

Sc. N. 74/7 B

THÈSES

présentées à

L'UNIVERSITÉ DE NANCY I

pour obtenir le grade de

DOCTEUR ÈS-SCIENCES PHYSIQUES

par

Jean-Paul HATON

Agrégé de l'Université

PREMIÈRE THÈSE

**CONTRIBUTION A L'ANALYSE, LA PARAMÉTRISATION
ET LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE.**

DEUXIÈME THÈSE

**PROPOSITION DONNÉE PAR L'UNIVERSITÉ : NOUVELLES
STRUCTURES DE MACHINES AUTOMATIQUES.**

Soutenues publiquement le 8 Janvier 1974, devant la Commission d'Examen

Membres du Jury :

Président : M. R. MAINARD

Examineurs : MM. B. PICINBONO

J.-C. SIMON

C. PAIR

M. LAMOTTE

Rapporteur : M. A. FRUHLING

1974

THÈSES

présentées à

L'UNIVERSITÉ DE NANCY I

pour obtenir le grade de

DOCTEUR ÈS-SCIENCES PHYSIQUES

par

Jean-Paul HATON

Agrégé de l'Université

**PREMIÈRE THÈSE : CONTRIBUTION A L'ANALYSE, LA PARAMÉTRISATION
ET LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE.**

**DEUXIÈME THÈSE : PROPOSITION DONNÉE PAR L'UNIVERSITÉ : NOUVELLES
STRUCTURES DE MACHINES AUTOMATIQUES.**

Soutenues publiquement le 8 Janvier 1974, devant la Commission d'Examen

Membres du Jury :

Président : M. R. MAINARD

Examineurs : MM. B. PICINBONO

J.-C. SIMON

C. PAIR

M. LAMOTTE

Rapporteur : M. A. FRUHLING

1974

A MES PARENTS

A MARIE-CHRISTINE

ET A NOS GARÇONS

AVANT - PROPOS

Ce travail a été réalisé au Laboratoire d'Electricité et d'Automatique, dans le cadre des activités du Groupe d'Etude sur la Parole.

Que Monsieur le Professeur A. FRUHLING, Directeur du L. E. A. , veuille trouver ici l'expression de ma profonde gratitude pour les conseils et le soutien qu'il n'a cessé de me prodiguer depuis mon arrivée au Laboratoire.

Monsieur le Professeur B. PICINBONO, Directeur du Laboratoire des Phénomènes Aléatoires et Président de l'Université de Paris-Sud, m'a fait le grand honneur de s'intéresser à ce travail. Je l'en remercie vivement, ainsi que d'avoir accepté de participer au Jury.

Je suis reconnaissant à Monsieur le Professeur R. MAINARD, Directeur de l'Institut Universitaire de Technologie de Nancy, de nous faire la confiance de présider la Commission d'Examen.

Je remercie Monsieur le Professeur J. -C. SIMON, Directeur de l'Institut de Programmation de l'Université de Paris VI, qui a très gentiment accepté de prendre connaissance de mon mémoire et de siéger à mon Jury.

Mes remerciements vont également à Monsieur le Professeur C. PAIR, Directeur de l'Institut Universitaire de Calcul Automatique de Nancy, pour l'intérêt qu'il porte à nos activités et pour sa présence dans ce Jury.

Je remercie aussi Monsieur M. LAMOTTE, Chargé de Recherches au CNRS, de la marque d'amitié qu'il m'a faite en acceptant de participer au Jury. Je tiens à l'assurer du plaisir que j'ai à collaborer avec lui au développement de notre groupe de recherches, dont il a été l'investigateur.

J'exprime enfin ma sincère gratitude à toutes les personnes qui ont contribué par leur gentillesse et leur compétence, à la réalisation de ce mémoire -en particulier Monsieur C. VIGNERON pour le dessin des figures, Mademoiselle J. BOGUET pour la frappe du manuscrit et Mademoiselle C. PRZYSBILSKI pour le tirage et la mise en page-, ainsi qu'à tous mes camarades du Groupe et du Laboratoire.

Nancy, le 30 Novembre 1973

TABLE DES MATIERES

	Page
INTRODUCTION	I
<u>PARTIE A - Introduction à la reconnaissance de la parole</u>	
INTRODUCTION	
<u>Chapitre I - Emission et perception de la parole</u>	
I-1 La chaîne de communication orale	1
I-2 Le signal de parole	2
I-2-1 Définition	2
I-2-2 Physique acoustique	4
I-2-3 Caractères principaux du signal vocal	4
I-3 L'appareil phonatoire humain	7
I-3-1 Description générale	7
I-3-2 Le larynx	10
I-3-3 Le canal vocal et ses articulateurs	10
I-4 Emission des sons élémentaires de la parole	10
I-4-1 Les sons élémentaires	11
I-4-2 Les voyelles	12
I-4-3 Les consonnes	17
I-5 Le mécanisme de l'audition	18
I-5-1 Anatomie de l'oreille	19
I-5-2 Essai d'interprétation du mécanisme de l'audition	23
a)	
b)	
I-6 La perception auditive	24
I-6-1 Perception des niveaux sonores	24
I-6-2 Perception des hauteurs sonores	25
I-6-3 Influence de la phase	26
I-6-4 Caractère temporel de la perception auditive	26
I-6-5 Hiérarchisation de la perception auditive	27
I-7 Conclusion	28

Chapitre II - Généralités sur la reconnaissance de formes

II-1 Définitions	30
II-1-1 Notion de forme	30
II-1-2 La reconnaissance de formes	30
II-1-3 Prétraitement et regroupement des formes	31
II-2 Organisation générale d'un système de reconnaissance de formes	31
II-3 Méthodes de discrimination	33
II-3-1 Méthode de corrélation	34
II-3-2 Méthodes paramétriques de partition	34
II-3-3 Méthodes statistiques	38
II-3-4 Méthodes syntaxiques	39
II-4 L'apprentissage	40
II-4-1 Définition	40
II-4-2 Apprentissage et performance d'un système	40
II-4-3 Autoapprentissage et génération de machines	41
II-5 Conclusion	42

Chapitre III - La reconnaissance de la parole

III-1 Définition	43
III-2 Applications de la reconnaissance de la parole	43
III-3 Aspects spécifiques de la reconnaissance de la parole	47
III-3-1 Caractères particuliers	47
III-3-2 Choix du niveau de reconnaissance	49
III-3-2 i)	
III-3-2 ii)	
III-4 Les travaux antérieurs	51
III-4-1 Les premières réalisations	51
III-4-2 Apparition de l'ordinateur	54
III-5 Etat actuel des recherches	57
III-5-1 Le point dans le monde	57
III-5-1 a)	
III-5-1 b)	
III-5-1 c)	
III-5-1 d)	
III-5-2 Les tendances nouvelles	62
III-6 Conclusion	65

PARTIE B - Analyse et paramétrisation du signal vocal

INTRODUCTION

Chapitre I - Quelques aspects de la redondance du signal vocal	65
I-1 Introduction	67
I-2 Aspect qualitatif	67
I-3 Aspects quantitatifs	68
I-3-1 Redondance et distributions linguistiques	69
I-3-2 Redondance et théorie de l'information	70
I-4 Conclusion	72

Chapitre II - Analyse du signal vocal

II-1 Introduction	73
II-2 Analyse spectrale	73
II-2-1 Généralités	73
II-2-2 La transformée de Fourier	74
II-2-2 a)	
II-2-2 b)	
II-2-2 c)	
II-2-3 Analyse par bancs de filtres	78
II-2-4 La représentation temps-fréquence de la parole	82
II-2-5 Analyse spectrale binaire	83
II-2-5 a)	
II-2-5 b)	
II-2-6 Autres méthodes d'analyse spectrale	88
II-3 Analyse temporelle	88
II-4 Conclusion	89

Chapitre III - Paramétrisation de la parole

III-1 Introduction	90
III-2 Différents types de paramètres	91
III-2-1 Paramètres physiques	91
III-2-2 Paramètres mathématiques	92

III-3 Compression de données et théorie de l'information	94
III-3-1 Introduction	94
III-3-2 Formulation du problème	95
III-3-3 Recherche de la base optimale	95
III-3-4 Amélioration en vue de la reconnaissance	98
III-4 Exemples de résultats obtenus	99
III-4-1 Introduction	99
III-4-2 Représentation des sons dans un sous-espace optimal	100
III-4-3 Application à la reconnaissance	105
a)	
b)	
III-4-4 Présélection de paramètres	107
III-5 Conclusion - Perspectives	108
 <u>PARTIE C - Reconnaissance de la parole au niveau acoustique</u>	
INTRODUCTION	110
 <u>Chapitre I - Généralités</u>	
I-1 Introduction	111
I-2 Evolution des travaux	111
I-3 Description générale du système	112
I-3-1 Implantation	112
I-3-2 Le choix du temps réel	113
I-3-3 Les conditions expérimentales	116
I-4 Choix du niveau de reconnaissance	118
I-4-1 Système de reconnaissance et modèle de perception	118
I-4-2 Reconnaissance globale et reconnaissance analytique	118
 <u>Chapitre II - Expériences de reconnaissance globale de mots</u>	
II-1 Introduction	119
II-2 Les problèmes de normalisation de la parole	119
II-2-1 Normalisation fréquentielle	119
II-2-2 Normalisation temporelle	120

II-3 Normalisation temporelle par comparaison dynamique	121
II-3-1 Principe	121
II-3-2 Algorithme	122
a)	
b)	
c)	
II-4 Organisation du système de reconnaissance	129
II-4-1 Programme de service conversationnel	129
II-4-2 Programme de reconnaissance	131
II-4-3 Organisation du lexique	133
II-5 Résultats expérimentaux	133
II-5-1 Performances pour différents vocabulaires	134
II-5-2 Influence de la taille du vocabulaire	137
II-5-3 Influence du nombre de locuteurs	137
II-5-4 Influence de la dérive de la voix	140
II-5-5 Temps moyen de calcul	142
II-5-6 Le problème de l'apprentissage	144
II-6 Perspectives d'avenir et extensions possibles	144
II-7 Conclusion	145

<u>Chapitre III - Segmentation et reconnaissance analytique de la parole</u>	
III-1 Introduction	147
III-2 Segmentation du signal vocal	148
III-2-1 Position du problème. Les segments	148
III-2-2 Principe de la segmentation utilisée	148
III-2-3 Description du programme	150
III-2-4 Résultats obtenus	150
III-3 Reconnaissance analytique de la parole	154
III-3-1 Présentation	154
III-3-2 Le système de reconnaissance	154
a)	
b)	
c)	
III-3-3 Résultats obtenus	157
III-4 Recherche lexicale d'un mot à partir de la chaîne phonémique	161
III-4-1 Position du problème	161
III-4-2 Les méthodes de recherche lexicale	162
III-4-3 Méthode de cheminement dans un tableau	164
III-4-4 Méthode de comparaison dynamique	166
III-5 Conclusion. Reconnaissance globale et reconnaissance analytique	168

PARTIE D - Utilisation de contraintes linguistiques,
syntaxiques et sémantiques en reconnais-
sance vocale

INTRODUCTION 170

Chapitre I - Statistique phonétique du français parlé

I-1 Introduction	172
I-2 Principe de l'étude	173
a)	
b)	
I-3 Etude des phonèmes	175
I-4 Etude des diphonèmes	177
I-5 Etude des triphonèmes	183
I-6 Etude des mots	185
I-6-1 Longueur des mots	185
I-6-2 Commencements et terminaisons	186
I-7 Validité des résultats	190

Chapitre II - Utilisation de contraintes linguistiques
en reconnaissance analytique

II-1 Introduction	192
II-2 Principe d'utilisation des contraintes linguistiques	193
II-2-1 Position du problème	193
II-2-2 Utilisation des diphonèmes	194
II-2-3 Utilisation des triphonèmes	194
II-2-4 Retour sur le système de reconnais- sance acoustique	195
II-3 Expériences préliminaires	195
II-4 Résultats expérimentaux pour le système complet	199
II-4-1 Schéma général du système	199
II-4-2 Interaction niveau acoustique - niveau linguistique	199
II-4-3 Exemples de résultats	201
II-5 Poursuite de ces travaux	203
II-6 Conclusion	204

Chapitre III - Reconnaissance au niveau syntaxique

III-1 Introduction	205
III-2 Utilisation d'une grammaire en recon- naissance automatique	206

III-3 Un exemple pratique en commande numérique de machines-outils	207
III-3-1 Présentation	207
III-3-2 Syntaxe et sémantique du langage étudié	209
a)	
b)	
III-3-3 Le programme de reconnaissance	212
a)	
b)	
III-3-4 Résultats obtenus	216
III-4 Retour sur les langages oraux de communication homme-machine	217
III-5 Conclusion	220

CONCLUSION 221

BIBLIOGRAPHIE 224

INTRODUCTION

La reconnaissance automatique de la parole s'inscrit dans le cadre général des recherches sur le dialogue entre l'homme et la machine. Née de l'informatique et largement tributaire de ses moyens, cette science récente n'en possède pas moins ses caractères propres. Elle est en particulier pluridisciplinaire, et elle amène ainsi le chercheur à aborder, au risque de se disperser, les domaines, apparemment éloignés, de l'acoustique, l'automatique, l'électronique, la phonétique, la linguistique, la physiologie, etc...

Jusqu'à présent les réalisations ont été modestes, comme cela a été constaté à plusieurs reprises [1-1] à [1-4]. On assiste maintenant à un approfondissement des recherches, mais les solutions proposées demeurent encore très partielles. Il ne s'agit point là d'un constat d'échec : la reconnaissance de la parole est simplement un problème très difficile, en général mésestimé au départ, et ces difficultés sont dues en partie au caractère particulier du signal vocal.

Ce mémoire présente l'état actuel de nos recherches. Ce n'est pas du tout un point final, mais seulement une étape dans un travail commencé en 1968 et dont l'aboutissement devrait être la reconnaissance en temps réel du discours continu. Dans ce sens, les systèmes de reconnaissance que nous avons mis au point ne sont pas figés, mais évoluent sans cesse et les performances atteintes sont, dans certains cas, déjà intéressantes.

Notre étude comporte quatre parties.

Dans une première partie sont rassemblées les notions de phonation et d'audition-perception nécessaires à la compréhension de la suite du travail. On y trouve aussi des généralités sur la reconnaissance de formes, dont la reconnaissance vocale ne constitue qu'un aspect et enfin un bilan actuel sur la reconnaissance de la parole.

La deuxième partie concerne l'analyse et la paramétrisation de la parole. La partie analyse est surtout consacrée à la transformée de Fourier et à son équivalent analogique, l'analyse par banc de filtres. C'est une méthode d'analyse spectrale qui est utilisée dans nos expériences de reconnaissance. Ensuite, on expose une méthode d'extraction de paramètres pertinents par compression d'information, que nous avons adaptée à la reconnaissance de la parole.

La reconnaissance vocale, au niveau acoustique, fait l'objet de la troisième partie. Toutes les expériences décrites fonctionnent en temps réel sur l'ordinateur T 2000 du Laboratoire d'Electricité et d'Automatique. Il s'agit à la fois d'expériences globales de reconnaissance de mots et d'expériences analytiques de segments obtenus par segmentation acoustique des mots. A cette occasion, les problèmes de recherche lexicale sont abordés. Enfin une étude comparée des deux approches, globale et analytique, permet de tirer quelques conclusions.

La quatrième partie, enfin, traite des niveaux supérieurs de reconnaissance : linguistique et syntaxique. Dans un premier temps, on étudie l'optimisation d'une chaîne de phonèmes par utilisation de contraintes linguistiques. Ces contraintes sont constituées de résultats statistiques sur les associations phonémiques, déterminés dans une étude préliminaire du français parlé. Ces études ont été effectuées en temps différé sur 10070 pour des questions de taille mémoire. Ensuite, on effectue la reconnaissance, en temps réel, de phrases de commande numérique de machines-outils, avec une analyse syntaxique rudimentaire, qui constitue une approche intéressante de la reconnaissance syntaxique.

PARTIE A

INTRODUCTION A LA RECONNAISSANCE

DE LA PAROLE

INTRODUCTION A LA PARTIE A

Cette première partie est destinée à situer le problème de la reconnaissance automatique de la parole et à introduire un certain nombre de notions nécessaires pour les développements ultérieurs.

De par sa nature même, la reconnaissance vocale se réclame de disciplines aussi variées que l'acoustique, l'informatique, la phonétique, la physiologie, etc... On trouvera ainsi, rassemblées ici, des notions sommaires concernant l'émission et la perception de la parole et des généralités sur les méthodes de reconnaissance de formes.

Pour terminer, la reconnaissance vocale est définie, ainsi que ses buts et ses limitations. Enfin, une revue des travaux antérieurs, dans le monde, depuis 1950 environ, est proposée, que nous avons voulu suffisamment complète sinon exhaustive.

CHAPITRE I

EMISSION ET PERCEPTION DE LA PAROLE

Le langage est une des facultés essentielles de l'homme, et c'est peut-être pour cela que l'on tient en général comme connus les phénomènes d'émission et de perception de la parole. Il nous a cependant paru nécessaire de rassembler ici des notions variées allant de l'acoustique physique à la physiologie de l'audition. Ces notions ont été réduites au minimum nécessaire à l'analyse et la reconnaissance de la parole.

Nous commençons par l'étude de la phonation pour finir par l'audition, suivant une gradation des difficultés : les contraintes phonétiques interviennent de façon moins critique que les contraintes auditives dans la communication orale. L'expérience confirme d'ailleurs cette constatation : la synthèse de la parole a progressé plus rapidement que la reconnaissance automatique.

I-1 La chaîne de communication orale.

Une communication entre un locuteur et un auditeur est une communication de cerveau à cerveau, par l'intermédiaire de la bouche et des oreilles. Cette communication a lieu à différents niveaux -sémantique, linguistique, phy-

siologique, acoustique - comme le schématise la figure I-1.

A partir d'une idée, le locuteur effectue une sorte de "codage" qui fournit une chaîne d'éléments phonétiques résultant en commandes neuro-motrices pour l'appareil phonatoire. Une chaîne de retour par l'intermédiaire de l'oreille permet au locuteur de contrôler ce qu'il dit. Les sons émis sont captés par l'oreille de l'auditeur et transformés en impulsions nerveuses. Ces impulsions sont transmises au cerveau, à nouveau décodées et résultent enfin en une idée, qui n'est pas forcément celle que le locuteur a voulu formuler...

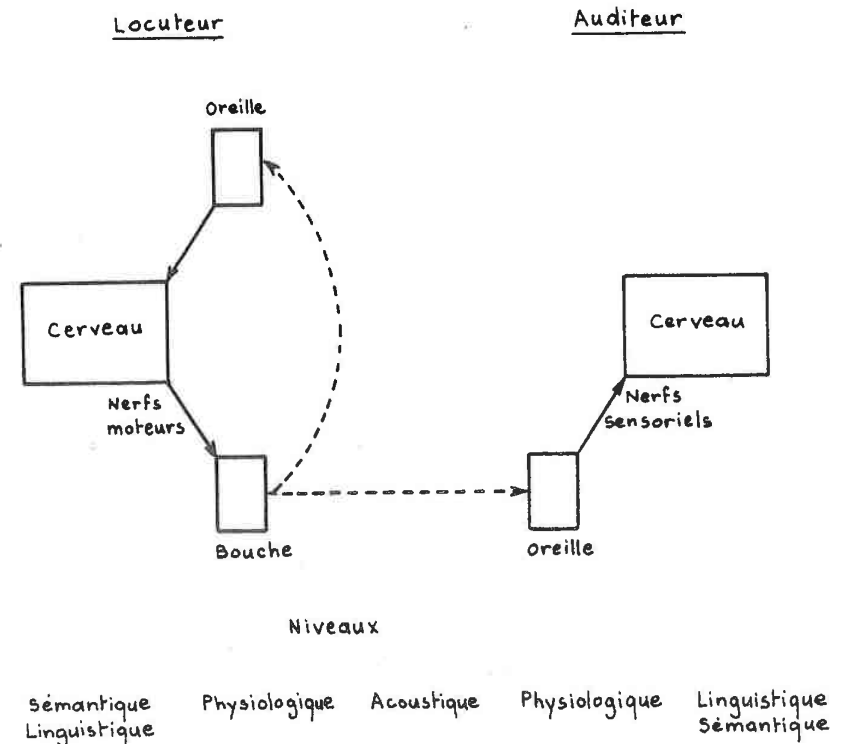
Cette description demeure assez vague, car de nombreux points de ce processus restent du domaine de l'hypothèse. Seules l'émission des sons vocaux et, à un degré moindre, la réception de ces sons par l'oreille, commencent à être connues.

1-2 Le signal de parole.

1-2-1 Définition :

La parole est un mouvement vibratoire résultant de l'interaction d'une source d'énergie acoustique et d'un système passif multirésonant, le canal vocal. On l'étudie presque toujours sous sa forme électrique, donnée par un microphone qui assure la transformation des variations de pression en variations de courant.

Avant d'étudier ses caractères originaux, nous rap-
pelons quelques résultats d'acoustique.



La chaîne de communication orale

Fig I-1

I-2-2 Physique acoustique :

L'oreille ne perçoit de sensation sonore que pour des fréquences comprises, en moyenne, entre 16 Hz et 17 000 Hz, ces limites variant selon les individus. Parmi tous les sons audibles, la parole normale n'occupe qu'un petit domaine, comme le montre la figure I-2/A-17. On constate que la dynamique de la voix est considérable : de l'ordre de 40 db en parole normale; elle peut même passer à 80 db en parole criée.

On peut classer grossièrement les sons en deux catégories :

- les bruits, suites de vibrations incohérentes, correspondant à des spectres plus ou moins continus ;
- les sons musicaux, périodiques, simples ou composés, et correspondant alors à des spectres de raies.

I-2-3 Caractères principaux du signal vocal :

La parole est une fonction d'une très grande complexité, tant du point de vue émissif que perceptuel. Le signal de parole possède par ailleurs des caractères qui en font un signal "à part" et qui expliquent en partie les difficultés rencontrées en reconnaissance vocale.

