi, Vervins (Aisne). Enfants ne présentant pas, dans chacune des deux écoles, de troubles pathologiques ou de retard de langage.

(3) Résultats de l'enquête effectué en Octobre 1979. Toutes les articulations ont été étudiées au détecteur de mélodie/voisement (Mélographe, CNET/FIES). Ces chiffres sont respectivement de 4,8%, 2,9% et 4,8% (Aix), 23,5%, 14,3% et 43,4% (Vervins) en positions initiale, intervocalique et finale (GOUDAILLIER, 1981b).

(4) Lors de l'enquête, chaque enfant doit décoder une série d'images, qui lui sont présentées, et doit en indiquer le nom.

(5) Les fonctions de stockage et de visualisation sont assurées par un oscilloscope à mémoire numérique GOULD ALLCO type 4000.

(6) Simultanément, le tracé oscillographique du signal (voie 1), celui des variations de Fo et d'intensité (déduites de l'analyse de ce signal) étaient observés sur un oscilloscope TEKTRONIX type 5000.

(7) Cf. planches 1 et 2 (réduction à 2fois 70% des originaux) (EG = électroglottographe).

(8) Ouvrier agricole, petit artisan, employé bureau, magasinier, etc... étaient les professions du père et/ou de la mère des enfants retenus pour l'analyse.

(9) Cf., entre autres, SERNICLAES/BEJSTER, 1978, auteurs, pour lesquels le VOT est l'indice de différenciation le plus sûr pour l'aspect perceptuel (prévoisement) vs. délai d'établissement du voisement (D.E.V.)). Rappelons que le trait de voisement peut avoir une réalisation pluri-indiciable (on se reportera, à cet effet, aux divers travaux publiés à ce sujet dans les Rapports d'activités de l'Institut de Phonétique de l'Université libre de Bruxelles entre 1975 et 1978 par Pierre BEJSTER, Renaud BEECKMANS, Willy SERNICLAES et Max WAJSKOP, ainsi qu'à l'étude d'ABRY Christian & BOË Louis-Jean, 1979).

(10) $\bar{x}$ *pour* / p t k / *est particulièrement élevé, car il s'agit principalement de réalisations de type* [## $k^{(h)}$]. *Si l'on compare* [$k^{(h)}a\overset{\leftarrow}{\smile}ʁt^{(h)}$] *et* [$ga\overset{\leftarrow}{\smile}ʁsõ$] *d'une part,* [$k^{(h)}\underset{\rightarrow}{a}z$] *et* [$gat^{(h)}ǫ$] *d'autre part (cf. planche 2), on constate que la différence due au VOT est particulièrement apparente sur les tracés.*

(11) *Pour l'enfant D.H. n'ont été retenues que les articulations à VOT positif, puisque celles-ci sont majoritaires dans le système de l'enfant (de même pour le Tableau 5).*

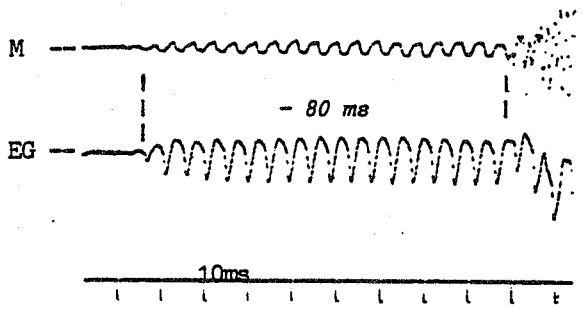
(12) *Les indices en sont la différence de VOT (négatif pour* / b d g / *vs. positif pour* / p t k /*), celle de durée des VOT positifs (*/ p t k / > / b d g /*) et celle de durée de "tenue" (= phase d'occlusion) (*/ p t k / > / b d g /*) ou de durée totale (*/ p t k / > / b d g /*).*

(13) *Les petits aixois ne réagissent pas, tout comme les adultes de la région, de la même manière, tant s'en faut.*

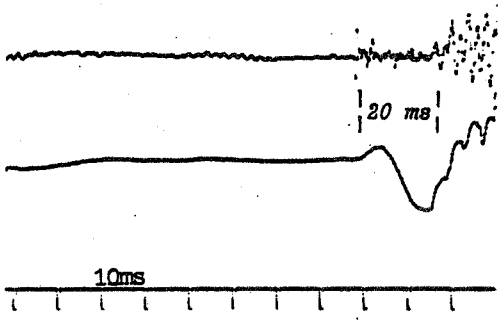
Jean-Pierre GOUDAILLIER, Exemple de traitement de l'opposition de "sonorité" par des enfants de Cours Préparatoire, XII° J.E.P., Montréal, Mai 1981.

Planche 1

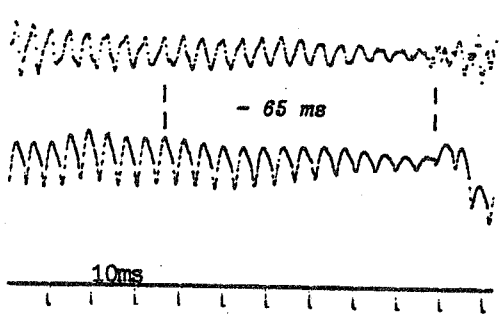
Enfant O.O. : [bɔs]



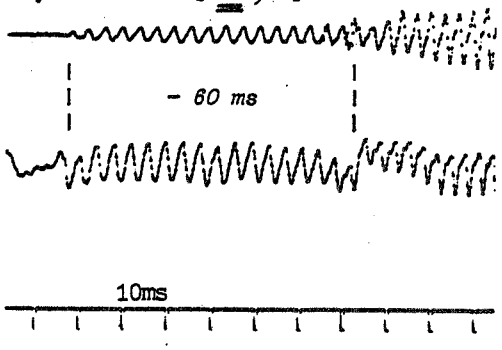
Enfant O.O. : [$y^{n}mɥt^{(h)}ɥ$]



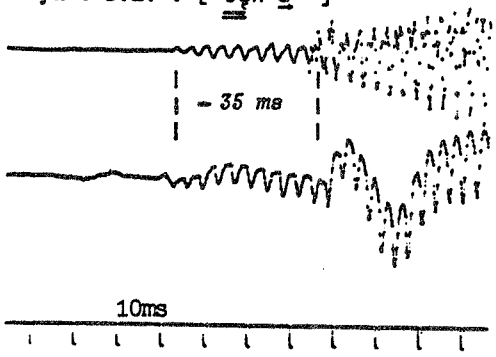
Enfant O.O. : [dexidɥ]



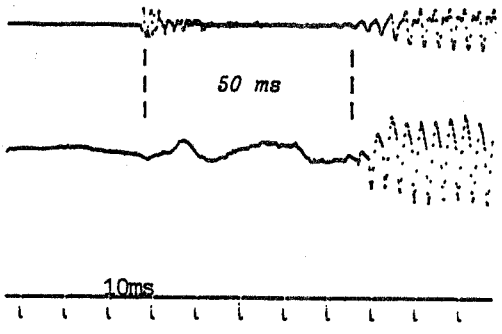
Enfant D.H. : [buʒi]



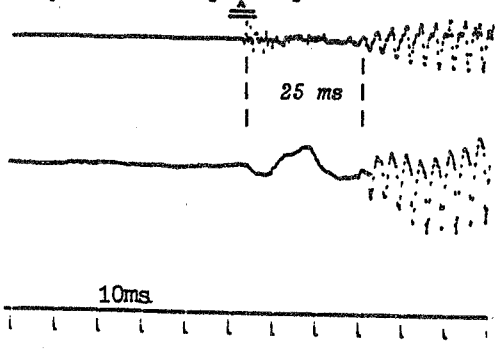
Enfant D.H. : [$bɑ̃^{w}a^{ʁ}$]



Enfant D.H. : [$t^{(h)}wɑ$]



Enfant D.H. : [$d^{w}ɑ$]

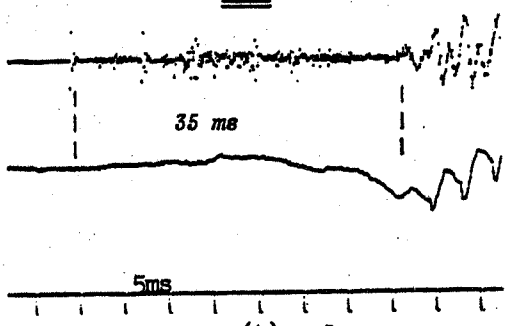


Document Laboratoire de Phonétique
Université René Descartes / Paris V

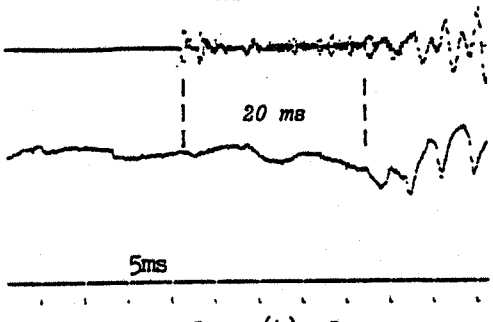
Jean-Pierre GOUDAILLIER, Exemple de traitement de l'opposition de "sonorité" par des enfants de Cours Préparatoire, XII° J.E.P., Montréal, Mai 1981.

Planche 2

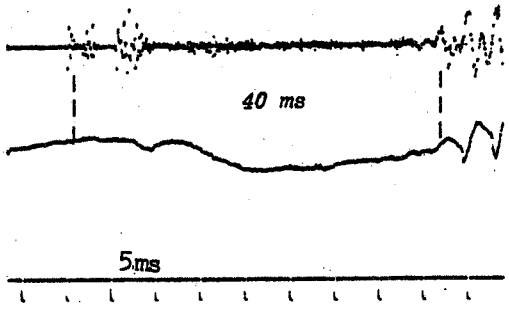
Enfant O.S. : [k(h)aʁt(h)]



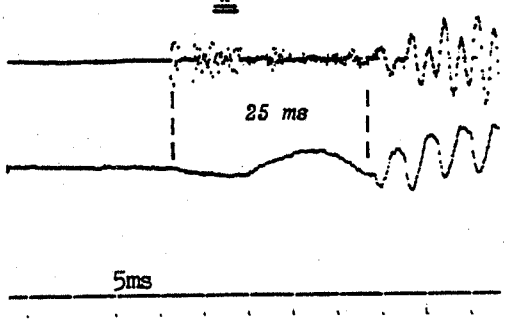
Enfant O.S. : [gaʁsõ]



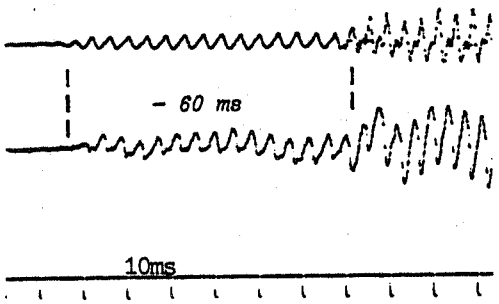
Enfant O.S. : [k(h)aʒ]



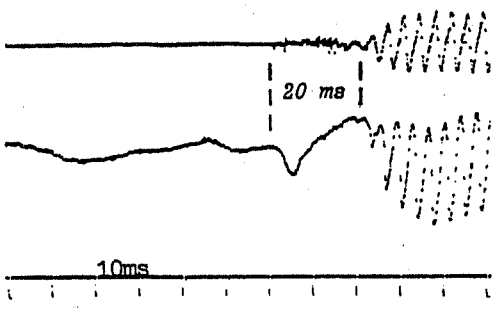
Enfant O.S. : [gat(h)ɐ]



Enfant O.S. : [bik(h)ɛt(h)]



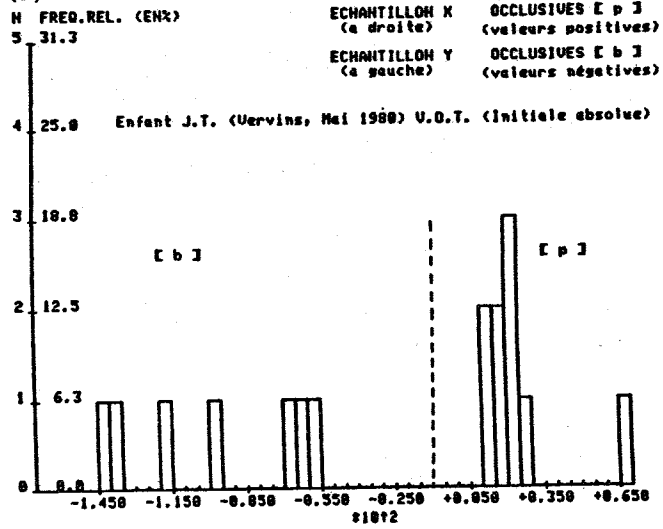
Enfant O.S. : [bi∫]



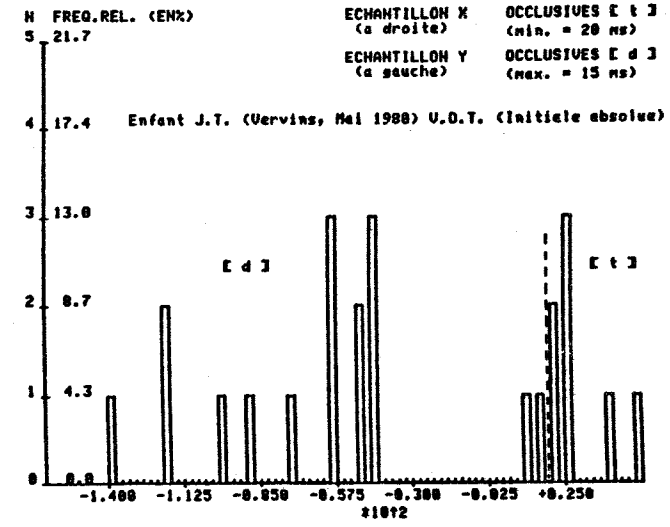
Document Laboratoire de Phonétique
Université René Descartes / Paris V

Jean-Pierre GOUDAILLIER, Exemple de traitement de l'opposition de "sonorité" par des enfants de Cours Préparatoire, XII° J.E.P., Montréal, Mai 1981.

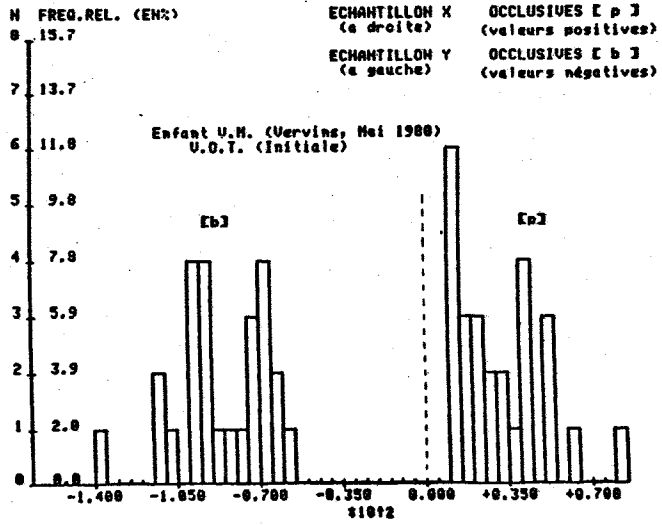
(1)



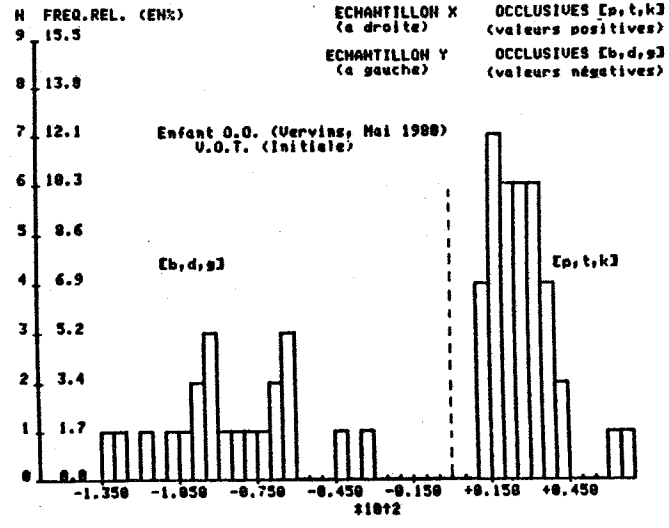
(2)



(3)



(4)



12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

INFLUENCE du SYSTEME LINGUISTIQUE ORIGINEL sur la PERCEPTION : CAS DU VOISEMENT

E. LHOTE
M. STRICKLAND
P.W. LEE

Laboratoire de Phonétique de BESANCON

.RESUME

Les auteurs ont tenté, dans deux tâches différentes, de faire apparaître comment des individus utilisant le français comme langue seconde (canadiens anglais, allemands, chinois) procèdent pour percevoir l'opposition de voisement en français par comparaison avec des francophones. La première tâche, de répétition, montre que certains auditeurs-locuteurs utilisent une référence au système phonologique de la langue maternelle. La seconde tâche, tâche d'identification, fait apparaître un regroupement des auditeurs qui possèdent l'opposition de voisement dans les occlusives de leur langue L_1 et une identification correcte chez ces derniers de la plupart des phonèmes, même si les indices de l'opposition sont différents en L_1 et L_2 ; elle met également en évidence la tendance à la neutralisation de l'opposition chez les auditeurs chinois, neutralisation qui atteste une mauvaise perception d'une opposition absente dans le système de L_1.

. ABSTRACT

Two tests were used in order to determine how foreign language speakers (English Canadian, German, Chinese) perceived the voicing opposition in French as opposed to French native speakers for the series of voiced / unvoiced occlusives. In a repetition task, it was shown that some interference of L_1 on L_2 appeared to be present in the productions of certain of the non-native French speakers. In an identification task, it was shown that those subjects possessing the voicing opposition in L_1 obtained higher scores in identification in L_2. The poor scores obtained by the Chinese subjects point to a neutralisation of an opposition which does not exist in the L_1 of those subjects.

INTRODUCTION

Il est communément admis que la nature de la langue maternelle L_1 est un élément important dans l'acquisition d'une langue seconde L_2 et la plupart des travaux effectués dans ce domaine mettent l'accent sur le fait que les erreurs de production sonore dans L_2 s'expliquent généralement par référence à la structure sonore de L_1.

Il est également devenu banal de dire que l'on produit bien ce que l'on perçoit bien ; et on est alors vite tenté de penser qu'il suffit de bien percevoir pour bien produire. . .

On sait de plus que la pratique d'un réel bilinguisme entraîne chez le bilingue l'organisation d'un système d'interférences entre L_1 et L_2.

Un certain nombre d'indices acoustiques semblent contribuer à la perception de voisement. DELATTRE (1958) a dégagé de ses expériences sur des syllabes synthétiques la présence de 10 indices participant à l'opposition de sonorité chez les consonnes françaises. LISKER et ABRAMSON (1964) ont montré, sur plusieurs langues, que des changements de V. O. T. suffisent à distinguer les occlusives voisées des occlusives non voisées. STEVENS et KLATT (1974) ont expérimenté sur des stimuli [ta] - [da] l'effet d'une variation dans la transition de formant (F_1 et F_2) lorsque celle-ci est associée aux variations de V. O. T., et ils ont observé qu'une adaptation à un indice peut influencer l'interprétation d'un autre indice relevant de la même distinction de voisement. Sachant par ailleurs que l'opposition de voisement est réalisée de façon spécifique dans une langue donnée, on peut se demander quels sont les indices utilisés par un bilingue lors de la production et la perception de L_2.

Comment procèdent au niveau phonologique les bilingues qui ont à utiliser des oppositions phonologiques différentes dans les deux langues ?

Comment perçoivent-ils les différences acoustiques de L_2 en fonction de la langue-mère L_1 ? et comment les reproduisent-ils ?

ANALYSE du PROBLEME

On a le plus souvent jugé la façon dont un locuteur de L_1 perçoit une langue L_2 par sa faculté de produire des sons, des mots, des phrases de L_2.

CARAMAZZA et al. (1973) ont quant à eux essayé d'apprécier la perception du V.O.T. chez des monolingues et des bilingues canadiens en testant à la fois la perception et la production ; il se dégage de leur travail l'idée intéressante que le système de production du bilingue s'adapte plus facilement aux mécanismes de L_2 que le système de perception, ce qui tendrait donc à différencier les deux processus.

ALBERT et OBLER (1978), s'appuyant sur des observations d'aphasiques bilingues, vont plus loin et avancent l'idée d'une dissociation possible entre le système de la production et le système de la perception.

Ces dernières observations projettent un éclairage nouveau sur l'hypothèse de deux systèmes interdépendants qu'ont longtemps défendue deux théories ; d'une part la théorie motrice de la perception selon laquelle, rappelons-le, nous interprétons les formes sonores de la parole en termes de mouvements articulatoires, mouvements nécessaires pour produire les formes perçues ; d'autre part le modèle d'analyse par synthèse (Analysis-by-Synthesis) proposé par STEVENS (1967-1973).

Il est donc permis de se poser la question : comment perçoit-on une langue seconde ?

Nous avions montré dans des travaux antérieurs (LHOTE, 1980), que la perception tonale varie avec la langue maternelle des auditeurs et nous souhaitions faire apparaître plus clairement de quelle nature est l'influence jouée par la langue maternelle.

Nous avons donc cherché à analyser de plus près la façon dont on perçoit une langue seconde en limitant le champ de l'observation aux unités segmentales, et plus particulièrement aux consonnes occlusives françaises reliées entre elles par l'opposition de voisement. Nous avons pour cela choisi de combiner deux modes d'approche, l'un faisant appel à la répétition, l'autre à la faculté d'identifier des auditeurs, ceci afin de pouvoir comparer les mécanismes perceptifs utilisés dans deux tâches différentes, l'une d'elles, la répétition, faisant appel à la fois à la perception et à la production.

Nos sujets appartiennent à quatre groupes linguistiques dont les langues L_1 sont le français, l'allemand, l'anglais canadien et le chinois.

PROCEDURE EXPERIMENTALE

L'expérience a été faite à partir de 26 stimuli naturels, représentant soit des mots de la langue française, soit des logatomes dont la structure dissyllabique peut être :

CVCV ou VCVC

Ce contexte syllabique place la consonne dans les positions initiale, médiane et finale. Cette consonne est une occlusive orale sourde ou sonore, c'est-à-dire [p, b, t, d, k, g] ; la voyelle ne peut être que [a, o, e].

Exemples : [agad] , [tako] , [agat] , [dako] , [etap] ...

L'enregistrement du corpus a été effectué en studio d'enregistrement par un locuteur francophone de Besançon, à l'aide de la chaîne glottométrique* et d'un microphone : chaque stimulus est capté simultanément à deux niveaux, buccal et glottique, il est enregistré sur bande magnétique, puis analysé et transcrit à l'aide d'un enregistreur graphique de type OSCILLOMINK. L'enregistrement magnétique sert de modèle pour la tâche de répétition.

Les auditeurs sont au nombre de 11 : trois sont francophones, trois anglophones canadiens, trois germanophones, deux chinois (l'un cantonais, l'autre mandarin). Tous les non francophones ont une bonne maîtrise du français et sont étudiants à l'université.

La première tâche consiste à faire répéter aux auditeurs 12 mots (le même mot revenant deux fois) prononcés par le modèle qu'ils entendent sur bande magnétique uniquement. La répétition est captée chez chaque auditeur à deux niveaux également, puis analysée. Il est alors possible de comparer sur les tracés de mélodie laryngée le temps de vibration des cordes vocales chez le modèle et chez chaque auditeur.

La deuxième tâche consiste à faire subir aux auditeurs un test d'identification en laboratoire de langues : on fait entendre les 26 stimuli et on demande aux sujets d'identifier les consonnes (52 en tout), dans un exercice écrit où des blancs doivent être remplacés par des consonnes

Exemple : _ a _ o

Précisons que l'auditeur n'a pas à nommer mais à écrire les consonnes qu'il croit reconnaître.

* chaîne comportant glottomètre et analyseur de mélodie, mise au point par le laboratoire d'Automatique de Besançon.

RESULTATS et DISCUSSION

1. Tests de répétition.

Les tracés glottographiques font clairement apparaître trois tendances dans les répétitions des auditeurs :

a) une présence totale de voisement durant toute la durée des sonores dans les trois positions qui s'oppose à une absence de voisement durant toute la durée de l'occlusive caractérise tout le groupe de francophones (Figure 1) ;

b) une tendance à réaliser différemment du modèle l'opposition de voisement chez 2 germanophones et chez deux anglophones canadiens

- chez les germanophones : désonorisation partielle des sonores intervocaliques, désonorisation totale des sonores initiales et finales (Figure 2)

- chez deux anglophones canadiens : on constate une désonorisation partielle des sonores initiales et finales (Figure 3).

Il est intéressant de remarquer que la désonorisation partielle se produit chez germanophoneset anglophones dans la phase finale de l'occlusive.

Mais il faut se garder de généraliser et de croire que les écarts par rapport au modèle français sont systématiques : chaque auditeur présente toujours une seconde tendance, plus ou moins accentuée chez chacun d'entre eux, qui consiste à répéter de façon identique au modèle. Il arrive qu'un germanophone et un anglophone répètent de façon très proche du modèle, du moins pour ce qui concerne la réalisation du trait de voisement ;

c) la troisième tendance est celle qui caractérise les deux auditeurs chinois. On observe : chez l'un à la fois une tendance à sonoriser des sourdes et une tendance contraire à assourdir des sonores ; chez l'autre un assourdissement quasi général des sonores, principalement à l'initiale.

Interprétation

La différence essentielle entre les deux groupes de non francophones ne provient pas de la nature des écarts dans les réalisations, mais des relations qui les régissent ; en d'autres termes les écarts que nous avons énumérés ci-dessus ne sont que les reflets d'une structure sous-jacente dans le cas des germanophones et anglophones, et au contraire la manifestation d'une absence de structure dans le cas des chinois.

En effet dans le premier groupe l'opposition de voisement subsiste ; elle est simplement réalisée différemment du modèle.

Exemple : chez un germanophone

l'occlusive voisée s'oppose à l'occlusive non voisée de façon différente selon la position de l'occlusive :

. à l'intervocalique, alors que la sourde présente durant toute la durée de l'occlusive (16 centisecondes en moyenne) une absence de vibration, la sonore présente une absence de vibration durant les 5,3 centisecondes précédant la voyelle finale (durée moyenne de la consonne : 9 cs)

. en finale, la voyelle qui précède la sonore est plus longue que celle qui précède la sourde (24 cs / 16 cs)

. à l'initiale, l'indice essentiel de l'opposition est dans ce cas la chute de hauteur qui accompagne la voyelle qui suit la sourde (1 ton 1/2), alors que celle qui suit la sonore ne présente pas de variation de hauteur (MASSARO et COHEN, 1976)

Dans le cas des chinois on peut être embarrassé pour interpréter des variations aussi grandes d'un locuteur à l'autre* : ce qui est révélateur, c'est le fait que l'indice principal de voisement, qui n'est pas utilisé comme en français, n'est pas pour autant remplacé par un autre indice, comme c'est le cas chez les anglophones et chez le germanophone : l'auditeur chinois qui assourdit systématiquement les sonores ne marque aucune différence dans les réalisations des deux catégories, ni par la durée, ni par l'aspiration, ni par la hauteur. Quant à l'autre auditeur qui semble mieux répéter, il faut reconnaître qu'il réalise parfois au hasard, l'une à la place de l'autre (Figure 4).

On peut partiellement expliquer ces deux types de comportement chez les auditeurs non francophones en se référant au système phonologique des langues L_1 : l'anglais et l'allemand possèdent une série d'occlusives orales qui s'opposent par le voisement, mais les indices de l'opposition ne sont pas les mêmes qu'en français ; en revanche le chinois (mandarin et cantonais) ne possède pas de série d'occlusives s'opposant par le voisement alors qu'il en possède qui s'opposent par l'aspiration. La tâche n'est donc pas la même pour les chinois qui ont à construire une nouvelle opposition phonologique, à l'intégrer à l'ensemble du système et à la différencier de l'opposition d'aspiration qui existe dans leur système initial. En fait les erreurs que nous avons constatées, différentes chez les deux auditeurs, attestent dans les deux cas une tendance à neutraliser l'opposition de voisement ; et cette tendance à la neutralisation est une preuve que l'opposition phonologique est mal perçue.

* on ne peut évoquer le fait que l'un parle le cantonais et l'autre le mandarin, la structure phonologique des occlusives étant la même dans les deux cas

On peut en conclure que ces auditeurs ont utilisé en répétition, en partie ou totalement, une référence à leur système phonologique initial dans une tâche qui faisait pourtant appel uniquement à la mémoire immédiate.

2. Les tests d'identification.

D'une façon générale, on peut dire que les consonnes sont assez bien identifiées. Quant aux lieux d'articulation, ils n'ont pas représenté de difficulté particulière aux auditeurs. Il est à noter toutefois que [p] est la consonne du modèle la moins bien perçue, celle qui a été le plus fréquemment à l'origine d'erreurs d'identification.

L'erreur la plus généralement commise, par tous les groupes linguistiques, a consisté à remplacer une consonne sourde par la sonore correspondante. La cause n'en est pourtant pas la même selon les auditeurs : quand un francophone remplace [pato] par [bato] et [kato] par [gato] , on voit tout de suite qu'en fait il y a attirance vers le mot de la langue, et que l'erreur est simplement due à l'ambiguïté que représente pour un francophone la présence de logatomes à côté de mots de la langue. Quand un anglophone remplace [apag] par [abag] et [tako] par [dako] il n'en est pas de même : on peut simplement dire que les indices français de l'opposition ne suffisent pas forcément à l'auditeur canadien pour identifier la sourde par rapport à la sonore.

Les tableaux 1 et 2 résument dans des histogrammes les résultats des deux tests en fonction des origines linguistiques des auditeurs.

3. Comparaison des deux types de tests

On peut être surpris de constater que des auditeurs non francophones qui, dans les répétitions, s'écartaient souvent du modèle français, obtiennent un score aussi performant que les francophones dans le test d'identification. Ceci montre de façon bien claire que la tâche d'identification n'est pas du tout de même nature que celle qui est mise en oeuvre dans la répétition : dans le premier cas (identification) l'auditeur met en corrélation la structure de relations de la langue L_2 qui lui est proposée avec celle qu'il a acquise durant l'apprentissage de L_2 par référence à L_1 ; dans le second cas (répétition) l'auditeur n'a pas le temps de remonter à l'analyse phonologique et utilise le système de relations qui lui est immédiatement disponible, ce qui, suivant les cas, l'amène à se servir soit de celui de sa langue maternelle, soit de celui de L_2, soit encore, comme c'est le cas pour les chinois, d'une forme hybride.

Si l'on admet que la tâche d'identification est plus spécifiquement une tâche de perception et que celle de répétition est plutôt une tâche de production (liée à la perception), on est tenté de penser avec ALBERT et OBLER (1978) que chez les bilingues il existerait un seul système de type composé (compound system) pour la perception, tandis qu'il y en aurait deux de type coordonné pour la production.

Le tableau 3 établit un parallèle entre les résultats des deux tests ; il fait plus clairement apparaître un lien entre les deux : quand l'auditeur ne possède pas de bon score dans le test de répétition, il n'est pas très bien placé dans le test d'identification, et c'est le cas de G_2, C1 et C2. Inversement quand l'auditeur a une bonne perception dans le test d'identification, il présente en général une bonne production dans le test de répétition. Un grand écart est à signaler chez A3 entre les résultats des 2 tests (T.R : 75 %, T.I : 100 %) : afin d'examiner cette divergence, on a fait entendre les répétitions de A3 aux trois auditeurs français. Ces derniers ont perçu, à notre grande surprise, comme des sonores les consonnes prononcées assourdies par A3. Cette observation met en évidence que l'auditeur A3 s'est servi dans la production d'un indice différent de celui qui est utilisé en priorité par les francophones pour distinguer des consonnes voisées des non-voisées. Or, cet indice est pertinent pour les oreilles francophones.

CONCLUSION

Nous avons tenté dans deux tâches différentes, l'une de répétition, l'autre d'identification, de faire apparaître comment des bilingues procèdent pour percevoir l'opposition de voisement en français.

Nous avons constaté que la tâche d'identification regroupe en fait les auditeurs qui possèdent l'opposition dans leur langue maternelle, même si les indices de l'opposition sont différents, indices que l'on retrouve assez souvent différents dans la tâche de répétition.

Il y a toutefois convergence des résultats obtenus dans les deux types de tests pour les auditeurs (chinois) qui n'ont pas l'opposition de voisement dans leur système d'occlusives.

Ces observations invitent à ouvrir la réflexion sur d'autres modalités d'expériences permettant de mieux cerner le processus de perception ; elles ouvrent également la voie à des applications spécifiques dans l'apprentissage des langues secondes.

REFERENCES

ALBERT M.L., OBLER L.K. (1978), The bilingual brain ; neuro-psychological and neurolinguistic aspects of bilinguism
Academic Press.

CARAMAZZA A., YENI-KOMSHIAN G.H., ZURIF E.B., CARBONE E. (1973), The acquisition of a new phonological contrast : the case of stop consonants in French-English bilinguals
Journ. Acoust. Soc. Amer. 54, n° 2, 421-428.

CARAMAZZA A., YENI-KOMSHIAN G.H. (1974), Voice Onset Time in two French dialects
Journal of Phonetics, 2, 239-245.

DELATTRE P. (1958) : Les indices acoustiques de la parole.
Phonetica 2 : 108-118, 226-251.

LHOTE E. (1980), Analyse et synthèse de faits de langue au niveau du larynx.
Thèse de doctorat d'état, Strasbourg, mai 1980.

LISKER L., ABRAMSON A. (1964)
A cross-language study of voicing in initial stops : acoustical measurements
WORD 20, 384-422.

MASSARO D., COHEN M. (1976) : The contribution of fundamental frequency, and voice onset time to the /zi/ - /si/ distinction
J.A.S.A. vol. 60, n° 3, 704-717.

STEVENS K.N., HALLE M. (1967), Remarks on analysis by synthesis and distinctive features
in Wathen-Dunn W. (Ed) : Models for the perception of speech and visual form
Cambridge, M.I.T. Press

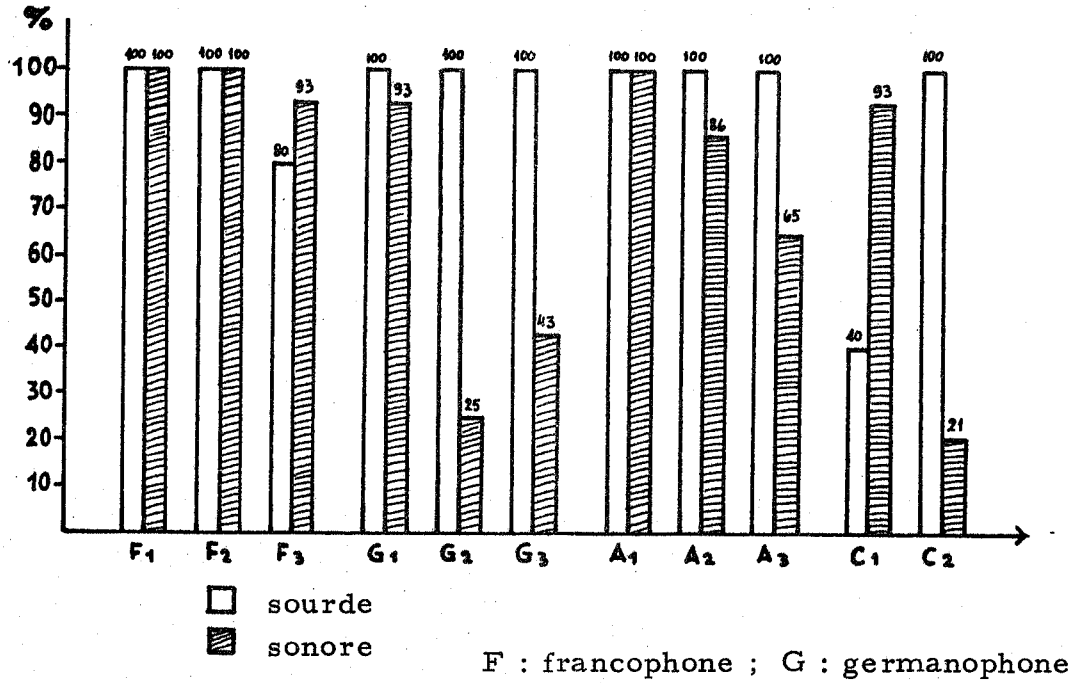
STEVENS K.N. (1973), The potential role of property detectors in the perception of consonants
Symposium on Auditory Analysis and Perception of Speech
Leningrad, URSS.

STEVENS K., KLATT D. (1974) : Role of formant transitions in the voiced - voiceless distinction for stops
J.A S.A. 55, 653-659.

PISONI D.B., SAWUSCH J.R. (1975), Some stages of processing in Speech Perception
in "Structure and Process in Speech Perception"
Communication and Cybernetics 11
Ed. by A. COHEN and S.G. NOOTEBOOM
Springer Verlag.

Tableau 1 :

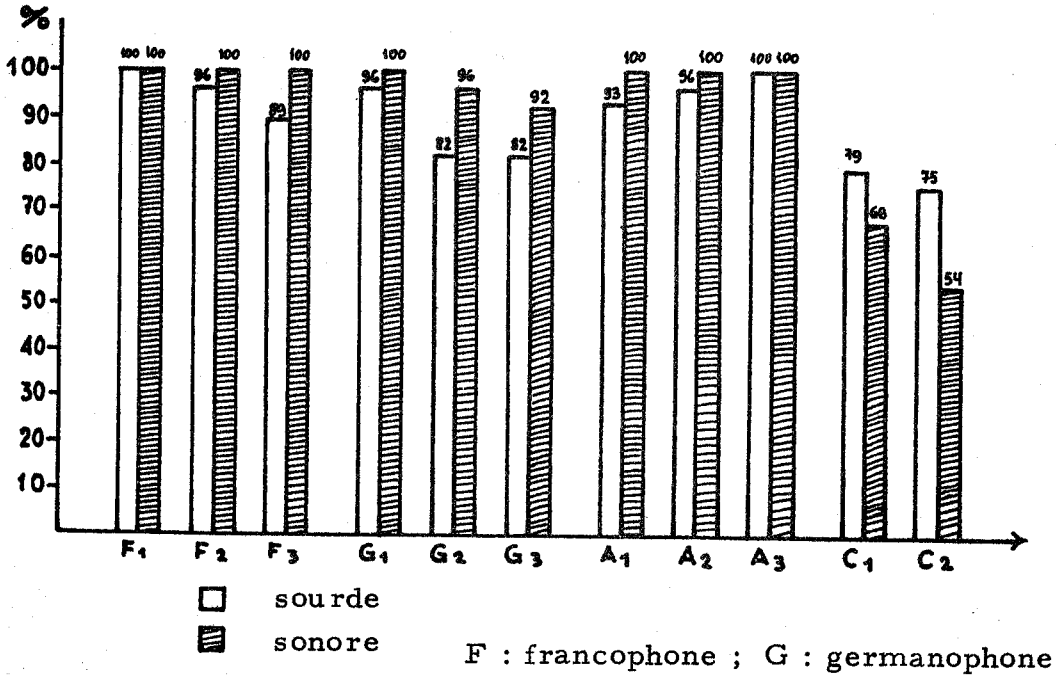
Pourcentages des répétitions correctes selon les auditeurs



F : francophone ; G : germanophone ;
A : anglophone ; C : chinois.

Tableau 2 :

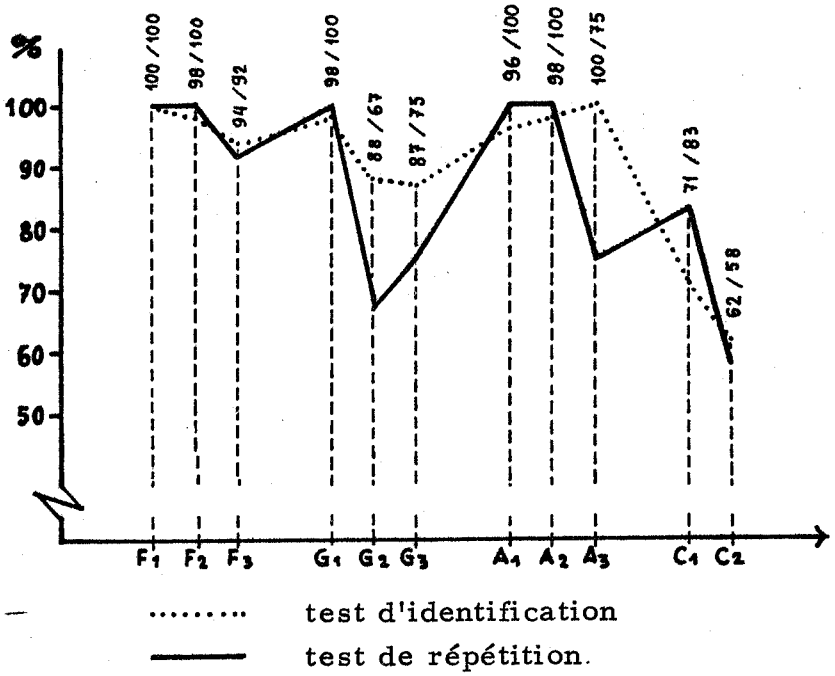
Pourcentage des réponses correctes au test d'identification

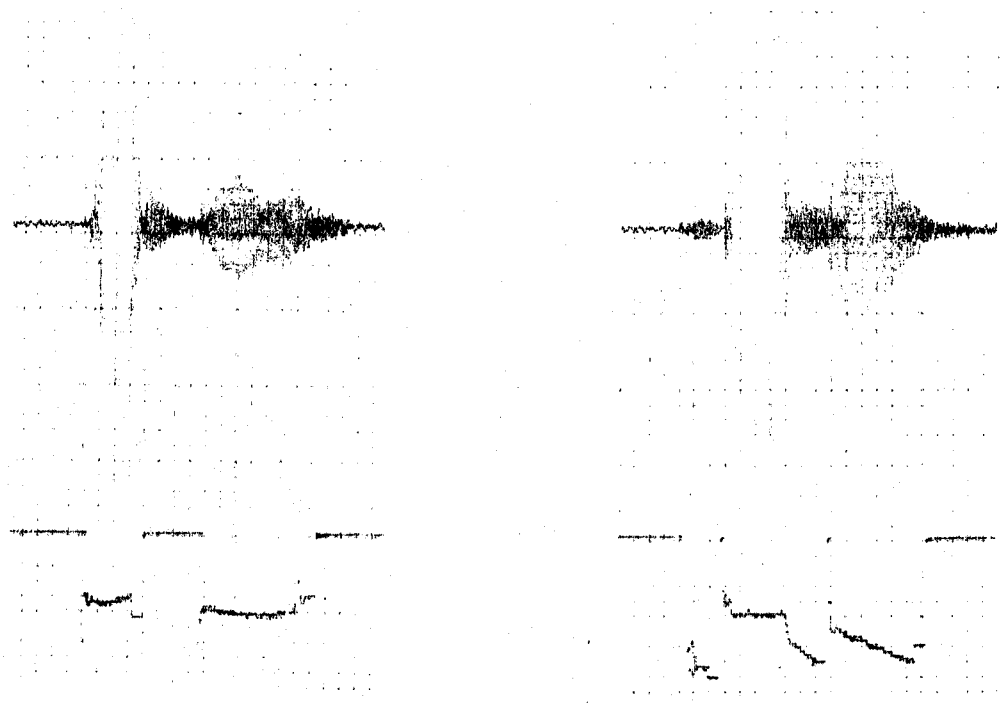


F : francophone ; G : germanophone ;
A : anglophone ; C : chinois.

Tableau 3 :

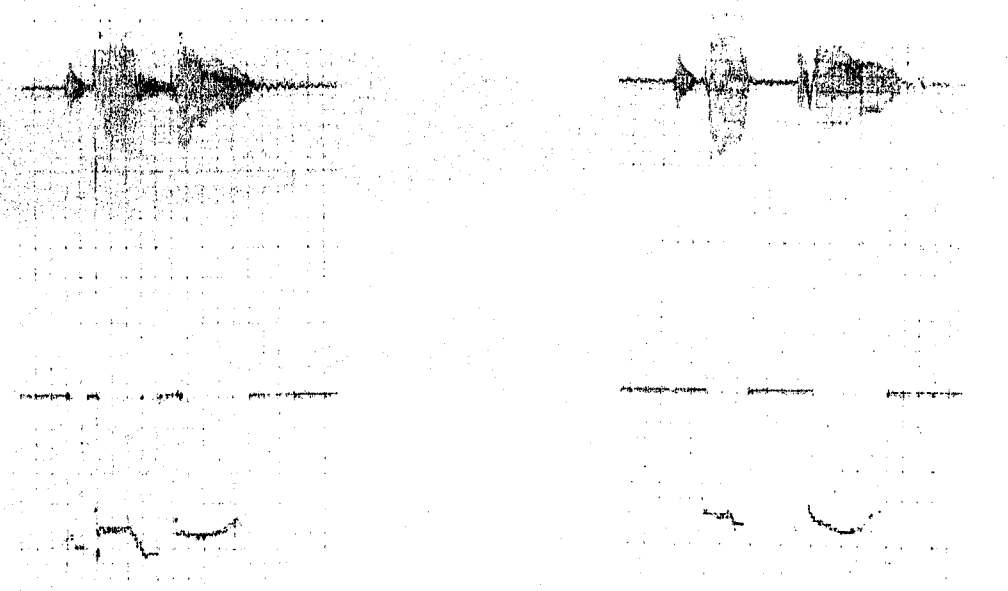
Pourcentages des réponses correctes selon les auditeurs





t a k o d a g o

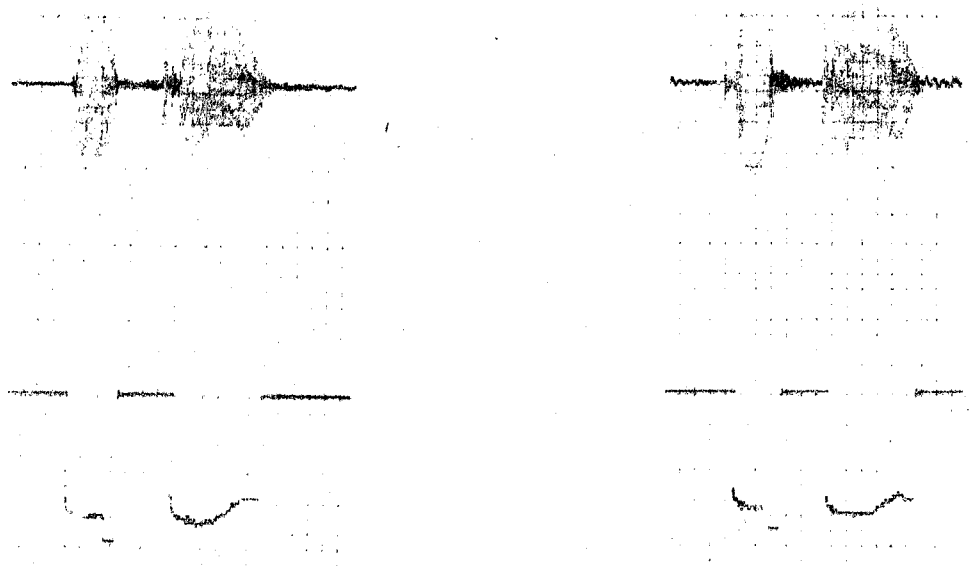
Figure 1 : répétition de [t a k o] - [d a g o] par un francophone
(1 division : 1 ton)



b a d o d a k o
° ° °

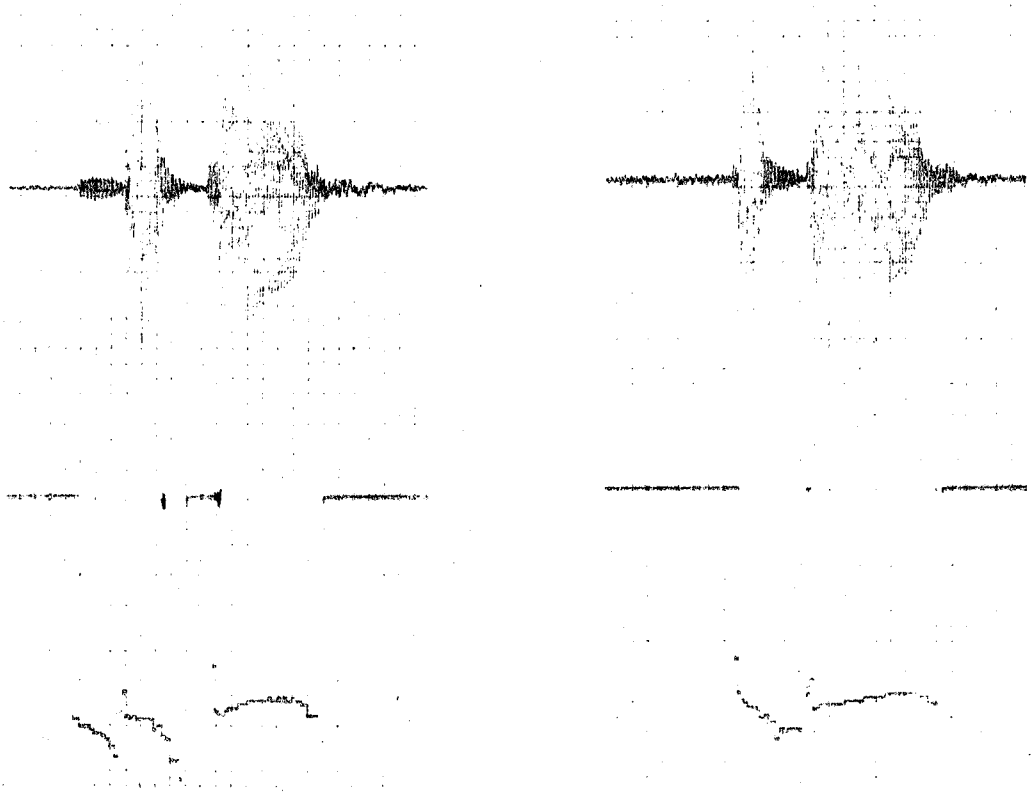
Figure 2 : répétition de [b a d o] et [d a k o] par un germanophone

les indices de voisement sont ici la durée de la tenue, l'absence / présence d'aspiration, la présence / absence de vibrations durant la tenue, la chute de hauteur qui suit la sourde.



d a k^{h} o t^{h} a k^{h} o

Figure 3 : répétition de [d a k o] - [t a k o] par un canadien anglophone.
L'indice visible de l'opposition à l'initiale est la présence / absence d'aspiration.



d a k o t a k^{g} o

Figure 4 : répétition de [d a k o] - [t a k o] par un chinois.
On observe sur ces tracés une sonorisation partielle ou totale des sourdes intervocaliques, ce qui n'est pas le cas à l'initiale.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ANALYSE MULTIDIMENSIONNELLE DE L'ESPACE PERCEPTIF DES SONS VOCALIQUES

F. LONCHAMP

INSTITUT DE PHONETIQUE

UNIVERSITE DE NANCY II

.**RESUME**

Dans les diverses analyses multidimensionnelles de l'espace perceptif des voyelles publiées, le nombre de dimensions interprétables varie de 2 à 6 selon les auteurs. Pour tenter de préciser ce nombre, 40 sujets français ont fourni des estimations numériques des dissimilarités entre les voyelles /i, e, ɛ , a, ɔ , o, u, y, ø , oe/. Ces données, ainsi que celles fournies par les travaux de HANSON (1967), MOHR & WANG (1968), SINGH & WOODS (1971) et TERBEECK (1977) ont été analysées par une méthode métrique (INDSCAL) et non-métrique (KRUSKAL). Pour les moyennes de groupe, la quantité de variance expliquée peut dépasser 90 % dans un espace bidimensionnel, ce qui semble exclure la présence de plus de 3 dimensions. Le plan des deux premières dimensions et le plan F1-F2 (ou F1-F'2) sont fortement corrélés. Le meilleur indice d'une troisième dimension semble être la stabilité, pour les deux méthodes, de l'arrangement des stimuli lors d'analyses répétées à partir de configurations initiales arbitraires. La voyelle rétroflexe /ə / et les voyelles antérieures arrondies /ø / et /oe/ se placent systématiquement à une extrémité du 3ème axe. On montre que "l'effet de plafonnement" des dissimilarités en fonction des distances ne peut être invoqué. Le trait acoustique commun à /ə , ø / et /oe/, c'est-à-dire la proximité de F2 et F3, est proposé comme corrélat physique de cette troisième dimension.

. ABSTRACT

MULTIDIMENSIONAL ANALYSIS OF VOCALIC PERCEPTUAL SPACE

In published sources, the number of recovered dimensions ranges from 2 to 6. In our experiment, 40 French subjects assessed similarities between 10 vowels / i e ɛ a ɔ o u y ø œ / on a 7 point scale. From this study and a reanalysis of 4 other studies (HANSON (67), MOHR & WANG (68), SINGH & WOODS (71), TERBEECK (77)) using both metric (INDSCAL) and non-metric (KRUSKAL) methods, the following conclusions can be drawn: a) the variance accounted for can reach 90% for two-dimensional group data, leaving little room for more than a third dimension; b) the first two unrotated dimensions and the F1-F2 (or F1-F'2) plane are highly correlated. The most reliable criterion for the presence of a third dimension seems to be the stability of stimuli ordering for repeated analyses from random starting configurations with both methods. Retroflex /ɚ/ and front non-high rounded /ø , œ/ vowels consistently end up at one end of an axis. No "ceiling effect" of similarities vs distances can explain this ordering for the non-metric approach. Tentatively, the common acoustical feature between / ɚ, ø , œ /, namely the closeness between F2 and F3, is suggested as the third dimension correlate. No evidence for a 4th dimension has been found.

Ce travail présente une revue critique des principales études portant sur l'analyse multidimensionnelle (AMD) de l'espace perceptif dans le domaine des sons vocaliques. Nos résultats personnels sont confrontés à une réanalyse des données publiées par différents auteurs afin de tenter de répondre aux questions suivantes :

- l'analyse multidimensionnelle des indices de (dis)similarité subjective d'un ensemble de voyelles permet-elle d'évaluer le nombre de dimensions de l'espace perceptif ?
- les dimensions correspondent-elles à des paramètres physiques du signal ?
- cette technique, ainsi que le modèle de perception implicite, permettent-ils une approche féconde des problèmes de perception des sons vocaliques ?