Il est d'abord caractérisé par sa continuité. En effet, dans le langage, les mots ne sont pas séparés comme dans l'écriture et les seuls arrêts correspondent aux pauses de respiration, aux pauses grammaticales et aux occlusions précédant les consonnes plosives. La segmentation est

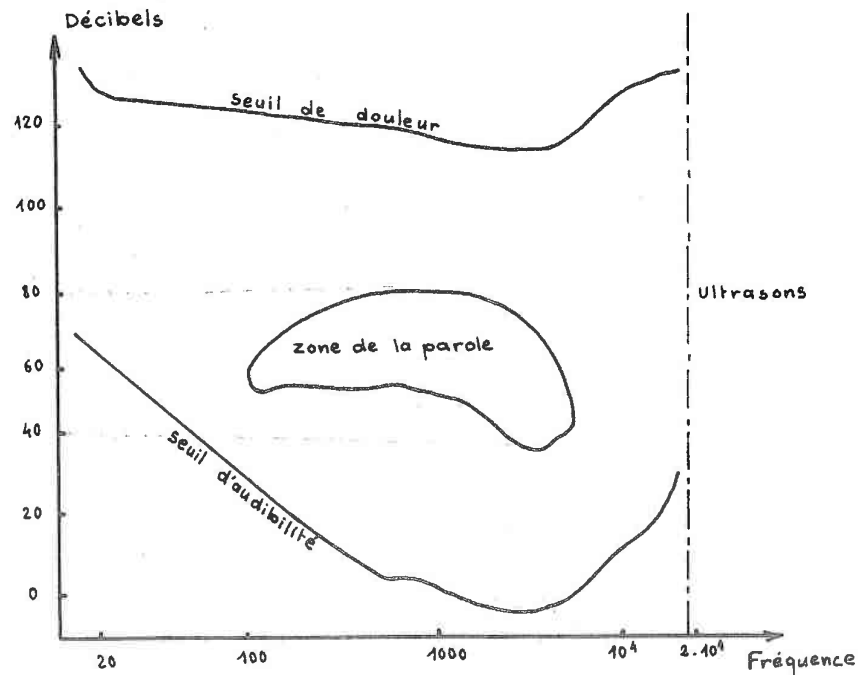


Diagramme d'audibilité de l'homme

Fig I.2

effectuée par le cerveau, en fonction du contenu sémantique du message et des contraintes linguistiques et syntaxiques de la langue. Cette segmentation constitue une des difficultés majeures en reconnaissance automatique ; nous reviendrons en détail sur ce problème.

D'autre part, le signal de parole présente une grande variabilité dans sa forme. Cette variabilité apparaît dans les différences de prononciation d'un locuteur à un autre, et même d'un seul locuteur, selon son état psychique et physique. En bref, il est impossible de prononcer plusieurs fois une phrase de façon absolument identique.

Enfin le signal de parole est extrêmement redondant. En effet, la parole transporte deux sortes d'informations : une information sémantique contenant le sens du message, et donc seule nécessaire à sa compréhension, et une information esthétique [A-2], spécifique de la communication orale, et caractérisant plus spécialement le locuteur. Or, le codage de la parole en ses phonèmes constitutifs (codage préservant l'intelligibilité de la parole) et la transmission de ce code requiert un débit d'information de 50 bits/seconde environ. Au contraire, pour transmettre une parole de qualité "haute-fidélité", il faut au moins un débit de 200 000 bits/s (30 000 bits/s environ pour la transmission téléphonique) [A-3].

Cette énorme redondance accroît en partie l'intelligibilité de la parole. Elle explique par exemple le fait que la compréhension d'un message soit bonne, même avec un niveau de bruit parasite important, malgré la grande variabilité du signal de parole que nous avons mentionnée. Nous reviendrons sur cet aspect dans la partie B.

Une dernière caractéristique du signal de parole provient de ce que ce signal est destiné à transmettre de l'information. Comme seules des variations peuvent effectuer cette transmission, il en résulte qu'il est essentiellement transitoire.

I-3 L'appareil phonatoire humain.

I-3-1 Description générale :

L'appareil phonatoire de l'homme se compose de deux parties distinctes, une source et un canal vocal (figure I-3).

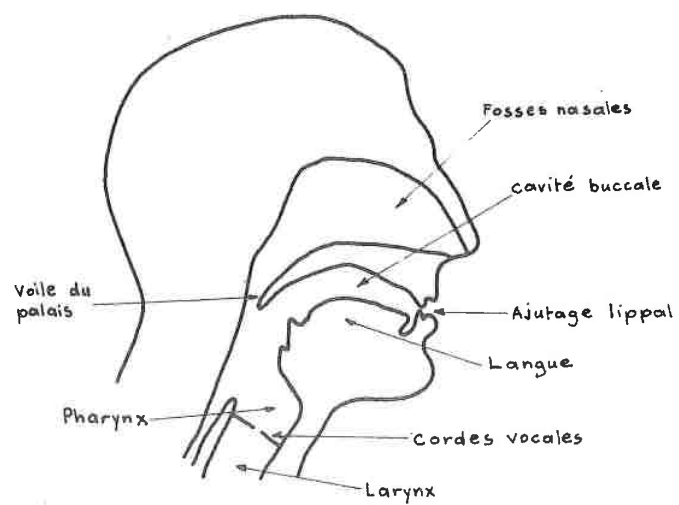
La source d'énergie acoustique utilisée pour produire la voix est l'air contenu dans les poumons. Le flux d'air sous pression arrive, à travers la trachée, jusqu'au canal vocal.

Le canal vocal a une longueur moyenne de 17 cm.

Il comprend :

- le larynx, instrument à anches musculaires, les cordes vocales ;
- des cavités résonantes déformables : pharynx, bouche, nez, cette dernière pouvant ou non être mise en communication grâce au voile palatal ;
- des organes d'articulation : langue, lèvres, mâchoires... , permettant d'obturer le canal vocal ou d'en modifier la forme.

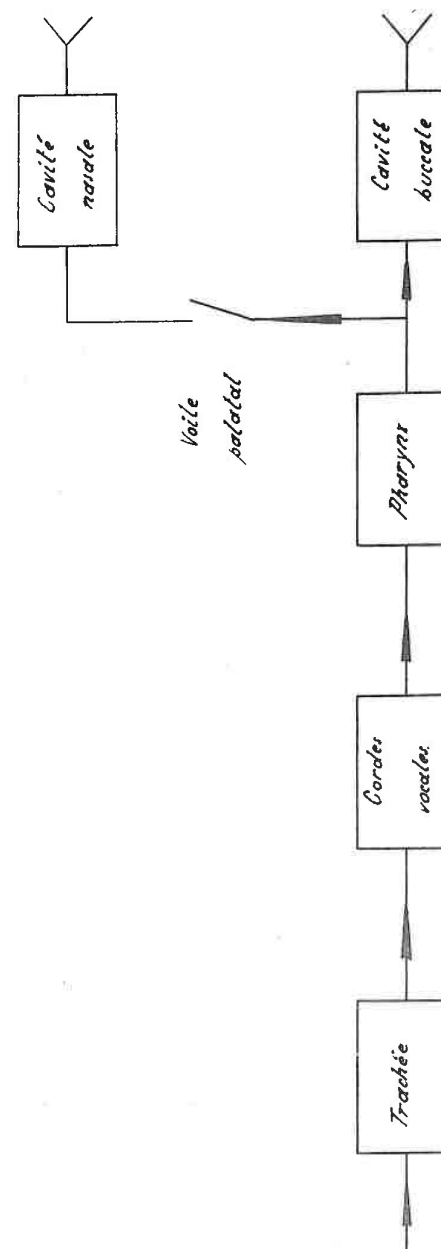
La figure I-4 donne un schéma de principe du canal vocal.



L'appareil phonatoire humain (d'après I.3)

Fig I-3

8.



9.

Schéma du canal vocal
Fig. I-4

I-3-2 Le larynx :

Le larynx est un assemblage assez complexe de cartilages, reliés par des ligaments et des muscles [A-5]. Il est situé au sommet de la trachée et contient les cordes vocales, replis de tissu élastique situés de part et d'autre du passage de la trachée. En respiration normale, les cordes vocales sont pures et laissent passer librement l'air. Par contre, durant la phonation, elles peuvent se rapprocher et entrer en vibration sous l'effet du flux d'air qui les traverse, découpant ainsi ce flux en une suite d'impulsions qui donnera un son sonore, ou voisé (sorte de vibrations de relaxation).

Signalons aussi que les cordes vocales peuvent être incomplètement rapprochées ; le frottement de l'air venant des poumons produit alors un son chuchoté.

I-3-3 Le canal vocal et ses articulateurs :

Les phonéticiens distinguent traditionnellement quatre cavités résonantes dans le canal vocal : le pharynx, les fosses nasales, la bouche et l'ajutage lippal. Leurs configurations peuvent varier énormément sous l'action des articulateurs : lèvres, dents, palais, langue, qui disposent d'une quantité considérable de muscles et de nerfs.

I-4 Emission des sons élémentaires de la parole.

Nos connaissances sur l'émission de la voix ont beaucoup avancé grâce à l'utilisation de méthodes modernes d'investigation (radiographie, cinéma) et au développement de modèles de l'appareil phonatoire (tels ceux de Flanagan [A-4] ou de Fant [A-6]).

I-4-1 Les sons élémentaires :

Trois variables essentielles permettent de classer les sons : le voisement (activité du larynx), le mode d'articulation (type du mécanisme) et le point d'articulation (endroit de resserrement maximal du conduit vocal). Cependant le classement articulatoire des phonéticiens apparaît insuffisant en reconnaissance. Tout d'abord, les positions articulatoires ne sont pas stables, et pas toujours caractéristiques d'un son (phénomène de compensation). Par ailleurs, nous avons déjà vu que la parole, consistant en un envoi d'information, nécessite des variations, ce qui fait que, lorsque l'on parle, l'appareil phonatoire n'est quasiment jamais au repos.

L'unité minimale d'une langue est le phonème. On peut en donner une définition reposant sur la notion de paire minimale : si deux sons peuvent définir une paire minimale, ils constituent deux phonèmes différents (par exemple /p/ et /t/ dans la paire pou-tout). Cette définition, purement formelle, fait du phonème une abstraction linguistique, mais pas une unité de travail utilisable en reconnaissance. L'image longtemps admise de phonèmes rigides "alignés les uns à côté des autres comme les perles d'un chapelet" [A-7] est tout à fait fautive. Il existe une dépendance très étroite entre les phonèmes successifs, certains traits caractéristiques demeurent mais les spectres sont fortement influencés par leur environnement. Ces traits pertinents font des phonèmes des classes de sons phonétiquement semblables, cette similitude n'empêchant pas qu'un grand nombre de réalisations d'un même phonème diffèrent toutes plus ou moins.

Il faut remarquer que les traits caractéristiques dont il est question ici ne sont pas les traits distinctifs ("distinctive features") introduits par Jakobson, Fant et Halle [A-8]. Ces traits, purement articulatoires, permettent d'effectuer un classement très intéressant des phonèmes, chaque phonème étant caractérisé par seulement quelques traits distinctifs.

Cependant la traduction sur le plan acoustique de ces traits est trop vague pour que ceux-ci soient vraiment utiles en reconnaissance automatique. Nous verrons d'ailleurs au chapitre III que les expériences de reconnaissance fondées exclusivement sur cette méthode conduisent à des performances limitées.

Le caractère dynamique des unités fondamentales du langage apparaît ainsi par opposition au caractère trop statique des descriptions articulatoires. C'est lui qui a présidé au choix des méthodes de reconnaissance que nous avons adoptées et qui a également orienté notre choix de "segments minimaux" pour la reconnaissance (cf. partie C).

I-4-2 Les voyelles :

Les voyelles sont des sons sonores, ou voisés, mettant en jeu la vibration des cordes vocales. Ces sons correspondent à une configuration relativement stable (par rapport aux consonnes) du canal vocal.

Considérons l'émission d'une voyelle soutenue (figure I-5). Les cordes vocales, fermées, sont mises en vibration par le flux d'air provenant des poumons, qui est ainsi découpé en une suite d'impulsions, de fréquence fondamentale égale à la fréquence de vibration des cordes (variable de 60 à 300 Hz environ selon l'individu, l'état de tension des cordes vocales, la pression subglottique). Cette fréquence représente aussi approximativement le fondamental du son émis, ou mélodie.

La vibration des cordes vocales est assimilable à un phénomène de relaxation ; le spectre de l'onde de pression qu'elles modulent est très riche en harmoniques (figure I-5-a). Cette onde est ensuite filtrée par le canal vocal et ses cavités résonantes, que l'on peut caractériser par sa fonction de transfert. Le spectre du son émis possède un certain nombre de maxima d'énergie correspondant aux pôles de cette fonction de transfert (figure I-5-b), appelés formants vocaliques.

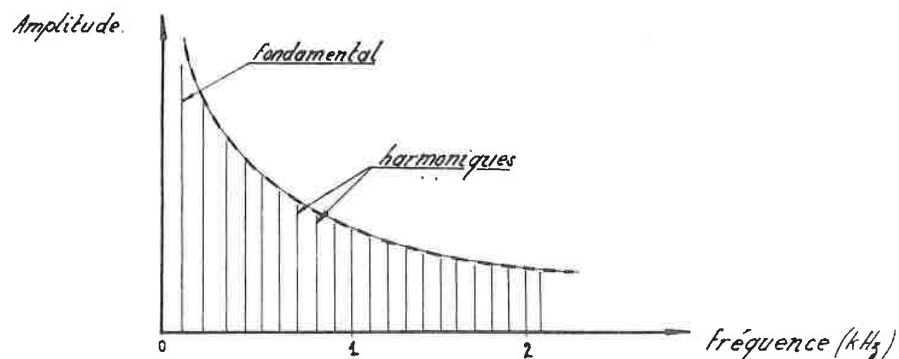
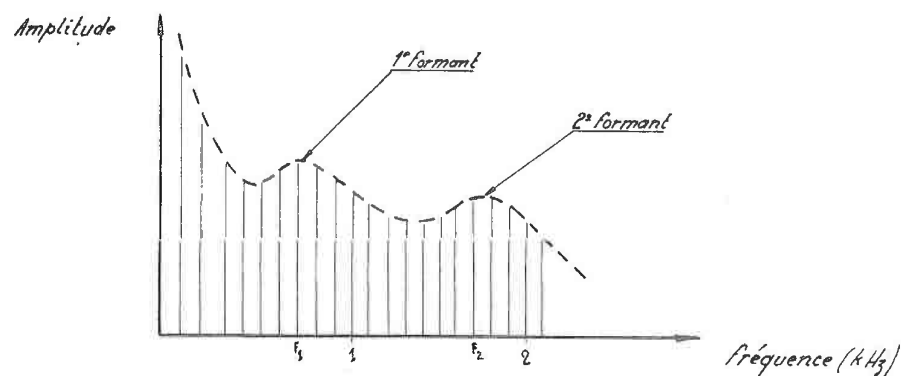
On connaît depuis longtemps, par les travaux sur la perception et la synthèse de la parole, l'importance des deux premiers formants dans la description acoustique des voyelles [A-9]. Le troisième formant est plus spécialement responsable du timbre, de la "couleur" de la voyelle.

La langue française utilise douze voyelles orales (c'est-à-dire émises exclusivement par la bouche) :

/a/, /ɑ/, /i/, /e/, /ɛ/, /ə/, /ø/, /œ/, /ɔ/, /o/, /u/, /y/

Emission d'un son sonore fig. I-5

14.

fig. I-5a Spectre de l'onde de pression.fig. I-5b Spectre du son émis.

Signalons à ce propos que la notation phonétique utilisée ici est celle du sous-ensemble de l'alphabet phonétique international correspondant au français, qui est donné au chapitre I, partie D.

En plus des voyelles orales, il existe des voyelles nasales correspondant à la mise en parallèle de la cavité résonante nasale sur la cavité buccale. La théorie montre que l'on obtient des résonances plus floues, il en résulte en pratique un élargissement des zones formantiques. De plus, le premier formant reste voisin de 500 Hz.

Le français utilise quatre voyelles nasales :

/s/, /z/, /ɛ/, /œ/

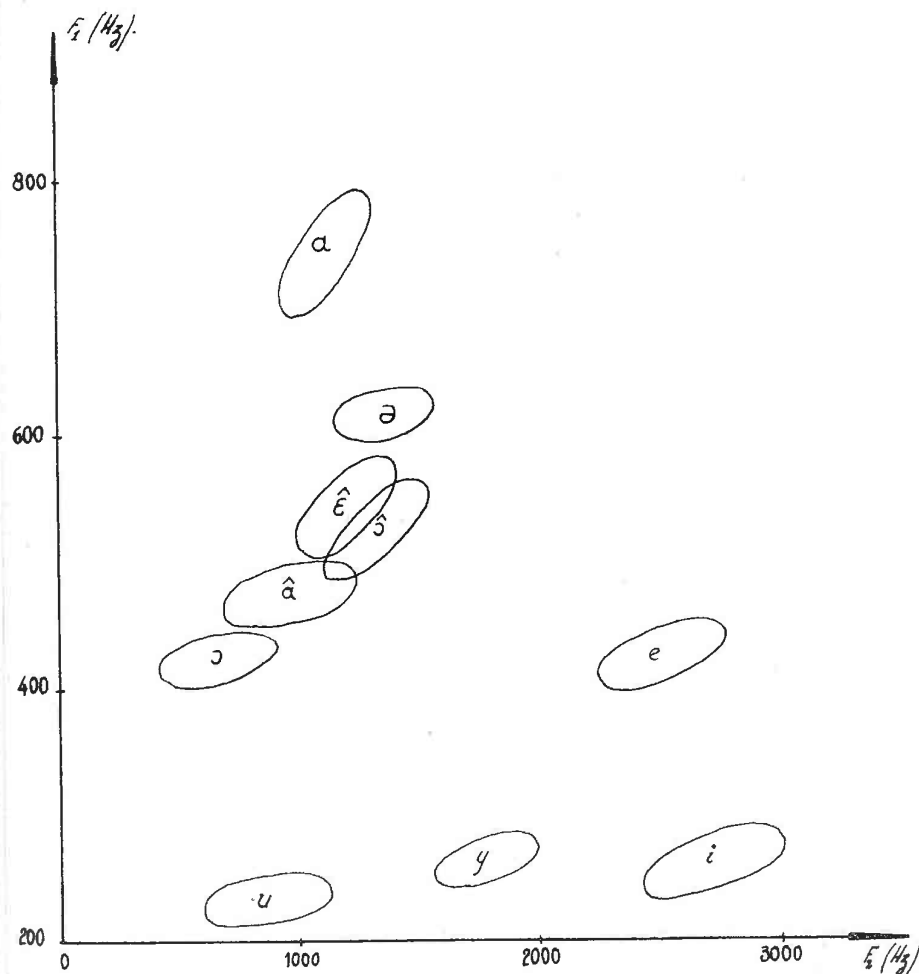
La dispersion des formants est importante, même pour un seul locuteur. La figure I-6 donne la répartition des deux premiers formants F_1 et F_2 , pour des voyelles prononcées par différents locuteurs. On remarque que c'est le second formant qui est le plus caractéristique.

Les limites de dispersion sont les suivantes, en moyenne pour toutes les voyelles françaises :

$$250 \text{ Hz } (/i/) \leq F_1 \leq 750 \text{ Hz } (/a/)$$

$$700 \text{ Hz } (/o/) \leq F_2 \leq 2350 \text{ Hz } (/i/)$$

La figure I-6 met en évidence l'interpénétration des zones correspondant à différentes voyelles : les formants sont parfois insuffisants pour différencier deux voyelles.



Représentation des voyelles françaises dans le plan

(1^{er} formant, 2^e formant)

Fig. 1.6

Nous avons considéré jusqu'à présent le cas de voyelles soutenues, émises par exemple isolément. Dans la parole normale, qui est, répétons-le, essentiellement mouvement, les voyelles deviennent inutiles, au point de vue informationnel, dès l'instant où elles sont reconnues. On conçoit ainsi qu'une importante partie de la redondance du signal de parole provienne de ce fait. Nous reviendrons sur la redondance de la parole en partie B.

I-4-3 Les consonnes :

Les consonnes se prononcent avec une fermeture ou un rétrécissement local du canal vocal, contrairement aux voyelles qui correspondent à un libre passage du flux d'air. Il en résulte que le spectre d'une consonne sera toujours plus ou moins un spectre de bruit.

Il existe des consonnes sonores et des consonnes sourdes, selon que le flux d'air provenant des poumons est ou n'est pas modulé par les cordes vocales.

On peut classer les consonnes en deux grands groupes :

- les occlusives correspondent à une fermeture momentanée du canal vocal suivie d'une ouverture brusque. Elles sont donc essentiellement transitoires et provoquent une évolution des formants du son sonore qui les précède ou qui les suit. Delattre a systématiquement étudié ces transitions formantiques [A-10]. La théorie du locus (point de convergence des différents formants affectés par une occlusive) a été déduite de cette étude. Cette théorie est surtout intéressante en phonétique théorique. Par contre, toutes ces recherches ont permis de mettre en évidence le rôle primordial des transitions dans la parole.

Les occlusives françaises sont :

/p/, /t/, /k/	sourdes
/b/, /d/, /g/	sonores

- les fricatives sont caractérisées par un bruit de friction dû à l'écoulement de l'air dans une section rétrécie du canal vocal. Ce sont donc des sons soutenus, sourds ou sonores, et qui présentent un spectre de bruit plus ou moins coloré.

Les fricatives françaises sont :

/f/, /s/, /ʃ/	sourdes
/v/, /z/, /ʒ/	sonores

I-5 Le mécanisme de l'audition.

L'appareil auditif reçoit les vibrations sonores qui se propagent dans l'air, les transmet et assure leurs transformations en sensations auditives. La fonction de l'ouïe est ainsi assurée par une suite de processus physiques, physiologiques et enfin nerveux.