I - L'ANALYSE MULTIDIMENSIONNELLE

L'AMD a pour but de mettre en évidence les dimensions psychologiques qui sous-tendent les mesures de (dis)similarité subjective entre des stimuli. Le modèle du système perceptif considéré transforme l'image auditive d'une voyelle en un point dans un espace multidimensionnel de manière que les distances entre les points correspondent "au mieux" aux dissimilarités subjectives. Les hypothèses suivantes sont nécessaires ou utiles pour définir concrètement les données expérimentales et le modèle lui-même :

- L'espace est relativement invariant par rapport au choix des sujets et des conditions expérimentales.
- Les fluctuations temporelles des indices sont aléatoires.
- Le nombre de dimensions est petit comparativement au nombre des stimuli.
- Les indices de (dis)similarité sont fonction des différences sur chacune des dimensions utilisées normalement par le sujet, et non d'un sous-ensemble arbitrairement choisi pour la durée de l'expérience.
- Les dimensions sont indépendantes (orthogonales).
- Les distances de l'espace sont des distances de Minkowski (avec les distances euclidiennes comme cas particulier important).

Nous détaillerons tour à tour les problèmes posés par la collecte des données, le choix de la méthode (métrique ou non-métrique), la détermination du nombre de dimensions et l'interprétation des dimensions extraites.

II - LA COLLECTE DES DONNEES

3 méthodes sont couramment utilisées pour obtenir les indices de (dis)similarité.

A - Estimation à partir d'une matrice de confusion

Deux stimuli sont d'autant plus semblables qu'ils sont plus souvent confondus dans une tâche d'identification. Si la tâche exigée est simple et naturelle, les inconvénients sont importants :

+ les matrices de confusion ne sont pas symétriques. Comme l'a noté JANSON (1977) le problème n'est pas celui d'un biais de réponse. L'asymétrie découle naturellement d'une déviation physique de certains stimuli par rapport à la norme.
+ La présence de nombreuses "cases vides", liée à l'absence de confusions entre des voyelles très différentes rend difficile la transformation du pourcentage de confusions en indice de dissimilarité. Seules des expériences de la taille de celle de PETERSON & BARNEY (1952), analysée par SHEPARD (1972), permettent d'utiliser cette technique. Une dégradation de la qualité des stimuli permet de remplir les cases vides, mais rend la matrice de confusion très asymétrique. Nous étudierons principalement les AMD dont les données ont été recueillies par les méthodes suivantes.

B - Méthode triadique

Pour chaque triplet de stimuli, les sujets doivent indiquer le couple le plus semblable et le plus différent. Le tableau des dissimilarités est formé par l'addition des scores pour chaque couple (par ex. 2 points pour la plus grande paire, 0 pour la petite et I pour la paire intermédiaire). On montre expérimentalement que les distances les plus grandes sont légèrement sous-estimées, ce qui peut être gênant dans une analyse métrique.

C - Echelle d'intervalle ("rating scale") et estimation numérique directe ("free magnitude estimation")

Pour ces deux techniques voisines, le sujet doit :
- indiquer sur une échelle à 7 positions (par ex.) le degré de similarité (I : très semblable ; 7 : très différent)
- ou fournir une estimation numérique, étant donnée une valeur arbitraire attribuée à l'une des paires.

On constate expérimentalement une variabilité assez forte ("bruit") liée à la difficulté de la tâche, ainsi qu'un effet de plafonnement ("ceiling effect") qui biaise négativement les estimations des distances les plus grandes. Cet effet doit être pris en compte dans les analyses métriques. (cf. fig. 4).

III - METHODES D'ANALYSE MULTIDIMENSIONNELLE

Deux méthodes ont été utilisées dans notre domaine : l'analyse non-métrique de KRUSKAL (1964) et l'analyse métrique individuelle de CARROLL & CHANG (1970) "INDSCAL". Une méthode récente, due à TAKANE & al. (1978), n'a jamais été employée. Un examen de leur procédure d'optimisation montre une pondération forte des distances les plus longues, qui sont le plus soumises à l'effet de plafonnement mentionné plus haut.

A - Méthode de KRUSKAL

Elle fait dériver une configuration spatiale multidimensionnelle de n points des n (n-1) /2 indices de dissimilarité entre tous les stimuli pris deux à deux. Ces derniers sont liés aux distances par une fonction monotonement croissante. Cette méthode est non-métrique dans la mesure où seul l'ordre des dissimilarités est utilisé pour le calcul itératif, par une méthode du gradient, de la distance inter-points. On cherche à minimiser le critère suivant, appelé "stress" :

$$s = \left(\sum_{ij}^{n} (d_{ij} - \hat{d}_{ij})^2 / \sum_{ij}^{n} (d_{ij} - \bar{d})^2 \right)^{1/2}$$

avec d_{ij} : distance entre les stimuli i et j ; $\bar{d}$: moyenne des distances
$\hat{d}_{ij}$: valeur numérique d'une série monotonement croissante par rapport aux indices de dissimilarité.

Cette méthode non-métrique ne nécessite pas le calcul d'une constante additive pour transformer les indices en distances. L'orientation des axes dépend arbitrairement de la configuration de départ retenue.

B - Méthode INDSCAL

Elle ajuste itérativement pour chacun des sujets, la matrice B des produits scalaires entre les vecteurs-distances et une estimation de cette même matrice entermes de produit de 3 matrices : $X_1 * X_r$ (produit scalaire des vecteurs de l'estimation) et W (matrice des "poids" individuels qui pondèrent les coordonnées des points sur chacun des dimensions). Cette procédure est métrique car les distances sont calculées dans un espace euclidien. Les indices de dissimilarité doivent donc être transformés en "distances" respectant les axiomes métriques (en particulier $d_{ik} \leq d_{ij} + d_{jk}$) par l'estimation d'une constante additive. Un intérêt de cette méthode est l'orientation unique des axes, si les sujets pondèrent différemment les dimensions qu'ils

utilisent. On démontre de plus que, si les projections des points sur les axes sont indépendantes, la somme des carrés des poids est égale à la variance expliquée par la configuration. Cette analyse individuelle suppose que les données de chaque sujet sont robustes, sans bruit excessif. Il est possible d'analyser les moyennes du groupe, mais l'orientation des axes sera arbitraire.

Comme nous le verrons, l'analyse des données peut être effectuée avec profit par les deux méthodes, afin de cumuler les avantages.

IV - DETERMINATION DU NOMBRE DE DIMENSIONS

C'est, avec l'interprétation des axes, un problème crucial. Plusieurs critères ont été proposés, bien que le problème ne puisse pas être considéré comme résolu.

+ Pour la méthode de KRUSKAL, l'introduction d'une dimension supplémentaire doit faire chuter la valeur du stress ; mais on ne connaît pas le seuil significatif. On dispose seulement de tables, obtenues par simulation, indiquant la valeur du stress en fonction du nombre de points et de dimensions pour des indices de dissimilarité aléatoires (cf. SPENCE & YOUNG 1978 pour une revue récente). Les valeurs obtenues ici sont très largement inférieures au seuil déterminé pour des configurations ne comportant que du bruit.

+ Pour la méthode INDSCAL, la quantité de variance expliquée semble être un critère plus pertinent. Tout tableau de distances peut être représenté exactement (variance expliquée = 100 %) en n-1 dimensions, n étant le nombre de points. Soit Vm la variance expliquée associée à n points en m dimensions. Si la variance résiduelle Em n'est due qu'au bruit, une dimension supplémentaire expliquera (100 - Vm) /n-m-1 % de la variance résiduelle. L'introduction d'une dimension de rang m + 1 doit conduire à une variance expliquée supérieure à Vm + Em. La figure 1 permet de constater que les données de BEECKMANS (1974) ne permettent pas d'extraire plus de 3 dimensions significatives. Malheureusement, le calcul de la variance expliquée par la somme des carrés des poids n'est exact que si les dimensions sont indépendantes ; une corrélation, même faible, fait décroître cette quantité. L'introduction d'une dimension significative peut conduire à un accroissement non-significatif de la variance expliquée, mais l'inverse n'est pas vrai.

En outre, la variance expliquée dépend de la constante additive (CA) choisie, même lorsque les axiomes métriques sont satisfaits (fig. 2). La fig. 3 permet d'éclairer cette liaison. Si 3 points sont alignés avec une constante additive x (schéma a), un choix différent de la CA (par ex. $y = x + 1$) donne la configuration (b). Si en (a) la variance sur D1 est de 100 %, elle est inférieure en (b) car la variance sur D2 n'est plus nulle. Pour une configuration quelconque de n points, une CA très grande place les points sur une hypersphère à n-1 dimensions. Le nombre important de dimensions extraites par TERBEECK (1977) s'explique en partie par le choix de larges CA. A l'inverse, si la configuration "réelle" est celle du schéma (b), l'utilisation de la CA x conduira à la perte d'une dimension. Ce problème délicat doit inciter à comparer toute analyse métrique à une analyse non-métrique.

+ L'ajustement d'un nombre important de dimensions avec INDSCAL fait apparaître des poids négatifs dont Mac CALLUM (1976) a montré qu'ils résultent d'une non-indépendance des dimensions. L'apparition de nombreux poids négatifs est un signal de danger qui signale une violation des hypothèses de l'analyse métrique.

+ L'effet de plafonnement peut conduire à l'extraction d'une dimension supplémentaire artefactuelle dont le seul rôle est de "linéariser" la fonction liant les indices de dissimilarité et les distances euclidiennes de l'espace. Un contrôle par une méthode non-métrique est à nouveau souhaitable.

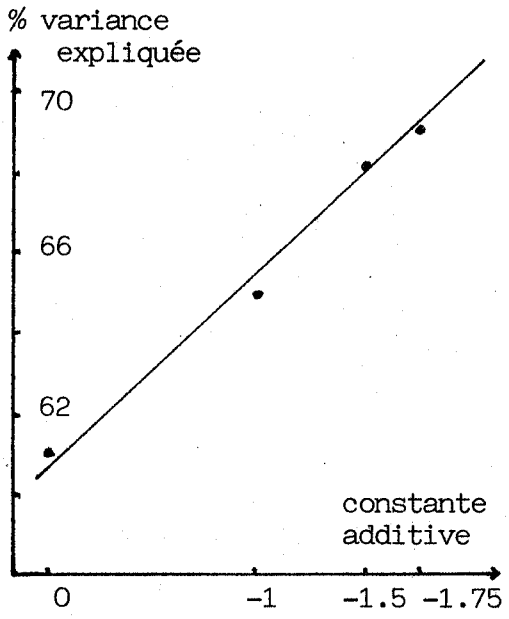


Fig. 2: Variation du % de variance expliquée global en fonction de la constante additive retenue (données de SINGH & WOODS - INDSCAL 3 dimensions)

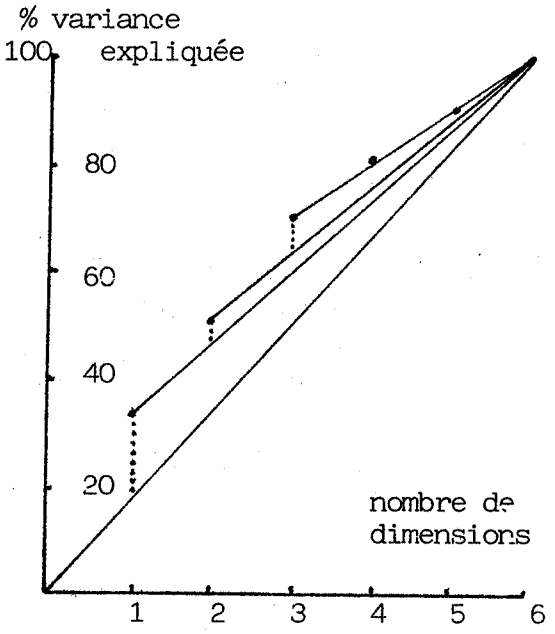


Fig. 1: ..... accroissement significatif de la variance expliquée
------ accroissement théorique de la variance en fonction du nombre de dimensions introduites. Les résultats suggèrent une analyse tridimensionnelle (données de BEECKMANS 1974)

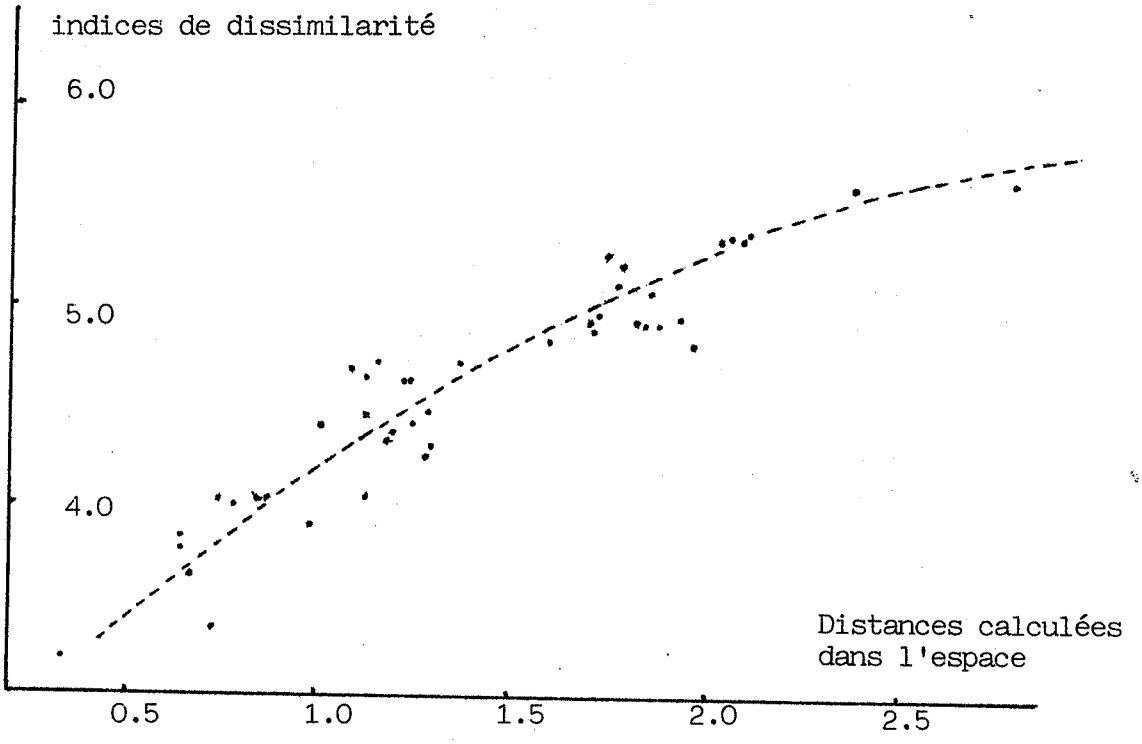


Fig 4: Relation non-linéaire entre les distances spatiales calculées et les indices de dissimilarité ("effet de plafonnement"); données de LONCHAMP, méthode de KRUSKAL en 2 dimensions, stress = 4.9 %

+ Pour les deux méthodes, on réduira les plans où les stimuli apparaissent placés en "fer à cheval" à une seule dimension. (SHEPARD 1974). Nous avons montré dans notre thèse que 6 dimensions extraites par TERBEECK devaient, par ce seul critère, être ramenées à 3 dimensions.

+ Le critère le plus utile nous semble être le critère de stabilité : on acceptera une dimension comme significative si des analyses répétées, à partir de configurations initiales différentes par deux méthodes, placent les stimuli dans un ordre voisin sur cette dimension.

V - EXAMEN COMPARE DE 5 ANALYSES MULTIDIMENSIONNELLES

Les caractéristiques essentielles des 5 études retenues sont présentées à la table I. On notera que l'étude de MOHR & WANG n'est pas une analyse multidimensionnelle. Nous avons retenu les études dont les résultats bruts étaient publiés et dont les données ne provenaient pas de stimuli dégradés ou de l'utilisation de matrices de confusion. Les résultats essentiels figurent également à la table I. Pour les analyses en 3 dimensions, on a effectué 15 tentatives pour la méthode de KRUSKAL et 10 pour la méthode INDSCAL avec des estimations initiales aléatoires de la troisième dimension. Les variances expliquées sont les valeurs les plus élevées obtenues pour les moyennes de groupe. Les rangs moyens pour DIII proviennent des solutions sans poids négatifs pour INDSCAL, et des 4 analyses où le stress était le plus faible pour la méthode de KRUSKAL. Les constantes additives choisies maximisent la quantité de variance expliquée en deux dimensions. Les figures 5 et suivantes présentent graphiquement quelques configurations obtenues. La taille limitée de cet article ne permet pas de présenter graphiquement tous les résultats. Pour les configurations obtenues par INDSCAL, les variances sur chaque axe sont égales et ne reflètent pas les poids donnés par les sujets.

A - LES DIMENSIONS I & II

Dans le plan formé par les deux premières dimensions, la quantité de variance expliquée peut atteindre plus de 90 % (SHEPARD (1972) : 97 % ; TERBEECK sujets américains : 91 % ; LONCHAMP 90 %). Toutes les études publiées à ce jour montrent que la position des voyelles dans le plan DI-DII correspond à leur position dans le plan physique des deux premiers formants. POLS & al. (1969), BEECKMANS (1974) et d'autres ont illustré ce résultat, que l'on peut constater également sur les figures 5 et suivantes. L'analyse des données individuelles par INDSCAL permet d'affirmer que l'orientation des dimensions DI et DII est très proche de celle des axes formantiques FI et F2. Il faut noter cependant avec BEECKMANS (1977) qu'on ne peut affirmer que FI et F2 sont extraits par le système auditif et utilisés dans la dérivation des indices de dissimilarité. Toute combinaison de paramètres physiques aboutissant au même ordre sur les dimensions est un candidat possible pour l'interprétation. Pour DII, par exemple, on peut songer, à la suite des travaux de CARLSON, FANT & al. (1975) à F'2, combinaison complexe de F2, F3 et F4. Comme on pouvait s'y attendre, les correspondances entre DI-DII et FI-F2 sont moins nettes pour les études utilisant des stimuli dégradés ou des matrices de confusion (BEECKMANS 1977, WILSON & BOND 1977, Klein & al. 1969, POLS & al. 1969, Van der KAMP & POLS 1971) en raison de l'asymétrie des tableaux de données.

Comme l'avait déjà noté HANSON (1967), les points semblent également se répartir en une structure quasi-circulaire. En particulier, la voyelle /y/ ne se trouve pas sur une droite joignant les voyelles /i/ et /u/, comme le laisseraient prévoir les fréquences formantiques. En AMD, on obtient une structure circulaire lorsque les indices de dissimilarité ont tendance à se répartir en deux groupes : indices faibles pour les couples de points contigus, indices

	LONCHAMP	MOHR & WANG	HANSON	SINGH & WOODS	TERBEECK
Type de données	Echelle intervalle 7 positions	Echelle intervalle 6 positions	Estimation numérique directe	Echelle intervalle 7 positions	Méthode triadique
sujet : nombre et langue	36 français 4 étrangers	80 américains	20 suédois (4 replications)	18 américains	8 américains 7 allemands 6 thais 7 turcs
méthode analyse de l'auteur		-----	+ composantes principales + KRUSKAL	KRUSKAL	PARAFAC = INDSCAL
résultats auteur pour D III		-----	contraste perceptif correl. F 3	rétroflexe/non rétroflexe	3 à 5 dimensions selon les langues. 9 labels différents
données individuelles	oui	non	oui	non	oui
structure circulaire 2 & 3 dimensions	oui	oui	oui	oui	oui
orientation DI & DII fixée	FI - F2	---	FI - F2	---	FI - F2
orientation après rotation		FI - F2		FI - F2	
KRUSKAL stress 2 dimensions	4.9 %	3.3 %	6.1 %	8 %	---
KRUSKAL stress 3 dimensions	1.2 %	< 1 %	< 1 %	4.2 %	3.5 % sujets américains
INDSCAL variance expliquée 2 dimensions	90 %	86 %	71 %	79 %	84, 89, 82, 88 % (91 % sujets américains seuls)
INDSCAL variance expliquée 3 dimensions	93 %	85 %	77 %	84 %	85, 90, 89, 89 %
Variance expliquée - variance théorique 3 dimensions	1.5 %	(- 3.5 %)	1 %	2.5 %	---
nombre solutions poids négatifs 3 dimensions	1/10	0/10	1/10	0/10	5/10
constante additive	-3	-1.25	-7	-1.5	0
INDSCAL poids 3 dimensions	0.64 0.70 0.17	0.81 0.38 0.23	0.80 0.29 0.20	0.75 0.46 0.26	0.58 0.71 0.12/0.63 0.70 0.13/0.37 0.81 0.32/0.81 0.47 0.09/
D III rang moyen INDSCAL	eioɛuaɔøyœ	æi-ueoyɔœø	øɛɯyueoia	ɞɷʌæɛɑ-ɪɔueɨ-o	idɑuyeoæʌɨø ɾ
D III rang moyen KRUSKAL	eaouɛiɔøyœ	iæuɔeøyœo	øɛoueɑɯyi	ɞɔæʌɛoɷɑ-iɪue	eoyuɨɑæʌɨøɑˈɾ

TABLE I

élevés pour les autres couples. Cette constatation, qui n'est pas incompatible avec l'effet de plafonnement illustré à la figure 4, peut être interprétée comme l'action d'un "effet de proximité". La méthode triadique, qui exige une estimation plus grossière que la technique de l'échelle d'intervalle, semble de ce fait préférable. La conservation de cette structure circulaire lorsque l'on introduit une troisième dimension indique que le nombre de dimensions n'est pas en cause. Il apparaît que les indices de dissimilarité sont largement fonction de deux paramètres qui pourraient être la fréquence du premier formant et F'2. Mais les indices reflètent également un effet de plafonnement et/ou un effet de proximité qui modifient quelque peu la configuration spatiale calculée.

B - LA DIMENSION III

Si l'unanimité se fait sur l'existence et l'interprétation des deux premières dimensions, il n'en va pas de même pour les dimensions de rangs supérieurs, et en particulier pour la troisième dimension. TERBEECK propose de 3 à 5 dimensions selon les langues, et même 6 pour la matrice globale ; WILSON & BOND et ANGLIN en extraient 4, HANSON 3 ...

Le pourcentage de variance expliquée est faible, particulièrement sil'on retranche la quantité théorique Em de l'accroissement : il ne dépasse pas dans ce cas 3 %. Il ne reste d'ailleurs dans certains cas que 10 % de variance à expliquer. Même si on analyse des moyennes de groupe, la présence plausible de "bruit", lié à la difficulté de la tâche, risque de masquer l'existence ou de gauchir l'interprétation de cette dimension.

Néanmoins toutes les études dont l'inventaire comprend la voyelle (américaine) retroflexe / ɚ / concluent à l'existence d'une troisième dimension (WILSON & BOND, ANGLIN, SHEPARD (1972)et les travaux réanalysés ici). La voyelle / ɚ / apparaît systématiquement à une extrémité de l'axe.

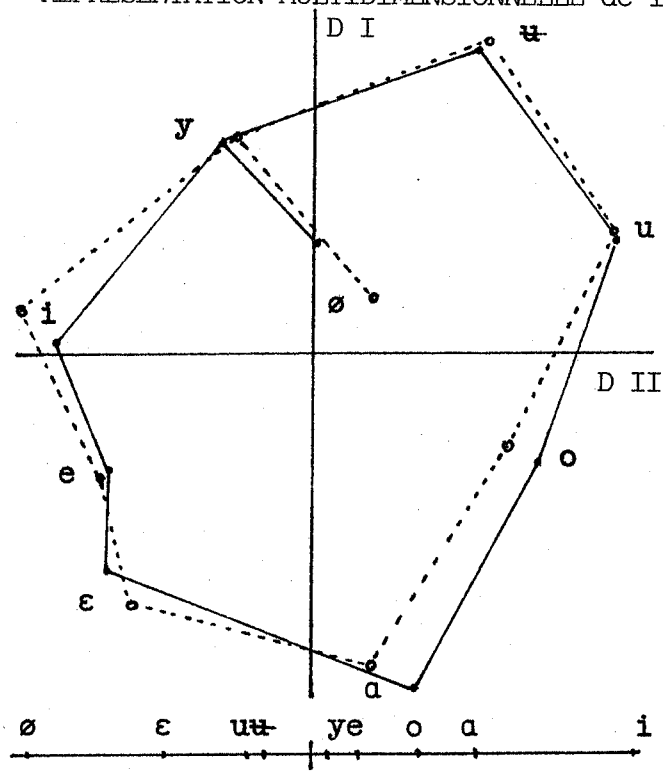
Notre analyse des données de SINGH & WOODS et TERBEECK confirme ce point, quelle que soit la méthode utilisée. A l'autre extrémité de l'axe se place un groupe de voyelles (**i, o , ɑ , u** et **e**) dont l'ordre est assez variable. Aucune explication en termes de paramètres physiques, à l'exception d'une dichotomie rétroflexe/non-rétroflexe, n'est suggérée par cet ordre. On remarquera que le critère de variance expliquée ne suggère pas clairement la présence d'une troisième dimension. En revanche, la stabilité des résultats pour les deux méthodes est un indice moins équivoque. Il n'est pas possible d'incriminer l'effet de plafonnement ou la violation des hypothèses métriques. Rappelons que l'ordre des voyelles semble indépendant des conditions d'initialisation.

Pour les trois études ne comportant pas de voyelles rétroflexes nous avons défendu dans notre thèse l'hypothèse de la bi-dimensionnalité de l'espace perceptif. Or il nous faut constater, à la lumière d'une analyse plus détaillée, que plusieurs voyelles montrent une stabilité certaine sur la troisième dimension. Il s'agit des voyelles antérieures arrondies/ø, œ/et/y/, qui se regroupent à une extrémité de l'axe. De plus, l'accroissement de la variance expliquée, même s'il est faible, mérite une explication. La position des autres voyelles est très variable.