La capacité informationnelle de l'oreille est de l'ordre de plusieurs dizaines de milliers de bits/s, ce qui peut paraître important par rapport à la quantité d'information sémantique transmise par la parole. Mais cette disproportion s'explique facilement car, d'une part, l'oreille perçoit des informations supplémentaires concernant le locuteur, et, d'autre part, la quantité d'information dépend non seulement de ce qui est transmis mais aussi de tout ce qui est susceptible d'être transmis. Or l'oreille peut percevoir bien d'autres informations que celles véhiculées par la parole.

I-5-1 Anatomie de l'oreille :

L'oreille se compose, à la suite du pavillon, d'un ensemble complexe de cavités creusées dans le rocher, et comporte trois parties : l'oreille externe, l'oreille moyenne et l'oreille interne (figure I-7).

L'oreille externe comporte le pavillon, sorte de cornet acoustique qui dirige les ondes vers l'oreille moyenne par le canal auditif externe.

L'oreille moyenne est une cavité contenant de l'air, reliée à l'oreille externe par le tympan, membrane vibrante de 1 cm² environ, et s'ouvrant dans l'oreille interne par deux petits orifices : la fenêtre ronde et la fenêtre ovale. Pendant la déglutition, l'oreille moyenne est en communication avec le pharynx, par l'intermédiaire de la trompe d'Eustache. Ceci permet une uniformisation des pressions de part et d'autre du tympan.

D'autre part, le tympan est relié à la fenêtre ovale par la chaîne des osselets (marteau, enclume, étrier), articulés entre eux de façon complexe.

L'oreille interne, de structure très complexe, est logée dans le labyrinthe osseux, qui est rempli d'un liquide appelé périlymphe. Cette cavité osseuse contient un sac à paroi très mince, de forme identique, entièrement fermé et rempli d'un liquide appelé endolymphe (de propriétés voisines de celles du périlymphe), le labyrinthe membraneux.

La partie essentielle de l'oreille interne est la cochlée, dans laquelle les vibrations mécaniques se transforment en impulsions nerveuses. La cochlée, enroulée sous forme d'une spirale hélicoïdale, est séparée en trois parties par la membrane basilaire et la membrane de Reissner : les rampes vestibulaire, tympanique et cochléaire (cette dernière partie est appelée

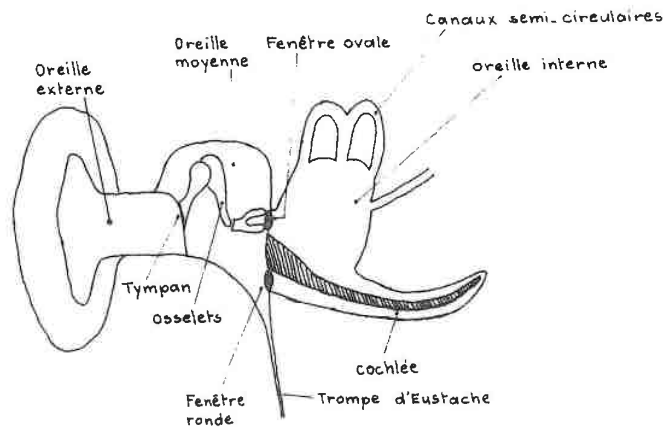


Schéma général de l'oreille (la cochlée est ici déroulée)

Fig I.7

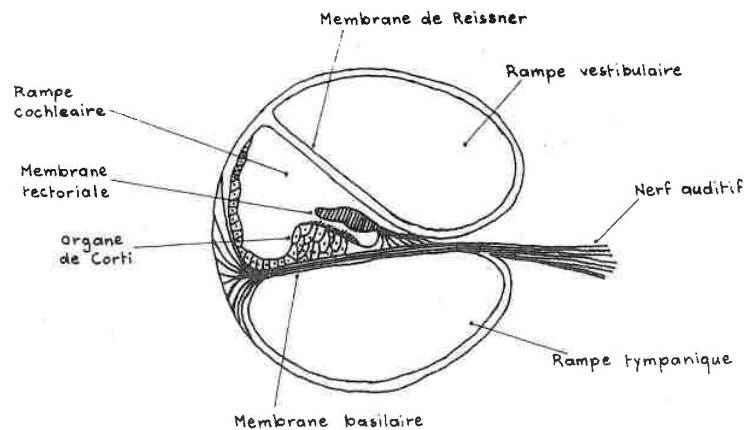
généralement canal cochléaire). La figure I-8 donne une coupe schématique de la cochlée suivant l'axe de la columelle.

Le canal cochléaire est rempli d'endolymphe, tandis que les deux autres rampes, en communication au sommet de la cochlée, sont remplies de périlymphe.

La membrane basilaire joue un rôle très important dans l'audition. Longue de 32mm environ, elle a une largeur qui croît régulièrement depuis la base jusqu'au sommet de la cochlée. Elle est constituée d'un grand nombre de fibres transversales, dont la longueur varie régulièrement. La membrane basilaire porte d'autre part l'organe de Corti, ensemble complexe comportant en particulier de nombreuses cellules ciliées (20 000 externes et 3 500 internes environ) en contact avec une troisième membrane, la membrane tectoriale, et sensibles au déplacement de la membrane basilaire.

De nombreuses connexions neuroniques relient les cellules de Corti au nerf auditif, ou nerf cochléaire, situé dans la columelle.

L'oreille interne, et en particulier la cochlée et l'organe de Corti, possède ainsi une structure à la fois complexe et régulière. Cette régularité est trop remarquable pour qu'on ne la lie pas au mécanisme de l'audition comme nous allons le voir.



Coupe schématique de la cochlée

Fig 1.8

1-5-2 Essai d'interprétation du mécanisme de l'audition :

Si l'anatomie de l'oreille est maintenant bien connue, le mécanisme de l'audition reste controversé. Ici encore, les modèles auditifs, plus ou moins partiels, qui ont été proposés (Békésy/A-11], Flanagan/A-4], Molès/A-12], Leipp/A-13]) permettent de faire progresser nos connaissances.

a) Rôle de l'oreille moyenne.

L'oreille moyenne n'intervient pas dans le codage des ondes sonores mais elle n'en joue pas moins un double rôle primordial :

- adaptation d'impédance - par ailleurs non linéaire - entre le milieu extérieur (air) et le milieu liquide de l'oreille interne. La vibration du tympan, sous l'effet des ondes sonores, est ainsi transmise à la fenêtre ovale avec une amplification de l'ordre de 35, par l'intermédiaire de la chaîne ossiculaire,
- asservissement de l'oreille au niveau sonore, avec une certaine constante de temps. Ce processus réflexe, sorte de compression dynamique du son, protège l'oreille interne, surtout la cochlée, très fragile, contre les bruits trop violents/A-14].

Diverses expériences ont également montré que la transmission du son par l'oreille moyenne dépend de la fréquence. Ce phénomène est encore quantitativement mal connu. Il dépend en particulier des individus, car ce sont en partie les caractéristiques physiques de l'oreille qui interviennent. Quoiqu'il en soit, les résultats obtenus concordent pour que l'on puisse affirmer que l'oreille moyenne se comporte comme un filtre passe-bas amorti/A-11]. Ceci constitue une protection supplémentaire de l'oreille interne contre les sons aigus et les ultrasons.

b) Rôle de l'oreille interne : codage cochléaire.

C'est dans l'oreille interne, et plus précisément au niveau de l'organe de Corti que s'effectue, comme nous l'avons dit, la transformation des vibrations acoustiques en impulsions nerveuses. Les travaux de Békésy ont permis de mieux connaître le comportement de la membrane basilaire. La vibration qui se propage depuis la fenêtre ovale provoque une déformation de cette membrane, avec une amplitude maximale qui évolue depuis le haut jusqu'à la base de la cochlée lorsque la fréquence augmente. Cette déformation est enregistrée par les cellules ciliées qui la transmettent au nerf auditif. En réalité, si les phénomènes sont beaucoup plus complexes, il n'en reste pas moins vrai que la membrane basilaire joue le rôle d'un analyseur spectral très grossier à large bande, assez amorti (avec un coefficient de surtension de l'ordre de 1). Ce modèle rend bien compte du comportement de l'oreille face aux signaux transitoires, mais, par contre, il semblerait indiquer que l'analyse en fréquence d'un signal stationnaire est grossière, ce qui est loin d'être le cas.

Depuis la théorie des résonateurs d'Helmholtz et la théorie téléphonique de Rutherford, de nombreuses hypothèses ont été avancées pour tenter de lever cette contradiction. Si aucune n'est entièrement satisfaisante, c'est cependant la théorie de Licklider [A-15] qui est la plus plausible. Selon lui, à l'analyse fréquentielle grossière effectuée par le cerveau, succède une analyse temporelle d'autocorrélation sur les impulsions nerveuses, assurée par les neurones.

I-6 La perception auditive.

I-6-1 Perception des niveaux sonores :

Selon la loi de Weber-Fechner, la sensation sonore est proportionnelle au logarithme de l'excitation qui lui donne naissance, autrement dit le seuil différentiel relatif d'intensité ($\frac{\Delta I}{I}$)

reste constant. Cette loi a donné naissance à la notation des intensités sonores en décibels :

$$I_{db} = 10 \log \frac{P}{P_0}$$

Cette notation permet de réduire à 120db environ l'intervalle séparant le seuil d'audibilité ($P_0 = 10^{-16} \text{ W/cm}^2$) et le seuil de douleur [A-16] (cf. figure I-2).

En fait, la loi de Weber-Fechner ne fournit qu'un ordre de grandeur pour les fréquences moyennes. En effet, bien que douée d'une grande sensibilité différentielle (de l'ordre du décibel), l'oreille apprécie difficilement l'intervalle entre deux sons de niveaux très différents, et la loi n'est en général pas vérifiée.

La perception par l'homme d'un niveau sonore, comme le repérage d'une température, est par ailleurs très subjective. L'oreille est incapable de donner une mesure du niveau sonore absolu ; elle ne peut fournir qu'une impression et semble de toute façon beaucoup plus adaptée pour effectuer des comparaisons.

I-6-2 Perception des hauteurs sonores :

Nous avons déjà vu que le champ auditif humain couvrait un domaine de fréquences allant de 16 à 17 000 Hz environ. La sensibilité de l'oreille varie avec la fréquence, elle est maximale aux environs de 2 000 Hz, comme l'ont vérifié les travaux de Fletcher et Munson. On peut noter que la plage de sensibilité maximale de l'oreille correspond précisément à la zone fréquentielle de la parole normale.

Nous avons déjà mentionné également le paradoxe de l'analyse fréquentielle de l'oreille, capable de distinguer aussi bien des transitoires rapides que des sons stationnaires de

fréquences très proches (le seuil différentiel relatif en fréquence atteint 0,3 % vers 2 000 Hz). Nous reviendrons sur cet aspect de la perception auditive au paragraphe I-6-4.

I-6-3 Influence de la phase :

La loi d'Ohm indique que l'oreille est insensible aux différences de phase. En réalité ceci n'est valable qu'en première approximation, et pour des différences de phase suffisamment faibles, comme le montrent les techniques de téléphonie à très grandes distances. Cette restriction s'explique très bien si l'on admet que l'oreille perçoit un spectre à court terme (cf. partie B chapitre II).

Il reste cependant que l'on peut modifier, dans un domaine restreint, les phases des composantes d'un son complexe sans modifier le timbre du son. Cette propriété est caractéristique de l'audition, et la distingue par exemple d'une autre fonction essentielle comme la vision.

I-6-4 Caractère temporel de la perception auditive :

L'inertie mécanique de l'oreille introduit un seuil de perception en temps, durée minimale d'un son perçu ou intervalle minimal séparant deux sons perçus séparément. Ce seuil est de l'ordre de quelques millisecondes, en moyenne [A-17].

Cette valeur est trop faible pour pouvoir rendre compte des phénomènes d'intégration temporelle qui interviennent dans l'audition. En effet, l'audition ne se réduit pas à une suite de sensations séparées : une mémoire à court-terme, appelée aussi mémoire phosphorescente [A-12], permet de rendre compte de la cohérence du déroulement au cours du temps des sons perçus. On s'accorde en général à penser qu'elle s'étend sur quelques secondes et que sa capacité est de l'ordre de 7 objets : chiffres, notes, etc... Sa capacité semble d'ailleurs mieux définie que sa durée.

La constante de temps introduite ainsi est due à des phénomènes nerveux et non plus mécaniques. De plus, pour lever le paradoxe de l'analyse fréquentielle (cf. paragraphe I-6-2), on est amené à supposer que cette constante de temps peut varier selon le type de son à analyser [A-18], conformément à la relation d'incertitude d'Heisenberg, à moins que plusieurs analyses, avec des constantes de temps différentes, aient lieu simultanément.

Nous avons simulé ce processus en reconnaissance par une intégration avec fuites des données spectrales, dans les expériences de reconnaissance par matrice d'apprentissage [C-1], [C-2].

I-6-5 Hiérarchisation de la perception auditive :

Le processus humain de compréhension de la parole est d'une très grande complexité, sans commune mesure avec celle des systèmes artificiels de reconnaissance vocale mis au point jusqu'à présent. Il se caractérise en particulier par une hiérarchisation des niveaux de perception : niveaux acoustique, linguistique, syntaxique, sémantique. Le passage d'un niveau à un autre se fait très rapidement, en général inconsciemment, lorsqu'une ambiguïté apparaît à un niveau. Tout cela donne un système très souple, qui utilise au mieux la redondance de la parole et peut s'affranchir ainsi de distorsions importantes des signaux.

Nous ne faisons ici qu'esquisser cette question qui sera reprise en détail en partie D.

I-7 Conclusion.

Nous avons examiné, dans ce chapitre, à la fois l'aspect émissif et l'aspect perceptuel de la communication parlée.

Après que le signal de parole a été défini, l'étude de l'émission vocale, sous ses divers aspects, a abouti à une description des sons élémentaires du français. De l'examen de la notion de phonème, nous déduirons des méthodes de segmentation et de reconnaissance à caractère dynamique.

L'étude de l'audition et de la perception auditive a mis en évidence la complexité de ces mécanismes. L'analyse et le prétraitement utilisés par la suite s'inspirent en partie des conclusions que nous avons tirées. Enfin, la hiérarchisation de la perception auditive apparaît dès à présent comme fondamentale et sera longuement reprise.

CHAPITRE II

GENERALITES SUR LA RECONNAISSANCE DE FORMES

Un ordinateur effectue beaucoup plus rapidement que le cerveau humain des opérations arithmétiques ou logiques. Par contre, il faut des programmes extrêmement complexes pour simuler, même avec un succès limité, sur un ordinateur, des processus de perception ou de raisonnement que l'homme effectue sans cesse, et souvent inconsciemment.

Cette branche de recherches en informatique a reçu le nom d'intelligence artificielle [A-19], et la reconnaissance des formes en fait partie : c'est donc une science récente liée au développement des ordinateurs, d'à peine plus de vingt ans d'âge.

La reconnaissance de la parole est un cas particulier du problème général de reconnaissance de formes, avec, en plus, des difficultés dues aux caractéristiques du signal de parole. Il est donc nécessaire d'introduire dans cette étude les méthodes générales de reconnaissance de formes, sans toutefois insister dans les détails, car les difficultés se situent plutôt au niveau de l'analyse (cf. partie B), dans le cas de la reconnaissance vocale.

Après avoir donné quelques définitions, nous passerons ainsi en revue rapidement les différentes méthodes de reconnaissance de formes, du moins celles qui sont applicables en reconnaissance de la parole.

II-1 Définitions.

II-1-1 Notion de forme :

Nous appellerons "forme" toute entité, statique ou dynamique, pouvant être décrite par un certain nombre de traits, et donc communicable d'homme à homme, ou d'homme à machine. Une forme peut ainsi être, suivant les cas, aussi bien une photographie, qu'une lettre manuscrite, qu'un son de la parole, ou qu'une trajectoire de particule dans une chambre à bulles.

Dans un problème donné, l'espace de représentation des formes, ou plus simplement espace des formes, est un sous-espace de l'espace physique.

II-1-2 La reconnaissance de formes :

Reconnaître une forme, c'est effectuer une partition de l'espace des formes en un nombre fini de classes, et décider de l'appartenance d'une forme présentée à l'une de ces classes. Les formes d'une même classe peuvent être identiques, mais le plus souvent les formes à reconnaître sont entachées d'un bruit parasite ; il n'y a plus alors identité mais seulement voisinage, similitude, au sens d'une mesure qui dépend du problème étudié. Ce dernier cas est en particulier celui rencontré en reconnaissance de formes vocales. Une classe de formes apparaît alors comme un "nuage" dans l'espace des formes.

La reconnaissance de formes couvre un grand nombre de domaines, tous plus ou moins liés à la communication homme-machine. Citons :

- la lecture optique,
- la reconnaissance vocale,

- la détection d'un objet sur une image (applications en météorologie, en physique nucléaire, dans le domaine militaire),

- l'aide au diagnostic, en médecine par exemple.

II-1-3 Prétraitement et regroupement des formes :

Dans l'espace des formes, une forme est décrite par des paramètres physiques fournis par des capteurs. Mais cette description n'est pas toujours directement adaptée à la reconnaissance ; on procède ainsi souvent à un traitement sur les formes brutes données par l'analyse physique, pour obtenir un nouvel espace des formes dans lequel la reconnaissance est facilitée. Cette opération est quelquefois complétée par un regroupement ("clustering") consistant, d'une part à condenser les nuages correspondant à une classe donnée, et d'autre part à éloigner les uns des autres les nuages des classes différentes.

Ces deux opérations sont fondamentales en reconnaissance de formes et nous reviendrons en détail sur ce problème en partie B.

II-2 Organisation générale d'un système de reconnaissance de formes.

Tout système de reconnaissance de formes se compose de trois parties, comme l'indique le schéma de base de la figure II-1 :

- un capteur, destiné à saisir les objets et à en donner une représentation ; ce sera, par exemple, un microphone dans le cas de la parole ;
- un étage de paramétrisation, susceptible d'effectuer un lissage de l'information donnée par le capteur, ou une extraction de paramètres pertinents, ou ces deux opérations simultanément ;

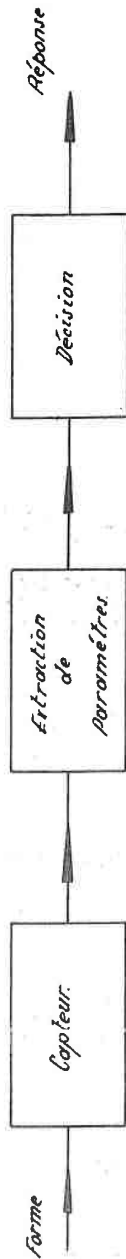


Schéma général d'un système de reconnaissance de formes

Fig. II.1

- un étage de décision, qui effectue la séparation dans l'espace des formes, grâce à des critères de classification des objets étudiés.

Ce schéma ne donne évidemment qu'une idée très générale sur l'organisation d'un système de reconnaissance de formes. En particulier, un même étage peut assurer plusieurs fonctions et c'est ainsi le système compresseur-analyseur ^{que} nous décrivons en partie B, effectuée à la fois un lissage et une extraction de paramètres. De plus, le caractère séquentiel de ce schéma n'exclut pas la mise en parallèle de plusieurs traitements, ni des retro et préactions d'un étage sur un autre (cf. partie D).

II-3 Méthodes de discrimination.

Il n'existe pas, pour l'instant, de théorie unifiée de la reconnaissance de formes. En fait, chaque application pose des problèmes spécifiques, et ceci est particulièrement vrai pour la paramétrisation (cf. partie B). Seules les méthodes de décision ont fait l'objet d'études systématiques [A-20] [A-21] [A-22].

Très schématiquement, les méthodes de décision peuvent être divisées en quatre groupes :

- méthodes de corrélation ;
- méthodes paramétriques de partition de l'espace des formes ;
- méthodes statistiques ;
- méthodes syntaxiques.

Cette subdivision est, à vrai dire, un peu artificielle. Certaines méthodes participent ainsi à plusieurs groupes, mais nous la conserverons pour la simplicité de l'exposé.

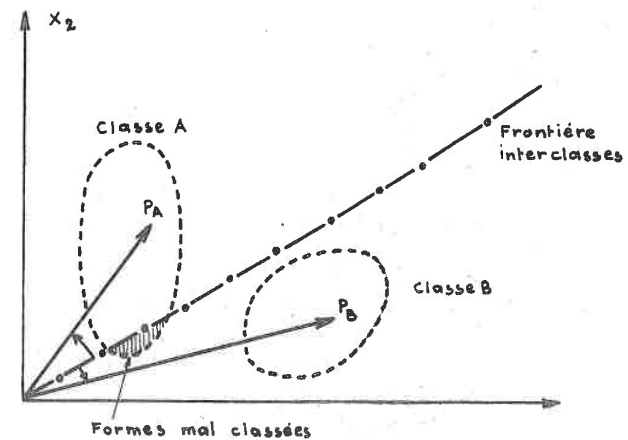
Nous avons déjà fait remarquer que, dans le cas de la parole, les difficultés se situaient plus au niveau du prétraitement que de la décision. Aussi, nous n'insisterons pas sur l'aspect théorique de ces méthodes, mais nous allons simplement passer en revue les principes généraux de chaque groupe.

II-3-1 Méthode de corrélation :

C'est la plus simple que l'on puisse utiliser. Elle consiste à calculer le taux de corrélation entre la forme inconnue et des formes prototypes normalisées, stockées en mémoire. La forme prototype choisie est celle qui donne la meilleure corrélation. La méthode est très restrictive, en ce sens qu'elle nécessite d'avoir des formes normalisées, ou du moins facilement normalisables. Cependant, si cette condition n'est pas remplie, une méthode de comparaison dynamique peut être utilisée (cf. partie C). Il n'existe pas de différences fondamentales entre les techniques de corrélation et les techniques de taxonomie que nous allons aborder maintenant (figure II-2, dans le cas d'un espace à deux dimensions).

II-3-2 Méthodes paramétriques de partition de l'espace des formes (taxonomie) :

Dans la mesure où elle peut être décrite par un ensemble de N paramètres, chaque forme, ou configuration, peut être représentée par un point dans l'espace des formes à N dimensions. Une classe d'appartenance est ainsi figurée par un nuage de points. Les méthodes de décision ont alors pour but de trouver des fonctions discriminantes définissant des hypersurfaces séparatrices, qui isolent les nuages les uns des autres.