Deux hypothèses explicatives peuvent être émises :

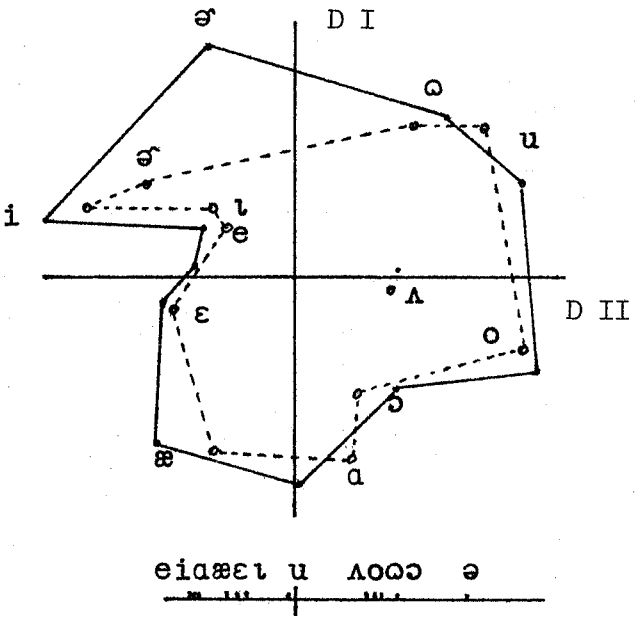
+ Les voyelles/ø, œ/et /ɚ/sont situées près du centre de la structure circulaire. Si l'effet de plafonnement joue un rôle, les "diamètres" seront sous-estimés par rapport aux "rayons". Le "moyeu" de cette configuration se trouvera légèrement en dehors du plan formé par les deux premières dimensions, pour permettre que deux rayons soient supérieurs au diamètre. Des simulations sur des données exactes incorporant un fort effet de plafonnement réfutent cette interprétation. La méthode de

REPRESENTATION MULTIDIMENSIONNELLE de l'ESPACE PERCEPTIF des VOYELLES



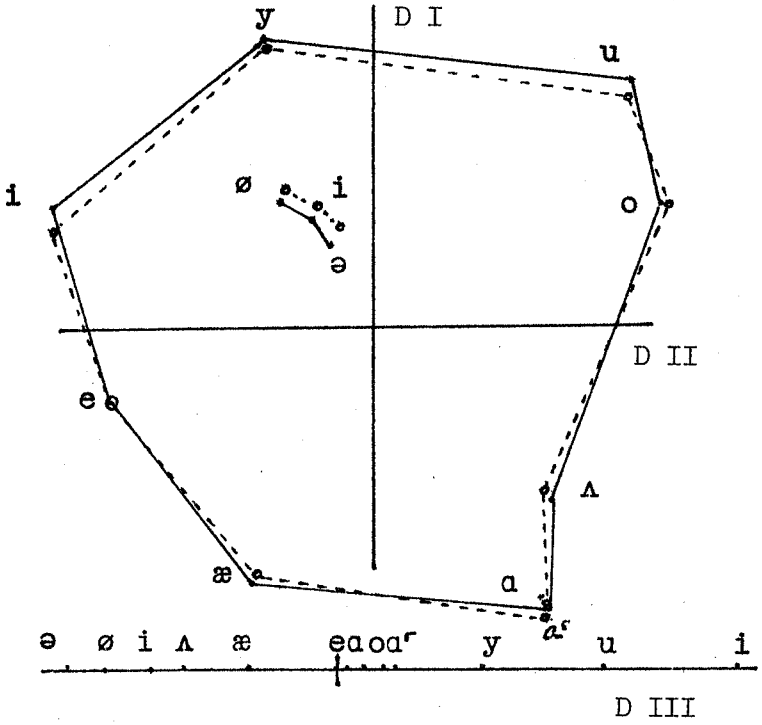
..... 2 dimensions
------- 3 dimensions

Fig. 7: données de HANSON (INDSCAL)



..... 2 dimensions
------- 3 dimensions

Fig. 8: données de SINGH & WOODS (KRUSKAL)



..... 2 dimensions
------ 3 dimensions

Fig. 9: données de TERBEECK (INDSCAL)

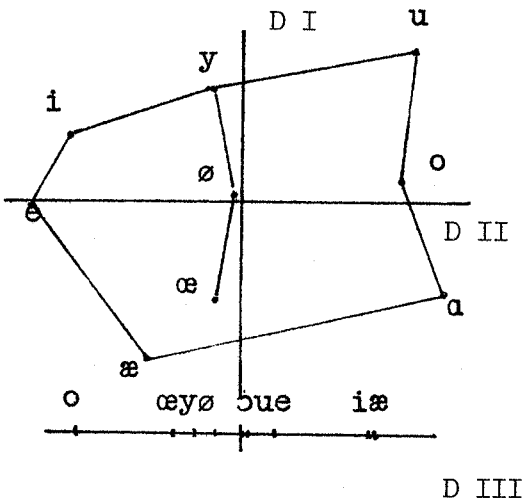


Fig. 10: données de MOHR & WANG (KRUSKAL 3 dimensions)

KRUSKAL n'est en effet sensible qu'à l'ordre des distances et non à leurs valeurs absolues. Or l'effet de plafonnement ne modifie pas radicalement l'ordre des indices de dissimilarité.

+ Les voyelles antérieures labialisées et/ɶ/ ont une caractéristique acoustique commune : la proximité de F2 et F3. Il est donc possible que cette caractéristique conduise les sujets à majorer quelque peu les dissimilarités perçues entre ces voyelles centrales et les voyelles "périphériques". Il convient assurément de procéder à des expériences complémentaires pour vérifier cette hypothèse. On notera cependant la proximité de /ɶ/et/ø/dans les résultats de TERBEECK.

Avant de conclure, les points suivants méritent d'être notés :
- l'analyse des comportements individuels, qui ne semble pas mériter de remarques particulières, est difficile en raison du caractère bruyant des réponses.
- un essai d'analyse des données de SINGH & WOODS en 4 dimensions pour réduire encore la valeur du stress a donné des résultats très instables qui interdisent toute interprétation.
- l'examen des interprétations proposées par différents auteurs pour la troisième dimension indique que celles-ci ne sont pas compatibles avec les résultats de notre analyse.

VI - CONCLUSIONS

L'espace perceptif des sons vocaliques ne semble pas comporter plus de trois dimensions. Mais il ne faut pas perdre de vue que les indices de dissimilarité d'une douzaine de points ne permettent guère de "supporter" plus de trois dimensions. On ne peut exclure absolument que les sujets utilisent en fait un plus grand nombre de dimensions. Il est également théoriquement envisageable que les sujets utilisent des dimensions "personnelles" dont les méthodes utilisées ici ne peuvent rendre compte. La limite des AMD serait à chercher principalement dans les méthodes de collecte des données trop frustres ou bruyantes, ainsi que dans des algorithmes peu robustes, qui spécifient mal (analyse métrique) ou trop peu (analyse non-métrique) la fonction liant les indices de dissimilarité aux distances inter-points.

Néanmoins, ces résultats nous paraissent plausibles pour les raisons suivantes :

+ On sait que les deux ou trois premières fréquences formantiques permettent de caractériser les voyelles de façon adéquate.
+ Le nombre des phonèmes vocaliques est souvent moins important que les inventaires analysés ici. Un grand nombre de langues n'ont que 5 voyelles et d'autres moins encore. Il faudrait expliquer pourquoi ces langues négligent des possibilités d'opposition sur des dimensions potentielles indépendantes.

Les expériences réalisées en vue d'AMD devraient, à nos yeux, tester les conditions suivantes :
+ nombre plus important de voyelles naturelles et synthétiques
+ nombre plus important de sujets et replication des observations pour chaque sujet, pour permettre une analyse individuelle
+ analyse d'espaces perceptifs plus restreints, en particulier dans la zone des voyelles antérieures.

Le but de ces expériences serait de vérifier le nombre de dimensions réellement utilisé par les sujets.

Comme l'écrit SHEPARD (1974) "J'ai pris conscience peu à peu que la plupart des découvertes faites à l'aide d'analyses multidimensionnelles auraient pu l'être sans elles". Mais les progrès de nos connaissances dans le domaine de la perception des sons du langage ne sont pas si rapides qu'on puisse négliger une source potentielle d'informations.

o o o o o o o o o o

ANGLIN M. (1971) cité in SINGH S. (1976) Distinctive features : Theory and validation, University Park Press.

BEECKMANS R. (1974) "Influence de la fréquence fondamentale sur l'espace perceptif des voyelles" RA 8-I/2 Inst. de Phon. U.L. Bruxelles, 109-128.

BEECKMANS R. (1977) "Une approche du processus d'identification des voyelles en français" RA 11/I Inst. de Phon. U.L. Bruxelles 69-82.

CARLSON R., FANT G., GRANSTÖM B. (1975) "Two-formant models, pitch and vowel perception" in FANT G. & TATHAM M. Auditory analysis and perception of speech, Academic Press, 55-82.

CARROLL J.D. et CHANG J.J. (1970) "Analysis of individual differences in MD scaling via an N-way generalization of Eckhart-Young decomposition" Psychometrika 35, 283-319.

HANSON G. (1967) "Dimensions in speech perception" Ericsson Technics I, I-75

JANSON T. (1977) "Asymmetry in vowel confusion matrices", J. of Phon. 5, 93-96.

KLEIN W. PLOMP R. POLS L.C.W. "Vowel spectra, Vowel space and vowel identification" JASA 48, 999-1009

KRUSKAL J.B. (1964) "Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis" Psychometrika 29, 1-27 ; "Nonmetric multidimensional scaling : a numerical method" 29, 115-129.

LONCHAMP F. (1978) Recherches sur les indices perceptifs des voyelles orales et nasales Thèse de 3ème Cycle, Nancy.

Mac CALLUM R.C. (1976) "Effects on INDSCAL of non-orthogonal perceptions of object space dimensions" Psychometrika, 41, 177-188

MOHR B. et WANG W.S.Y. (1968) "Perceptual distance and the specification of phonological features" Phonetica 18, 31-45.

PETERSON G.E. et BARNEY H.L. (1952) "Control method used in a study of the vowels" JASA 24, 175-184.

POLS L.C.W. Van der Kamp L.J. Th. PLOMP R. (1969) "Perceptual and physical space of vowel sounds" JASA 46, 458-467.

SHEPARD R.N. (1972) "Psychological representation of speech sounds" in DAVID E. D. et DENES P.B. Human communication : a unified view, Mc-Graw-Hill, 67-113.

SHEPARD R.N. (1974) "Representation of structure in similarity data : problems and prospects" Psychometrika 3, 373-421.

SINGH S. et WOODS D.R. (1971) "Perceptual structure of 12 American English vowels" JASA 49, 1861-1866.

SPENCE I. et YOUNG F.W. (1978) "Monte Carlo studies in nonmetric scaling" Psychometrika 43, 115-117.

TAKANE Y., YOUNG F.W., de LEEUW J. "Nonmetric individual differences multidimensional scaling : an alternating least squares method with optimal scaling features" Psychometrika (1977) 42, 7-67.

TERBEECK D. (1977) A cross-language multidimensional scaling study of vowel perception W.P.P. UCLA N° 37.

Van der KAMP L.J. Th. et POLS L.C.W. (1971) "Perceptual analysis from confusions between vowels" Acta Psychologica 35, 64-77.

WILSON H.F. et BOND Z.S. (1977) "An INDSCAL analysis of vowels excerpted from 4 phonetic context", J. of Phon. 5, 361-367.

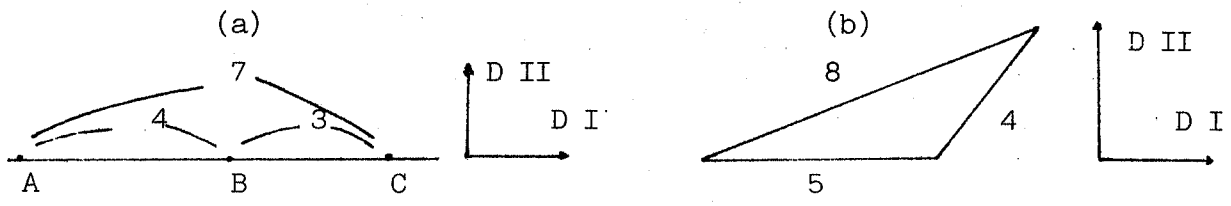


Fig. 3: Variation du % de variance expliquée sur D I en fonction de la const. additive retenue. a) const. add. = x, variance expl. (D I) = 100 %
b) const. add. = x+1, variance expliquée (D I) 100%

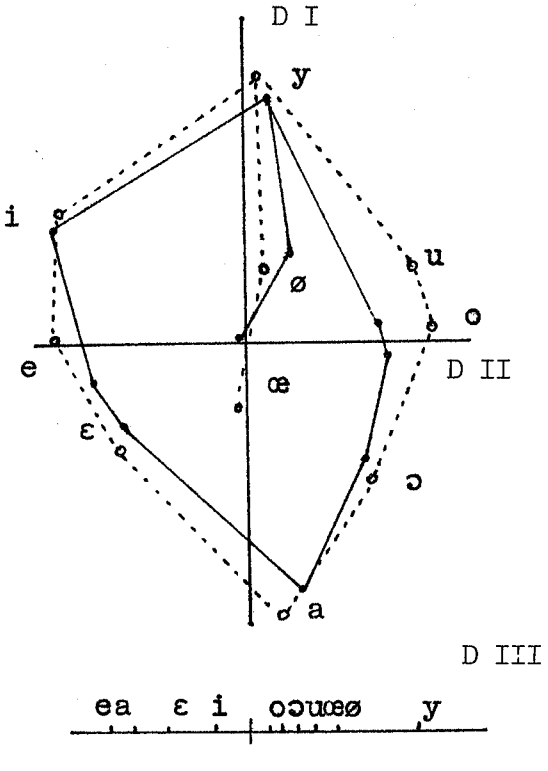


Fig. 5 LONCHAMP (KRUSKAL)

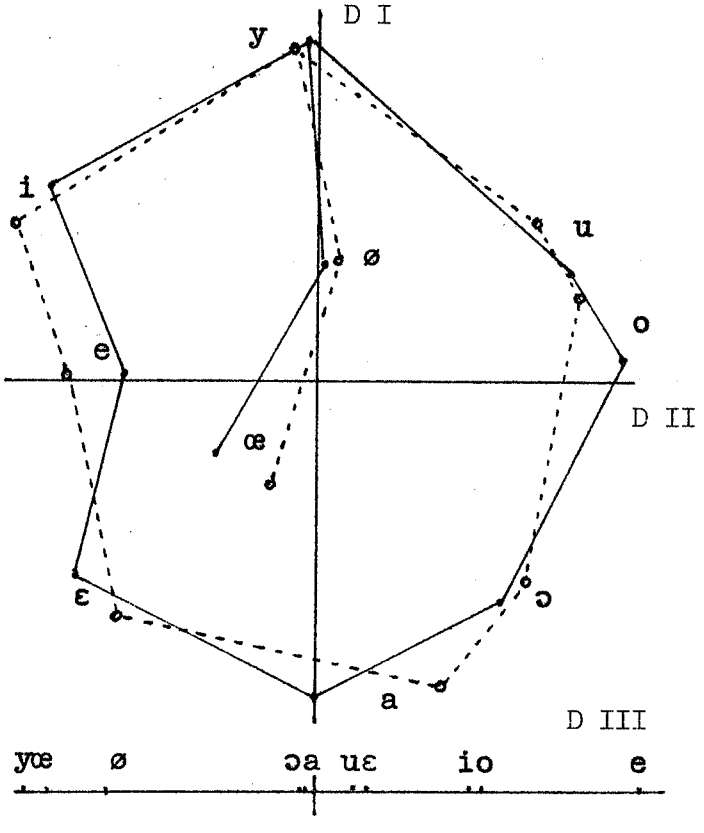


Fig. 6 LONCHAMP (INDSCAL)

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

ESTIMATION DE LA DUREE INTRINSEQUE DES VOYELLES

Y. NISHINUMA — Institut de Phonétique (CNRS LA 261)
S. BARBER — Université de Provence
D. J. HIRST — France

.RESUME

Afin d'obtenir une méthode simple d'estimation de la durée intrinsèque des voyelles, nous avons étudié des données expérimentales portant sur trois langues (américain, anglais et japonais) ainsi que d'autres publiées par divers auteurs portant sur 6 langues différentes (allemand, américain, anglais, français, hollandais et italien). L'analyse de regression multiple a été effectuée principalement avec les paramètres acoustiques : Fo, F1, F2 et F3. Les résultats montrent que les F1, F2 ainsi que la variable calculée Dq donnent une bonne estimation.

. ABSTRACT

With the intent of formulating a simple method to estimate the intrinsic duration of vowels, we examined 3 groups of experimental data (American and British English and Japanese) and 10 groups of published data from various works concerning 6 languages - Dutch, English (American and British), French, German, and Italian. By adopting a linear model hypothesis, we carried out a multiple regression analysis using mesured vowel duration as the dependent variable, and its corresponding Fo, F1, F2 and F3 as independent variables. We also introduced a computed variable Dq representing sub-categorical mean duration, which allowed us, in the estimation formula, to specify the mean size of the calculated duration. Results show that, due to the intercorrelations of variables, Fo and F3 were eliminated, and F1, F2 and Dq gave a quite good estimation of the duration for all data analysed. Therefore we think that the intrinsic duration of vowels can be satisfactorily estimated by means of these three variables, at least for the languages considered.

INTRODUCTION

L'objectif de la présente étude porte essentiellement sur la détermination prévisionnelle de la durée intrinsèque des voyelles. Cette étude fait partie intégrante du projet en cours de recherche sur la prosodie. Nous avons, en effet, développé un modèle d'analyse prosodique pour le français (NISHINUMA et ROSSI, 1979), dans lequel les voyelles sont corrigées, entre autres, en fonction de la durée intrinsèque et co-intrinsèque (effet consonantique sur la voyelle). Les coefficients nécessaires avaient été déterminés au préalable expérimentalement (DI CRISTO, 1976). Or, en vue d'une éventuelle application du modèle à une nouvelle langue, il nous a paru indispensable de trouver une méthode d'évaluation de ces coefficients pour la langue concernée, à partir des données obtenues de l'analyse acoustique relativement simple.

Notre démarche repose sur les faits suivants bien connus. On sait, en nous référant aux systèmes phonétiques jusqu'à présent étudiés, que la durée intrinsèque a une distribution statistique propre à chaque voyelle. De même, les voyelles se caractérisent spectralement par les paramètres acoustiques relativement bien définis, notamment par les formants. A partir de ces deux faits, il est statistiquement possible d'établir une relation entre la durée et les autres paramètres acoustiques et de faire une estimation de celle-ci.

PROCEDURE

Nous cherchons la fonction pour laquelle la durée estimée $\hat{D}$ s'approche de la durée observée Do avec la plus petite marge d'erreur. Confrontés au problème de trouver la méthode et le modèle adéquats, nous avons procédé à priori à l'analyse de regression multiple, dont la structure mathématique est linéaire :

$$\hat{D} = A_1P_1 + A_2P_2 + \ldots\ldots + A_nP_n + C \qquad (1)$$

Rappelons que, dans notre cas, $\hat{D}$ est la variable dépendante, les P sont les variables indépendantes (valeurs observées des paramètres acoustiques à n dimensions), les A sont les coefvicients attribués à ces dernières et C représente une constante d'ajustement.

DONNEES

Notre analyse s'est effectuée principalement sur trois corpus expérimentaux dont la composition est la suivante. Le premier comportait 6 voyelles anglaises britanniques /i, ɑ, u ; I, ʌ, ʊ/ précédées de la consonne /m/ et suivies des consonnes /m,p,b,s,z/. Cette séquence CVC a été enchassée dans une phrase porteuse sous forme de logatome à mono- et bi-syllabe. Un des auteurs de l'article a répété à trois reprises l'ensemble des fiches contenant l'exemple. Nous avons obtenu 270 échantillons. Le corpus de l'anglais américain d'un nombre de

voyelles réduit soit /i, ɑ, u/ et 7 consonnes postposées /m,p,b,k,g,s,z/ a été réalisé par une locutrice entraînée (originaire de Californie, co-auteur de la présente étude). Le nombre de voyelles s'élève à 189 au total. Le troisième corpus, concernant le japonais, contenait 5 voyelles combinées avec 3 consonnes pré- et post-posées /b,d,g/. Cinq locuteurs de Tokyo ont lu le corpus 4 fois, soit au total 500 voyelles.

Sur ces corpus réalisés, nous avons effectué l'analyse acoustique assistée par ordinateur ; à savoir l'acquisition des données (16 bits, 10 KHz), la segmentation et l'analyse de cepstre. Sous l'éditeur de signal, la durée a été segmentée et simultanément vérifiée perceptuellement par la reproduction synthétisée (ESPESSER, 1978). Nous avons mesuré les valeurs de fréquence : Fo, F1, F2 et F3 et d'intensité : Io, I1, I2 et I3 au centre de la voyelle avec une fenêtre de 25 ms.

Afin de comparer nos résultats avec ceux obtenus à partir des données de langues différentes, nous retenons, parmi les données publiées exploitables, les travaux de DOHALSKA-ZICHOVA (1974) pour le fraçais, ceux de FERRERO et al (1975) pour l'italien, enfin ceux de KOOPMANS-van BEINUM (1980) en hollandais. En ayant pour but de combler partiellement cette lacune et bien que le procédé soit criticable, nous avons synthétisé quelques ensembles des données "hybrides" en rassemblant deux sources d'information différentes (durée et formants) pour les langues suivantes : anglais, français et allemand (voir Tableau 1).

RESULTATS D'ANALYSE

Pour l'analyse des données expérimentales de l'anglais britannique, nous obtenons 90 observations : pour chaque voyelle une observation représente la moyenne des 3 répétitions. Les variables indépendantes de départ sont les 8 paramètres acoustiques de fréquence et d'intensité ci-dessus mentionnés : on y ajoute une variable artificielle que nous appelons Dq. Du fait de la distribution bimodale des voyelles brèves et longues, celle-ci a été calculée sur la variable dépendante (Do) afin de différencier ces deux catégories des voyelles. Le tableau 2 récapitule les indices statistiques, soit : R (coefficient de corrélation multiple) et R^{**2} (coefficient de corrélation multiple doublement ajusté). La multi-corrélation entre la durée et l'ensemble des 8 paramètres acoustiques est assez bonne (R = 0.87), et ceux-ci expliquent 76 % de la variation des données en durée ($R^2 = 0.87^2 = 0.76$). Toutefois, R ne bouge pas et R^{**2} (indicateur réel de la précision de l'estimation) augmente significativement quand les variables de l'intensité sont supprimées, étant donné que celles-ci n'ont pas dépassé le seuil de validité du paramètre t (t = 2.0). Donc, l'apport des variables de l'intensité est négligeable. Par ailleurs, les deux variables Fo et F3 entravent la corrélation des Dq, F1 et F2 avec la durée. Du fait de ces interactions entre les variables indépendantes, seules les variables Dq, F1, F2 et F3 sont retenues pour l'analyse définitive, dont les résultats figurent au tableau 2 (I). Ces derniers montrent que la combinaison Dq-F1 donne le meilleur rapport de

corrélation multiple, mais au point du vue de la précision d'estimation, la combinaison Dq-F1-F2 est la meilleure.

Le tableau 2 (II) indique que pour les données expérimentales de l'américain, le rôle des paramètres des Fo et Io ne contribue pas efficacement à l'estimation de la durée, car l'indice R^{**2} ne varie pratiquement pas à cause de leur interaction avec le F1.

En ce qui concerne les données expérimentales du japonais, nous avons introduit comme variable indépendante Dq, la durée moyenne de chaque locuteur. Le tableau récapitulatif montre qu'entre les cas 1 et 5, le coefficient R^{**2} ne varie pas de façon significative, en raison des intercorrélations des variables indépendantes. Finalement la combinaison des variables Dq, F1 et F2 donne une meilleure estimation de la durée.

De ces trois ensembles de données, nous retiendrons F1 et F2 comme paramètres acoustiques les plus utiles à l'estimation de la durée intrinsèque des voyelles. La formule (1) prend donc la forme suivante :

$$\hat{D} = a_o D_q + a_1 F_1 + a_2 F_2 + C \qquad (2)$$

La figure 1 représente graphiquement l'estimation par l'équation résultante de l'analyse pour les trois cas expérimentaux : les points reliés indiquent les valeurs moyennes observées, ceux non reliés, les valeurs estimées.

Pour les données hybrides de l'anglais (de 1) à 4) du tableau 1), nous avons obtenu la meilleure estimation avec les variables Dq-F1-F2 pour lesquelles le t calculé a dépassé le seuil adopté (t = 0.2). Par contre en allemand et en italien, le meilleur résultat a été obtenu par la combinaison des variables Dq-F1. Il en est de même pour le français et le hollandais. Si l'on analyse les trois premières langues comportant le F3, il s'est avéré que, d'une part, l'apport du F3 est non-significatif, d'autre part, la combinaison des variables Dq-F1 fournit toujours le meilleur résultat statistique. D'après une analyse globale des 5 langues sur 3 variables Dq-F1-F2, la contribution du F2 n'a pas été statistiquement significative. En conclusion, l'estimation de la durée peut s'effectuer à partir des variables Dq-F1-F2 ou Dq-F1. Nous illustrons à titre d'exemple les résultats d'estimation pour l'anglais (Figure 2).

Pour le français, nous avons extrait des données de DOHALSKA-ZICHOVA (1974) les moyennes des : Do, F1 et F2 pour 16 voyelles réalisées par 4 locuteurs et introduit la variable Dq afin de différentier les catégories orale/nasale. La corrélation obtenue est de l'ordre de 0.91. Par contre, les données de DI CRISTO combinées aux valeurs formantiques de DELATTRE (1965) fournissent une meilleure corrélation pour les variables combinées de : Dq, F1, F2 (Figure 3).

Des résultats analogues ont été observés pour les trois autres langues : allemand, hollandais et italien (Figure 4).

A titre d'essai, nous avons procédé à l'analyse globale des données des 5 langues. Les variables indépendantes utilisées sont : Dq, F1 et F2. Les résultats de l'analyse montrent une bonne corrélation avec un score de R = 0.97 ainsi que R**2 = 0.94 pour la meilleure combinaison des données. Dans ce dernier cas, l'équation d'estimation de la durée (2) prend les coefficients suivants :

$$\hat{D} = 0.973\ D_q + 0.059\ F_1 - 0.00848\ F_2 - 11.07 \quad (3)$$

L'écart-type sur erreur d'estimation que donne la formule (3) est de l'ordre de 16 ms, ce qui représente un peu moins de 10 % de la durée moyenne générale. Signalons que le t calculé pour le F1 est significatif au niveau de 0.01 et le F2 au niveau de 0.05.

DISCUSSION

Nous venons d'observer dans plusieurs cas l'intercorrélation de la Fo avec le F1 ; et celle-ci n'a pas obtenu un apport substantiel dans l'analyse statistique concernée. Le problème analogue se pose entre le F3 et F2, bien que le coefficient d'intercorrélation n'est pas aussi élevé que dans le cas précédent. On sait que dans l'analyse de régression multiple, l'intercorrélation provoque un effet négatif pour la corrélation entre la variable dépendante et l'ensemble des variables indépendantes. Cet effet évoqué a été statistiquement explicité par notre étude antérieure (HIRST et al, 1979) dans laquelle nous avons démontré qu'une bonne corrélation multiple existe entre la Fo et les formants. Egalement l'intercorrélation entre le F3 et le F2 coïncide avec les travaux de LADEFOGED et HARSHMAN (1979) qui montrent une corrélation multiple entre le F3 et deux autres formants.

L'introduction de la variable indépendante créée Dq qui représente la moyenne de la variable dépendante des catégories opposantes ne semble pas être pleinement justifiable sur le plan statistique théorique. Toutefois, cette variable Dq peut assumer une fonction pratique. En effet, nous avons accès, dans la formule d'estimation, à un terme correspondant à la variable en question. Par le biais de ce terme, il nous est possible de faire varier le centre de distribution de la durée intrinsèque, à savoir : voyelles longues/brèves, orales/nasales etc.

Il est maintenant bien accepté que la durée intrinsèque des voyelles a un rapport physiologique direct avec l'ouverture (mouvement mandibulaire) qui se traduit acoustiquement par le F1. Effectivement, de notre analyse il a été démontré que le F1 est une variable très dominante dans l'estimation de la durée.

DI CRISTO (1978) examine le problème en terme d'analyse de variance : il démontre que dans le sens de l'évolution du F1, la différence en durée entre les voyelles (e.g. /i, ɛ, ɑ/ du français) se trouve significative ; par contre la durée ne diffère pas significativement entre les voyelles avec leur F2

similaire (e.g. /i, y, u/). Ce qui coïncide pour nos résultats à la faible contribution du F2, sauf pour le japonais et l'anglais (données hybrides). Notons que dans les autres ensembles des données, bien que l'indice statistique t du F2 ne se trouve pas significatif, la présence de celui-ci provoque, dans bien des cas, une diminution de l'écart type sur l'erreur d'estimation ainsi que de l'écart entre les valeurs minimales et maximales d'erreur. Il peut apporter donc une légère amélioration dans l'estimation.

L'erreur d'estimation se trouve assez faible, soit un écart type allant de quelques millisecondes à un peu moins de 20 ms, en moyenne une dizaine de millisecondes ; ce qui nous laisse prévaloir une fiabilité de précision dans les résultats escomptés par rapport aux normes explicitées (NOOTEBOOM et DOODMAN, 1980).

CONCLUSION

En résumé, dans le but de concevoir une méthode simple afin d'estimer la durée intrinsèque des voyelles, nous avons étudié 3 corpus expérimentaux et 9 corpus publiés dans différentes revues et ouvrages, portant sur 6 langues différentes. En adoptant l'hypothèse d'un modèle linéaire, l'analyse de regression multilinéaire a été effectuée avec les variables du domaine fréquentiel : F1, F2 et F3 ainsi qu'une variable indépendante calculée Dq qui représente les durées moyennes sous-catégorielles. Selon l'analyse effectuée, nous avons constaté un apport nul du F3 dans la corrélation multiple, en raison de son intercorrélation avec le F2. D'autre part, nous avons obtenu la meilleure multi-corrélation sur critère R**2 avec les variables Dq-F1-F2 pour les données expérimentales du japonais et celles hybrides de l'anglais, ainsi que Dq-F1 pour les autres cas (anglais britannique et américain avec les données expérimentales, français, allemand, hollandais et italien). Du point de vue de la précision dans l'estimation, la combinaison Dq-F1-F2 nous semble la meilleure d'après la vérification. Ainsi pensons-nous avoir pu nous assurer que la durée intrinsèque des voyelles est effectivement prédictible d'une manière satisfaisante au moyen de variables peu nombreuses.

REFERENCES

DELATTRE, P., 1965, Comparing the Phonetic Features of English, French German, and Spanish : An Interim Report, Julius Grous Verlag, Heidelberg.

DI CRISTO, A., 1978, De la microprosodie à l'intonosyntaxique, thèse d'Etat, Université de Provence.

DI CRISTO, A., 1980,"La durée intrinsèque des voyelles du français", Travaux de l'Institut de Phonétique d'Aix, (sous presse).

DOHALSKA-ZICHOVA, M., 1972, Analyse spectrographique des voyelles françaises, Universita Karlova, Prague.

ESPESSER, R., 1978, EDIS : éditeur de signal sur T1600, Institut de Phonétique d'Aix-en-Provence.

FERRERO, E.F., MAGNO-CALDOGNETTO, E., VAGGES, K. & LAVAGNOLI, C., 1975, "Some acoustic and perceptual characteristics of the Italian vowels", 8th International Congress of Phonetic Sciences, (Leeds).

HILL, D.R., WITTEN, I.H. & JASSEM, J., 1978, "Some results from a preliminary study of British English speech rhythm", Man-Machine System Report, (University of Calgary) 78/26/5.

HIRST, D.J., DI CRISTO, A. & NISHINUMA, Y., 1979, "Estimation of the intrinsic Fo : A comparative study, Proceedings of the 9th International Congress of Phonetic Sciences, p. 381.

JORGENSEN, H.P., 1969, "Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache mit einer spezifischen Untersuchung der Struktur ihrer formanterfrequenzen", Phonetica, 19, pp. 217-245.

KLATT, D.H., 1976, "Linguistic uses of segmental duration in English : Acoustic and perceptual evidence", J. Acoust. Soc. Am., 59, pp. 1208-1221.

KOOPMANS-van BEINUM, F.J., 1980, Vowel Contrast Reduction, Academische Pres, Amsterdam.

LADEFOGED, R., HARSHMAN, R., 1979, "Formant frequencies and movements of the tongue", Frontiers of Speech Communication Research, (eds. B. LINDBLOM & OHMAN), Academic Press, London, pp. 25-48.

LEHISTE, I., 1964, Acoustical Characteristics of Selected English Consonants, Mouton, The Hague, pp. 25-34.

LINDBLOM, B.E.F., 1975, "Some temporal regularities of spoken Swedish", Auditory Analysis and Perception of Speech, (eds. G. FANT, M.A.A. TATHAM), Academic Press, London, pp. 387-396.

LISKER, L., 1974, "On 'explaining' vowel duration variation", Haskins Laboratories : Status Report on Speech Research, SR-37/38, pp. 225-232.

NEWEKLOWSKY, G., 1975,"Spezifisch Dauer und spezifische Tonhöhe der Vokale", Phonetica, 32(1), pp. 38-60.

NISHINUMA, Y., & ROSSI, M., 1979, "Essai d'automatisation de l'analyse prosodique du français", Travaux de l'Institut de Phonétique d'Aix-en-Provence, 6, pp. 207-227.

NISHINUMA, Y., 1979, Un modèle d'analyse automatique de la prosodie, Edition du C.N.R.S., Paris.

NOOTEBOOM, S.G., & DOODEMAN, G.J., 1980, "Production and perception of vowel length in spoken sentences", J. Acoust. Soc. Am., 67, pp.276-287.

PETERSON, G.E., & BARNEY, H.L., 1952, "Control method used in a study of the vowels", J. Acoust. Soc. Am., 24, pp. 175-184.

PETERSON, G.E., & LEHISTE, I., 1960, "Duration of syllable nuclei in English", J. Acoust. Soc. Am., 32, pp. 693-703.

SHARF, D.J., 1962, "Duration of post-stress intervocalic stops and preceding vowels", Language and Speech, 5, pp. 26-30.

UMEDA, N., 1975, "Vowel duration in American English", J. Acoust. Soc. Am., 58, pp. 434-445.

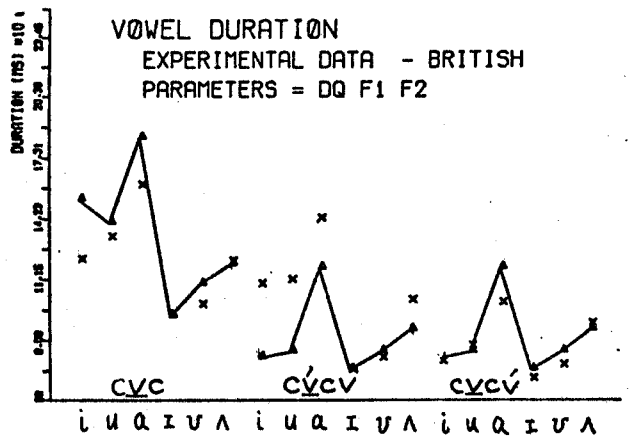
WELLS, J.C., 1962, A Study of Formants of the Pure Vowels of British English, M.A. Thesis, University of London.

Langue	Sources des données	
	Durée	Formants
. Anglais	1) PETERSON & LEHISTE (1960) 2) SHARF (1962) 3) HARRIS & UMEDA (1974) 4) HILL et al. (1978)	LEHISTE (1964) LEHISTE (1964) PETERSON & BARNEY (1952) WELLS (1962)
. Français	5) DOHALSKA-ZICHOVA (1974)* % 6) DI CRISTO (1978)	DELATTRE (1965)%
. Allemand	7) NEWEKLOWSKY (1975)	JØRGENSEN (1969)
. Italien	8) FERRERO et al. (1975)*	
. Hollandais	9) KOOPMANS-van BEINUM (1980)*%	

Tableau 1 - Données homogènes (*) et hybrides analysées. Le symbole (%) désigne les données comportant 2 formants.

Données	Variables indépendantes	Coefficients R	Coefficients R^{**2}	Ecart type sur erreur	Nombre observations
I	Dq F1 F2 F3	0.936	0.712	19.3	90
	Dq F1 F2	0.916	0.718	19.2	
	Dq F1	0.905	0.722	19.3	
II	Dq F1 F2 F3	0.894	0.764	16.2	63
	Dq F1 F2 F3	0.873	0.771	14.5	
	Dq F1	0.862	0.772	14.7	
III	Dq F1 F2 F3	0.931	0.800	14.1	125
	Dq F1 F2	0.930	0.816	14.1	
	Dq F1	0.892	0.628	16.5	

Tableau 2 - Récapitulatif de l'analyse de regression multilinéaire sur les données expérimentales de l'anglais britannique (I), américain (II) et du japonais (III). Les indices les meilleurs sont soulignés. Le nombre d'observation représente le total divisé par le nombre de répétitions.



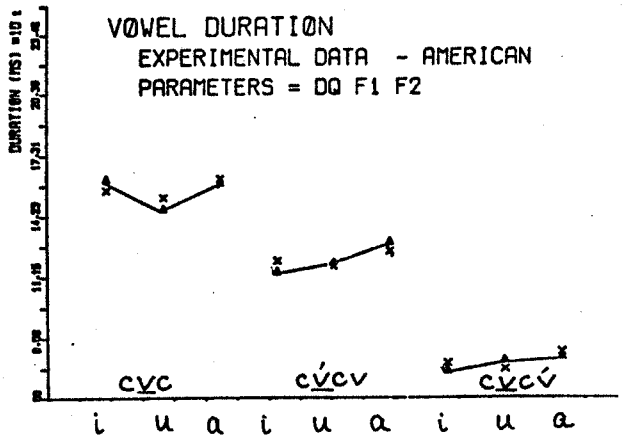


Figure 1 - Durée intrinsèque des voyelles : anglais, américain et japonais.
Les points reliés sont les valeurs observées et ceux non reliés, les valeurs estimées. Un point représente la valeur moyenne sur 5, 21 et 20 valeurs respectivement pour l'anglais, l'américain et le japonais.

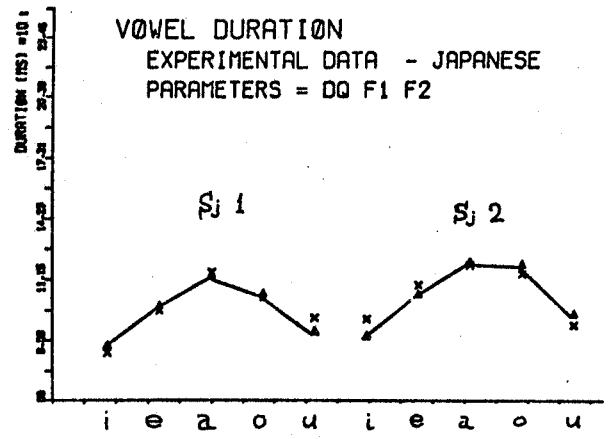
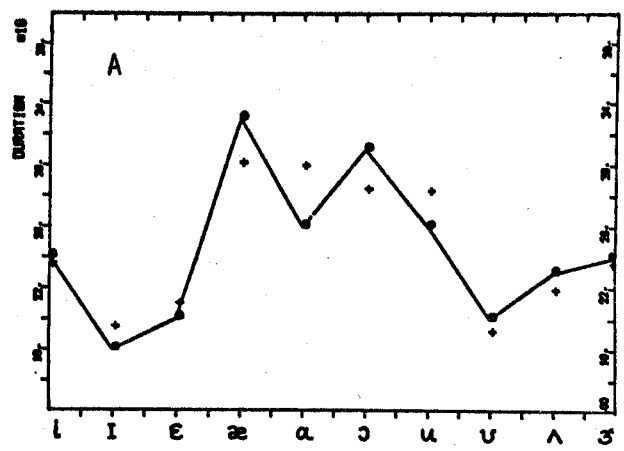
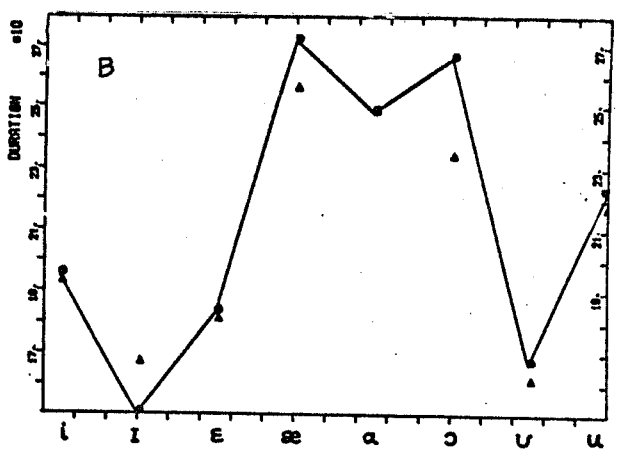
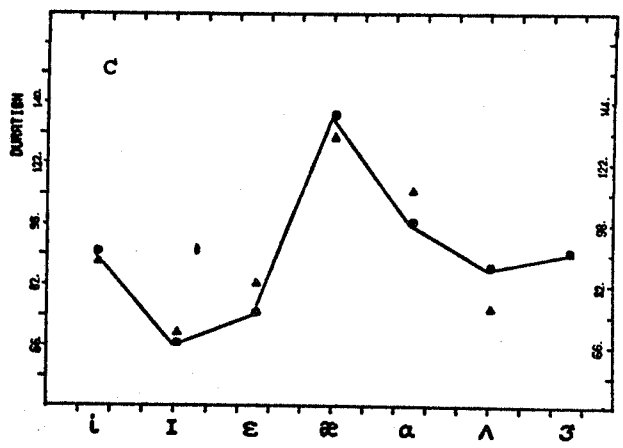
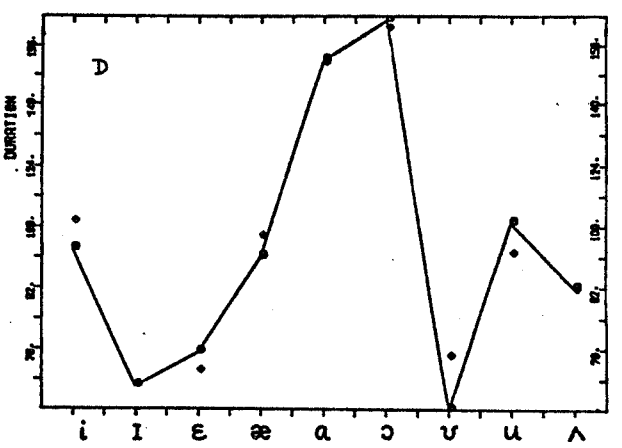


Figure 2 - Durée intrinsèque des voyelles : anglais
A : d'après les données de PETERSON et LEHISTE (1960)
B : SHARF (1962)
C : HARRIS et UMEDA (1976)
D : HILL et al. (1978)
Les symboles reliés représentent les valeurs observées et ceux non reliés, les valeurs estimées.









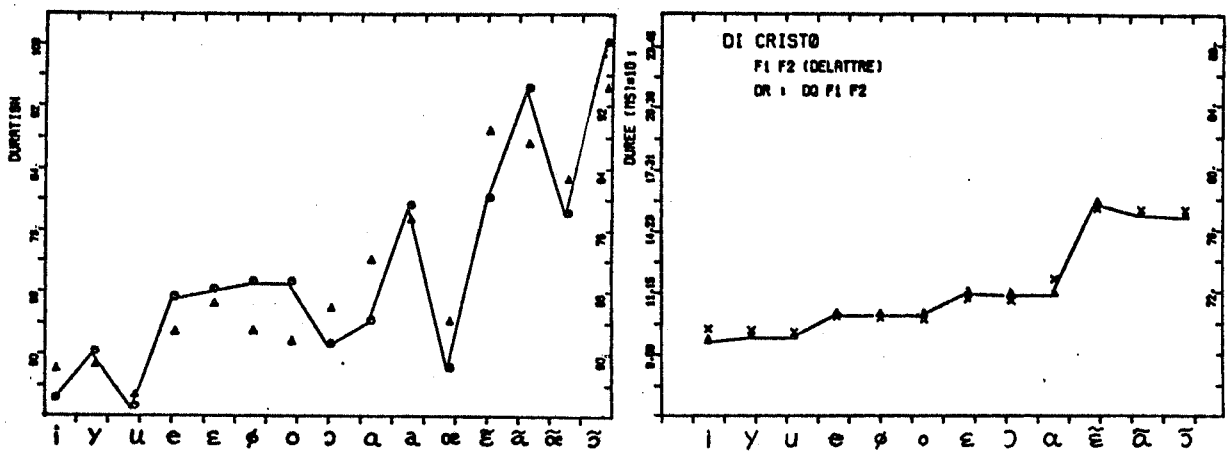
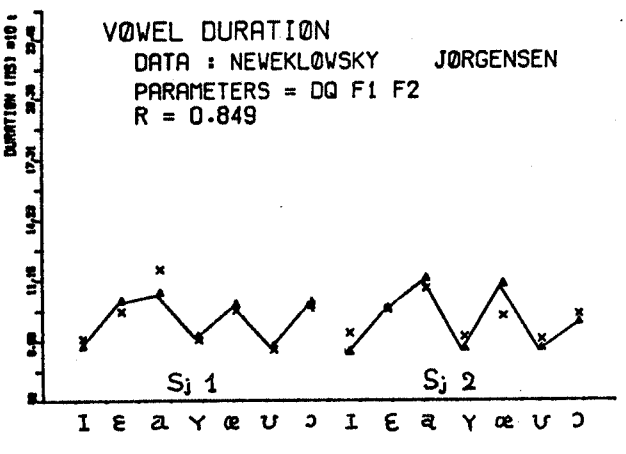
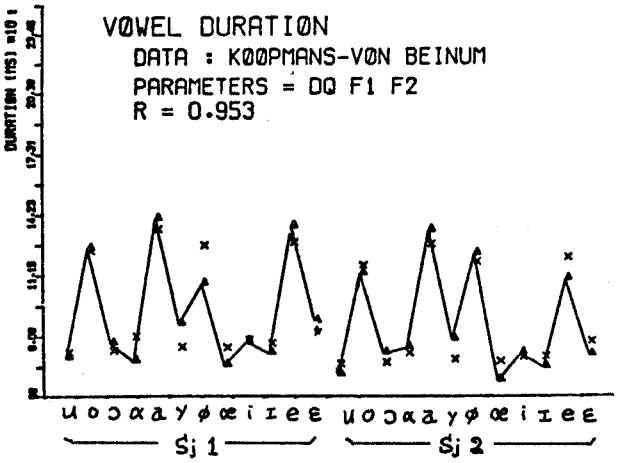


Figure 3 - Durée intrinsèque des voyelles : français
A gauche, d'après les données de DOHALSKA-ZICHOVA (1972)
A droite, d'après celles de DI CRISTO (1978)
Les symboles reliés représentent les valeurs observées
et ceux non reliés, les valeurs estimées.





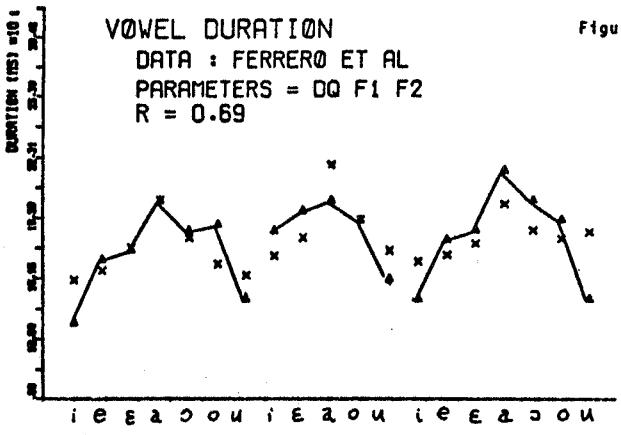


Figure 4 - Durée intrinsèque des voyelles : allemand, hollandais et italien
D'après NEWEKLOWSKY (1975), KOOPMANS-van BEINUM (1980) et FERRERO et al. (1975)
Le symbole triangle représente les valeurs observées
Le symbole croix, celles estimées.

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

RENCONTRE DE CONSONNES IDENTIQUES A LA JOINTURE DE MOT EN NEERLANDAIS.

Jean-François P. BONNOT (Université d'Amsterdam)
André BOTHOREL (Université de Strasbourg)

.RESUME

Quels sont les phénomènes de type phonétique qui se produisent en néerlandais lorsqu'il y a rencontre de consonnes identiques à la jointure de mots ? Pour répondre à cette question, nous avons étudié successivement les variables suivantes : durée de consonnes, durée des voyelles précédente et suivante, hauteur du fondamental, intensité et spectre des voyelles. L'examen des résultats obtenus montre que la marque différenciative la plus nette reste la durée consonantique. En ce qui concerne les autres variables, on constate qu'il n'y a pas un comportement uniforme chez tous les sujets, cependant chez un même sujet on relève un comportement cohérent. En conclusion, nous indiquons que, si nous n'avons pas déterminé l'existence de tendances spécifiques, c'est que ce type de phénomène n'est pas un fait de langue très important et que, n'étant pas très structuré, il permet une grande liberté dans les réalisations individuelles.

. ABSTRACT

This study seeks to identify the phonetic properties of occurrences in Dutch of two identical consonants across word boundaries. The following parameters are examined : consonant duration, pre-consonantal and post-consonantal vowel duration, pitch, intensity and spectral features of vowels. Though the data obtained show considerable subject to subject variation, individual subjects are nonetheless found to be highly consistent. This would appear to support the view that the phenomenon examined is not linguistically important, its lack of integration into the system accounting for the lattitude left to individual speakers in this particular area.

RENCONTRE DE CONSONNES IDENTIQUES A LA JOINTURE DE MOTS EN NEERLANDAIS

J.F.P. BONNOT (Université d'Amsterdam)
André BOTHOREL (Université de Strasbourg)

Introduction

Dans des langues comme le néerlandais ou le français où l'on ne rencontre pas de consonnes phonologiquement longues (à l'inverse du finnois ou de l'arabe par exemple), le phénomène de gémination -entendue comme succession de deux consonnes identiques comportant "un certain fléchissement de la tension au milieu de la tenue" (Chlumsky 1933:84)- ne peut s'appliquer qu'à la rencontre accidentelle de deux consonnes identiques à la jointure de mots. En restreignant notre étude au cas du néerlandais, nous allons tenter de déterminer quels sont les phénomènes de type phonétique qui se produisent lorsque, accidentellement, deux consonnes identiques se rencontrent à la jointure de mots. Le Néerlandophone prononce-t-il une seule consonne, introduit-il une pause ou réalise -t-il une géminée dans des exemples comme "kan naaien" vs. "kan aaien" ; "liep pover" vs. "liep over" ?

Zwaardemaker et Eijkman (1928:260-264) se plaçant essentiellement dans une perspective de phonétique descriptive, estiment qu'il peut y avoir des géminées en néerlandais. Cependant, illustrant leur propos de nombreux exemples, ils font ressortir que les réalisations présentent une grande variété (ainsi que le soulignera également Branco van Dantzig, 1935) ; ce sont surtout les indices de durée, de frontière de pression et de coupe syllabique qui retiennent leur attention. Considérant la question d'un point de vue fonctionnel, van Wijk (1939:60) note que dans des composés comme "vissoep", "eettafel" ou à la frontière de deux mots : "hij is ziek", "ik eet thuis", "... on prononce souvent une consonne simple appartenant à une ou deux syllabes, selon que la prononciation de la voyelle précédente est à coupe faible ou ferme. Mais si l'on désire faire apparaître clairement chacun des deux membres... phonologiquement on obtient alors deux consonnes... même si la réalisation phonétique est plutôt une géminée". Nous citerons encore les opinions plus catégoriques de van den Berg (1974:60-61) et de Bichakjian (1980) pour qui il n'y a pas de géminées en néerlandais moderne : "en néerlandais, les géminées étymologiques se sont simplifiées à l'intérieur des morphèmes et les géminées ad hoc se simplifient à la rencontre de deux mots" (Bichakjian 1980:216).

Mais c'est sans doute Cohen et al. (1972:52) qui, pour ce qui concerne notre approche, posent le mieux le problème : "Il est possible de distinguer "een Noor" (un Norvégien) de "een oor" (une oreille) -même lorsque le contexte n'indique pas quelle est l'intention du locuteur- sans que le [n] soit prolongé. En fait, on entend dans "een oor" une certaine interruption entre [n] et [o], appelée "open juncture"... Dans le cas contraire, "een Noor", on parle de "close juncture". Jusqu'à présent, on n'est pas

parvenu à établir de façon précise la nature phonétique de ce phénomène, mais il ne fait pas de doute qu'il est audible et joue un rôle distinctif".

Nous voudrions apporter quelques éléments de réponse à ce problème dont la complexité vient de ce qu'il ne se traduit pas sous des formes analogues, mais par des séries de manifestations relativement diverses, dont il serait intéressant de connaître la proportionnelle répartition.

PROTOCOLE EXPERIMENTAL

Afin de disposer d'un corpus assez vaste présentant des exemples significatifs et permettant de relever le type de stratégie utilisée par les locuteurs, nous avons fait établir par M. et M.-R. Beck une série de 35 phrases relativement courtes (cf annexe). Les consonnes retenues ont été les suivantes : /p m f t d n s l k r/. Le corpus a été enregistré à l'Institut de linguistique appliquée de l'Université d'Amsterdam par trois sujets : un sujet féminin (F) et deux sujets masculins (Ha et Hb). Les documents expérimentaux (mingogrammes, sonagrammes) ont été réalisés à l'Institut de phonétique de Strasbourg. Les sujets n'étaient pas prévenus du but de la recherche et avaient pour consigne de lire les phrases -présentées en désordre et, de plus, séparées par des phrases "gratuites"- avec naturel. Nous avons veillé à ce que les conditions accentuelles soient comparables ; nous n'avons cependant pu éviter que les locuteurs insistent plus ou moins sur tel aspect, en rapport avec l'accent de phrase dont la localisation dépend du désir de privilégier certains éléments par rapport à d'autres (incidence de la signification voulue).

La transcription phonétique a été faite par J.J. Spa.

RESULTATS DE L'ANALYSE

Une comparaison globale des enregistrements permet de constater que la variabilité inter-locuteurs est plutôt réduite : à part quelques différences dues à une modification du schéma intonatif, les réalisations de nos locuteurs sont dans l'ensemble très proches les unes des autres. Une comparaison systématique des lignes mélodiques permet de constater que les accents de phrase se situent en règle générale sur les mêmes syllabes chez les trois locuteurs, des variations de hauteur (secondaires en quelque sorte) peuvent affecter une autre syllabe et la marquer d'une nuance sémantique, ainsi qu'on peut le relever, par exemple, chez Ha qui présente un second sommet sur "gesloten" (phrase 32), alors que la culmination la plus importante se trouve sur "acht" comme c'est le cas pour les deux autres sujets. Du fait de la similitude des réalisations entre les sujets, la comparaison reste possible.

D'autre part, il semble que la facture de certaines phrases soit responsable de la présence de coups de glotte devant voyelle initiale de syllabe, comme pour deux sujets sur trois dans "die komm eet" (28) ou pour les trois sujets dans "veel eer" (32). Nous n'avons relevé que la présence d'un seul [ə] entre deux consonnes : dans le cas de "hij liep_pover" (4) dit par le sujet F.

Il apparaît donc que le corpus, tel qu'il a été réalisé par nos trois locuteurs peut se prêter favorablement à une analyse expérimentale. (Les résultats seraient sans doute légère-

ment différents si nous analysions les faits en parole spontanée ; mais ceci est un autre problème...)