Classification par corrélation avec des formes prototypes

Fig II-2

Ces fonctions $g_1, g_2, \dots, g_n$ des formes X , sont choisies de telle sorte que, pour toute forme X appartenant à la classe C_i :

$$g_i(X) > g_j(X) \quad \text{pour} \begin{cases} i, j = 1, 2, \dots, n \\ i \neq j \end{cases}$$

La surface séparatrice des classes C_i et C_j est alors donnée par :

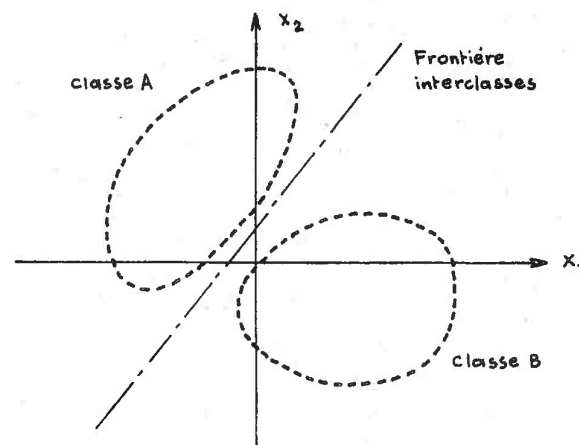
$$g_i(X) - g_j(X) = 0$$

Dans le cas le plus simple de fonctions discriminantes linéaires, ces surfaces sont des hyperplans. Si la séparation linéaire est insuffisante, on est amené à utiliser des surfaces séparatrices "linéaires par morceaux", ou même des surfaces polynomiales de degré supérieur ou égal à 2. La figure II-3 résume ces différentes possibilités dans un problème à deux dimensions.

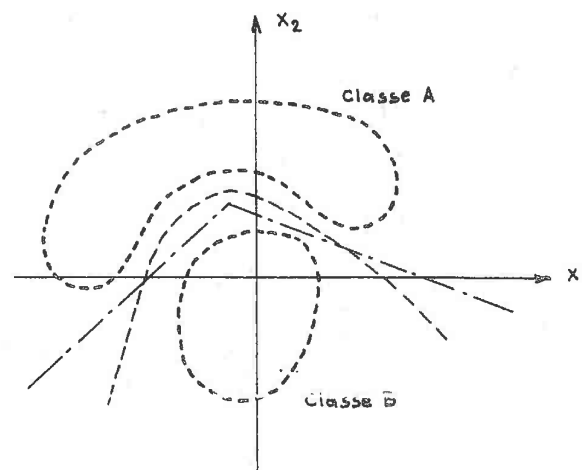
Les méthodes linéaires restent les plus utilisées, surtout pour des raisons de commodité de calcul.

Deux machines les représentent typiquement : le perceptron de Rosenblatt/A-23] et la matrice d'apprentissage. Nous reviendrons un peu plus en détail sur la seconde que nous avons étudiée en vue de la reconnaissance vocale.

Dans certains problèmes -c'est le cas en reconnaissance vocale- il peut se produire une interpénétration des différents nuages. La recherche d'hypersurfaces séparatrices se fait alors par optimisation par rapport à un critère, par exemple le taux de mauvaise classification.



Séparation linéaire



Séparations linéaire par morceaux et non linéaire

Séparation par hypersurfaces

Fig II-3

Des méthodes apparentées à ces méthodes de taxonomie sont celles qui consistent à définir une distance dans l'espace des formes (distance euclidienne, distance de Hamming, ou autre mesure) et de classer une forme, soit dans la catégorie dont le centre d'inertie est le plus proche (au sens de la mesure choisie), soit dans la catégorie des k plus proches voisins (règle $kNN/A-24$).

Ces dernières méthodes reviennent finalement à utiliser des hypersurfaces linéaires par morceaux.

II-3-3 Méthodes statistiques :

Les problèmes de reconnaissance de formes présentent souvent un aspect statistique qui a été mis à profit pour la recherche des fonctions discriminantes. Bénéficiant des résultats acquis en probabilités, en théorie de la décision, etc..., les méthodes statistiques de reconnaissance sont parmi les plus avancées. Nous nous bornons à envisager le principe des techniques les plus classiques dans ce domaine.

Tout d'abord, il est nécessaire de connaître :

i) les probabilités a priori de chacune des classes de formes :

$$p(C_i), i = 1, 2, \dots, n$$

ii) les probabilités conditionnelles :

$$p(X / C_j)$$

pour qu'une forme X appartienne à une classe C_j

iii) une fonction coût (ou perte) :

$$\lambda(C_i / C_j)$$

représentant le coût associé à la décision $X \in C_i$, quand, en fait, $X \in C_j$.

Cette fonction dépend bien entendu du problème étudié, et son choix est primordial. Il s'agit alors de trouver les fonctions discriminantes permettant de minimiser la valeur moyenne du coût λ . Cette règle de décision est appelée "règle de décision de Bayes".

En général, on ne connaît pas les expressions analytiques des probabilités utilisées. Il faut alors les "apprendre" sur des échantillons de données, qui doivent être très nombreux pour avoir une bonne approximation. De plus, toutes les valeurs de ces probabilités doivent être gardées en mémoire, ce qui représente un stockage d'information important.

Ces deux dernières contraintes limitent l'utilisation pratique de ces méthodes qui n'en demeurent pas moins très riches, pour les problèmes à caractère probabiliste bien entendu.

II-3-4 Méthodes syntaxiques :

Les problèmes d' "analyse de scène" ("scene analysis"), pour lesquels les méthodes de décision précédentes sont peu efficaces, ont amené le développement de nouvelles méthodes de reconnaissance, appelées méthodes syntaxiques ou linguistiques $[A-25]$. Plus récentes que les précédentes, et d'un principe tout différent, ces méthodes s'adressent à la structure des formes. Une forme apparaît alors comme l'assemblage d'un certain nombre d'éléments de base, décrit par une phrase d'un langage formel. Structure de formes et syntaxe d'une langue se correspondent ainsi étroitement. La caractérisation d'une forme présentée se fait alors par une analyse syntaxique de la phrase qui la décrit.

Jusqu'à présent, les applications ont été limitées à l'étude de cas simples [A-26], mais les résultats sont prometteurs, compte-tenu des progrès effectués en analyse des langages formels [A-27].

II-4 L'apprentissage.

II-4-1 Définition :

Plutôt que de fournir globalement au système, de façon exhaustive, les règles de classement des formes, on procède, en général, à un ajustement progressif des paramètres des fonctions discriminantes, par présentation, supervisée ou non, d'un grand nombre de formes types. Cette phase est appelée apprentissage. Elle nécessite des calculs importants, le plus souvent incompatibles avec le temps réel. De plus, comme l'apprentissage a lieu sur un nombre limité de formes, la connaissance acquise par le système demeure forcément incomplète.

En général, l'apprentissage est un processus séquentiel. Le système se corrige peu à peu au cours du temps, au fur et à mesure que de nouvelles formes se présentent : il acquiert ainsi une sorte d'expérience.

L'apprentissage est une phase très importante de la reconnaissance, et il est intéressant de réduire sa durée au maximum pour la commodité d'emploi, surtout dans le cas d'un système à utilisateurs nombreux et occasionnels.

II-4-2 Apprentissage et performances d'un système :

Deux points pratiques importants, sur lesquels nous reviendrons par la suite, sont à noter dès à présent. Ils concernent la relation existant entre l'apprentissage et l'évaluation des performances d'un système de reconnaissance :

i) le corpus de formes servant à tester un système de reconnaissance doit être différent du corpus ayant servi à l'apprentissage du système. Sinon on obtient une estimation biaisée des performances de ce système ;

ii) l'apprentissage apparaît beaucoup plus comme une adaptation réciproque homme-machine que comme une simple adaptation de la machine. Dans ces conditions, les performances d'un système pour un utilisateur familier risquent également d'être biaisées.

Ces deux points fixent en quelque sorte les limites de l'apprentissage.

II-4-3 Auto apprentissage et génération de machines :

L'apprentissage tel que nous l'avons défini précédemment suppose la présence d'un moniteur, ou d'un superviseur, capable de juger les performances du système qui apprend. L'idéal serait un autoapprentissage de la machine, sans moniteur, ce qui suppose de la part de cette machine une faculté de s'auto évaluer, donc une certaine réflexion si l'on veut. Un tel apprentissage serait très intéressant en reconnaissance vocale où, nous le verrons, les performances d'un système peuvent se dégrader dans le temps, sous l'effet de phénomènes tels que la dérive de la voix (cf. partie C chapitre II).

Ce domaine reste encore en grande partie à explorer. Ces problèmes rejoignent le problème connexe de la génération automatique de machines à reconnaître, elle-même liée aux travaux sur la recherche automatique d'opérateurs de reconnaissance [A-28].

II-5 Conclusion.

La reconnaissance de la parole, cas particulier de la reconnaissance de formes, utilise les techniques de cette dernière. Cependant, dans notre cas, les difficultés se situent plus au niveau du traitement du signal et de sa paramétrisation qu'au niveau de la décision. C'est pourquoi nous nous sommes contentés dans ce chapitre de rappeler brièvement les grandes méthodes de décision en reconnaissance de formes, susceptibles d'être utilisées en reconnaissance vocale.

Nous avons aussi abordé l'important problème de l'apprentissage, qui est fondamental en reconnaissance vocale, et sur lequel nous reviendrons souvent.

CHAPITRE III

LA RECONNAISSANCE DE LA PAROLE [A-29]

III-1 Définition.

Une machine à reconnaître la parole peut être définie comme un système capable de "comprendre" un son, un mot ou un message parlés, selon son degré de complexité. Cette compréhension, sans correspondre exactement au processus humain, permet cependant de transformer la parole en une action - s'il s'agit d'un ordre - ou en un message écrit (principe de la "machine à écrire phonétique").

III-2 Applications de la reconnaissance de la parole.

Historiquement, le premier but de la reconnaissance de la parole a été la recherche d'un mode de communication efficace et économique [A-30]. En effet, nous avons vu que la reconnaissance automatique d'un message parlé et la transmission codée de ses sons constitutifs réduirait la quantité d'information à transmettre dans un rapport d'au moins 1 000 à 1. Il est certain que dans l'état présent d'avancement des techniques de télécommunications, ce gain ne présente plus qu'un intérêt limité. Mais dans le cas de communications extra-terrestres à très longue distance, cette méthode semble être une des meilleures que l'on puisse envisager [A-31].

Mais, actuellement, la principale application d'un système à reconnaître la parole serait la communication orale entre l'homme et l'ordinateur, ou plus généralement entre l'homme et la machine. Cette idée, déjà ancienne [A-32], correspond à un réel besoin. En effet, si l'utilisation des ordinateurs tend à se généraliser, le langage constitue encore un des obstacles majeurs dans la communication avec ces derniers [A-33]. Dans la mesure où un ordinateur pourrait comprendre ne serait-ce qu'un langage simple et bien structuré, on pourrait alors parler de dialogue puisque, dès à présent, le problème de la réponse orale des ordinateurs est en voie de résolution, des unités à réponse orale étant déjà commercialisées [A-34]. De plus, la parole présente de nombreux avantages par rapport aux systèmes classiques de communication avec les ordinateurs (lecteurs de cartes, télétypes, ...) : c'est un mode de communication universel, rapide et qui ne nécessite pas la présence de l'opérateur à proximité de l'appareil.

Avec l'accroissement constant des capacités de mémoire des ordinateurs, on peut facilement imaginer des services de renseignements par téléphone, dans des domaines aussi variés que les banques, les assurances, les réservations de places ou la gestion de stocks.

Cependant, malgré tous ses avantages [A-36], la parole n'est sûrement pas destinée à remplacer les autres modes de communication avec les ordinateurs, mais seulement à les compléter. Ce serait, entre autres, une méthode idéale pour une utilisation occasionnelle d'un ordinateur par des non-spécialistes.

En fait, les applications possibles d'un système de reconnaissance de la parole dépendent beaucoup de la complexité et de la généralité de ce système. Nous appelons généralité la faculté, pour un système, de s'adapter à un vocabulaire important et à un nombre élevé de locuteurs.

Le tableau III-1 donne une idée générale des diverses applications possibles:

	Nombre de locuteurs		Taille du vocabulaire		Mode d'élocution		type de reconnaissance	
	grand	petit	grand	petit	parole continue	mots isolés	analytique	globale
A		X		X		X		X
B	X			X		X		X
C		X		X	X		X	
D	X			X		X		X
E	X		X		X		X	
F	X		X		X		X	

A : contrôle oral de machines, tri postal, etc. . .

B : composition orale d'un numéro de téléphone (ou équivalent)

C : mode conversationnel oral avec un calculateur spécialisé (type commande de processus)

D : entrée de programmes en ordinateur

E : applications plus générales : consultation de banques de données

F : machine à écrire phonétique.

Tableau III-1

La communication orale avec un ordinateur implique déjà un système de reconnaissance très évolué, que les plus optimistes n'espèrent pas avant 1980. Mais un système plus limité, tant au point de vue du vocabulaire que du nombre de locuteurs, rendrait déjà de grands services dans de nombreux domaines :

- transmission d'ordres simples à des machines [A-36], pour laquelle l'expérience montre d'ailleurs qu'elle entraîne moins d'erreurs qu'une transmission "presse-boutons". Le robot "Hand-Eye-Ear" de Vicens [B-15] est un des premiers exemples. Cette application serait très intéressante lorsque l'opérateur a les mains déjà occupées, comme par exemple un pilote de vaisseau spatial [A-37]. Dans le même ordre d'idées, les laboratoires BELL ont présenté un appareil expérimental capable de composer un numéro de téléphone à partir de l'énoncé oral des chiffres composant ce numéro. L'utilisation du téléphone par des handicapés moteurs est donnée comme application possible de ce système [A-38] ;

- tri postal : aux U. S. A. , les Postes ont déjà passé un contrat à ce sujet avec les laboratoires R. C. A. , et de même, un système oral est utilisé au service de manutention des bagages dans un aéroport ;

- etc...

En plus de ces applications, il faut ajouter que les recherches effectuées en reconnaissance de la parole conduiront sans doute à une meilleure connaissance des mécanismes humains de la perception ; elles permettront ainsi de progresser dans des domaines tels que l'intelligence artificielle ou la rééducation des handicapés. Cette dernière application fait partie de nos recherches actuelles. Grâce à un moyen très simple de communication avec un ordinateur, tel qu'une console de visualisation graphique, on peut en effet utiliser un système conversationnel de reconnaissance en temps réel, comme celui décrit en partie C, pour la rééducation intensive des mal-entendants.

Par ailleurs, ces applications s'avèreront de plus en plus nombreuses, à partir du moment où des systèmes vraiment performants et abordables seront proposés.

Car, pour l'instant, l'inconvénient le plus sérieux d'un système de reconnaissance de la parole est son coût élevé. Mais avec les progrès réalisés en équipement hardware, on peut espérer que le prix de revient d'une unité de reconnaissance vocale pourra être amené à une limite raisonnable. Le "Voice Command System", encore rudimentaire, proposé par Glenn et Hitchcock [A-39], qui a été le premier système de reconnaissance commercialisé, en est un exemple.

III-3 Aspects spécifiques de la reconnaissance de la parole.

III-3-1 Caractères particuliers :

Tout en ne constituant qu'un aspect du problème général de la reconnaissance des formes, la reconnaissance de la

parole n'en possède pas moins des caractères bien particuliers, dus surtout à la nature très spéciale du signal vocal. Ces caractères expliquent en partie les difficultés énormes que l'on rencontre dans ce domaine ; on peut principalement distinguer deux caractères :

- l'originalité du signal vocal, dont nous avons déjà parlé au chapitre I, et que l'on peut résumer en trois mots : variabilité, continuité et redondance ;

- l'adaptation homme-machine, qui est un phénomène que l'on retrouve dans toutes les expériences de reconnaissance vocale. L'adaptation de la machine intervient dans les performances globales d'un système. En effet, ce n'est pas seulement le nombre de locuteurs compris par une machine qu'il importe de connaître, mais aussi le fait que les voix de ces locuteurs sont assez proches ou non ; autrement dit il est important de savoir si une machine est adaptée à un type particulier de voix ou pas.

L'adaptation du locuteur à une machine apparaît surtout au cours des expériences menées en temps réel. On remarque en effet qu'un locuteur a tendance à modifier, parfois inconsciemment, sa voix de façon à obtenir une compréhension optimale. Ceci risque également de fausser les performances d'un système de reconnaissance, encore que ce soit un phénomène tout à fait normal et courant : on parle fort à une personne sourde, ou lentement à un étranger. Nous reviendrons sur cet aspect lors de l'étude de notre système de reconnaissance en temps réel.

III-3-2 Choix du niveau de reconnaissance :

Le caractère continu du signal de parole permet d'envisager deux approches bien différentes, tant dans leur principe que dans leurs applications, de la reconnaissance vocale. On peut les nommer :

- reconnaissance "analytique", ou par "sons élémentaires",
- reconnaissance "globale", ou par "mots".

Ces deux approches correspondent à deux processus de compréhension concurremment utilisés par le cerveau. En effet, un mot inhabituel, ou n'existant pas dans la langue, est compris grâce à une reconnaissance fragmentée au niveau d'éléments phonétiques ; par contre, dans une conversation courante, où l'on "prévoit" en quelque sorte ce qui va être dit, la compréhension d'un mot consiste simplement en une comparaison globale avec un nombre limité de mots, correspondant au sujet de la conversation.

Nous allons souligner rapidement les caractères propres de chacune de ces deux approches :

i) Reconnaissance "analytique", ou par "sons élémentaires" :

Ces sons élémentaires peuvent être des segments minimaux (constitutifs des phonèmes), des phonèmes, des diphonèmes, des syllabes, etc... Un message est ainsi représenté par une suite ininterrompue d'éléments phonétiques.

Cette méthode est générale, en ce sens qu'elle ne nécessite la mise en mémoire que des éléments phonétiques choisis (toujours en nombre limité, cf. partie D chapitre II) et n'introduit que peu de limitation sur le vocabulaire utilisé.

Elle impose par contre de bien prononcer les mots, de façon à ce que les éléments phonétiques soient tous présents, encore que les erreurs puissent être corrigées aux niveaux d'analyse supérieurs.

Une bonne segmentation du message est absolument indispensable, puisqu'il s'agit d'extraire du flux continu de parole les éléments phonétiques, avant de les reconnaître. Nous verrons que c'est un des problèmes les plus délicats à résoudre.

Cette méthode de reconnaissance est idéale pour effectuer la transmission d'un message. Rappelons que la transmission d'un message correspond, dans ces conditions, à un débit moyen de 50 à 100 bauds.

ii) Reconnaissance "globale" ou par "mots" :

Ces "mots" peuvent être aussi bien des mots au sens courant du terme, que des expressions ou même des phrases complètes [B-15].

Jusqu'à présent une prononciation "mot" par "mot" a seule été envisagée, le problème de la segmentation en mot de la parole continue n'ayant encore été qu'effleuré. Si elle s'affranchit ainsi du problème de la segmentation, la méthode nécessite par contre la mise en mémoire de tous les mots du vocabulaire.

De plus, un même mot peut présenter des différences considérables d'une prononciation à l'autre, et il est parfois nécessaire de stocker plusieurs représentations d'un même mot. On conçoit que l'on soit rapidement limité par la taille du vocabulaire utilisé.

La méthode globale, moins générale mais aussi moins complexe que la méthode analytique aboutira la première à des réalisations opérationnelles. Nous verrons qu'elle a été d'ailleurs plus souvent employée jusqu'à présent. Il n'en reste pas moins que la solution générale du problème réside dans l'utilisation d'une reconnaissance analytique.

Nous utiliserons, pour notre part, les deux approches (partie C), ce qui nous permettra d'en effectuer la comparaison.

III-4 Les travaux antérieurs.

III-4-1 Les premières réalisations :

Les premiers systèmes de reconnaissance de la parole, sous forme d'appareillages électroniques, furent proposés aux environs de 1950. Ces appareils, bien que de capacités limitées, ont cependant contribué à faire avancer nos connaissances dans le domaine de la communication parlée.

La première réalisation fut, sans doute, celle de J. Dreyfus-Graf en 1950 [A-40]. Dans cet appareil, nommé phonétographe par son inventeur, le signal vocal était analysé par six filtres répartis de 200 Hz à 3 600 Hz (correspondant à la bande téléphonique). Les sorties des filtres contrôlaient le mouvement d'un stylet enregistreur fournissant une représentation graphique des sons étudiés. Ces diagrammes se révélaient similaires à de très anciens signes égyptiens ou chinois. Poursuivant le principe des compresseurs sélectifs d'amplitude, Dreyfus-Graf a produit différentes versions améliorées de son appareil ; une version récente [A-41] permet la reconnaissance de 18 mots très courts prononcés par plusieurs locuteurs.

L'appareil de C. P. Smith (1951)/A-42], assez rudimentaire, utilisait pour la première fois une technique de corrélation, pour comparer les formes acoustiques fournies par l'analyse (banc de 32 filtres) à différents masques, et détecter ainsi les phonèmes.

En 1952, Davis, Biddulph et Balashek, des laboratoires Bell, présentaient un appareil capable de reconnaître les 10 chiffres/A-43]. Cet appareil comptait les passages à zéro du signal vocal dans deux bandes de fréquence (en deçà et au-delà de 900 Hz) et comparait ensuite à des formes standard. Le taux de reconnaissance atteignait 99 % pour un seul locuteur. Une version plus évoluée a été proposée par Dudley et Balashek en 1958/A-44]. La reconnaissance atteignait alors 90 % pour différents locuteurs, à condition qu'ils modifient leurs voix. Notons que la reconnaissance était effectuée aux niveaux successifs du phonème puis du mot.

Il faut attendre ensuite 1956 pour noter une réalisation intéressante, la "machine à écrire phonétique" de Olson et Belar (Laboratoires R. C. A.)/A-45] capable de reconnaître 10 mots très courts prononcés par une seule personne. L'unité de reconnaissance utilisée est la syllabe. Un banc de huit filtres (de 250 à 20 000 Hz) fournit une représentation binaire de la syllabe qui est ensuite appliquée à un décodeur préalablement ajusté au locuteur.

Des versions plus évoluées de cette machine peuvent traiter un vocabulaire d'environ 50 mots anglais et français, avec des performances du même ordre/A-46].

Un système de reconnaissance d'un principe tout à fait différent fut proposé, également en 1956, par Wiren et Stubbs/A-47]. Ces derniers, reprenant la notion de "caractères distinctifs" émise par Jakobson, Fant et Halle en 1952/A-48], recherchaient les phonèmes constitutifs du signal vocal par une suite de tests binaires réalisés par des circuits électroniques (sons sonores-sourds, turbulents-non turbulents, etc...). Bien que partiellement réalisé, le système donnait des résultats encourageants dans la reconnaissance des voyelles (94 % pour 21 locuteurs).

En 1958, une nouvelle étape était franchie avec l'expérience de Fry et Denes/A-49]. Mettant en pratique l'idée de Fry selon laquelle l'homme fait appel à des contraintes syntaxiques et sémantiques pour comprendre ce qu'il entend, ces auteurs construisirent un système de reconnaissance à deux "étapes" dans lequel une reconnaissance préliminaire purement acoustique était affinée par l'utilisation de données linguistiques. Ces données, très simples, étaient les probabilités d'apparition des diphonèmes. Pour un vocabulaire limité à 14 phonèmes, le pourcentage de reconnaissance atteignait 72 %, soit une amélioration d'environ 20 % grâce à l'utilisation des données linguistiques. L'expérience, dont le but était de prouver la validité de la méthode, fut en ce sens concluante.

Une autre "machine à écrire phonétique", imposante et ambitieuse, fut proposée au Japon, en 1961, par Sakai et Doshita (NEC et Université de Kyoto)/A-49]. Elle devait effectuer la segmentation de la parole continue et sa transcription en phonèmes, en se fondant principalement sur les deux premiers formants des voyelles. Ce projet semble actuellement en partie abandonné.

III-4-2 Apparition de l'ordinateur :

Les systèmes de reconnaissance entièrement électroniques que nous venons de passer en revue, sont surtout limités par leur faible capacité de mémoire. Les ordinateurs, apparus vers 1958, donnent une nouvelle dimension aux recherches, leur utilisation en reconnaissance de la parole devenant quasi-générale à partir de cette date.

Les premières expériences furent réalisées aux Etats-Unis, par F.D. et C.W. Forgie en 1959, Smith et Klem en 1960, Denes et Mathews en 1960, Shultz en 1961, Sholtz et Bakis en 1961.

Tous ces chercheurs effectuaient une analyse spectrale de la parole à l'aide d'un système hardware sauf Shultz dont le programme comportait une analyse simulée du signal vocal.

Des travaux de F.D. et C.W. Forgie au M.I.T., on peut retenir d'abord, en 1959, une reconnaissance de 10 voyelles anglaises par utilisation des deux premiers formants avec un pourcentage de 93 % pour une vingtaine de locuteurs/A-50/, ensuite, en 1962, une tentative de reconnaissance de deux consonnes fricatives anglaises/A-51/. Les résultats quoique moins bons, sont intéressants si l'on considère la similitude entre les spectres de ces consonnes.

Au M.I.T., toujours, Smith et Klem réalisèrent une autre tentative de reconnaissance de voyelles/A-52/, en appliquant les résultats de la décision statistique. Avec une fonction de décision quadratique, ils obtenaient un taux de reconnaissance de 94 % pour des mots monosyllabiques prononcés par 21 locuteurs.

La tentative de Denes et Mathews/A-53/est différente des précédentes dans son principe. Elle poursuit celle, déjà citée, de Fry et Denes/A-48/en ce sens que l'on fait encore appel à des contraintes linguistiques en restreignant considérablement le vocabulaire. Le but était la reconnaissance des 10 chiffres par comparaison à des formes "standard" stockées en mémoire.

Les auteurs concluaient à l'utilité indiscutable de l'ordinateur en reconnaissance de la parole.

Le programme de Shultz (IBM)/A-54/permettait aussi la reconnaissance des 10 chiffres à 97 % pour 50 locuteurs, hommes et femmes.

Sholtz et Bakis, également d'IBM, reconnaissaient les 10 chiffres/A-55/. Une série de tests portant sur les énergies du signal vocal dans 40 bandes de fréquence (de 100 Hz à 8 000 Hz), segmentait les mots prononcés en leurs phonèmes constitutifs. Les séquences de phonèmes obtenues étaient ensuite comparées à des séquences-types stockées en mémoire. Le pourcentage de reconnaissance atteignait 96 % pour 10 locuteurs des deux sexes.

Des expériences similaires furent menées également au Japon. Par exemple, en 1962, Suzuki et Nakata proposaient un système de reconnaissance de 300 mots monosyllabiques japonais prononcés par deux locuteurs/A-56/. Une analyse spectrale (banc de 26 filtres) permettait la détermination des fréquences des deux premiers formants par la méthode des moments/A-57/. Comme il était prévisible, les résultats étaient surtout bons pour les voyelles.

D'autres tentatives, reposant parfois sur des principes tout à fait différents, sont également à noter. Ainsi, à l'Université de Stanford, l'équipe de Talbert utilisait des éléments adaptatifs du type neurones, les adalines [A-58] pour simuler la reconnaissance de quelques chiffres et mots très courts [A-59]. Les résultats obtenus restaient modestes ; cependant cette voie de recherche peut être intéressante.

Dans une tout autre direction, Hemdal et Hughes, reprenant la notion de caractères distinctifs déjà utilisée [A-47], reconnaissaient 20 voyelles et diphtongues dans différents environnements vocaliques [A-60]. Quatre tests, portant sur les fréquences des deux premiers formants, étaient par exemple suffisants pour différencier entre elles dix voyelles, par formation d'un arbre de reconnaissance.

D'autres types d'analyses que l'analyse fréquentielle ont été utilisés. Ainsi Bezdel et Chandler reconnaissaient des voyelles avec une analyse "zero-crossings" [A-61]. Leur méthode consiste à calculer une distance euclidienne entre les formes étudiées et celles placées dans un dictionnaire.

Tout en restant encore très limités, ces programmes de reconnaissance représentaient déjà un gros progrès puisqu'ils étaient valables, sans aucun ajustement, pour un nombre plus important de locuteurs.

Au fur et à mesure de l'avancement des connaissances et du développement des ordinateurs, des expériences plus importantes furent tentées. Citons celles de King et Tunis (IBM) en 1965 [A-62] et de Gold (MIT) en 1966 [A-63]. Toutes deux

comportent une analyse fréquentielle de la parole. King et Tunis expérimentèrent plusieurs vocabulaires, pour un total de 30 mots tirés de langages de programmation, et plusieurs méthodes de décision. La reconnaissance atteignait 98 % pour plusieurs locuteurs masculins. Le programme de Gold permettait de reconnaître un vocabulaire de 54 mots anglais.

Parallèlement à ces recherches, le problème, en quelque sorte complémentaire, de l'identification d'un locuteur a été également abordé [A-64] [A-65]. Les paramètres intéressants dans ce cas ne sont pas forcément ceux utilisés en reconnaissance de la parole [A-66], mais les conclusions obtenues dans un domaine peuvent éventuellement être utiles dans l'autre. Certains chercheurs accordent à la méthode de reconnaissance de locuteurs une fiabilité intermédiaire entre celles de la graphologie et de la dactyloscopie [A-67].

En fait, il s'agit de systèmes de vérification d'identité d'un locuteur connu (utilisables par des banques, des organismes de crédit, etc...), et non pas d'identification d'un locuteur inconnu. Un brevet pour un tel système a été déposé récemment par les laboratoires Bell [A-68].

III-5 Etat actuel des recherches.

III-5-1 Le point dans le monde :

Nous avons déjà fait remarquer que la reconnaissance vocale n'en est encore qu'à ses débuts. La reconnaissance globale, en temps réel, d'un vocabulaire restreint est chose à peu près acquise, pour un nombre limité de locuteurs du même sexe. Mais le problème de la reconnaissance de vocabulaires importants et de phrases (domaine de la reconnaissance analytique) est loin d'être résolu, surtout en temps réel, seul contexte intéressant du point de vue pratique.

Les Etats-Unis et le Japon restent en tête dans ce domaine. Mais il ne faut pas négliger les résultats des chercheurs soviétiques. En Europe, les recherches sont moins avancées, avant tout parce que leur prix de revient est très important, mais il y a cependant des résultats notables.

a) En Amérique :

Aux Etats-Unis, la reconnaissance de la parole possède deux pôles géographiquement opposés : Stanford (Université) et Cambridge (M.I.T. et Université Harvard). A Stanford, Reddy s'intéressait à la segmentation et à la reconnaissance de la parole continue (il fut l'un des premiers à aborder ce problème) [B-16]. Le travail de Vicens [B-15] représente le système de reconnaissance le plus évolué à l'heure actuelle. Il est donc intéressant de citer quelques résultats et conclusions : par exemple "92 % de reconnaissance après 16 secondes environ pour une liste de 561 mots et courtes phrases prononcés par un locuteur" ; l'utilisation de méthodes propres à l'intelligence artificielle semble ouvrir des perspectives intéressantes dans le domaine de la reconnaissance de la parole. Deux applications pratiques de son système sont proposées par Vicens : une machine à calculer "de bureau", Descal, et un robot manipulateur, Hand-Eye-Ear, tous deux commandés à l'aide d'un langage à syntaxe très simple. Reddy poursuit maintenant ses recherches sur la reconnaissance en temps réel d'un langage à syntaxe simplifiée, à Pittsburgh [A-69]. Les recherches menées à Cambridge concernent à la fois l'analyse et la reconnaissance. En 1968, Bobrow et Klatt ont présenté un système capable de reconnaître 100 mots ou courtes phrases [A-70]. A partir de l'analyse effectuée par un banc de filtres, ces auteurs se sont attachés à trouver des paramètres intéressants pour la reconnaissance.

D'autres travaux, de moindre importance semble-t-il, ont été réalisés. Ils utilisent des méthodes classiques d'analyse fréquentielle [A-71] [A-72] [A-73] ou temporelle [A-74] pour la reconnaissance de chiffres ou de monosyllabes. Miller, Ross et Wine ont proposé de leur côté un programme permettant la reconnaissance en temps réel de quelques mots, avec un ordinateur utilisé en temps partagé [A-75]. On peut noter également un système hardware de reconnaissance des 10 chiffres par 10 locuteurs de Teacher et al. [A-76], remplaçant les 3 premiers formants par un "formant équivalent" unique, de fréquence variable [A-77]. Une version plus évoluée, reconnaissant un vocabulaire de 30 mots a aussi été proposée [A-79]. Le "Voice Command System" de Glenn et Hitchcock [A-39] a été le premier appareil de reconnaissance vocale commercialisé. C'est un appareil autonome, comportant un petit calculateur. L'analyse est effectuée par un banc de 16 filtres. L'appareil reconnaît un vocabulaire de 24 mots isolés prononcés par un locuteur, après 5 cycles d'apprentissage. On peut remarquer que les utilisations proposées concernent le contrôle aéronautique et l'enseignement par ordinateur. Depuis, un autre système (VIP 100), un peu plus performant, a également été mis au point [A-123].

Au Canada, des études sont menées à Vancouver par Donaldson sur la segmentation et sur une approche analytique de la reconnaissance de mots isolés. Ces recherches sont fondées sur les propriétés de l'analyse zero-crossings [A-78].

b- Au Japon :

Si la machine à écrire phonétique a été abandonnée pour l'instant/A-80/, la NEC n'en possède pas moins un système de reconnaissance des 10 chiffres en temps réel. De plus, des études partielles ont été menées, concernant diverses méthodes de reconnaissance de la parole/A-81/A-82/, ainsi que l'utilisation de contraintes linguistiques/A-83/. Plus récemment, Chiba et Sakoe ont proposé une méthode de reconnaissance de mots dans la parole continue, dans laquelle les petites variations de durée d'un mot sont compensées par programmation dynamique/A-84/.

c- En U. R. S. S. :

Il existe peu de publications soviétiques sur le sujet. Néanmoins, les Russes s'intéressent de près à la commande orale d'ordinateur/A-85/ et les résultats obtenus sont intéressants : reconnaissance de 168 mots (tirés du langage Algol) et chiffres par Velichko/A-86/, reconnaissance de 50 instructions par l'équipe de Vysotkij/A-87/. Ces derniers utilisent une analyse spectrale de la parole et notent l'intérêt des informations linguistiques en reconnaissance de la parole.

Par ailleurs, des études ont été menées sur la recherche de paramètres/A-88/A-89/, même dans le domaine des infra-sons/A-90/.

Actuellement, le nombre des équipes soviétiques s'intéressant à la reconnaissance de la parole est en augmentation, comme l'a montré le dernier congrès spécialisé (ARSO VI). Ici aussi, des méthodes de programmation dynamique ont été employées. Nous reviendrons plus en détail, en partie C, sur

ces méthodes, que nous avons également utilisées.

d- En Europe :

Des recherches importantes sont effectuées sur la parole en Suède (analyse, synthèse). Elles devraient prochainement déboucher sur la reconnaissance vocale.

Aux Pays-Bas/A-94/ Pols a mené plusieurs expériences de reconnaissance, de voyelles/A-91/ d'abord puis d'un vocabulaire limité à 20 mots/A-92/. Des études sur la reconnaissance des chiffres parlés ont également débuté à l'Université technique de Delft/A-93/.

En Allemagne, diverses tentatives de reconnaissance de quelques mots ont été menées. En particulier, Zwicker et al./A-95/ reconnaissent les 10 chiffres à 95 % par comparaison à des formes standard. Des systèmes électroniques existent aussi, comme ceux de Anke et Hoeschele/A-96/ ou de Tillmann (Bonn Université)/A-97/ qui reconnaissent respectivement les 10 chiffres et 20 nombres ou mots courts. En Italie, on peut noter diverses expériences de reconnaissance de voyelles sur ordinateur/A-98/A-99/. Un système entièrement électronique à circuits logiques a également été proposé/A-100/, lui aussi limité à la reconnaissance des voyelles.

Des expériences de reconnaissance de mots, utilisant une analyse spectrale ou une analyse zero-crossings ont été également décrites/A-101/.

Les travaux effectués en Grande Bretagne concernent des systèmes de reconnaissance sur ordinateur : Shearme et Leach, par exemple, reconnaissent 32 mots prononcés par 10 locuteurs masculins à l'aide d'une analyse spectrale simulée (90 % de reconnaissance)/A-102/. Un appareil entièrement électronique permettant de reconnaître quelques mots simples et les 10 chiffres a également été mis au point par Gilmour/A-103/.

Actuellement, quatre ou cinq équipes travaillent sur le sujet/A-104/. Les derniers résultats intéressants concernent la reconnaissance de mots isolés/A-105/A-106/, et l'utilisation de contraintes linguistiques dans la reconnaissance phonémique de parole continue/A-107/.

En France, une dizaine d'équipes travaillent actuellement sur la reconnaissance de la parole. Les premières études concernaient la reconnaissance de chiffres ou de mots simples/A-108/A-109/A-110/B-29/A-113/, à l'aide d'une analyse spectrale du signal vocal. Le C. N. E. T. a mis au point une machine à calculer "de bureau" du type de celle décrite par Vicens/A-111/. La thèse de Tubach présente le système le plus évolué de reconnaissance de phrases. A l'aide de contraintes syntaxico-sémantiques, Tubach reconnaît un programme Algol, prononcé par un locuteur, quasiment à 100 %. Au niveau acoustique, les paramètres fondamentaux sont les deux premiers formants, déterminés par programme/A-112/.

Une approche originale est celle d'Alinat/A-114/ qui a simulé analogiquement une cochlée, afin de reconnaître les phonèmes par une série des tests sur critères acoustiques.

III-5-2 Les tendances nouvelles :

Après l'intérêt porté à la reconnaissance de formes en général au cours des dernières années, on a assisté à une diminution des recherches sur la reconnaissance de la parole. Cela peut s'expliquer par l'ampleur du problème posé et par la faible connaissance que l'on a des mécanismes de perception et de production de la parole. C'est pourquoi les chercheurs se tournent de plus en plus vers les travaux fondamentaux sur la parole. Il s'agit maintenant de rechercher - à tous les niveaux : acoustique, syntaxique, ... - des invariants du langage. Cette

étude conduit à la notion de règles ("cues") de production de la parole. En ce sens, le modèle de reconnaissance "d'analyse par synthèse" proposé par Halle et Stevens/A-115/ reste prometteur, quoiqu'il n'ait jamais été entièrement réalisé.

Une autre voie, intéressante, consiste à simuler le comportement du cerveau par des réseaux de "neurones" électroniques/A-116/ dont de nombreux modèles ont été proposés/A-117/. Malheureusement, cette méthode conduit à des systèmes de reconnaissance extrêmement complexes. Sparkes/A-118/ décrit un schéma de reconnaissance de la parole également calqué sur le processus physiologique. C'est le modèle de Hill/A-119/ qui se rapproche le plus de ce schéma, mais il n'est pas certain qu'un système de reconnaissance automatique doive copier le processus humain correspondant, lui-même d'ailleurs encore mal connu.

Dans le domaine de l'analyse de la parole, il faut noter l'apparition des méthodes optiques de traitement du signal, principalement en lumière cohérente : filtrage, holographie/A-120/. Jusqu'à présent, les réalisations sont limitées à la reconnaissance de quelques mots, comme celles de Leverington/A-121/, Wilson/A-122/ qui utilise en fait une méthode de corrélation. A notre avis, le recours à un procédé optique de comparaison ne présente actuellement qu'un intérêt limité, car la comparaison entre deux mots est "anisotrope", en ce sens que des variations locales, non linéaires, apparaissent au sein des mots, comme nous le verrons en partie C.

Depuis peu, on assiste à une concertation plus grande des équipes et à un nouveau départ des recherches en reconnaissance vocale : Congrès ARSO VI en U. R. S. S., projet ARPA aux U. S. A. [A-124]. Ce renouveau est de plus caractérisé par une interpénétration de nombreuses disciplines, apparemment très éloignées.

On a vu que la plupart des systèmes de reconnaissance de la parole réalisés jusqu'à présent étaient limités au niveau acoustique ; il est maintenant certain que cela est insuffisant et que le traitement de la parole doit être effectué successivement à plusieurs niveaux : acoustique, mais aussi linguistique, syntaxique, sémantique. Les travaux de Vicens [B-15], Tubach [A-109], Alter [A-125] sur le langage Fortran et Reddy [A-126] sur le décryptage d'une séquence de phonèmes, constituent les premiers pas dans cette direction.

De façon générale, les méthodes utilisées sont plus souples et tiennent plus compte du caractère dynamique de la parole, comme celles que nous avons utilisées pour notre part (cf. partie C).

Quoiqu'il en soit, de nombreux problèmes restent à résoudre avant de concevoir un système suffisamment général, puisque l'on peut remarquer que, même pour un seul locuteur, aucun système n'est capable de reconnaître à la fois la voix normale, chuchotée ou criée [A-127].

III-6 Conclusion.

Les premières réalisations en reconnaissance de la parole ont été proposées vers 1950 sous la forme d'appareillages électroniques à performances limitées. L'utilisation des ordinateurs donne, dès 1958, une nouvelle dimension aux recherches. Cependant, malgré l'évolution des ordinateurs et des méthodes d'analyse, cette science en est encore à ses débuts. En effet, si la reconnaissance en temps réel de vocabulaires très limités est chose à peu près acquise, on est encore loin d'un système capable de comprendre une conversation continue. L'utilité d'un tel système est dès à présent indiscutable ; en particulier cela changerait considérablement les rapports entre l'homme et la machine.

Il est permis d'espérer que ces systèmes verront le jour dans la décennie qui commence. C'est maintenant un problème plus théorique que technique, lié à l'avancement de la connaissance des processus humains d'émission et de perception de la parole, et à l'utilisation de contraintes évoluées concernant le langage.

PARTIE B

ANALYSE ET PARAMETRISATION
DU SIGNAL VOCAL.

INTRODUCTION A LA PARTIE B.

En parlant des méthodes d'analyse utilisées par différents chercheurs, nous avons anticipé, dans le chapitre précédent, sur l'objet de cette seconde partie. Nous exposerons ici les méthodes d'analyse et d'extraction de paramètres utilisées en traitement du signal de parole, sur le plan général et sur le plan de nos propres réalisations.

Il s'agit donc d'explicitier le second étage d'un système de reconnaissance de formes (cf. figure A-II-1) dans le cas particulier de la parole. Le but de l'analyse est de rechercher un codage aussi condensé et efficace que possible de l'information transportée par la parole. Si les ordinateurs étaient capables de traiter directement cette information, l'analyse deviendrait superflue. Pour l'instant ce n'est pas le cas et, de toute façon, la redondance du signal vocal est telle que l'on a tout intérêt à en tirer parti. Nous revenons, dans un premier chapitre, sur cette redondance, en particulier à la lumière de la théorie de l'information. Ensuite les principales méthodes d'analyse du signal vocal sont décrites, ainsi que le système d'analyse que nous utilisons. Enfin, nous parlons des méthodes mathématiques de paramétrisation, en particulier de la méthode de Karhunen-Loeve que nous avons adaptée à notre problème.

CHAPITRE I

QUELQUES ASPECTS DE LA REDONDANCE DU SIGNAL VOCAL/B-1/

I-1 Introduction.

On a vu précédemment (partie A chapitre I) que la redondance du signal de parole, c'est-à-dire la surabondance de l'information transmise par la parole, est une caractéristique fondamentale de la communication parlée. Cette propriété remarquable est très intéressante en traitement automatique, et en particulier en reconnaissance, car elle permet d'effectuer une réduction importante de l'information à traiter, sans pertes importantes d'information utile. Cette compression d'information se traduit à la fois par une réduction du temps de traitement et un gain en place mémoire.