Nous nous sommes attachés à étudier successivement les variables suivantes :

1) durée des consonnes
2) durée de la voyelle 1 (V1CV2)
3) durée de la voyelle 2
4) hauteur du fondamental
5) intensité des voyelles
6) spectre des voyelles.

1) Durée des consonnes

Pour presque toutes les réalisations, la durée du segment consonantique est plus importante dans le cas de rencontre de deux consonnes identiques (dorénavant CI) par rapport aux consonnes simples (dorénavant CS). Le rapport CI/CS varie de façon considérable. Il n'est jamais inférieur à 1 (cette valeur est obtenue pour deux oppositions). Il peut atteindre 2.8 (1 cas). Les sujets présentant des schémas différents, il est difficile de discerner un ordre de progression des rapports.

Durées relatives de CI

Ha	Hb	F	
1.37	1.55	2.55 R.	/p/
2.11	2.25	1.75	/t/
1.66	1.71	2.11	/k/
1.66	1.87	2.33	/m/
1.33	1.12	1.11	/m/
1.	1.5	1.	/m/
1.5	1.57	2.4	/n/
8.75 R.	2.	1.5	/l/
7.	2.5	1.31	/l/
1.33	1.57	2.	/f/
- R.	1.55	- R.	/s/
1.	1.	2.	/r/
1.16	1.5	1.	/d/
2.8	2.12	2.68 R.	/d/
2.14	2	3.57 R.	/d/

Nb. R. indique une réarticulation évidente.

Ainsi, au plan de la durée consonantique, une marque différenciative est presque toujours présente : il s'agit le plus souvent d'une différence de durée des articulations proprement dites. Dans les cas de réarticulation, cette différence de durée reste bien entendu présente, mais elle est accompagnée d'une marque supplémentaire, sous la forme d'une pause, d'une explosion ou d'un [ə].

2) Durée de la voyelle 1 (V1CV2).

Le rapport V1 (CI)/V1 (CS) ne fait pas apparaître de nette tendance chez les informateurs Ha et Hb. Par contre, chez F, nous constatons que dans 10 cas, la voyelle précédant CI est plus

brève ; dans 5 cas seulement, elle est plus longue. Il semble donc bien que nous soyons fondés à parler d'une tendance chez cette informatrice.

3) Durée de la voyelle 2

Comme précédemment, en ce qui concerne Ha et Hb, il est difficile de dégager une systématique. A nouveau, cependant, nous recueillons une indication intéressante pour ce qui est de F: dans 9 cas, la voyelle suivant CI est plus brève, dans 1 cas, elle a la même durée, et dans 4 cas elle est plus longue que la voyelle suivant CS.

Ha	Hb (V1/V1)	F	Ha	Hb (V2/V2)	F	Voyelles V1 - V2	N° phrase
0.83	1.2	1.33	1.	1.15	1.35	/i/ - /o/	4/10
1.42	1.16	0.66	0.85	0.75	1.16	/œ/ - /œ/	1/30
1.2	1.16	0.66	1.	0.9	1.	/ɪ/ - /o/	33/17
1.	1.5	0.5	1.	1.22	0.9	/o/ - /e/	13/21
0.66	1.	1.4	1.07	1.25	1.13	/ɔ/ - /e/	35/14
0.5	0.83	0.63	1.87	1.66	0.7	/ɔ/ - /e/	35/28
1.	0.58	1.14	0.8	0.91	0.64	/œ/ - /a/	11/2
0.83	0.83	0.88	1.2	1.25	1.9	/e/ - /e/	5/20
1.	1.25	1.14	0.6	0.58	0.55	/e/ - /e/	5/32
1.	0.81	0.85	0.85	1.	0.75	/e/ - /i/	9/35
1.1	0.88	0.57	-	-	-	/y/ -	6/13
1.25	0.5	1.23	1.42	1.83	0.88	/ə/ - /œ/	25/16
0.64	0.52	0.68	0.4	0.4	0.22	/e/ - /ə/	17/23
0.78	0.52	0.75	1.6	0.35	0.22	/e/ - /ə/	12/23
1.3	1.08	0.82	1.6	0.7	0.8	/e/ - /u/	19/29

Durées relatives de la voyelle précédant/suivant CI (V-CI/V-CS)

4) Hauteur du fondamental

L'intérêt de la mesure de la valeur du fondamental, en particulier pour les voyelles qui précèdent les consonnes (CI ou CS) n'est pas négligeable, si l'on considère qu'il pourrait y avoir là un indice susceptible de se mettre au service d'une éventuelle différence au niveau du passage voyelle-consonne (voyelle + CI et voyelle + CS). Pour Ha et Hb, les résultats sont assez décevants, car on ne peut relever aucune régularité. Par contre, et de même que pour la durée, nous avons constaté une tendance assez nette chez l'informatrice F : dans 11 cas, Fo est plus bas devant CI, les variations se situant entre - 5 et - 90 Hz, étalement 85 Hz ; pour 4 réalisations, Fo est plus élevé (+ 20 à + 45 Hz, étalement 25 Hz).

Ce paramètre n'est pas révélateur pour ce qui est des voyelles suivant les consonnes.

Il convient de préciser que la mesure utilisée a été effectuée à l'extrémité de la voyelle en contact.

5) Intensité des voyelles

Nos calculs ont porté sur la chute d'intensité (V1) et la montée d'intensité (V2) exprimées en valeurs absolues et en

valeurs relatives. Ces mesures permettent de déterminer le degré d'adhérence phonique -forte ou faible- dans le passage de la voyelle à la consonne (CI et CS), au sens que lui donne Goudaillier (1977:171-203) : -plus la surface ("absolue" ou "relative") est petite, plus l'adhérence phonique est grande (autrement dit, plus la coupe est ferme) ; -plus la surface est grande, plus l'adhérence phonique est réduite (plus la coupe est lâche).

Si nous appliquons cette méthode à l'étude de l'opposition CI vs. CS, il apparaît que le sujet Ha présente plutôt une tendance à l'adhérence phonique forte, au passage voyelle-CI (9 cas sur 15), alors que l'informatrice F présente une préférence nette pour l'adhérence phonique lâche (11 cas sur 14). Les résultats de "surface absolue" et de "surface relative" sont en accord dans la très grande majorité des cas. De ce point de vue, Hb n'a pas de stratégie précise.

En examinant l'adhérence phonique au passage CI-voy. 2, nous constatons que les surfaces sont plus importantes pour Hb et surtout pour F, ce qui nous conduit à admettre que la montée d'intensité est plus lente. Le sujet Ha présente une tendance inverse.

Nous admettrons donc en conclusion qu'en ce qui concerne ce paramètre de l'intensité vue sous l'aspect d'une adhérence phonique plus ou moins forte ou plus ou moins faible, les résultats sont également assez divergents. Il est cependant possible de mettre en lumière un comportement relativement stable pour un même sujet, ce qui est l'indication de l'existence d'une stratégie personnelle et constante pour chaque locuteur vis-à-vis du degré d'adhérence phonique.

6) Analyse spectrale des voyelles

La recherche d'un comportement similaire des locuteurs les poussant à différencier la qualité de la voyelle devant CI par rapport à CS nous a conduits à effectuer l'analyse spectrale des voyelles précédentes, et ceci afin de déterminer si, à ce niveau, il se manifestait une régularité quelconque.

Les différences de valeurs fréquentielles des formants 1 à 3 pourles voyelles précédant CI et CS ne sont guère significatives. Nous n'avons pu relever l'existence d'une tendance déterminée, permettant de définir une stratégie commune.

Quant aux mesures de fréquences effectuées à la transition, elles ont été soumises à des calculs de variabilité statique aussi bien que dynamique. La "variabilité statique" a été calculée de la façon suivante : pour chacun de nos sujets, nous avons mesuré les valeurs fréquentielles des formants 1/2/3/ à l'extrémité de la transition en contact, puis nous avons effectué la différence entre les valeurs les plus éloignées. Par exemple pour la voyelle /ɑ/, dans "kan naaien", nous obtenons :

	F	Ha	Hb	Différence
F1	409	614	573	205
F2	1475	1147	1065	410 (en Hz)
F3	2213	2295	2459	246

Il suffit ensuite de prendre les valeurs figurant dans la colonne "différence" et d'effectuer la moyenne pour les valeurs des transitions au contact de CI d'une part et de CS d'autre part. Ceci confond les timbres vocaliques en ne laissant subsister que le paramètre "proximité de CI ou CS" : la comparaison est rendue possible du fait du parallélisme du corpus. (En ce qui concerne F3, trop souvent nous n'avons qu'une donnée, ce qui empêche un calcul fiable).

	CI	CS	Variabilité statique :
F1	104,8	92	Calcul effectué à partir des
F2	295,5	319,9	données des 3 inf.

Pour CS, c'est F1 qui a le "coefficient" le plus faible et qui présente donc la variabilité statique la plus réduite ; pour F2, nous notons l'inverse.

D'autre part, connaissant les valeurs médianes et extrêmes (en contact) des F 1/2/3, nous avons la possibilité d'en déduire la différence grâce à une simple soustraction. On obtient alors pour chaque formant, et suivant le sens de la transition, une valeur nulle, positive ou négative. Nous avons considéré les valeurs absolues (donc sans tenir compte du signe), et nous avons effectué la moyenne. Nous avons retenu ce résultat comme indice de la variabilité dynamique. Comme précédemment, le calcul est uniquement arithmétique, mais il serait facile, avec plus de données, de chercher le "coefficient de variation". Il faut souligner que ce paramètre ressortit au mouvement, puisque nous prenons en compte :

'fréqu. milieu F1/2/3' - 'Fréqu. extrémité en contact F1/2/3'

---

nombre d'occurrences en contact avec CI ou CS

Les résultats sont les suivants, chaque locuteur étant individualisé :

	F		Ha		Hb	
F1	20	23,5	29	13,9	35	2,9
F2	94,9	70,6	35,7	98	70,7	101
	CI	CS	CI	CS	CI	CS

Les données de Ha et Hb vont dans le même sens, celles de F en sens contraire. Toutefois, si l'on calcule la "variabilité dynamique" pour l'ensemble des locuteurs, on trouve que F2 varie plus au contact de CS qu'à celui de CI, l'inverse étant observé pour F1. Ceci peut être mis en rapport avec la variabilité "statique".

	CI	CS	
F1	28	13,7	Variabilité dynamique
F2	66,4	92,2	3 locuteurs.

Lorsque la moyenne générale est effectuée, la stratégie particulière de F se trouve gommée. Cela constitue un problème puisque nous n'avons que 3 sujets (la situation serait toute différente si nous avions 30 informateurs). Toutefois, tout en nous gardant de ne rien affirmer de définitif, et puisque l'on sait que le F2 est particulièrement sensible aux variations articulatoires, on pourrait en déduire que les variations articulatoires de la voyelle précédant CI sont plus "modérées" que celles de la voyelle précédant CS.

DISCUSSION

A l'examen des résultats de notre analyse, nous nous apercevons qu'il y a assez peu de convergences dans le comportement de locuteurs devant opposer une "consonne double" à une "consonne simple" à la jointure de mots.

1 - La constatation d'une convergence dans les résultats apparaît principalement dans la différence de durée entre CI et CS. Dans la majorité des cas, la durée du segment consonantique est plus importante pour CI. Le rapport CI/CS n'est en tout cas jamais inférieur à 1. Il existe donc généralement une marque différenciative nette au plan de la durée. Nous avons calculé, toutes consonnes confondues, mais en excluant les occurrences réarticulées, les rapports moyens pour les 3 sujets :

Ha = 1.58 Hb = 1.72 F = 1.65

Comme on peut s'en rendre compte, ces résultats sont très comparables.

2 - La durée de la voyelle précédant la consonne est généralement considérée comme un indice important, en ce sens que cette voyelle subit un abrègement au contact d'une consonne double ou géminée. Or, nous n'avons trouvé de différences notables ni en ce qui concerne le sujet Ha, ni pour l'informateur Hb. Cependant, chez le sujet F, dans dix cas contre cinq, le segment vocalique est plus bref lorsqu'il précède CI.

3 - Le praramètre "durée de la voyelle subséquente" est habituellement considéré comme étant d'importance bien moindre que le précédent, pour ce qui est des différenciations au plan consonantique. Cette conception se vérifie pour les informateurs Ha et Hb ; mais à nouveau, pour F, il apparaît que dans 9 cas sur 14, cette voyelle est plus brève après CI : il semble donc y avoir là un indice pouvant être retenu.

4- Quant à l'intensité, quel que soit le point de vue sous lequel on la considère, force est d'admettre qu'elle ne se manifeste pas d'une façon régulière dans une stratégie permettant de différencier consonnes doubles (ou géminées) et consonnes simples. La seule régularité qu'on puisse lui reconnaître serait d'ordre individuel puisque chez un même sujet ses manifestations apparaissent assez homogènes.

5 - Nous avons vu qu'en ce qui concerne fo de la voyelle antéposée, on ne pouvait déceler aucune systématique pour les sujets Ha et Hb. Par contre, dans 11 cas sur 14, chez l'informatrice F, le fo est plus bas pour la voyelle suivie de CI.

6 - Nous avons tenté de déterminer s'il existait une relation entre la hauteur de fo et la durée de la voyelle. En fait, nous n'avons trouvé aucune symétrie significative entre les variations de durée de la voyelle et la hauteur de son fondamental. Si, comme nous le pensons, les variations de hauteur sont plus directement liées aux nuances sémantiques que la durée, les différences de durée peuvent être plus directement attribuées à l'influence du voisinage de CI et CS.

CONCLUSION

Les résultats que nous avons obtenus ne nous permettent pas de dire que la rencontre de deux consonnes identiques à la

frontière de mots en néerlandais se traduit par un comportement uniforme chez tous les sujets. Les moyens utilisés pour distinguer deux consonnnes identiques (se suivant accidentellement) d'une consonne unique correspondante, comme dans "ik kook / ik ook ; ik kan naaien / ik kan aaien ; hij liep pover / hij liep over", sont de plusieurs sortes : dans certains cas, des manifestations de réarticulation peuvent être observées ; ces marques, dont la fréquence d'occurrence est faible, sont au nombre de trois : pause, élément vocalique de type [ə], explosion dans le cas de certaines réalisations occlusives ; la plupart du temps, cependant, la marque différenciative la plus nette se situe au plan de la durée du segment consonantique, majoritairement la plus importante pour les consonnes identiques. Nous confirmons donc ainsi l'analyse déjà ancienne, mais semble-t-il toujours valable, de Zwaardemaker et Eijkman (1928), qui admettent que la différence de durée constitue un des indices pouvant entrer, concurremment avec d'autres, dans la distinction CI/CS. On pourrait joindre à cette manifestation la stabilité plus grande du F2 de la voyelle précédant les consonnes identiques par rapport à la simple.

Il convient d'insister sur le comportement spécifique de l'informatrice F, qui met en oeuvre une stratégie dont nous pouvons dire, au vu des indices dont nous disposons, qu'elle paraît plus élaborée, plus homogène que celle des sujets masculins. Cette régularité dans le comportement se manifeste non seulement au plan des durées de CI et CS, mais également à celui des voyelles anté et postposées qui montrent une tendance très nette à l'abrègement devant et après CI, alors que ce n'est le cas ni pour Ha ni pour Hb. Cette observation est à rapprocher de la conclusion à laquelle arrive Koopmans-van Beinum (1980:71) : "we found dialectal deviations being more explicit in male than in female speakers, durations of vowels being systematically longer in female than in male speakers..., all tendencies pointing to a more careful articulation in females resulting in more generally understandable prononciation and larger acoustic contrasts".

D'autre part, et nous rejoignons ici Cohen et al. (1972), il semble bien qu'il se produise des manifestations au niveau de la jointure voyelle-consonne : c'est ainsi que chez F, le fondamental des voyelles suivies de CI est majoritairement plus bas que lorsque ces mêmes voyelles sont suivies de CS (la mesure a été effectuée à la transition). Finalement, on peut également noter que les sujets manifestent un comportement individuel cohérent, s'agissant de l'adhérence phonique voyelle+consonne. Peut-être parviendrait-on à préciser les stratégies utilisées et même à dégager une certaine unité au niveau des commandes motrices les plus périphériques en procédant à une analyse EMG. Les deux organes les plus indiqués pour ce genre de recherche nous paraissent être les lèvres et le voile. Nous avons d'ailleurs appliqué cette technique à la question des rencontres de consonnes identiques et des groupes de consonnes labiales en français. Il s'est avéré que l'opposition CI/CS s'accompagne d'une restructuration de l'activité EMG. (Bonnot et al. 1980).

Cependant, si nous n'avons pas relevé, au plan interpersonnel, de tendances très spécifiques, nous pensons que cela est dû au peu de rendement de ce type d'oppositions, qui n'est phonologique qu'accidentellement et dont les locuteurs s'acquittent de plusieurs manières, jusques y compris l'admission de la confu-

sion lorsque le contexte peut valablement prendre le relais. Ce phénomène, qui n'est finalement pas un fait de langue très important et qui ne recouvre pas l'ensemble du système, n'est pas structuré (au sens fonctionnel) et entraîne autant de réalisations individuelles. C'est en considérant l'économie du système tout entier que nous pouvons déterminer l'interprétation et la place qu'il convient de lui donner (Bothorel, 1971:199).

Une analyse phonétique menée avec soin sur un corpus dit par trois sujets a permis de relever les différences de comportement des locuteurs aux prises avec un phénomène de parole dont ils ont le sentiment qu'il est réalisé de façon uniforme (une production en style soigné confirmerait ce sentiment). Or à l'analyse, cette sensation se révèle parfois trompeuse, mais, pour communiquer avec autrui, la "persuasion" n'a-t-elle pas un rôle de grande importance ? C'est, en tout cas, une variable qui dépasse notre champ d'investigation.

REFERENCES

BERG van den, B., 1958, Foniek van het Nederlands, van Goor zonen, La Haye, 7ème tirage 1974.

BICHAKJIAN, B., La lénition romane et l'hypothèse d'une lénition hollandaise, in Liber Amicorum Wijnen, 210-219, van Gorcum, Assen 1980.

BONNOT, J.F.P., GREINER, G.F., MATON, B. et CHEVRIE-MULLER, C., 1980, Etude électromyographique de l'organisation temporelle (timing) des consonnes simples, géminées et des groupes de consonnes labiales en français, Séminaire international sur la labialité, GALF, 7 et 8 Février 1980, Lannion/CNET, 22 p.

BOTHOREL, A., 1971, A propos du breton parlé à Argol : quelques observations sur les consonnes géminées, Travaux de l'Institut de Phonétique de Strasbourg 3, 195-233.

CHLUMSKY, J., 1935, Analyse de traité de phonétique de M. Grammont, Archives néerlandaises de phonétique expérimentale, XI, 73-106.

COHEN, A., EBELING, C.L., FOKKEMA, K. et HOLK, van A.G.F., Fonologie van het Nederlands en het Fries, inleidingtot de moderne klankleer, Martinus Nijhoff, La Haye, 3è éd. 1972.

DANTZIG, Branco van, 1935, Petite contribution à l'examen expérimental de la gémination dans la langue néerlandaise, Archives néerlandaises de phonétique expérimentale XI, 119-132.

GOUDAILLIER, J.P., 1977, Mesures de la chute d'intensité au-delà d'un maximum vocalique à partir de l'étude d'une ligne d'intensité -les surfaces de chute d'intensité- le concept d'adhérencephonique, Travaux de l'Institut de Phonétique de Strasbourg 9, 171-203.

KOOPMANS-van BEINUM, F.J., 1980, Vowel contrast reduction, An acoustic and perceptual study of Dutch vowels in various speech conditions, Academische Pers B.V. Amsterdam.

WIJK, van B., 1939, Phonologie ; een hoofdstuk uit de structurele taalwetenschap, Martinus Nijhoff, 's-Gravenhage, réimp. 1965.

ZWAARDEMAKER, H. et EIJKMAN, L.P.H., 1928, Leerboek der Phonetiek, de Erven F. Bohn N.V., Haarlem.

ANNEXE

CORPUS (présentation systématique regroupée).

30 - [t] Met z'n vieren in de flat ? Dat acht ze uitgesloten
[dɑt>ɑxtsə]
1 - Ze zei dat tachtig servetjes genoeg was.
[dɑtɑxtəx]
2 - [n] Ik will een hond die je kan aaien, en die je niet meteen vergast op een beet. [jəkɑn>ajə(n)]
11 - Ik heb een butler di kan koken en kan naaien, en altijd klaar staat. [kɑnajə(n)]
29 - [d] Jan-Willem en André doen mee.
[ɛn>ɑndredũme]
19 - Je vertrek zou me leed doen.
[məledun]
23 - [d] De drank van de griekse goden was Mede, een honingdrank.
[wɑsmedə]
12 - Ik meed de waterkant zoveel ik kon.
[ɪkmedəwaterkɑnt]
17'- Ik deed de deur open.
[ɪkdedədør]
10 - [p] Ik liep over de dam toen ik hem zag.
[ɪklip>over]
4 - Hij liep pover gekleed, maar keek erg opgewekt.
[lipovər]
14 - [m] De Halley komeet verschijnt weer.
[dəhɛlikǫmet]
28 - Die kom eet jij niet leeg.
[dikǫm etjɛj]
35 - Geef iets en kom mee.
[xevits ɛŋ kǫme]
21 - Ik ga zo mee.
[ɪkxazome]
13 - Loes en Truus gaven me een kerstboom mee.
[bome]
17 - [k] Jan schrok zich een hoedje, ik ook overigens.
[ik>ok]
33 - Ik kook erg lekker.
[ɪkokɛrxlɛkər]
35 - [f] Geef iets en kom mee.
[xevits]
9 - Edelachtbare, geef fietsendieven geen milde straf.
[xefitsəndivən]
16 - [r] Ze liet haar hondje in een crematorium verassen.
[vər>ɑsən]
25 Zeg vader niets, ik wil hem verrassen.
[vərɑsən]
13 - [s] Loes en Truus gaven me een kerstboom mee.
[trys>xavən]
6 - Dirk en Guus schaven de zijkant, jij het bovenvlak.
[xysxaven]
32 - [l] Met heel veel eer en genoegen deel ik U mee...
[vel>er]
20 - O ja ? Je zou veeleer opgelucht moeten zijn.
[jezɔw vel>er]
5 - Te duur ? Vergis je niet dame, in zo'n zitkuil is nogal veel leer verwerkt.
[veler] (> "implosif" ; < "explosif")

Surfaces en valeurs relatives de la chute d'intensité

SURFACE	ADHERENCE + ou - forte	VOYELLE	NATURE DE LA SEQUENCE	Extraits LOCUTEURS
850	+	[ɑ]	dat tachtig	Ha
1328		"	dat acht ze	Ha
823	+	"	"	Hb
1766		"	"	Hb
2051	-	"	"	F
1366		"	"	F
431	+	[ɑ]	kan naaien	Ha
470		"	kan aaien	Ha
514	-	"	"	Hb
131			"	Hb
				F
				F
1362	+	[ɪ]	liep pover	Ha
2059		"	liep over	Ha
1502	+	"	"	Hb
1913		"	"	Hb
909	-	"	"	F
641		"	"	F
390	--	[e]	veel leer	Ha
135		"	veeleer	Ha
175		"	veel eer	Ha
259	++	"	"	Hb
792		"	"	Hb
469		"	"	Hb
851	--	"	"	F
431		"	"	F
383		"	"	F
1591	-	[e]	Geef fietsen	Ha
700		"	Geef iets	Ha
1482	-	"	"	Hb
700		"	"	Hb
721	-	"	"	F
412		"	"	F
1483	-	[ɪ]	ik kook	Ha
485		"	ik ook	Ha
1071	-	"	"	Hb
303		"	"	Hb
1613	-	"	"	F
708		"	"	F
1022	+	[y]	Guus schaven	Ha
1201		"	Truus gaven	Ha
853	+	"	"	Hb
1606		"	"	Hb
367	+	"	"	F
580		"	"	F
880	-	[o]	kerstboom mee	Ha
242		"	zo mee	Ha
576	-	"	"	Hb
390		"	"	Hb
216	+	"	"	F
472		"	"	F
596	+-	[ɔ]	kom mee	Ha
734		"	kom eet	Ha
404		"	komeet	Ha

SURFACE	ADHERENCE + ou - forte	VOYELLE	NATURE DE LA SEQUENCE	Locuteur
318	++	[ə]	komeet	Hb
550		"	"	Hb
1047		"	"	Hb
1254	+-	"	"	F
1818		"	"	F
864		"	"	F
etc.				

Ces surfaces ont été obtenues selon la formule de Goudaillier (1977:180) :

$$\frac{\text{chute en msec}}{\text{durée en msec de la voyelle}}\,\% \times \text{chute en sones maximum d'intensité}\,\%$$

$$2$$

12èmes JOURNEES D'ETUDE SUR LA PAROLE

MONTREAL 25.26.27 mai 1981

FONCTIONS, ROLES ET TRATIEMENT DE L'INTENSITE DANS LA PAROLE

CHRISTEL SORIN - CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS
LANNION A - ROUTE DE TREGASTEL - B.P. 40
22301 - LANNION

.**RESUME**

Dans la première partie de cet article, nous tentons de passer en revue les différents rôles de l'intensité dans la perception de parole. Il apparaît que le traitement auditif de l'intensité est fortement dépendant des facteurs temporels. Si la durée d'observation est trop brève, l'importance de la fonction linguistique de l'intensité peut être sous-estimée et son traitement auditif mal modélisé. De plus, les interactions entre intensité et fréquence fondamentale sont différents aux niveaux segmental ou suprasegmental de la perception de parole.

Dans la deuxième partie, nous comparons et discutons les résultats de ROSSI (1978) sur la perception de pentes d'intensité sur des voyelles stationnaires et nos résultats (SORIN, 1978) sur la discrimination de pentes d'intensité en fin de phrases. Le but est de déterminer à quel traitement auditif l'intensité est soumise dans les deux cas.

Nous montrons qu'un modèle "purement" psychoacoustique (intégration syllabique) peut expliquer les résultats observés. Nous mettons l'accent sur l'importance qu'il y a à faire une étude purement psychoacoustique d'un paramètre physique en contexte avant d'examiner les fonctions et traitement linguistiques de ce paramètre.

FUNCTIONS, ROLES AND TREATMENT OF INTENSITY IN SPEECH

ABSTRACT

In the first part of this paper, we attempt to overview several roles of intensity in speech perception. It appears that the auditory treatment of intensity is highly dependent on temporal factors. If the observation period is too brief, the importance of the linguistic function of intensity may be underrestimated and its auditory treatment incompletely observed. Also, the interactions between the intensity and the fundamental frequency are different at the segmental and suprasegmental levels of speech perception.

In the second part of the paper, we compare and discuss the results of ROSSI (1978) on the perception of intensity glides of vowels with our results (SORIN, 1978) on the discrimination of intensity glides at the end of sentences. The aim is to determine which auditory treatment is used in both cases.

We show that a "simple" psychoacoustic model ("syllabic" integration) can explain the results obtained. We emphasize the importance of a purely psychoacoustic study of a physical parameter in its context before investigating the linguistic function and treatment of that parameter.