Après avoir rappelé rapidement l'aspect qualitatif de cette redondance, nous aborderons son aspect quantitatif, d'abord d'un point de vue purement statistique, puis en introduisant les concepts de la théorie de l'information.

I-2 Aspect qualitatif.

Il est facile de mettre en évidence la redondance de la parole. On peut, par exemple, écrêter énormément le signal sans que, pour cela, il cesse d'être compréhensible.

On constate donc une grande résistance aux perturbations (écrêtages, bruits, distorsions, ...) de l'information parlée, disons du "squelette sémantique" de la parole. La redondance sert en partie à corriger les erreurs à la réception : il est bien connu qu'un codage trop compact donne un message très vulnérable au bruit. On est même parfois amené à l'augmenter artificiellement, en répétant une phrase ou en épelant ses mots.

Mais la redondance provient aussi de la présence, dans le signal vocal, d'une information caractérisant le locuteur, inutile pour la compréhension du message. Ce type d'information est spécifique de la communication parlée. Parfois appelée "information esthétique" (cf. partie A chapitre I), elle renseigne l'auditeur sur l'intonation et sur l'état du locuteur. Nous verrons qu'il est difficile de chiffrer quantitativement cette information esthétique.

Nous avons vu aussi que, du point de vue informationnel, seules les parties transitoires ont un sens. Toutes les parties stationnaires sont ainsi redondantes dans la parole. Des mesures tendent à montrer que la partie redondante occupe près de la moitié de la durée d'un discours [B-2].

I-3 Aspects quantitatifs.

Pour tenter de chiffrer la redondance de la parole, nous utiliserons des résultats statistiques d'une étude sur le français parlé que nous avons menée [B-3], et qui est exposée en partie D.

I-3-1 Redondance et distributions linguistiques :

La redondance apparaît au niveau des distributions des éléments phonétiques du langage parlé. On trouve ainsi que les nombres de di- et de triphonèmes (association de deux et trois phonèmes) effectivement utilisés dans le français parlé, décrit par 33 phonèmes, sont limités :

- sur 1 089 (33^2) diphonèmes possibles, 615 seulement sont permis, soit 56 %,
- sur 35 937 (33^3) triphonèmes possibles, 2 509 seulement sont permis, soit 7 %.

La redondance chiffrée ainsi est due, entre autres, à des contraintes linguistiques et articulatoires, et les résultats obtenus sont remarquablement stables d'un corpus à l'autre (cf. partie D chapitre I).

La redondance apparaît aussi dans les lois de distribution des éléments phonétiques.

Ainsi, pour les diphonèmes, la probabilité du diphonème i-j, composé des phonèmes i (de probabilité P_i) et j (de probabilité P_j) serait :

$$P_{ij} = P_i \times P_j$$

si les phonèmes i et j étaient indépendants. En réalité, il existe une interdépendance entre les distributions phonémiques, et la relation précédente n'est pas vérifiée ; le phénomène est encore plus net au niveau des triphonèmes. Quelques exemples seront donnés au chapitre I partie D.

Toutes ces distributions obéissent à la règle des probabilités composées, ou règle de Bayes.

I-3-2 Redondance et théorie de l'information :

La théorie de l'information [B-4] permet de chiffrer la redondance de la parole en débit d'information. L'information moyenne, ou entropie, associée à la sélection d'un élément x_i parmi N , est :

$$S = - \sum_{i=1}^N p(x_i) \log_2 p(x_i).$$

Pour le français parlé, en supposant les 33 phonèmes équiprobables :

$$p(i) = \frac{1}{33} \quad \forall i = 1, 2, \dots, 33$$

l'entropie est :

$$S = \log_2 33 \approx 5 \text{ bits.}$$

Par contre, en tenant compte des probabilités d'occurrence réelles des phonèmes, on trouve :

$$S = - \sum_{i=1}^{33} p(x_i) \log_2 p(x_i) = 4,5 \text{ bits}$$

En tenant compte des probabilités des diphonèmes, on obtient :

$$S = 4,1 \text{ bits}$$

et avec les triphonèmes :

$$S = 3,8 \text{ bits.}$$

La redondance serait, bien entendu, encore plus importante si l'on considérait des associations de phonèmes d'ordre supérieur.

Si l'on admet un débit moyen de 10 phonèmes par seconde en conversation normale, on voit que le débit d'information, en ce qui concerne la seule transmission du contenu sémantique d'un message, est de l'ordre de 50 bauds.

Comparons ce chiffre à la capacité d'information d'un canal capable de transmettre la parole. Pour un tel canal, de type continu, le théorème de Shannon donne le taux de transmission maximal sans erreur :

$$C_{\max} = \Delta F \log_2 (1 + B)$$

où ΔF est la largeur de bande du canal et B le rapport signal/bruit.

Dans les conditions de transmission téléphonique on trouve :

$$C_{\max} = 30\,000 \text{ bauds}$$

Dans nos conditions expérimentales (cf. chapitre II)

$$C_{\max} = 7\,000 \log_2 (1 + 100) \approx 50\,000 \text{ bauds.}$$

Et dans les conditions de "haute fidélité" :

$$C_{\max} \approx 200\,000 \text{ bauds.}$$

Dans tous les cas, la redondance est énorme. Nous savons qu'elle est due en partie à la présence de "l'information esthétique", mais il est difficile de chiffrer cette dernière en débit d'information. On peut simplement noter que les expériences d'analyse et de synthèse de parole montrent qu'un débit de 2 000 à 3 000 bauds suffit à transmettre une parole compréhensible. L'analyseur spectral-codeur binaire que nous utilisons transmet, quant à lui, 24 bits chaque centième de seconde, soit un débit de 2 400 bauds.

I-4 Conclusion.

L'énorme redondance du signal vocal est bien connue, elle sert en partie à renforcer la résistance de l'information sémantique d'un message parlé aux diverses perturbations.

Cette redondance peut être chiffrée, tant en termes purement statistiques qu'en débit d'information. C'est de toute façon un aspect fondamental de la communication parlée, et les méthodes d'analyse et de codage dont nous allons parler maintenant ont pour but de diminuer cette redondance, sans perdre, autant que possible, d'information utile.

C H A P I T R E I I

ANALYSE DU SIGNAL VOCAL

II-1 Introduction.

Il existe de nombreuses méthodes d'analyse applicables à la parole. Sans faire d'exposé général, nous nous contenterons de rappeler ici les notions d'analyse spectrale nécessaires à l'exposé du système d'analyse que nous utiliserons dans les expériences de reconnaissance, et de citer pour mémoire les méthodes les plus couramment employées dans notre domaine.

Dans tout ce qui suit, le signal vocal en un point donné est désigné par $f(t)$. Cette fonction du temps représente les variations de la pression acoustique en ce point dues à la propagation de l'onde sonore ; on l'étudie le plus souvent sous forme électrique par l'intermédiaire d'un microphone.

II-2 Analyse spectrale.

II-2-1 Généralités :

La représentation dans le domaine fréquentiel de la parole présente un certain nombre d'avantages que montre l'étude de la phonation et de l'audition (cf. partie A chapitre I). En effet, il est possible de décrire les sons de parole en termes de fréquence, à partir du mécanisme d'émission vocale et, de plus, nous avons vu que l'oreille effectuait une analyse spectrale plus ou moins grossière des sons.

Mais, à strictement parler, une analyse spectrale simple est insuffisante pour la parole, phénomène non stationnaire où le temps joue un rôle primordial. Il faut donc rechercher un spectre évolutif, permettant d'introduire la dimension temporelle.

II-2-2 La transformée de Fourier :

a) L'intégrale de Fourier : définition, propriétés.

Une fonction du temps, apériodique, $f(t)$ et sa densité spectrale d'amplitude $F(\omega)$ se correspondent dans la transformée de Fourier :

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{j\omega t} d\omega \\ \text{et} \quad F(\omega) &= \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \end{aligned} \quad (1)$$

La transformée de Fourier n'existe que si

$$\int_{-\infty}^{+\infty} |f(t)| dt \quad \text{est finie.}$$

b) Spectre à court-terme.

En pratique, la condition précédente n'est pas réalisée pour la parole et, de plus, on ne connaît à l'instant t_0 que les valeurs passées du signal ($t \leq t_0$). Il est donc nécessaire de modifier $f(t)$ de façon à rendre calculable sa transformée de Fourier. Ceci est obtenu à l'aide d'une fenêtre temporelle $h(t)$ qui est telle que le produit

$$h(t) \cdot f(t)$$

possède une transformée de Fourier.

Mais, en considérant le signal $f(t)$ dans toute sa durée (par exemple la prononciation d'un mot), on lui fait correspondre un spectre unique par la transformée de Fourier. On perd alors l'information concernant les variations dans le temps du signal, c'est à dire une grande partie de l'information utile.

Nous retrouvons ici le rôle privilégié que joue le temps dans la parole : on cherche en fait le spectre "évolutif" de $f(t)$ à l'instant t , qui sera sans doute différent du spectre à l'instant $t + \Delta t$. Pour cela, il est nécessaire de ne tenir compte que du passé récent de $f(t)$, en choisissant une fenêtre $h(t)$ permettant d'isoler une portion de $f(t)$.

Le spectre instantané, ou plutôt "à court-terme" de $f(t)$, est alors :

$$F(\omega, t) = \int_{-\infty}^t f(\tau) h(t-\tau) e^{-j\omega \tau} d\tau \quad (2)$$

Il apparaît comme le produit de convolution :

$$f(t) e^{-j\omega t} \star h(t).$$

La quantité $|F(\omega, t)|^2 = F \cdot F^*$ est le spectre de puissance à court-terme.

La fonction $f(t)$ est ainsi représentée par une succession de spectres instantanés où les variations du signal apparaissent bien. En fait les spectres obtenus dépendent beaucoup de la fenêtre temporelle $h(t)$ que l'on a choisie. Parmi toutes les fenêtres possibles (triangulaire, carrée, etc...), c'est la fenêtre sinusoidale, ou de Hanning, qui introduit le moins de perturbations dans le signal $f(t)$.

Quoiqu'il en soit, le théorème de Plancherel (*) indique que le "filtrage temporel" effectué par $h(t)$ modifie le spectre fréquentiel du signal $f(t)$.

c) Traitement numérique.

Toutes les opérations de traitement du signal peuvent être effectuées par calculateurs digitaux. Un signal continu est alors représenté par une suite discrète de valeurs obtenues par une méthode d'échantillonnage [B-35].

Il est alors nécessaire de transposer sous forme discrète les relations de définition (1), pour les utiliser sur ordinateur.

La transformation des intégrales en sommes, ce qui revient à limiter les variations de pulsation ω , donc de fréquence, à un nombre discret de valeurs, fournit une solution :

$$X(f_n) = \Delta T \sum_{k=0}^{N-1} x(t_k) e^{-2\pi j f_n t_k} \quad (3)$$

avec $f_n = \frac{\omega_n}{2\pi}$, n pouvant prendre toutes les valeurs $0, \pm 1, \pm 2, \dots, \pm N/2$, en supposant que le signal $f(t)$ est représenté par la suite de N échantillons $x(t_k)$ pendant la durée T .
On peut simplifier en notant que :

$$\Delta T = \frac{T}{N}$$

et
$$\Delta f = \frac{1}{T}$$

Par suite, en posant :

$$t_k = k \Delta T$$

et
$$f_n = n \Delta f$$

(*) Rappelons que ce théorème très important établit que la transformée de Fourier d'un produit de convolution est un produit simple et réciproquement.

il vient :

$$X(n) = \Delta T \sum_{k=0}^{N-1} x(k) e^{-2\pi j \frac{n_k}{N}} \quad (4)$$

avec $n = 0, 1, 2, \dots, N-1$

On obtiendrait de même, pour la transformée inverse :

$$x(k) = \Delta f \sum_{n=0}^{N-1} X(n) e^{2\pi j \frac{n_k}{N}} \quad (4)$$

avec $k = 0, 1, 2, \dots, N-1$.

Les relations (3) et (4) forment la transformée discrète de Fourier (DFT).

Le spectre est ainsi défini de façon discrète : à N échantillons x de $f(t)$ correspondent N valeurs X du spectre calculables par un ordinateur.

Le problème semble ainsi résolu, à cela près que les relations précédentes nécessitent des temps de calcul beaucoup trop importants. Mais l'algorithme de transformée rapide de Fourier (FFT) permet de tourner cette difficulté. Proposé initialement par Cooley et Tukey [B-5], cet algorithme a fait l'objet de très nombreuses études [B-6]. La FFT n'est qu'une des méthodes permettant de calculer la DFT, mais avec une telle réduction du temps de calcul nécessaire qu'il est maintenant possible d'envisager une transformée de Fourier d'un signal de parole en temps réel, sur un gros calculateur, ou encore avec un processeur câblé.

Cependant, malgré cette nouvelle possibilité, la plupart des analyses spectrales en traitement de la parole sont effectuées analogiquement par des bancs de filtres. Cette façon de procéder fournit en effet une approximation de l'analyse de Fourier.

II-2-3 Analyse par bancs de filtres :

Le filtrage fréquentiel, ou filtrage linéaire [B-36], est l'équivalent, dans le domaine spectral, de l'opération qui a été appelée filtrage temporel lors de l'étude du spectre à court-terme. Un filtre passe-bande de fréquence ν_0 ne transmettra ainsi que les fréquences du signal $f(t)$ comprises dans sa largeur de bande autour de ν_0 , en effectuant le produit $F(\nu) \cdot H(\nu)$ où $H(\nu)$ représente la transformée de Fourier de la réponse impulsionnelle $h(t)$ du filtre.

Cette opération nécessite par conséquent que l'on connaisse le spectre fréquentiel $F(\nu)$ du signal $f(t)$, ce qui n'est pas le cas pour la parole. Mais le théorème de Plancherel permet de remplacer le produit simple $F(\nu) \cdot H(\nu)$ du domaine fréquentiel par le produit de convolution dans le domaine temporel $f(t) * h(t)$

$$f(t) * h(t) \iff F(\nu) \cdot H(\nu).$$

Ainsi, toute opération de filtrage effectuée sur un signal tel que le signal vocal est en fait une convolution.

Par changement de variable, le spectre à court-terme $F(\omega, t)$ de l'équation (2) peut encore s'écrire :

$$F(\omega, t) = \int_0^{\infty} f(t-\lambda) h(\lambda) e^{j\omega\lambda} e^{-j\omega t} d\lambda \quad (2bis)$$

soit :

$$\begin{aligned} F(\omega, t) &= e^{-j\omega t} \left[\int_0^{\infty} f(t-\lambda) h(\lambda) \cos \omega \lambda d\lambda + \right. \\ &\quad \left. + j \int_0^{\infty} f(t-\lambda) h(\lambda) \sin \omega \lambda d\lambda \right] \\ F(\omega, t) &= e^{-j\omega t} [a(\omega, t) + j b(\omega, t)] \end{aligned}$$

Le spectre à court-terme est alors :

$$|F(\omega, t)| = \sqrt{a^2 + b^2}$$

et l'on montre qu'il peut être approché par l'enveloppe de $a(\omega, t)$ (ou de $b(\omega, t)$). Cette enveloppe peut elle-même être obtenue par redressement et filtrage passe-bas de $a(\omega, t)$ (ou de $b(\omega, t)$).

Ainsi $|F(\omega, t)|$, autour de la fréquence $\frac{\omega}{2\pi}$, peut être obtenu en filtrant $f(t)$ par un filtre passe-bande calé sur cette fréquence, puis par redressement et filtrage passe-bas du signal obtenu (figure II-1).

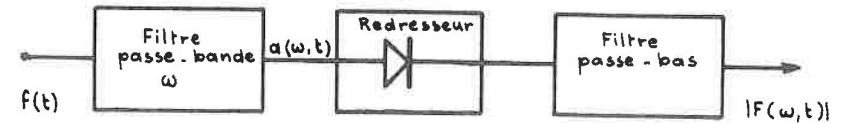


Schéma pratique de mesure du spectre à court-terme

Fig II. 1

Il est nécessaire d'utiliser autant de circuits identiques à celui de la figure II-1 que l'on désire obtenir de valeurs de $|F(\omega, t)|$, correspondant à des fréquences différentes. De plus, pour traduire l'évolution du spectre dans le temps, on

enregistre à une certaine fréquence d'échantillonnage les valeurs obtenues, ce qui donne une représentation comme celle de la figure II-2.

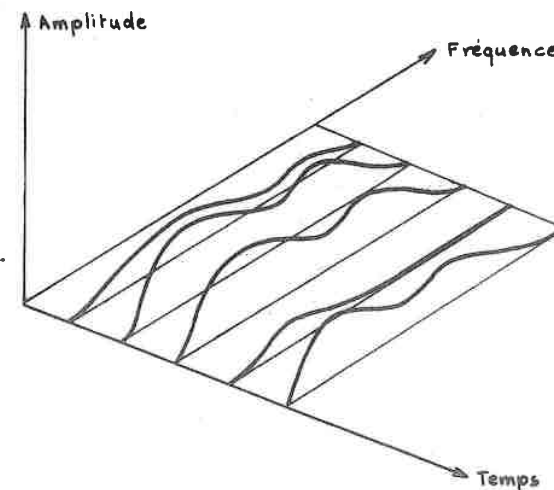
En fait, la quantité $a(\omega, t)$ (ou $b(\omega, t)$), fournie par le filtre passe-bande, dépend beaucoup des caractéristiques intrinsèques de ce filtre. Il en résulte que le spectre à court-terme ainsi obtenu diffère du spectre de Fourier précédent, et que des spectres donnés par différents systèmes de filtres diffèrent aussi entre eux.

Cette méthode de mesure du spectre à court-terme est utilisée dans la plupart des analyseurs spectraux, construits dans des buts de reconnaissance, de synthèse ou de transmission de la parole. C'est en particulier le principe de la plupart des Vocoders (Voice Coders), en particulier de leur "ancêtre" le Channel Vocoder de Dudley [B-7], du Sound Spectrograph dont nous allons parler en II-2-4, et de l'analyseur-codeur que nous utilisons en reconnaissance.

Les filtres utilisés dans ces analyseurs sont le plus souvent des filtres actifs analogiques. La mise au point de mémoires rapides en circuits intégrés permet d'envisager maintenant la réalisation de filtres numériques [B-37], dont les qualités de stabilité, précision, etc... sont supérieures, mais dont le prix encore élevé freine pour le moment la généralisation.

De nombreux travaux ont été effectués, sur les fréquences des filtres à utiliser pour la parole, et aussi les largeurs de bande optimales [B-9]. En effet, un système résonateur tel qu'un filtre passe-bande obéit à une loi d'incertitude

$$\Delta F \cdot \Delta t \approx 1 \quad (5)$$



Représentation de la parole par une suite de spectres à court-terme

Fig II.2

où ΔF est la largeur de bande et Δt le temps de montée du filtre. Ce temps de montée, de l'ordre de quelques millisecondes pour les fréquences usuelles de la parole doit être pris en considération dans le choix de la fréquence d'échantillonnage du spectre (cf. partie II-2-5). De plus, la relation (5) implique qu'à un Δt petit (grande résolution temporelle) correspond un ΔF grand (faible résolution fréquentielle), et réciproquement. On voit que le choix de la largeur de bande ΔF dépend du but que l'on recherche, et, le plus souvent, ce choix correspondra à un compromis entre les résolutions temporelle et fréquentielle.

II-2-4 La représentation temps-fréquence de la parole :

L'enregistrement des spectres instantanés donne une représentation du signal de parole dans l'espace à trois dimensions : temps-fréquence-densité spectrale d'amplitude ou de puissance. Cette représentation a beaucoup d'intérêt, mais ses trois dimensions la rendent peu pratique. On préfère effectuer une projection de "l'objet sonore" ainsi obtenu, sur le plan des deux dimensions les plus pertinentes, le temps et la fréquence. La troisième dimension, amplitude ou puissance, est ainsi représentée par un noircissement plus ou moins important dans le plan de projection. Il n'est alors plus possible de mesurer avec précision cette amplitude, mais ceci est un inconvénient mineur, que l'on retrouve d'ailleurs dans la perception humaine (cf. partie A chapitre I).

Ces spectrogrammes vocaux, ou sonagrammes, permettent une étude commode et très souple des sons vocaux, et ils jouent un grand rôle dans toutes les études sur la parole (un exemple de sonagramme est donné en II-2-5).

On les obtient en général à l'aide d'un seul filtre qui couvre tout le spectre par hétérodynage en plusieurs passages (principe du sonagramme). Il est aussi possible d'obtenir, en temps réel, un sonagramme à partir d'une batterie de filtres comme nous l'avons réalisé.

Les sonagrammes mettent en évidence le compromis entre résolutions temporelle et fréquentielle mentionné en II-2-3. Si l'on utilise un filtre à bande étroite, la résolution fréquentielle est bonne, les harmoniques apparaissent bien, mais les phénomènes transitoires mal. Au contraire l'utilisation d'un filtre à bande large estompe les harmoniques au profit des zones formantiques et des transitoires.

Là aussi, le choix dépend de l'étude que l'on désire effectuer.

II-2-5 Analyse spectrale binaire :

a) Principe.

Une variante de l'analyse spectrale par filtres consiste à ne faire intervenir la dimension amplitude, ou puissance, que sous forme binaire, 0 ou 1. Cette méthode, certes grossière, effectue une importante réduction de l'information à traiter, puisqu'un seul bit suffit pour coder l'amplitude, et nos expériences de reconnaissance prouvent que ce codage préserve cependant une bonne partie de l'information utile.

Ce codage binaire est obtenu à partir du spectre à court-terme donné par les filtres, à l'aide de comparateurs, disposés à la sortie de chaque paire de filtres consécutifs F_n et F_{n-1} . Si la tension de sortie du filtre F_{n-1} est supérieure

à celle du filtre F_n , l'information binaire est codée 0, et 1 dans le cas contraire. Cela revient à coder les pentes de la courbe amplitude-fréquence représentant le spectre à court-terme à un instant donné. De plus, l'information 0 peut correspondre au cas où les réponses des filtres sont inférieures à un seuil en tension V_0 , introduit pour s'affranchir des fluctuations aléatoires dues, par exemple, à des bruits parasites. La figure II-3 donne un exemple de codage d'un spectre à court-terme.

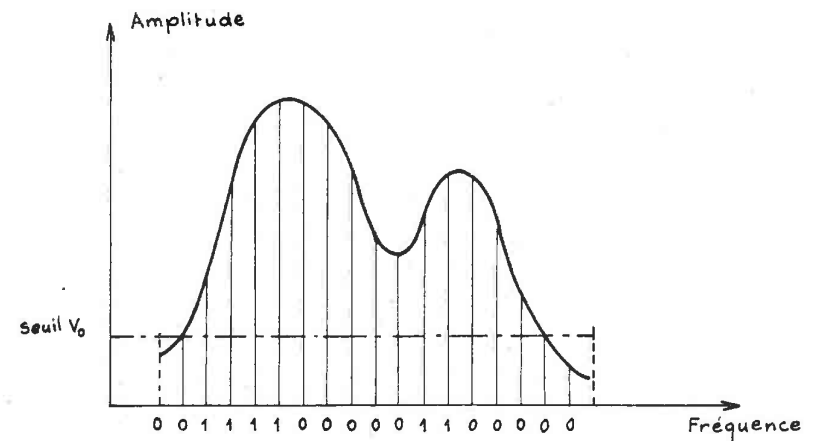
b) Description de l'analyseur-codeur.

Cet appareil a été décrit par ailleurs [B-10]/[B-11] et nous rappelons simplement ses principales caractéristiques. Il se compose de 24 filtres actifs passe-bande, répartis de façon à peu près logarithmique entre 110 Hz et 7 500 Hz, comme l'indique le tableau II-1.

Filtre N°	Fréquence centrale (Hz)	Filtre N°	Fréquence centrale (Hz)
1	110	13	1 200
2	160	14	1 350
3	230	15	1 500
4	310	16	1 730
5	395	17	1 970
6	470	18	2 200
7	515	19	2 500
8	600	20	2 900
9	670	21	3 600
10	740	22	4 700
11	860	23	6 000
12	960	24	7 500

TABLEAU II-1

Répartition des filtres de l'analyseur spectral.



Exemple de codage binaire du spectre à court-terme

Fig II.3

La fréquence d'échantillonnage choisie est de 100 Hz. Elle correspond précisément à un compromis et permet de rendre compte de bon nombre de phénomènes transitoires de la parole. L'analyseur fournit ainsi 24 données binaires chaque centième de seconde, soit un débit d'information de 2 400 bauds. Ces données peuvent être perforées sur ruban ou introduites en temps réel sur ordinateur comme nous le verrons. Dans ce dernier cas, la représentation temps-fréquence d'un mot prononcé (sorte de "sonagramme binaire") peut être visualisée instantanément sur écran cathodique, ou imprimée sur télétype.

La figure II-4 donne un exemple de représentation, pour le mot "SAPIN", ainsi que le sonagramme correspondant, effectué avec une largeur de bande $\Delta F = 300$ Hz.

c) Remarques sur l'utilisation de l'analyseur-codeur binaire.

L'analyseur-codeur dont nous venons de rappeler brièvement les caractéristiques est celui que nous avons utilisé pour toutes nos expériences de reconnaissance. Le débit d'information réduit, ainsi que la forme binaire des données en font un outil très intéressant, en particulier pour les expériences en temps réel, car, de plus, l'analyse obtenue est peu sensible aux variations d'amplitude du signal, donc aux variations dans l'intensité d'élocution.

Mais la réduction d'information réalisée se fait bien entendu avec perte d'une fraction de l'information utile. L'amplitude n'est certes pas une dimension négligeable. Aussi les performances de cet analyseur sont-elles forcément limitées. De plus, des difficultés peuvent apparaître dans certains cas, en particulier avec des sons nasalisés dont l'amplitude est faible, et qui ne donnent presque rien à l'analyse à cause de l'existence du seuil V_0 . Un exemple est donné sur la figure II-5, qui représente une mauvaise prononciation du mot "ANNEE", où le phonème /n/ n'apparaît que par quelques "1".

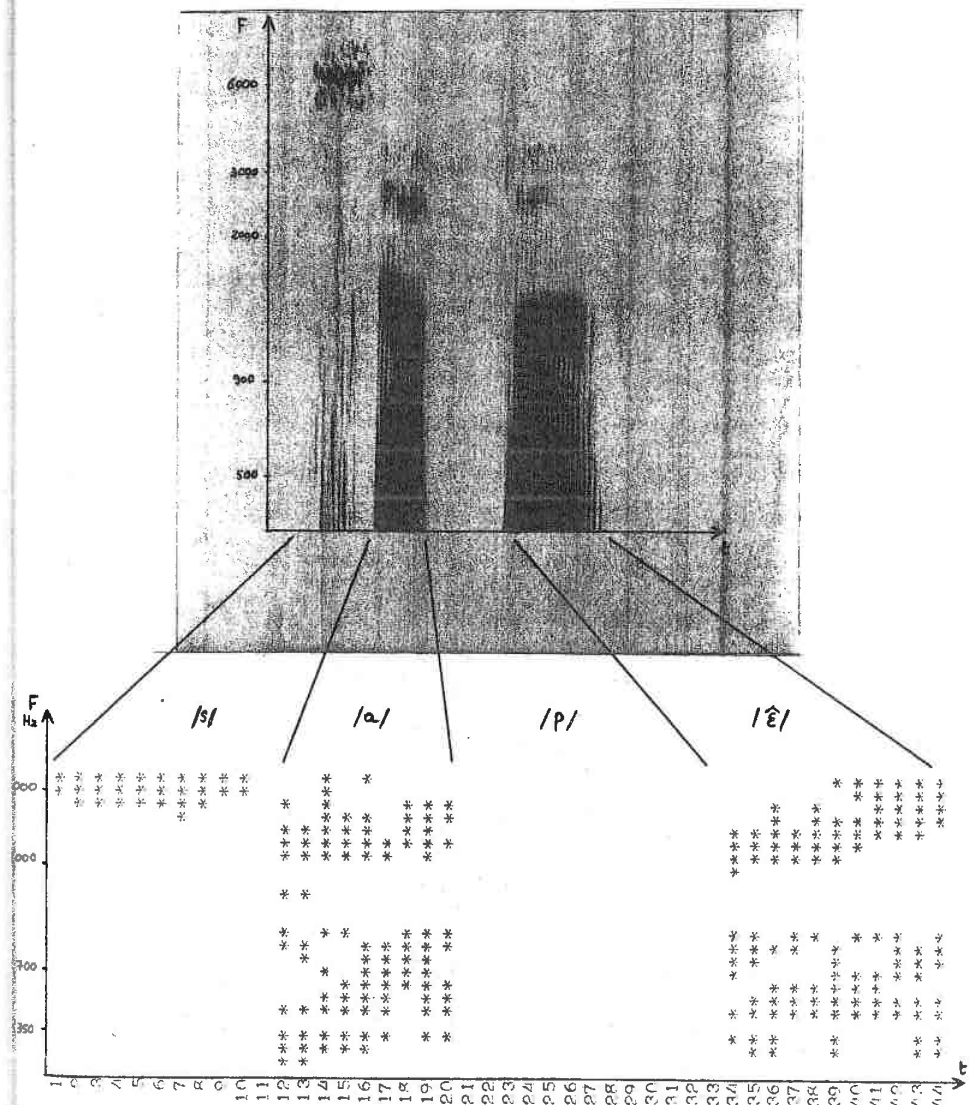


Figure II-4

Sonagramme et représentation binaire
du mot "SAPIN"

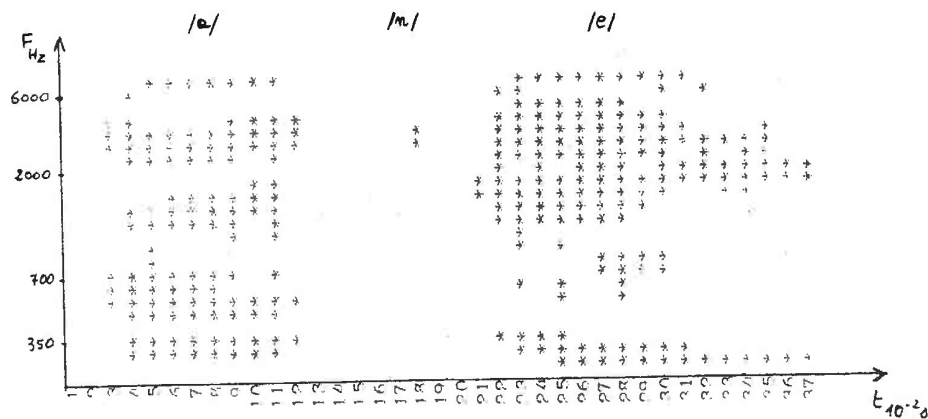


Figure II-2

Représentation binaire du mot "ANNEE"

Aussi prévoyons-nous maintenant une nouvelle analyse spectrale qui, en rendant compte plus complètement de l'amplitude, permettra de résoudre certaines ambiguïtés, au prix d'une augmentation de l'information à traiter.

II-2-6 Autres méthodes d'analyse spectrale :

Il existe d'autres méthodes d'analyse spectrale, correspondant au choix d'autres bases que celle de Fourier. On peut ainsi citer l'analyse par corrélation, l'analyse par transformée de Walsh [B-12], qui se prête très bien au traitement sur petit calculateur car elle ne fait intervenir que des additions et des soustractions.

Une autre méthode de grand intérêt est la méthode d'analyse par prédiction linéaire du signal vocal [B-13]. Cette "codage prédictif" part du principe que l'on peut prédire approximativement la valeur d'un échantillon de parole à partir d'un développement limité des échantillons précédents.

Appliquée depuis peu à l'analyse de la parole, la méthode de codage prédictif paraît très puissante, et c'est peut-être un outil d'avenir en traitement du signal vocal, car c'est la seule, pour l'instant, qui utilise le caractère anticipatif de la parole.

II-3 Analyse temporelle.

L'analyse de la parole peut se faire dans le domaine temporel, aussi bien que dans le domaine fréquentiel. Un type d'analyse temporelle consiste à compter le nombre de passages par zéro ("zero-crossings") du signal de parole.

Peterson [B-14] fut un des premiers à appliquer cette méthode à l'analyse de la parole, et plusieurs améliorations y ont été apportées depuis (en particulier le préfiltrage du signal avant le comptage).

L'intérêt réside dans le fait que les taux de zero-crossings du signal de parole et de sa dérivée sont des approximations respectivement des premier et deuxième formants, qui sont des paramètres intéressants en reconnaissance vocale.

Il en résulte que les paramètres donnés par la méthode de zero-crossings sont très séduisants, d'autant plus que l'on peut facilement les obtenir par un dispositif électronique [B-15].

Mais ces paramètres présentent une grande variabilité, de sorte qu'ils ne peuvent être utilisés que comme paramètres secondaires, en particulier pour la segmentation du signal vocal, pour laquelle ils sont très utiles [B-16] [B-17].

II-4 Conclusion.

Les méthodes d'analyse du signal vocal peuvent être grossièrement classées en méthodes fréquentielles et temporelles. Les analyses fréquentielles paraissent les plus intéressantes, comme le laissait prévoir l'étude de l'audition et de la perception ; mais des analyses des deux types peuvent très bien se compléter. Nous avons surtout insisté sur l'analyse de Fourier et sur son approximation électronique, l'analyse par filtres, mais il est également possible d'utiliser d'autres bases : Walsh, Haar, autres fonctions orthogonales.

Enfin ce chapitre a permis de préciser les caractéristiques principales et les limites de l'analyseur-codeur binaire que nous avons utilisé dans nos expériences.

CHAPITRE III

PARAMETRISATION DE LA PAROLE

III-1 Introduction.

Nous entendons par "paramétrisation de la parole" l'extraction d'un nombre limité de caractères pertinents ("features") à partir du signal vocal, ou de données fournies par une analyse préliminaire. Il s'agit donc encore d'une compression d'information, qui peut être mise en oeuvre par un dispositif câblé, aussi bien qu'à l'aide d'un algorithme. On pourrait dire qu'analyse et paramétrisation ne font qu'un, si ce n'est que la dernière opération fournit un nombre nettement plus restreints de paramètres, qui sont eux-mêmes plus élaborés. Le terme de prétraitement est également utilisé, à moins juste titre semble-t-il, puisque tout est traitement dans un système de reconnaissance vocale, depuis le micro qui capte le mot prononcé jusqu'à la réponse finale.

Les paramètres utilisés pour représenter la parole peuvent être de natures très différentes, depuis les paramètres purement mathématiques jusqu'aux paramètres physiques tels que formants, mélodie, etc... Ces derniers restent assez subjectifs, c'est une des raisons pour lesquelles il n'existe pas de théorie générale de la paramétrisation, quoique le problème soit mieux cerné en ce qui concerne les paramètres mathématiques.

Le problème de paramétrisation est fondamental en reconnaissance vocale, comme en reconnaissance de formes en général (cf. partie A chapitre II), car le nombre de paramètres retenus détermine en grande partie la rapidité du processus de reconnaissance, tandis que leur qualité en détermine les performances.

Après avoir défini plus précisément la notion de paramètres, nous exposerons une méthode mathématique de compression d'information. Les résultats obtenus avec nos données spectrales binaires seront ensuite examinés.

III-2 Différents types de paramètres.

Un paramètre, ou caractère, est une propriété de l'objet que l'on considère, transportant une certaine quantité d'information sur cet objet. Il est commode d'effectuer la distinction entre paramètres physiques et mathématiques [B-18], bien qu'elle ne soit pas toujours bien nette.

III-2-1 Paramètres physiques :

Ces paramètres ont toujours un aspect plus ou moins subjectif, même s'ils sont obtenus par des méthodes mathématiques. Deux paramètres physiques importants et très utilisés en reconnaissance vocale sont la fréquence fondamentale et les fréquences des formants.

Le problème d'extraction du fondamental ne possède pas encore de solution entièrement satisfaisante, malgré un certain nombre de méthodes nouvelles [B-19] [B-20], plus efficaces que le traditionnel filtrage passe-bas.

La recherche des fréquences des formants a fait l'objet d'études encore plus nombreuses, ces paramètres étant essentiels pour la caractérisation des voyelles, comme nous l'avons déjà vu. Les passages par zéro fournissent une méthode (cf. chapitre II), mais il est aussi possible de détecter les formants par un algorithme simple à partir des résultats d'une analyse spectrale. La figure III-1 montre l'évolution des trois premiers formants dans la syllabe "SO", obtenus à partir de l'analyse binaire.

III-2-2 Paramètres mathématiques :

La paramétrisation peut être aussi une opération mathématique de compression d'information, consistant à représenter la parole dans une certaine base, et fournissant des quantités sans réalité physique.

Diverses méthodes de compression d'information ont été utilisées, dans différents domaines. Elles donnent toutes des résultats du même ordre [B-21] et sont plus ou moins bien adaptées à chaque problème particulier. On peut les classer en gros en trois catégories : méthodes séquentielles, statistiques ou informationnelles.

La méthode que nous avons utilisée, à la fois statistique et informationnelle, est dérivée du développement de Karhunen-Loeve. Nous allons en rappeler les principes, avant de voir son utilité en reconnaissance automatique.

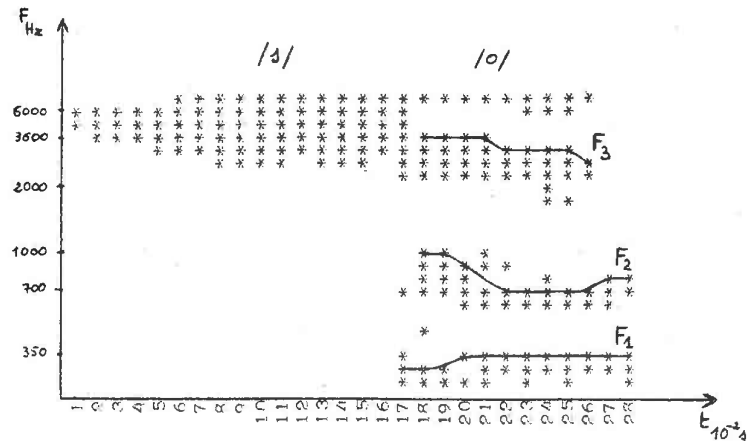


Figure III-1
Evolution de F_1 , F_2 et F_3 dans la syllabe "SO"
(locuteur masculin)

III-3 Compression de données et théorie de l'information [B-22].

III-3-1 Introduction :

Etant donnée une population d'objets, pour nous les objets sonores qui constituent la parole, le choix des paramètres caractéristiques revient au choix d'une base de représentation, d'un espace, dans lequel on décrit ces objets. Pour choisir cette base, nous effectuons le raisonnement suivant : les paramètres qui semblent a priori intéressants en reconnaissance sont ceux dont la dispersion est importante pour l'ensemble de la population étudiée. Par contre, un paramètre qui varie peu d'un objet à l'autre n'apporte pas d'information concernant la distinction entre objets et peut donc être omis. Cette omission constitue déjà une compression de données, et nous voyons apparaître la théorie de l'information. En effet la base de représentation qui permet d'effectuer cette compression est celle qui minimise l'entropie au sens de Shannon [B-4], de l'information transportée. Cette base est la base de Karhunen-Loeve, qui minimise aussi l'erreur moyenne commise dans la représentation d'un objet, pour une dimension de l'espace donnée [B-23]. Cette propriété a, par exemple, été utilisée pour les signaux transmis par satellite. On a montré également que, pour une certaine population d'objets, caractérisée par une distribution de probabilités, la base de Karhunen-Loeve était la plus efficace pour les cas les plus défavorables [B-24].

Des méthodes analogues ont déjà été utilisées en reconnaissance de chiffres écrits [B-25] et dans des problèmes de géographie humaine [B-26], entre autres. De plus, des variantes intéressantes ont été proposées [B-27] [B-28], mais elles ne sont valables que dans le cas de deux classes d'objets, ce qui en limite considérablement l'utilisation.

III-3-2 Formulation du problème :

Soit une population d'objets de même nature, appartenant à m classes distinctes connues (par exemple des phonèmes, des mots, ou simplement un ensemble de spectres à court-terme), chacune de probabilité a priori $p(\alpha)$, avec $\alpha = 1, 2, \dots, m$.

On a :

$$\sum_{\alpha=1}^m p(\alpha) = 1 \quad (1)$$

Chaque objet α est représenté par n caractéristiques x_i^α , $i = 1, 2, \dots, n$, soit encore par un vecteur $\{x^\alpha\}$ dans l'espace orthonormé à n dimensions des formes.

Soit Ψ une base orthonormée de cet espace, composée des n vecteurs $\{\Psi^\beta\}$, $\beta = 1, 2, \dots, n$. Tout vecteur forme $\{x^\alpha\}$ peut être décomposé sur cette base :

$$\{x^\alpha\} = \sum_{i=1}^n x_i^\alpha \{\Psi^i\} \quad (2)$$

On se propose de rechercher la base optimale, au sens défini précédemment, pour représenter la population considérée.

III-3-3 Recherche de la base optimale :

Les quantités $|x_i^\alpha|^2$ constituent une mesure de la validité du choix de la base pour représenter les formes étudiées. En particulier, les vecteurs coordonnés $\{\Psi^i\}$ correspondant à des x_i^α faibles peuvent être négligés sans que le vecteur forme $\{x^\alpha\}$ soit beaucoup perturbé.

Si l'on veut, de plus, tenir compte de la distribution de probabilité de la population étudiée, on peut chiffrer l'importance du vecteur coordonné $\{\Psi^i\}$, pour l'ensemble des formes étudiées, par la quantité :

$$p_i = \sum_{\alpha=1}^m p(\alpha) |x_i^\alpha|^2 \quad (3)$$

qui possède les propriétés d'une probabilité puisque l'on a :

$$p_i \geq 0 \quad \forall i \\ \sum_i p_i = 1$$

Effectuer une compression d'information consiste à trouver une base Φ_0 telle que, dans cette base, les quantités p_i soient concentrées sur un petit nombre de coordonnées de l'espace. Du point de vue de la théorie de l'information, cela revient à chercher la base Φ_0 telle que l'entropie de l'information transportée :

$$S(\Psi) = - \sum_i p_i \log p_i \quad (4)$$

soit minimale, soit :

$$S(\Phi_0) \leq S(\Psi) \quad \forall \Psi$$

Considérons donc une transformation unitaire de l'espace des formes, qui est alors caractérisé par la nouvelle base Φ de n vecteurs $\{\Phi^{\sigma}\}$, $\sigma = 1, 2, \dots, n$. Un vecteur forme $\{x^\alpha\}$ devient le vecteur $\{y^\alpha\}$ défini par :

$$\{y^\alpha\} = \sum_{j=1}^n y_j^\alpha \{\Phi^j\} \quad (5)$$

En appelant B la matrice de transformation on a la relation entre les composantes x_i^α et y_i^α :

$$y_i^\alpha = \sum_{j=1}^n B(i, j) x_j^\alpha \quad (6)$$

Avec la condition (4), la base optimale est alors formée des vecteurs propres $\{\varphi^k\}$ de la matrice d'autocorrélation A des formes considérées :

$$A(\xi, \eta) = \sum_{\alpha=1}^m x_\xi^\alpha x_\eta^\alpha p(\alpha) \quad (7)$$

La matrice A, hermitique en général, est symétrique dans notre cas où ses termes sont réels.

On a :

$$\sum_{\eta=1}^n A(\xi, \eta) \varphi_\eta^k = \lambda_k \varphi_\xi^k \quad (8)$$

en appelant λ_k , $k = 1, 2, \dots, n$ les valeurs propres de la matrice d'autocorrélation.

Si l'ensemble des objets $\{x^\alpha\}$ présentent n caractères indépendants, la matrice A possède n valeurs propres non nulles, en comptant, en cas de dégénérescence, une valeur propre d'ordre p comme p valeurs propres distinctes.