FONCTIONS, ROLES ET TRAITEMENTS DE L'INTENSITE DANS LA PAROLE

Ainsi donc, si tout le monde s'accorde à reconnaître dans l'intensité l'un des trois facteurs prosodiques essentiels, le consensus fut à partir des années 50 (1) presque aussi grand pour affirmer que le "rendement perceptuel" de l'intensité était nettement inférieur à celui de la fréquence et de la durée. Cette affirmation, souvent généralisée à tout le processus de perception de parole résulte en fait essentiellement d'études sur l'accent. A titre d'exemple, pour l'anglais, les travaux de FRY qui servent de référence mentionnent que "l'intonation de la phrase est le facteur dominant dans la perception de l'accent et qu'en ce sens l'indice de fréquence fondamentale peut l'emporter sur l'indice de durée" (FRY 1958), "le rapport des durées étant lui-même un indice plus effectif que le rapport d'intensité" (FRY 1955).

De même, BOLINGER (1958) indique que le corrélat acoustique de l'accent perçu des auditeurs américains est plus souvent la fréquence fondamentale que l'intensité. Les travaux de DENES (1959), MORTON et JASSEM (1965) vont dans le même sens. Cette constatation apparaît souvent comme justification à l'étude privilégiée de la mélodie (Fo) (voir par exemple HADDING-KOCH et STUDDERT-KENNEDY (1965)).

Pour le français, les études sur la réalisation de l'accent (et même de l'accent d'insistance) attribuent également à l'intensité un rôle très secondaire. DELATTRE (1951) écrit que "le seul élément physique qui soit toujours en excès dans la syllabe accentuée est la durée... l'intensité n'est pas plus forte dans la syllabe accentuée que dans les autres". RIGAULT (1961) affirme que "la fréquence est de loin le facteur physique dominant de l'accent". SEGUINOT (1976), dans une étude sur l'accent d'insistance, aboutit à la conclusion que "acoustiquement, ce sont surtout la hauteur et les ruptures mélodiques qui caractérisent ce phénomène".

Toutefois, il importe de remarquer que dès que les sujets ont à identifier ou à décrire la nature de l'accent (naturel) ou artificiellement introduit) ils font très souvent appel à la notion d'intensité, ou plus exactement de force sonore (2).

---

(1) Rappelons que jusque dans les années 50, on trouvait au contraire presque toujours dans les diverses définitions de l'accent une allusion à l'intensité, voir RIGAULT (1961) pour une revue de ses travaux.

(2) Dans le vocabulaire courant le terme "intensité" est aussi utilisé pour désigner la force sonore. Le terme exact serait "sonie". Nous entourerons le terme "intensité" de guillemets quand il a été employé par l'auteur dans son sens perceptuel.

Ainsi, RIGAULT rapporte, à l'issue de tests "descriptifs" qu'"une proéminence due essentiellement à une haute fréquence est ressentie et interprétée comme un accent d'intensité". De même SEGUINOT (1976) relève que si une nette majorité d'énoncés marqués par l'accent d'insistance ont été perçus comme marqués d'une augmentation de hauteur, une minorité significative d'entre eux ont été reconnus dans l'analyse perceptive comme marqués d'une augmentation "d'intensité". Ces deux auteurs expliquent l'utilisation par les sujets de la notion de force sonore, apparemment contradictoire avec l'analyse acoustique, en introduisant ici la notion d'effort physiologique. RIGAULT affirme qu'"un auditeur a donc tout naturellement tendance à juger de la proéminence d'une syllabe d'après l'effort physiologique qu'il fournirait lui-même pour la prononcer".

SEGUINOT constate, lui, la présence d'un coup de glotte très net dans les énoncés accentués jugés plus "intenses" et en déduit que "c'est vraisemblablement cet effort articulatoire qui incite l'auditeur à parler d'intensité". Sous l'influence de la théorie motrice, on a en effet souvent cherché une origine plus articulatoire que "purement" perceptuelle à la sensation de force sonore pour la parole. Ce fut le cas, par exemple, pour l'étude de l'intensité spécifique des voyelles. LEHISTE et PETERSON (1959) ont montré que des voyelles d'intensité différente peuvent être perçues comme isophones pour peu qu'elles soient produites avec le même effort articulatoire. LADEFOGED (1961) et LADEFOGED et MAC KINNEY (1963) tendent à montrer que le jugement des auditeurs est dans ce cas en relation directe avec l'effort physiologique fourni, lui-même proportionnel au carré de la pression sous-glottique. Toutefois, ROSSI (1971) démontre que la sonie des voyelles peut être déduite d'une mesure physique sur le signal acoustique, basée sur la répartition fréquentielle de l'énergie notamment au niveau des premiers et deuxièmes formants. Ainsi, la sensation de force sonore ne nécessiterait pas, pour se constituer, un recours au niveau articulatoire, puisqu'un traitement "purement" psycho-acoustique du signal permet d'en rendre compte.

On voit apparaître ici une modification essentielle, un tournant dans l'approche des recherches en perception de parole : les corrélats articulatoires des sensations ont des manifestations acoustiques qui permettent, à elles seules, si l'on tient compte des propriétés psychophysiques du système auditif, de rendre compte des phénomènes perceptuels observés.

Pour en revenir à l'accent, à titre d'exemple, même si l'indice le plus constant de la mise en relief dynamique est un changement brutal de la courbe de pression sous-glottique et qu'il constitue un invariant de l'accent d'insistance en français (MARCHAL 1976), c'est d'abord au niveau acoustique que l'on doit chercher s'il n'existe pas de paramètres "psychoacoustiquement" responsables de la perception, suivant les cas, d'une modification de hauteur ou de force sonore (3). Ceci nécessite alors de prendre en compte des notions comme

---

(3) En effet, la justification articulatoire souvent évoquée de cette apparente "confusion" entre hauteur et force sonore n'est pas une réelle explication : elle se contente de mentionner qu'une augmentation de la pression sous-glottique entraîne une augmentation simultanée de l'intensité et de la fréquence fondamentale, les deux paramètres étant donc souvent très corrélés jusqu'à ce qu'intervienne un ajustement de la tension des cordes vocales. L'"illusion" d'une modification de force sonore alors que la modification physique porte sur F_0 serait donc d'ordre linguistique. Mais les mécanismes la régissant ne sont pas précisés. Il serait en particulier intéressant de savoir pourquoi une modification de F_0 est plus souvent perçue comme une modification de force sonore alors qu'une modification de l'intensité semble rarement prise pour une modification de hauteur.

la sensibilité différentielle de l'ouïe, l'effet de masque fréquentiel et temporel, les bandes critiques, le temps d'intégration etc... Ce n'est que dans la mesure où cette approche psychoacoustique échoue que l'on doit faire appel à des contraintes linguistiques "explicatives" (comme par exemple, les contraintes articulatoires) ce qui revient alors à postuler l'existence de modes de perception spécifiques à la parole.

C'est dans le cadre de cette problématique que nous nous situons ici. Nous nous proposons de relever, dans une première partie, diverses influences du paramètre intensité dans certaines tâches perceptuelles sur le signal de parole (tâches de discrimination ou d'identification), mentionnées dans la littérature. Dans une deuxième partie, nous étudierons plus particulièrement la sensibilité différentielle de l'ouïe à des variations lentes d'intensité à partir d'une expérience que nous avons effectuée sur des phrases entières. En effet, les groupes terminaux et interrogatifs sont marqués en français par une décroissance de l'intensité, contrairement aux groupes centraux. Si, comme le suggère LIEBERMAN (1967)""les codes" linguistiques n'exigent pas un fonctionnement optimal du système auditif et travaillent avec une marge d'erreur", il importe tout particulièrement de connaître cette marge d'erreur pour interpréter le rôle et le traitement de tel ou tel paramètre acoustique dans la perception de parole. C'est ce que nous tenterons de faire pour l'intensité.

I - DU ROLE DE L'INTENSITE DANS CERTAINES TACHES PERCEPTUELLES SUR LE SIGNAL DE PAROLE

Les travaux dont nous nous proposons de rendre compte ici relèvent de deux axes de recherche en perception de parole bien distincts. Le premier concerne les relations entre syntaxe et prosodie, l'autre les rapports entre les résultats de la psychoacoustique et certains faits de perception de parole. C'est dans ce double cadre que nous observerons le rôle de l'intensité.

I.1. Syntaxe et prosodie :

Nous avons vu dans l'introduction qu'il existe un nombre important de données concernant le rôle des divers indices prosodiques dans la perception de l'accent.

Plus rares et plus contradictoires, sont les informations sur la manière dont tous ces indices sont utilisés dans la perception courante de la parole. Ainsi, si le problème de l'intéraction syntaxe - évolution de F_0 -durée a déjà été souvent abordé (voir Structure and Process in Speech Perception 1975, ROSSI et Di CRISTO, 1977), le rôle de l'intensité dans la détermination de la structure syntaxique n'a suscité que peu d'attention. Nous faisons ici le point sur les relations syntaxe-intensité mises en évidence dans la littérature.

Dans deux études très similaires SCHOLES (1971) et STREETER (1978) ont testé, sur des expressions ambigües, la perception des frontières intersyntagmatiques. L'expression choisie par SCHOLES était "good flies quickly", les variations des indices suprasegmentaux (I, F_0, D) étant obtenues en extrayant cette expression de huit phrases différentes se répartissant en deux classes du type "the good flies quickly past" ou du type "the good flies quickly passed". STREETER a utilisé, quant à elle, une expression algébrique ambigüe du type "A plus E fois O", dont les trois paramètres suprasegmentaux étaient

modifiés par analyse-synthèse par prédiction linéaire. La tâche des sujets était dans les deux cas identiques : il s'agissait de déterminer si l'expression présentée correspondait au schéma AB/C ou A/BC c'est-à-dire de localiser la frontière intersyntagmatique. Les résultats relevés par les deux auteurs sont toutefois assez différents. SCHOLES aboutit à la conclusion que l'amplitude relative des mots "good" et "flies" est le principal déterminant des jugements de localisation de frontière : la durée des pauses et les contours de F_0 ne permettent de rendre compte que d'un tiers de jugements alors que l'amplitude relative permet, selon lui, d'expliquer 90 % des jugements. Il propose la règle suivante : si l'intensité (VU) du noyau vocalique du terme A est supérieure ou égale à celle du terme B, la frontière sera perçue entre A et B. STREETER, par contre, constate que l'amplitude est l'indice le moins efficace pour lever l'ambiguïté et n'est utilisée que combinée à une information de durée. L'auteur s'étonne elle-même du faible "rendement" de l'amplitude et l'attribue au fait que l'amplitude était une variable inter-sujets, contrairement à F_0 et à la durée : chaque groupe de sujet effectuait tous ses jugements pour une seule condition d'amplitude. On peut penser, en effet, que le manque d'"ancrage", de valeur de comparaison ait nettement incité les sujets à négliger l'information fournie par l'intensité. Toutefois, une seconde expérience, où l'information d'amplitude était mise en conflit avec l'information de durée et d'intonation, confirme sa constatation. Mais il importe de noter que les expressions alors présentées aux sujets ne sont plus du tout "naturelles" : ce sont par construction, des formes "hybrides" combinant l'amplitude de la version AB/C et la durée et la mélodie de la forme A/BC. La tâche des sujets n'est donc plus exactement la même. On ne peut pas être certain que les stratégies perceptuelles qu'ils utiliseront resteront identiques (4) : tenteront-ils diverses "anamorphoses" pour retrouver une "naturalité virtuelle" ou vont-ils privilégier tel ou tel indice dont l'information qu'il véhicule est mémorisée comme "naturelle" ?.

Quant à la règle proposée par SCHOLES, elle doit être modifiée, au moins dans sa formulation, pour tenir compte de l'intensité spécifique des voyelles : deux noyaux vocaliques d'intensité différente peuvent être isophones. Quel indice joue alors : le niveau de pression acoustique ou la sonie ? Très probablement c'est la sonie. Ce problème étant mis à part, il semble que l'on puisse constater qu'effectivement l'amplitude chute plus après une frontière syntagmatique : la "règle" de SCHOLES rend bien compte des résultats observés par STREETER, tant que les expressions utilisées restent naturelles (figure 2 page 1587) et grâce au fait que les deux voyelles /a/ et /e/ sont dans le même ordre de classement en terme d'intensité spécifique que les voyelles de l'exemple de SCHOLES. Mais cette règle ne risque d'être généralisable que si elle s'exprime en terme de sonie. Seules des expériences complémentaires permettront de dire si oui ou non l'intensité corrigée ou la sonie remplit une fonction linguistique de démarcation syntagmatique.

---

(4) On sait, en effet, que la nature de la demande dans des tâches de mémorisation et/ou de compréhension influe sur les unités structurales utilisées et sur le niveau linguistique, le volume et les catégories de stimulus préservés par le sujet (AARONSON 1979)

I.2. Psychoacoustique et perception de parole :

En général en psychoacoustique, les tâches demandées aux sujets concernent soit l'évaluation globale d'une grandeur sensorielle (mesure absolue de sonie, de raucité, etc...) soit la discrimination d'une variation de grandeur physique (seuils différentiels, effet de masque, etc...). Par contre, les tâches demandées aux sujets dans les études de perception de parole sont beaucoup plus variées (5). Elles peuvent porter certes sur la discrimination des variations d'une grandeur physique (c'est par exemple le cas des expériences sur les pentes d'intensité que nous décrirons dans la deuxième partie) et sur l'évaluation globale d'une grandeur sensorielle (par exemple, les études sur la sonie des phrases FASTL 1977, PEDERSEN 1968, SORIN 1977, etc...) mais aussi sur l'identification de phonèmes (études de catégorisation), de frontières syntaxiques (par exemple, les études relevées ci-dessus), de formes linguistiques (par exemple, phrases interrogatives ou affirmatives GRAILLOT et BOE 1977, AUTESERRE et Di CRISTO 1972 ; présence ou absence d'emphase LIEBERMAN 1967, etc...). Pour les raisons que nous avons déjà évoquées plus haut (AARONSON 1976), il importe de bien distinguer les tâches d'identification et de discrimination (voir à ce sujet l'étude de POLLACK et PISONI 1971). Les études de discrimination sur sons de parole doivent nous permettre d'observer si la précision du traitement auditif est aussi grande que lorsqu'il s'agit de sons simples, sans signification linguistique. Les tâches d'identification, quant à elles peuvent aider à déterminer si dans un mode d'écoute "linguistique", le traitement perceptuel des paramètres physiques est ou non modifié par rapport au traitement "purement" psychoacoustique. Le but est de pouvoir faire la part, dans les effets perceptuels, de ce qui est résultat d'un traitement psychoacoustique observable pour tout signal acoustique complexe et de ce qui est résultat d'un traitement spécifique au signal de parole (SERNICLAES et SORIN 1978).

Nous limitant au rôle et traitement de l'intensité, nous prendrons, pour illustrer cette question, deux exemples : le premier concerne la discrimination de différences d'intensité sur des transitions de type formantique (DORMAN 1974, PASTORE et al. 1975, CUTTING et DORMAN 1976), l'autre concerne la perception de l'effet de déclinaison de la fréquence fondamentale (PIERREHUMBERT 1979).

DORMAN (1974) étudie la discrimination de variations d'intensité allant jusqu'à 9 dB sur les 60 ms de transitions de F_2 en présence et en l'absence de contexte syllabique, ainsi que sur les 60 premières millisecondes de voyelles stationnaires. Il observe que les différences d'intensité sont détectées avec un score du niveau du hasard quand elles sont appliquées aux transitions en contexte syllabique. Par contre, quand les différences d'intensité portent sur des transitions de F_2 isolées ou sur des voyelles stationnaires, elles sont parfaitement détectées. L'auteur en déduit que "dès que les indices acoustiques des plosives sont encodés dans une représentation

(5) Voir à ce sujet les remarques de LEVELT et FLORES d'ARCAIS in Structure and Process in Speech Perception (1975) sur le manque d'attention apportée en général à la structure des tâches dans les expériences de perception de parole ainsi que sur les différents usages du terme "perception" (tentant à englober les "jugements" sur certains faits de parole et la perception courante réelle). Les stades dont nous rendons compte ici relèvent en général du premier groupe ("jugements").

phonétique, toute l'information acoustique est stockée dans une mémoire auditive à court terme, relativement inacessible". PASTORE et al. (1975) contestent cette explication faisant appel à un encodage phonétique. Reproduisant l'expérience de DORMAN sur des sons purs et des variations linéaires de fréquence, ils relèvent en effet des résultats identiques à ceux de DORMAN. Un examen attentif de la tâche des sujets dans les trois cas (syllabe, voyelle stationnaire, transition isolée) les amène à mettre à part les résultats relevés sur voyelles stationnaires : dans ce cas, il s'agit, pour eux, d'une tâche de détection avec ancrage (référence), l'ancrage étant constitué par la partie constante de la voyelle ou du son pur continu. Or, comme nous l'avons déjà plusieurs mentionné, le type de tâche influe sur les performances du sujet. PASTORE et al rapportent que, d'après les travaux de TANNER (1958) et RABB et al. (1963), les performances dans des tâches de détection avec ancrage sont nettement supérieures aux performances dans des tâches de simple détection. Ceci leur permet de conclure que la principale raison des écarts observés entre la discrimination de variation d'intensité en contexte syllabique ou sur des transitions isolées réside en fait dans l'effet de masque antérieur.

CUTTING et DORMAN (1976) complètent l'expérience en faisant varier l'ampleur (en Hz) et la durée des transitions formantiques. Ils s'accordent avec PASTORE et al. pour conclure que l'effet observé est non-phonétique mais ils étudient plus précisément comment l'effet de masque intervient en perception de parole. Les différences de discrimination sont dues, d'après ces auteurs, plus à un effet de masque "de reconnaissance", tel qu'il a été défini par MASSARO (1975) qu'à un simple effet de masque périphérique.

Nous ne nous étendrons pas sur ce sujet, traité également par d'autres auteurs (BUDOHOSKA et SZYMANSKI 1974 par exemple). Nous nous contenterons de relever le fait suivant : CUTTING et DORMAN relèvent que si les stimuli syllabiques et les voyelles isolées sont présentés dans une même série de stimuli, la détectabilité des différences d'intensité augmente nettement pour les transitions en contexte syllabique et diminue pour les voyelles stationnaires. Ils en concluent que "une large proportion de l'effet est attribuable au contexte et non à l'effet de masque". Ce rôle du contexte mérite d'être souligné. Il reste toutefois à en expliquer les mécanismes.

D'autre part, ils remarquent que moins l'ampleur en Hz des transitions est grande, plus les différences d'intensité deviennent perceptibles. On rencontre ici de nouveau un fait souvent observé en perception de parole et que nous retrouverons dans l'étude sur la déclinaison : fréquence et intensité sont liées au niveau perceptuel. D'après CUTTING et DORMAN, ces deux paramètres pourraient même contribuer de manière interactive aux empreintes disponibles dans la mémoire à court terme. Nous nous garderons bien d'apporter une conclusion définitive à cet ensemble de remarques. Nous avons mentionné ces travaux pour mettre l'accent sur le rôle des tâches (et donc du protocole expérimental) sur les performances et stratégies des sujets, même dans une simple épreuve de discrimination et aussi pour montrer comment un résultat dû à première vue à un encodage phonétique spécial peut s'expliquer en terme de propriétés purement psychoacoustiques du signal.

Avant d'évoquer le rôle de l'intensité dans la perception de l'effet de déclinaison, nous mentionnerons brièvement, pour compléter ce tour d'horizon, le rôle de l'intensité relevé par SYRDAL-LASKY (1978) dans une tâche d'identification et de discrimination. Etudiant l'influence de l'intensité sur la perception catégorielle des plosives et de transitions isolées de F_2, l'auteur montre que l'identification se dégrade quand l'intensité augmente et que l'ampleur des pics de discrimination diminue, les causes de ce phénomène restant encore très obscures, comme l'auteur le conclut.

L'étude de PIERREHUMBERT (1979) concerne l'effet de déclinaison, c'est-à-dire la décroissance régulière de F_0 le long d'une phrase et son rôle linguistique éventuel. Les stimuli utilisés étaient des phrases sans signification du type "ma Ma ma ma Ma ma" dont les contours d'amplitude et de F_0 étaient modifiés. Les sujets avaient à juger laquelle des deux syllabes accentuées était la plus aigüe. La tâche comportait donc une identification (détermination des syllabes accentuées) et une comparaison de hauteur. Les résultats montrent que si les deux syllabes accentuées ont la même fréquence fondamentale, la deuxième paraît plus aigüe que la première. Quant à l'intensité, elle intervient de manière significative : dans la phrase naturelle, où l'amplitude de la deuxième syllabe est inférieure de 3,6 dB à celle de la première syllabe, la deuxième syllabe doit avoir un F_0 inférieur de 11 Hz à celui de la première syllabe pour être perçue à la même hauteur. Par contre, quand les deux syllabes ont la même intensité, l'écart doit atteindre 17 Hz. Autrement dit, une augmentation du niveau entraîne une augmentation de la fréquence perçue. Les résultats classiques de la psychoacoustique ne permettent pas d'expliquer ce phénomène. En effet, l'effet d'intéraction fréquence-intensité observé par divers auteurs sur des sons purs et plus récemment par TERHARDT (1974) prédirait, pour le domaine de fréquence concerné, un effet exactement opposé.

Il est tentant de comparer ces résultats à ceux observés par ROSSI (1978) concernant l'intéraction de glissandos de fréquence et de pentes d'intensité sur des voyelles isolées de 200 ms. En effet, ROSSI montre qu'une pente d'intensité négative associée à un glissando décroissant diminue la sensibilité auditive c'est-à-dire augmente l'ampleur du glissando de fréquence nécessaire pour être perçu. Or, c'est juste le contraire qu'observe PIERREHUMBERT : en présence d'une diminution d'intensité (cas naturel), un écart de 11 Hz est perçu comme un écart de 17 Hz sans diminution d'intensité. Toutefois, il faut rappeler que dans l'expérience de PIERREHUMBERT les comparaisons portent sur deux syllabes séparées d'environ 600 ms alors que chez ROSSI les pentes d'intensité et de fréquence portent sur 200 ms. Ces deux durées correspondent à des segments de parole sur lesquels les traitements linguistiques possibles sont très différents.

Les résultats sur la déclinaison peuvent être également utilement comparés aux observations de LIEBERMAN (1960) sur la perception de l'accent dans des mots bisyllabiques. Il mentionne l'existence d'un effet de compensation entre F_0 et l'intensité dans la perception de l'accent naturel : une variation de F_0 cohérente avec l'accent perçu compense une modification non-cohérente d'amplitude. Or, pour en revenir à l'expérience de PIERREHUMBERT, dans la phrase naturelle, la deuxième syllabe accentuée est moins intense que la première. Si l'on élève artificiellement son intensité, l'oreille exigera une diminution plus grande de F_0 que dans la parole naturelle pour "compenser" cet accent exagéré. Mais cette opération exige du sujet l'utilisation d'une connaissance de l'évolution prosodique normale d'une phrase. Ainsi, le traitement perceptuel qu'il utilise inclut une analyse linguistique contrairement au traitement psychoacoustique des signaux brefs.

L'on peut donc conclure que les traitements de l'intensité observés diffèrent suivant la longueur du signal de parole utilisé. En particulier, si la durée d'observation choisie est trop brève l'importance des fonctions de l'intensité dans la parole pourra être sousestimée. D'autre part, les intéractions perceptuelles entre F_0 et I au niveau intersyllabique semblent être différentes de celles existant au niveau intrasyllabique. Ceci renforce l'intél'intérêt d'études sur la perception de parole continue.

II - DISCRIMINATION ET IDENTIFICATION DE VARIATIONS LINEAIRES D'INTENSITE EN PAROLE CONTINUE

Nous venons de relever un certain nombre de phénomènes perceptuels liés au traitement segmental et suprasegmental de l'intensité. En français, comme nous l'avons indiqué dans l'introduction, l'intensité semble intervenir particulièrement en fin de groupe intonatif(6). Avant d'analyser quelle est, à ce niveau, l'importance de sa fonction linguistique, il semble utile de connaître avec quelle précision ce paramètre est perçu par l'ouïe, dans un contexte naturel de parole continue. Ainsi, l'expérience que nous allons décrire ne permet pas de conclure sur la fonction linguistique de l'intensité mais indique quel est le traitement auditif auquel ce paramètre est alors soumis et sa "marge d'erreur". Cette connaissance doit permettre d'étudier ultérieurement plus rigoureusement l'utilisation linguistique de ce critère.

Rappelons d'abord quelques valeurs de seuils différentiels d'intensité relevés sur des sons de parole. FLANAGAN (1955, 1957) met en évidence que le seuil différentiel d'intensité est de $\pm$ 3 dB pour le deuxième formant et de $\pm$ 1 dB pour une voyelle stationnaire, ce qui est également une bonne approximation du seuil différentiel sur le premier formant. Quant au troisième formant, le seuil différentiel d'intensité y serait de $\pm$ 5 dB. D'après les études de DORMAN (1974) citées plus haut, on peut retenir que le seuil différentiel d'intensité sur une transition formantique en contexte syllabique est de l'ordre de 9 dB. Rappelons, toutefois, que le contexte expérimental semble influer notablement sur les résultats. Quant au seuil différentiel d'intensité sur des phrases, la pratique courante des équipes de téléphonométrie des laboratoires de télécommunications nous en donne une idée assez précise. Les opérateurs de téléphonométrie comparent en effet la force sonore d'une phrase normalisée émise à travers une ligne de référence et une ligne à tester. La précision de leur mesure atteint $\pm$ 1 dB. Bien que cette phrase normalisée soit à l'usage devenue peu représentative de la parole courante (prononciation "compressée" en niveau et monotone en intonation), on peut cependant supposer que le seuil différentiel sur des phrases réelles sera du même ordre. C'est en effet la précision qui a été relevée dans des expériences faites antérieurement (SORIN 1976) concernant la comparaison en sonie de phrases naturelles présentées à l'endroit et à l'envers. Le seuil différentiel d'intensité sur des phrases est donc du même ordre que celui observé sur voyelles stationnaires.

Il est sans doute rare, dans la perception courante de parole continue, que l'auditeur soit mené à faire une réelle évaluation de seuils différentiels d'intensité. Par contre, il est amené en permanence à traiter des variations plus ou moins rapides d'intensité pour en tirer des informations tant pour la discrimination phonétique que pour des jugements prosodiques. La connaissance de la sensibilité auditive à des variations linéaires de pression acoustique doit donc être plus prégnante que celle sur les seuils différentiels si l'on vise une description des processus de traitement réel de la parole continue.

---

(6) A titre d'exemple, si l'on veut générer en français une phrase interrogative à partir d'une phrase énonciative, il faut modifier le contour de F_0 mais aussi augmenter l'intensité de la partie finale et diminuer celle de l'avant dernière syllabe (CONTINI et BOE 1975).

II.1. Discrimination de crescendos d'intensité en fin de phrase (7) :

Les expériences que nous avons menées visaient à déterminer la discrimination de variations d'intensité portant sur les 200, 400 et 600 dernières millisecondes de six phrases énonciatives et interrogatives, prononcées chacune par un homme et une femme. La modification apportée aux phrases naturelles était une modification linéaire de pression dont l'amplitude finale G était réglable par pas de 1 dB. Sept sujets avaient à déterminer, par la méthode des limites (seuil ascendant et descendant) le seuil de détection d'une modification en fin de phrase, sans avoir à préciser la nature de cette modification. Leur tâche était donc de nature purement psychoacoustique. Les résultats (figures 1.a) et 1.b)) montrent qu'une modification du signal est perçue dès que l'amplitude finale G du crescendo dépasse 5 dB lorsqu'il porte sur les 200 dernières millisecondes et que le seuil s'affine ensuite au fur et à mesure que la durée du crescendo augmente. Pour une durée de crescendo de 400 ms, le seuil est atteint pour G = 3 dB et s'abaisse jusqu'à G = 2 dB pour un crescendo portant sur les 600 dernières ms (dans une préexpérience, on avait pu observer que le seuil ne se modifiait plus pour des durées supérieures à 600 ms). Une analyse de variance (3(durée) x 6(phrase) x 2(sexe locuteur)) ne met en évidence aucun effet phrase significatif (l'intensité spécifique des voyelles n'intervient donc pas directement), mais confirme que l'effet durée est significatif ($F(2(216) = 157\ p < 0.01$).

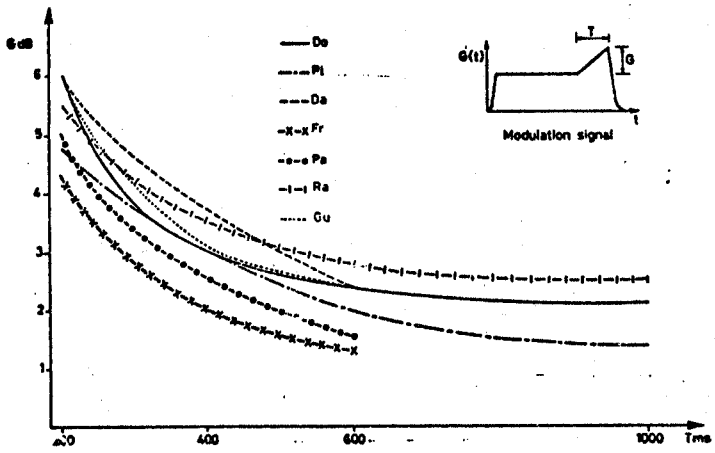


FIGURE 1.a) : Valeurs du gain G juste perçu en fonction de la durée T du crescendo pour 7 sujets (moyennés sur 12 phrases).

Values of the gain G just perceived as function of duration T of the rising glide for 7 (seven) subjects (average of 12 (twelve) sentences).

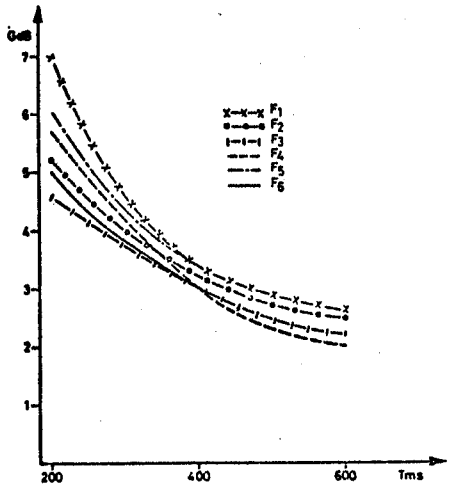


FIGURE 1.b) : Valeurs du gain G juste perçu en fonction de la durée T pour 6 phrases (voix de femme) moyennées sur 7 sujets. L'écart type inter-individuel pour chaque phrase (voix d'homme et de femme) est inférieur à :

Values of the gain G just perceived as function of the duration T for 6 sentences (female voice) average on 7 subjects. The intersubject standard-deviation for each sentence (male and female voice) is inferior to :

$\sigma = 1{,}4$ *dB for* $T = 200$ *ms*
$\sigma = 1{,}2$ *dB for* $T = 400$ *ms*
$\sigma = 1$ *dB for* $T = 600$ *ms.*

(7) Une partie des travaux évoqués ici a été exposée en détail dans SORIN (1978).

Afin d'étudier si ces résultats correspondaient à un traitement "spécifique" au signal de parole, l'expérience a été reproduite avec des signaux de bruit "à spectre moyen de parole" dont l'enveloppe était modulée par l'enveloppe des 12 phrases précédemment utilisées. En raison de difficultés expérimentales, seuls huit parmi les douze signaux ont été testés. Cinq sujets parmi les sept précédents ont pris part à l'expérience. Les résultats obtenus (figure 2) mettent en évidence le même phénomène : une décroissance significative de l'amplitude du crescendo juste détecté quand la durée de la modification augmente (analyse de variance 3(durée) x 8 (signaux) $F(2,96) = 11,6$ $p < 0,01$). Comme pour les phrases, le seuil atteint pour une durée du crescendo supérieure ou égale à 600 ms, le seuil différentiel d'intensité observé sur les phrases entières.

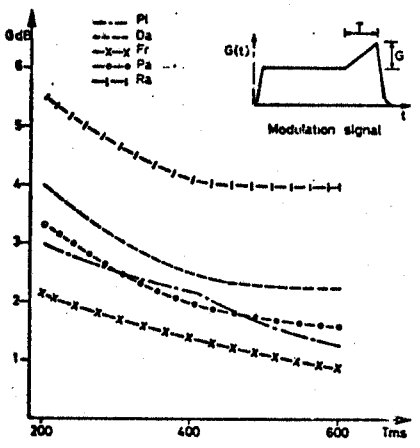


FIGURE 2 : Valeurs du gain G juste perçu en fonction de la durée T du crescendo pour 5 sujets (moyennées sur 8 signaux de bruit à enveloppe de phrase).

Value of the gain G just perceived as function of the duration T of the rising glide for 5 (five) subjects (average of 9 (nine) white noise signals with sentence envelope).

Une analyse de variance (3 (durée) x 8 (signaux) x 2 (phrase/bruit)) confirme que les résultats obtenus par les cinq sujets communs sur les phrases et les signaux de bruit modulé ne sont pas statistiquement significatifs. L'hypothèse nulle est confirmée par un test d'homogénéité de deux variances ($F(23,23) = 1,25$) à un niveau de signification $\alpha = 0.05$.

L'hypothèse que nous suggérons pour rendre compte de ce phénomène postule que l'oreille intègre la partie finale du signal à partir du maximum T_M d'intensité situé à plus de 200 ms de la fin T_F de la phrase. Cette intégration s'applique donc, au moins, sur toute la syllabe finale. C'est ensuite sur les valeurs de cette puissance moyenne que porte la comparaison. Une modification est détectée dès que l'augmentation ΔP de cette puissance moyenne (donnée ci-dessous) par rapport à la puissance moyenne correspondante de la phrase naturelle dépasse 1 à 2 dB.

$$\Delta P = 10 \log \frac{\int_{T_M}^{T_F} \overline{G(t)}^2 \, p^2(t) \, dt}{\int_{T_M}^{T_F} p^2(t) \, dt} \text{ dB}$$

Ce processus nous permet de prédire correctement l'amplitude observée des pentes d'intensité liminaires pour chaque phrase et signal de bruit (SORIN 1978).

II.2. Discussion :

Ces résultats méritent d'être comparés avec ceux obtenus par ROSSI (1978) dans une étude portant sur les glissements d'intensité sur des voyelles stationnaires de 200 ms. ROSSI conclut en effet qu'une pente d'intensité atteignant moins de 11 dB sur 200 ms n'est pas perçue, que cette pente soit positive ou négative. L'important écart entre cette valeur et les résultats de notre expérience doit être étudié en détail pour les conséquences qu'il pourrait avoir au point de vue méthodologique. En effet, si les résultats perceptuels diffèrent aussi notablement pour des tâches "purement" auditives lorsque l'on passe de voyelles stationnaires à des phrases entières, l'intérêt d'une approche "microscopique" du traitement de la prosodie dans la parole continue s'en trouverait nettement limité.

En fait, les deux expériences diffèrent non seulement par les stimuli utilisés mais par leur but même et donc par leur procédure. Pour ROSSI, il s'agissait avant tout de déterminer la sonie globale de glissement d'intensité, les valeurs de seuil étant déduites de mesures indirectes, comme l'auteur le mentionne lui-même. Quant à nous, le but de l'expérience était bien une mesure de discrimination d'une pente croissante d'intensité.

Examinons précisément les tâches demandées aux sujets dans l'expérience de ROSSI. Ceux-ci avaient, dans une série de tests, à indiquer si les deux stimuli d'une paire (constituée d'une voyelle stationnaire et d'une voyelle à intensité croissante ou décroissante, ou de deux voyelles identiques d'intensité croissante ou décroissante) étaient identiques ou différents. Dans une autre série de tests, les mêmes sujets devaient toujours indiquer si le deuxième stimulus d'une paire était égal, plus fort ou moins fort que le premier (évaluation globale et mesure de sonie). Pour mieux comprendre sur quel critère les sujets ont basé leur réponse dans chaque test, il nous semble important d'expliquer pourquoi l'on observe un pourcentage important de "fausse alarme" dans le premier test (identique - différent). En effet, deux voyelles identiques soumises à des pentes identiques sont percues comme identiques avec un pourcentage seulement de l'ordre du hasard, quand les sujets atteignent 15,8 dB et 20 dB (AV et BO figure 5 ROSSI, 1978). Tout se passe ici comme si les sujets, incapables d'"intégrer" la variation d'intensité pour en déduire une sonie globale, se déclaraient incapables de juger si oui ou non les stimuli étaient identiques. Le critère recherché semble donc avoir été un critère de force sonore globale et non de similitude de "nature" temporelle . Nous pensons que les sujets ont pu être incités à utiliser ce critère par le contexte expérimental : en effet, les paires étaient toutes constituées d'une voyelle stationnaire et d'un voyelle "à pente" mises à part les paires de contrôle constituées de voyelles "à pente" identiques sur lesquelles on observe justement un fort pourcentage de fausse alarme. On peut se demander si les résultats auraient été identiques si le "corpus" avait contenu également des paires constituées de deux voyelles "à pente", différentes, localisant ainsi l'attention du sujet sur la similitude d'évolution temporelle...

Ainsi, il nous semble que l'on peut déduire de l'expérience de ROSSI qu'une pente d'intensité peut être "intégrée" et donner lieu à une sensation de sonie globale tant que la pente ne dépasse pas 11 dB. Mais l'on ne peut en déduire que toute pente d'intensité inférieure à cette valeur n'est pas détectée.

En fait, l'ambiguïté réside dans l'expression "perception d'un glissement d'intensité" utilisée par ROSSI : s'agit-il d'une détection et identification du glissement en tant que tel ou d'une simple discrimination d'une modification ? Il aurait été plus précis de parler de seuil d'identification du glissement pour évoquer la valeur de la pente à partir de laquelle le sujet reconnaît à coup sûr la présence d'un crescendo ou d'un décrescendo d'intensité et de garder le terme "seuil de discrimination" pour la valeur de la pente à partir de laquelle une modification du signal est perçue par rapport à un signal stationnaire. Les résultats de ROSSI permettent d'affirmer que le seuil de discrimination est inférieur à 7,8 dB pour une pente croissante sur 200 ms puisque celle-ci donne lieu à un accroissement du niveau équivalent de plus de 5 dB par rapport au signal stationnaire ayant même intensité que l'intensité initiale du crescendo. Pour une pente décroissante, le seuil est de l'ordre de 4 dB puisqu'un tel décrescendo donne lieu à une diminution de niveau équivalent d'environ 3 dB, valeur qui compte tenu de la précision des résultats est très proche du seuil différentiel d'intensité sur voyelles stationnaires. Quant au seuil d'identification d'une pente en tant que telle, il serait notablement supérieur : les résultats de ROSSI laissent supposer qu'il est de l'ordre de 11 dB tant pour une pente positive que négative.

Cette mise au point permet de comparer plus rigoureusement les résultats que nous avons obtenus sur des phrases avec ceux de ROSSI sur les voyelles. Le seuil de discrimination d'un crescendo serait donc, suivant notre analyse, inférieur à 7,5 dB pour une voyelle de 200 ms et de 5 dB pour un crescendo de même durée en fin de phrase (8). L'écart est donc très faible.

Quant à l'interprétation des résultats observés, elle diffère également entre les deux auteurs. ROSSI explique que la sonie d'un glissement d'intensité est déterminée par l'intensité du signal aux 2/3 de sa durée pour une pente positive et au 1/3 pour une pente négative. Nous n'avons pas pu vérifier cette hypothèse sur les résultats de la discrimination de crescendos en fin de phrase : les valeurs d'intensité aux 2/3 du signal sont très différentes suivant les phrases et conduisent donc à des seuils calculés très différents. Par contre, notre hypothèse supposant que l'oreille intègre la pression acoustique sur une durée supérieure ou égale à 200 ms permet de prédire les résultats observés par ROSSI avec une précision inférieure à 1 dB.

En effet soient P_1 et P_2 les pressions acoustiques initiales et finales du crescendo et P_m la pression acoustique moyenne $(P_1 + P_2)/2$. Le "point d'égalisation" (PE) relevé par ROSSI correspond à la différence de niveau entre la fin du crescendo et le niveau de la voyelle stationnaire jugée isophone. Suivant notre hypothèse, il doit correspondre à la différence $D = 20 \log P_2 - 20 \log P_m$. Le tableau ci-dessous donne les valeurs relevées par ROSSI et les valeurs de D calculées suivant l'expression ci-dessus pour les différentes valeurs de pente.

---

(8) Si des données ultérieures venaient confirmer que le seuil de discrimination de pente est significativement plus petit lorsqu'il porte sur la fin d'une phrase que sur une voyelle stationnaire de même durée, il faudrait se demander dans quelle mesure cet avantage ne serait pas dû au fait que la tâche sur les phrases correspond à une expérience avec ancrage (dont nous avons mentionné les particularités plus haut) l'ancrage étant constitué par la première partie non modifiée de la phrase. Cela impliquerait alors que le sujet détecte le crescendo non pas en comparant les parties finales des phrases naturelles et modifiées mais en comparant la sonie de la partie finale modifiée à la sonie de la portion fixe de la phrase, ou à la sonie finale attendue (si l'on suppose que le sujet a en mémoire des tracés de l'évolution normale de l'intensité d'une phrase). La phrase de référence doit alors devenir inutile. Ceci reste à étudier.

	D	PE
20 dB	5,2 dB	6 dB
15,8 dB	4,6 dB	4,5 dB
9,6 dB	3,6 dB	3,5 dB
7,8 dB	3,1 dB	2,5 dB
16 dB	5,1 dB	6 dB
11 dB	4 dB	4 dB
8,8 dB	3,2 dB	3,5 dB
3,8 dB	1,8 dB	1 dB

TABLEAU 1 :

Comparaison des valeurs de PE relevées par ROSSI (1978) et calculées (D) par l'auteur pour des pentes de l'intensité croissantes et décroissantes.

TABLE 1 :

Comparison of the value PE observed by ROSSI (1978 b) and calculated (D) by the author, according the syllabic integration hypothesis, for rising and decreasing glides on vowels.

Nous pouvons déduire de ce bon accord que deux voyelles, l'une d'intensité constante, l'autre d'intensité croissante ou décroissante sont jugées également fortes si leurs niveaux de pression acoustique moyennée sur les 200 ms de leur durée sont identiques.

En conclusion, l'hypothèse suivant laquelle la dicrimination d'une pente d'intensité peut être effectuée à partir d'une intégration de la pression acoustique du signal sur une durée au moins égale à 200 ms s'applique aussi bien aux résultats obtenus en fin de phrases que sur une voyelle isolée. Ce mécanisme ne nécessite aucune identification linguistique préalable du stimulus : en effet des résultats identiques ont été obtenus en remplaçant les phrases par des signaux de bruit à bande large dont l'évolution temporelle de l'enveloppe était calquée sur celle des phrases. On peut penser qu'il s'agit là d'un mode de traitement intersyllabique de l'intensité valant pour des durées de modifications longues. En ce qui concerne la perception de variations d'intensité plus brèves, plusieurs travaux portent à croire que leur importance pour la discrimination phonétique (STEVENS 1973 et 1979 par exemple) mais aussi pour la perception catégorielle de stimuli non linguistiques (CUTTING et ROSNER 1974), nécessité l'existence d'une détection spécifique. Cette détection d'une variation brusque-lente de l'intensité porte sur des durées très brèves (de l'ordre de 50 ms). Elle peut donc relever d'un mode de traitement différent et indépendant de celui des modifications plus longues évoqué ci-dessus. Une certaine analogie peut être relevée avec le traitement de la fréquence : NABELEK et HIRSCH (1969) mentionnent qu'il existe deux mécanismes de discrimination de variation de fréquence. Le premier intervient pour des variations rapides sur un grand intervalle de fréquence, le deuxième pour des variations lentes et d'excursion faible. POLLACK (1968) et TSUMURA, SONE et NIMURA (1973) montrent que suivant l'information dont il dispose et suivant la tâche demandée, le sujet détecte une variation de fréquence soit en comparant les hauteurs initiales et finales soit en percevant directement un glissement de fréquence.

Nous sommes donc amenéesà faire l'hypothèse de l'existence de modes de traitement différents de l'intensité, suivant que l'information perceptuelle recherchée se situe à un niveau segmental ou suprasegmental. Cette hypothèse s'accorde bien avec le modèle de perception de parole proposé par COLE et SCOTT (1974). Ces deux auteurs proposent un modèle comprenant trois processus de traitement différents au niveau de l'analyse acoustique : détection d'indices invariants et transitionnels au niveau segmental et utilisation des propriétés de l'enveloppe au niveau suprasegmental. Ceci suppose que l'on puisse traiter à la fois des unités de parole de longueur très faible (transition) et de longueur beaucoup plus importante, de l'ordre de la syllabe, voir de plusieurs syllabes (KOZHEVNIKOV et CHISTOVITCH 1965 MILLER 1962). L'intéraction ou la hiérarchie de ces deux types de traitement est encore mal connue. MILLER (1978) montre, dans une étude récente portant à la fois sur l'identification de voyelles en contexte syllabique et sur leur discrimination en hauteur et sonie qu'il existe une intéraction symétrique entre les traitements segmentaux et suprasegmentaux contrairement à ce qui avait été trouvé par WOOD (1974) pour les consonnes. Celui-ci mentionnait en effet que l'identification des consonnes semblait dépendre de l'analyse suprasegmentale. Toutefois, il reste à déterminer si c'est l'analyse acoustique préliminaire ou la réponse de "l'étage" phonétique qui détermine le mode d'intéraction mis en jeu.

Les résultats de notre expérimentation ont permis de montrer que le traitement suprasegmental de l'intensité observé peut être efffectué indépendamment d'une identification linguistique du stimulus et correspond donc à un processus "purement" psychoacoustique. Il reste à savoir comment ce type de traitement est utilisé dans la perception courante de la parole, par exemple dans le cas où la tâche du sujet ne porte pas sur la perception de l'intensité elle-même mais sur l'identification de schémas prosodiques ou de frontières syntaxiques. LIEBERMAN (1967) rappelle en effet que si des sujets sont capables de discriminer très finement des différences d'intensité et même de catégoriser la sonie de sons purs en quatre classes, ils ne font qu'une distinction catégorielle binaire de la proéminence quand ils écoutent de la parole continue.

L'étude de la fonction linguistique de l'intensité en fin de phrase reste donc à faire. Nous n'avons rendu compte ici que de la manière dont l'intensité est traitée pour devenir indice.

CONCLUSION

Nous avons voulu montrer, par ce bref survol de travaux aux buts très divers :

- que l'intensité est une information utilisée à de nombreux endroits du processus de perception de parole
- qu'il existe un niveau psychoacoustique de traitement du signal de parole où la pression acoustique est intégrée sur une durée supérieure à 200 ms pour servir d'indice suprasegmental
- que les fonctions linguistiques de l'intensité restent encore à préciser, tant dans leurs mécanismes que dans leurs effets.
- que toute étude de la fonction linguistique d'un paramètre physique du signal gagne à être précédée d'une étude psychoacoustique de ce paramètre en contexte permettant de connaître la capacité auditive et les modes de traitement auditif mis en jeu par le sujet. Cette approche permet de mieux situer où, quand et comment interviennent les effets propres à la perception de parole ("purement" linguistiques).

BIBLIOGRAPHIE

AUTESERRE D. et Di CRISTO A., 1972, Recherches psychosémantiques sur l'intonation de la phrase française TIPA AIX - Vol. I , 61-98.

AARONSON D., 1976, Performance theories for sentence coding : Some Qualitative Observation, J. Of Exp. Psychol. : Hum. Percep. & Perf., Vol. 2 n° 1, 42 - 55.

BUDOHOSKA W., and SZYMANSKI L., 1974, Backward and Forward masking between two speech-like sounds, Acta Neurobiol. Exp., 34, 723 - 736.

BOLINGER D., 1958, A theory of pitch accent in English, Word, 14, 109.

COLE R.A. & SCOTT B., 1974, Toward a theory of Speech Perception, Psychological Review, Vol. 81, N° 74, 348 - 374.

CONTINI M., et BOE L.J., 1975, Etude quantitative de l'intonation en français, Proc VIII ICPS Leeds.

CUTTING J.E. & DORMAN M.F., 1976, Discrimination on intensity differencies carried on formant transitions varying in extent and duration, Perception & Psychophysics, 20, 101 - 107.

CUTTING J.E. & ROSNER B.S., 1974, Categories and Boundaries in speech and music Perception and Psychophysics Volume 16 (3) 524 - 370.

DENES P., 1959, A preliminary investigation of certain aspects of intonation Language and Speech, 2, 106.

DELATTRE P., 1951, Principe de phonétique française, Middlebury College.

DORMAN M., 1974, Discrimination of intensity differences on formant transitions in and out of syllable context, Perception and Psychophysics, 16, 84 - 86.

FASTL H., 1977, Loudness of running speech, J. Audiol. Technique, 16, 2 - 13

FLANAGAN J., 1955, A difference limen for vowel formant frequency, JASA, Vol. 27, 613 - 617.

FLANAGAN J., 1957, Estimates of the Maximum Precision Necessary in Quantizing Certain "Dimensions" of vowel sounds, JASA Vol. 29, N° 4, 533 - 534.

FRY D.B., 1958, Experiment in the perception of stress, Language and Speech 1, 126 - 152.

FRY D.B., 1955, Duration and intensity as physical correlates of linguistic stress JASA 27, 765 - 768.

GRAILLOT P. & BOE L.J., 1977, Relation entre l'évolution de la fréquence laryngienne et de l'intensité pour la phrase énonciative, GALF, 8ème JEP - Aix-en-Provence.

HADDING-KOCH K. & STUDDERT-KENNEDY M., 1965, An experimental study of some intonation contours, Phonetics, 2, 175 - 210.

KOZHEVNIKOV V. & CHISTOVITCH L., 1965, Speech : Articulation and Perception, MOSCOU, LENINGRAD.

LADEFOGED P., 1961, Subglottal activity during speech, Proc. 4 ICPS, Helsinki 73 - 91.

LADEFOGED P. & N.P. MAC KINNEY, 1963, Loudness, sound pressure level and subglottal pressure in speech, JASA, 35, 454 - 460.

LEHISTE I. & G.G. PETERSON, 1959, Vowel amplitude and phonemic stress in American English, JASA, 31, 428 - 435.

LIEBERMAN P., 1960, Some acoustic correlates of word stress in American English, JASA, 32, 451 - 454.

LIEBERMAN P., 1967, Intonation, perception and language, MIT Press, Cambridge.

MARCHAL A., 1976, Notions de physiologie pulmonaire appliquées à la description phonétique de l'accent d'insistance en français, Studie Phonetica, Didier, 12, 93 - 120.

MASSARO D. W., 1975, Backward recognition masking, JASA, 58, p. 1059 - 1065.

MILLER G.A., 1962, Decision units in the perception of speech, IEEE Transactions on Information Theory, IT. 8, 81 -83.

MILLER J.L., 1978, Interactions in processing segmental and suprasegmental features in speech, Perception & Psychophysics, Vol. 24(2) 175 - 180.

MORTON J. & JASSEM W., 1965, Acoustic correlates of stress, Language and Speech, 8, 159 - 181.

NABELEK & HIRSCH, 1969, On the discrimination of frequency transitions, JASA, 45, pp. 1510.

PASTORE et al. 1975, Discrimination of intensity differences on formant like transitions, Perception and Psychophysics, 18, 224 - 226.

PEDERSEN O.J. & PEDERSON S.B., 1968, Possibilities for the objective measurement of the loudness level of speech, Proc. 6. ICA Tokyo, A-65.

POLLACK I., 1968, Detection of rate of change of auditory frequency, J. of Psychol. 77, 535 - 541.

POLLACK I. & PISONI D.B., 1971, On the comparison between identification and discrimination test in speech perception, Psychon. Sci., Vol. 24 (6), 299 - 300.

PIERREHUMBERT J., 1979, The declination effect in perception, JASA, Vol. 66, N° 2, 363 - 370.

RABB D.H., OSMAN E. & RICH E. 1963, Effects of waveform correlation and signal duration in detection of noise bursts in continuous noise, JASA, 35, 1942 - 1946.

RIGAULT A., 1961, Rôle de la fréquence de l'intensité et de la durée vocalique dans la perception de l'accent en français Proc.4 ICA Helsinki 735-748.

ROSSI M. et Di CRISTO A., 1977, Propositions pour un modèle d'analyse de l'intonation - 8ème JEP LANNION - 323-332.

ROSSI M., 1971, L'intensité spécifique des voyelles, Phonetica, 24, 129 - 161.

ROSSI M., 1978 a), Interaction of intensity glides and frequency glissandos. Language and Speech, Vol. 21, part. 4, 384 - 396.

ROSSI M., 1978 b), The perception of non-repetitive intensity glides on vowels. J. of Phonetics, 6, 9 - 18.

SCHOLES R.J., 1971, On the spoken disambiguation of superficially ambiguous sentences, Language and Speech, 14, 1 - 11.

SEGUINOT A., 1976, L'accent d'insistance en français standard, Studia Phonetica, Didier, 12, 1 - 58.

SERNICLAES W. & C. SORIN, 1978, Le traitement perceptif des indices acoustiques et des traits phonétiques, 9ème JEP LANNION - Synthèse des rapporteurs, Revue d'Acoustique, Vol. 11, N° 47, 197 - 202.

SORIN C., 1976, Etude de la sonie de la parole. Quelques expériences préliminaires, Recherches acoustiques, CNET LANNION, Vol. III, 111 - 131.

SORIN C., 1978, Seuils différentiels et perception de crescendos d'intensité en fin de phrase, Proc. 9ème JEP LANNION, 57 - 69.

STEVENS K.N., 1973, Potential role of property detectors in the perception of speech, MIT RLE, Quaterly Progress Report, 155 - 168.

STEVENS K.N., 1979, The search for invariant acoustic correlates of phonetic features in "Perspectives in the study of speech", ed. by P.D. EIMAS & J. MILLER.

STREETER L.A., 1978, Acoustic determinants of phrase boundary perception, JASA, Vol. 64 (6), 1582 - 1592.

SYRDAL-LASKY A., 1978, "Effects of intensity on the categorical perception of stop consonants and isolated second formant transitions". Perception and Psychophysics, Vol. 23(5), 420 - 432.

TANNER W.P. Jr., 1958, What is Masking ? JASA, 30, 919 - 921.

TERHARDT E., 1974, Pitch of pure tones : its relation to intensity in Facts and Models in Hearing, ed. by E. ZWICKER and E. TERHARDT, Springer Verlag.

TSUMURA T., SONE T. et NIMURA T., 1973, Auditory detection of frequency transition JASA, 53, 17 - 25.

WOOD C.C., 1974, Parallel processing of auditory and phonetic information in speech discrimination. Perception & Psychophysics, 15, 501 - 508.