Numérotions les axes du nouvel espace des formes de telle sorte que les valeurs propres λ_k soient rangées par ordre décroissant :

$$1 \gg \lambda_1 \gg \lambda_2 \gg \dots \gg \lambda_n \quad (9)$$

On peut alors réduire la dimension de l'espace de n à n_0 ($n_0 < n$) en ne retenant que les n_0 premiers axes. On définit ainsi un espace à n_0 dimensions, défini par la base des vecteurs propres $\{\varphi^k\}$, $k = 1, 2, \dots, n_0$, qui fournit une représentation des formes de départ valable avec un "seuil de confiance" σ :

$$\sigma = \sum_{k=1}^{n_0} \lambda_k < 1 \quad (10)$$

III-3-4 Amélioration en vue de la reconnaissance :

La méthode précédente permet de déterminer les caractères importants d'une forme, autrement dit de tester la "qualité informante" des opérateurs qui ont permis d'appréhender ces caractères à partir des formes brutes. La notion d'opérateur est utilisée ici au sens général ; ce peut être aussi bien, dans le cas de la parole, un filtre d'analyseur spectral qu'un sous-programme de prétraitement.

Mais la qualité informante d'un opérateur varie selon que l'on se place du point de vue de la communication ou de la reconnaissance. Dans le premier cas les caractères intéressants sont les caractères d'importance statistique la plus grande. Dans le second cas les caractères intéressants sont les caractères distinctifs de chaque classe de formes, c'est-à-dire ceux qui différencient une classe donnée des autres classes.

Il est donc nécessaire d'aménager la méthode pour l'utiliser à la reconnaissance. Pour cela on définit un "objet moyen" $\{x^M\}$ tel que :

$$x_i^M = \frac{1}{m} \sum_{\alpha=1}^m x_i^\alpha \quad (11)$$

Puis on retranche $\{x^M\}$ des vecteurs formes étudiés $\{x^a\}$. On obtient alors des formes pondérées $\{x^a\}^P$:

$$\{x^a\}^P = \{x^a\} - \{x^M\} \quad (12)$$

constituées surtout des caractères distinctifs de chaque classe, correspondant à une faible corrélation avec les autres caractères.

La méthode de compression précédente, appliquée aux formes définies par (12), fournit alors les paramètres les plus utiles pour distinguer entre elles les formes de départ $\{x^a\}$.

III-4 Exemples de résultats obtenus.

III-4-1 Introduction :

Nous avons appliqué cette méthode de compression de données aux résultats fournis par l'analyse spectrale binaire, dans le cadre d'un système de reconnaissance par matrice d'apprentissage [B-29]/[B-30]. La simulation, puis la réalisation électronique [B-31] de ce système furent nos premières expériences de reconnaissance. Nous en avons maintenant abandonné le principe, à cause de la trop grande rigidité de la matrice d'apprentissage, au profit de méthodes plus souples (cf. partie C), mais les résultats n'en sont pas moins valables pour notre propos actuel.

Dans ces expériences, menées en temps différé sur calculateur 10070, les données binaires étaient intégrées dans le temps avec une certaine constante de fuite pour tenir compte de la structure continue de la parole, ce qui correspond un peu à la constante de temps de l'oreille (cf. partie A chapitre I).

Les sorties des 24 filtres, intégrées, peuvent être considérées comme des variables aléatoires, dont on peut calculer la variance pour un certain corpus de données de parole. Sur la figure III-2 qui donne les variances des 24 filtres, on peut constater que le filtre le plus efficace ne transporte qu'à peine 6 % de l'information totale. La compression d'information fournira un sous-espace bien plus efficace pour représenter nos données.

III-4-2 Représentation des sons dans un sous-espace optimal :

L'organigramme de la figure III-3 indique le déroulement des calculs pour la recherche d'un sous-espace correspondant à un seuil de confiance σ fixé. Les résultats qui suivent correspondent à un corpus de 60 mots prononcés par trois locuteurs masculins.

La variance des axes optimaux est portée sur la figure III-4. Nous nous sommes limités aux 12 premiers axes, les autres n'intervenant que pour moins de 1 % de la variance totale. On voit que le premier axe, à lui tout seul, transporte 38 % de l'information totale disponible, soit presque autant que les 8 meilleurs filtres de l'analyseur spectral (cf. figure III-2).

La figure III-5 donne la variation du seuil de confiance σ en fonction de la dimension n du sous-espace optimal : les huit premiers axes optimaux donnent une représentation valable avec un seuil de 95 %.

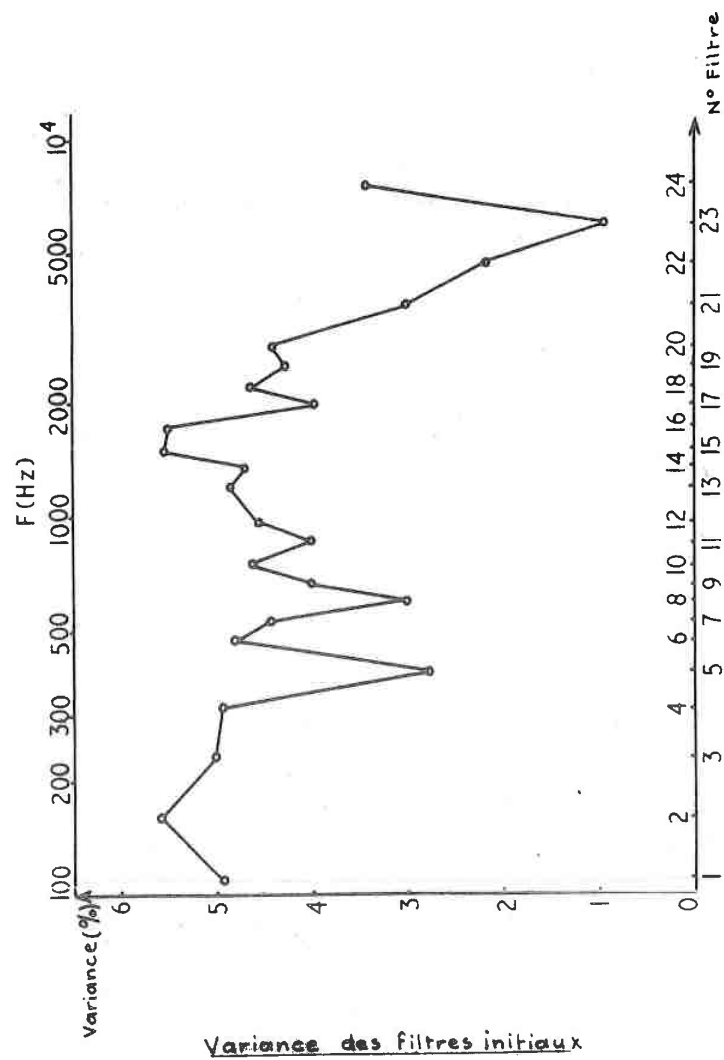


Fig III-2

101

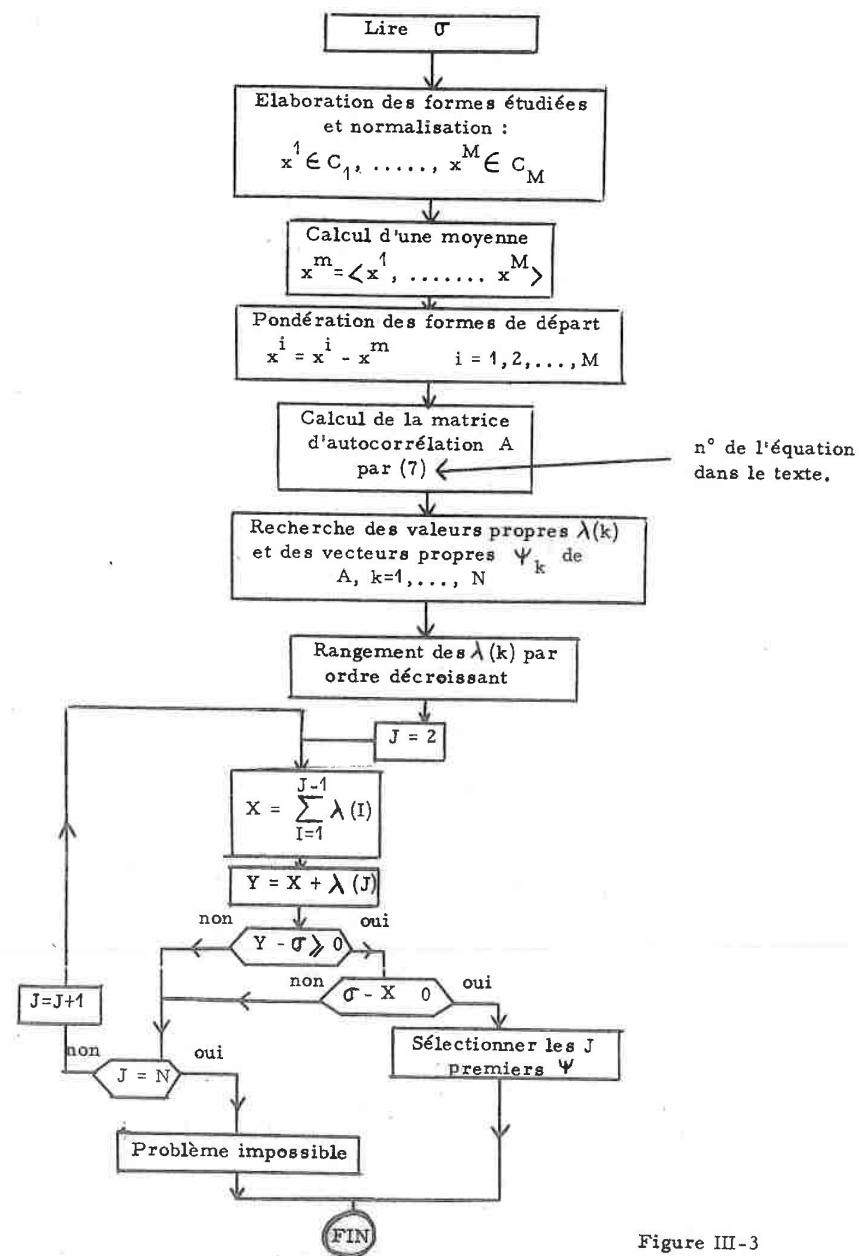
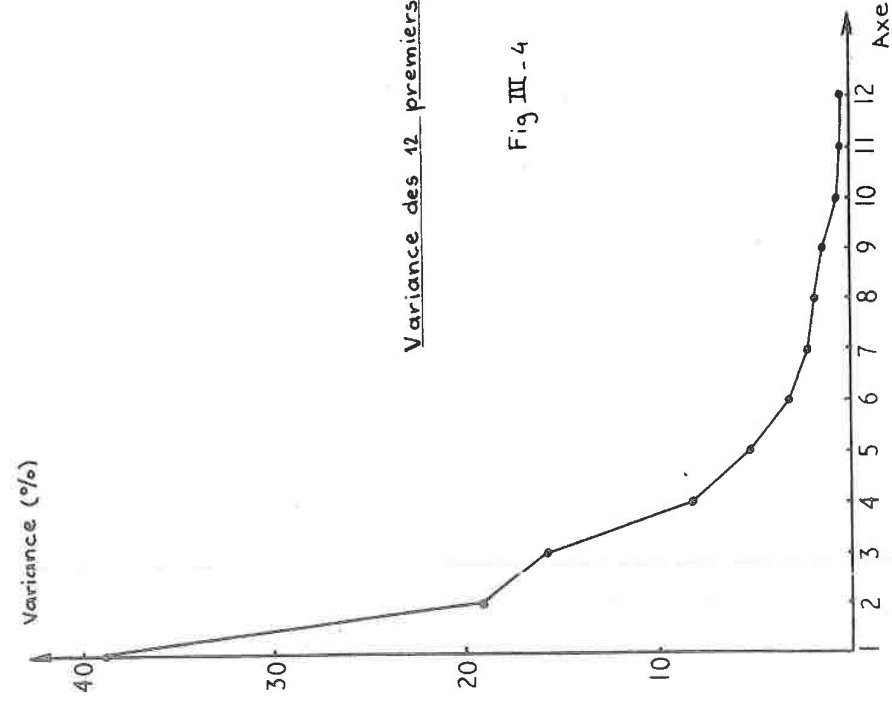
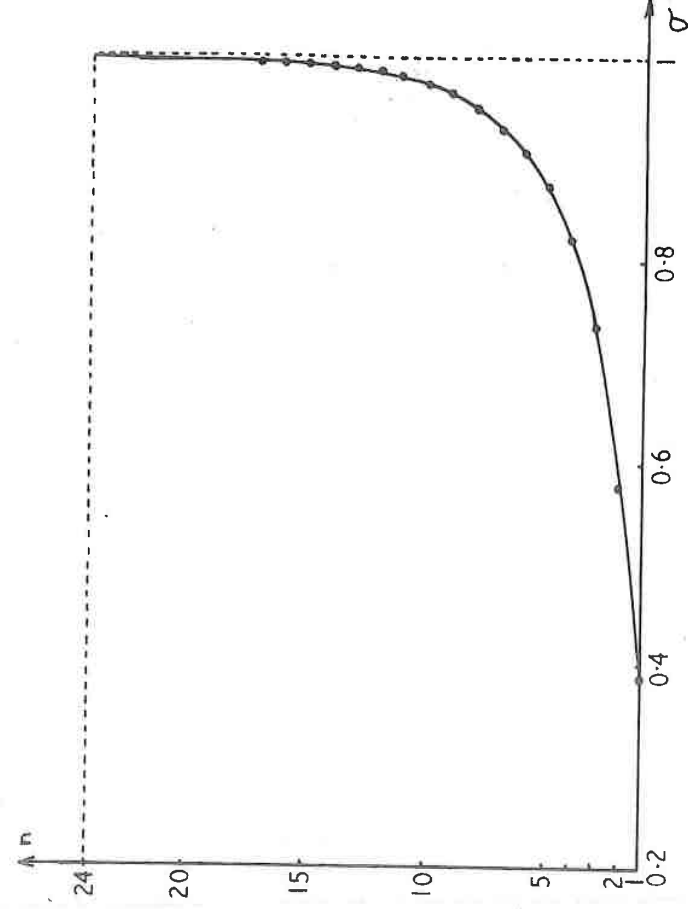


Figure III-3



103



104

A titre d'exemple, la distribution des centres d'inertie de sept voyelles dans le sous-espace des trois premiers axes ($\sigma = 74\%$), a été portée sur la figure III-6. Ces sept voyelles -/a/, /i/, /y/, /ə /, /o/, /e/, /ɛ /- n'ont pas été prononcées isolément mais prises dans le contexte des 60 mots de données. Ces points sont bien séparés, et, de plus, chaque nuage présente une dispersion raisonnable. On peut dire que ces voyelles sont déjà séparables à l'aide de ces trois paramètres mathématiques, de même qu'elles le seraient à l'aide des trois paramètres physiques que sont les fréquences formantiques.

III-4-3 Application à la reconnaissance [B-32]:

a) Choix de la dimension du sous-espace.

Les résultats précédents montrent qu'il est possible de représenter les formes initiales à 24 dimensions, dans un sous-espace optimal de dimension nettement moindre, avec une perte d'information faible.

Pour l'application en reconnaissance, il s'agit de prendre comme dimension du sous-espace le nombre minimal nécessaire pour décrire les différences spectrales des sons étudiés. Pour déterminer ce nombre, nous avons répété la procédure précédente de compression, en ne considérant que les valeurs moyennes de chaque classe de sons (phonèmes) et non plus tous les phonèmes des 60 mots utilisés. Nous avons ainsi constaté qu'environ six paramètres étaient plus que suffisants, les autres ne rendant compte que des fluctuations à l'intérieur d'une même classe, et des variations dues aux différents locuteurs.

Nous avons ainsi retenu 6 axes, correspondant à $\sigma = 91\%$.

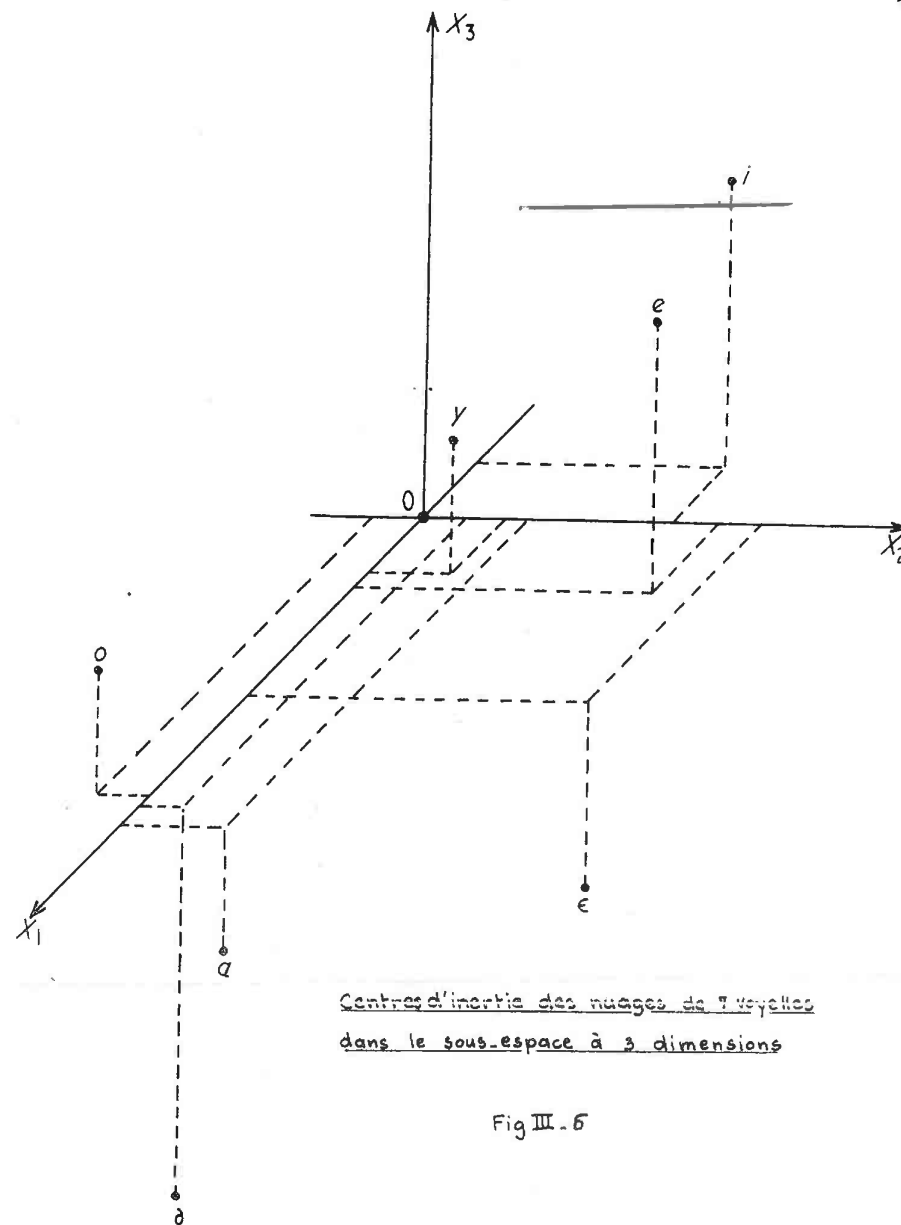


Fig III-6

b) Résultats obtenus.

Rappelons que les résultats de reconnaissance ci-dessous correspondent à l'utilisation du système simulé à matrices d'apprentissage que nous avons maintenant abandonné. Ces résultats sont intéressants à comparer entre eux, mais ne doivent pas être considérés dans l'absolu.

Avec les 24 paramètres initiaux (sorties des filtres intégrées), nous avons obtenu 76 % de reconnaissance sur les 20 mots tirés des données initiales (constituées de 60 mots par 3 locuteurs) et prononcés par le même locuteur.

Dans les mêmes conditions, avec 6 paramètres optimaux, le taux de reconnaissance était de 73 %, avec, bien entendu, une diminution du temps de calcul et du volume de données à stocker. La diminution du taux de reconnaissance (3 %) est plus que compensée par ces gains en temps et place mémoire qui permettent un traitement ultérieur utilisant des contraintes évoluées : linguistiques, syntaxiques (cf. partie D).

III-4-4 Présélection de paramètres :

Nous avons vu tout l'intérêt que peut présenter l'utilisation de paramètres mathématiques du type de ceux donnés par le développement de Karhunen-Loeve, ou par des transformations du même genre, telles que l'analyse factorielle [B-33] [A-92]. Mais, pour être efficaces, ces méthodes nécessitent de traiter d'importants corpus de données, et les calculs à effectuer sont complexes (par exemple diagonalisation et inversion de matrices 24 x 24, dans notre cas). Pour réduire le volume de ces calculs, nous avons cherché une méthode rapide et suffisamment sûre permettant d'éliminer a priori les paramètres initiaux manifestement peu intéressants. Trois méthodes

ont été retenues et testées avec nos données : deux méthodes statistiques (variance et redondance) et une méthode informationnelle (information mutuelle). Nous n'insisterons pas ici sur leurs résultats donnés dans [B-34].

III-5 Conclusion - Perspectives.

La paramétrisation est une opération préliminaire très utile en reconnaissance de la parole, puisqu'elle permet d'effectuer une compression de l'information à traiter, avec le moins possible de pertes d'information utile. Il en résulte un gain en temps de calcul et en encombrement mémoire. Les paramètres extraits peuvent être soit physiques, soit purement mathématiques, comme ceux donnés par la méthode dérivée du développement de Karhunen-Loeve, que nous avons mis au point pour notre cas particulier.

Les résultats obtenus montrent la validité de ces paramètres mathématiques. Ainsi, les pourcentages de reconnaissance sont presque identiques, selon que l'on utilise seulement quelques paramètres optimaux, ou la totalité des données spectrales. Par ailleurs, l'expérience montre que les voyelles sont séparables dans un sous-espace à trois paramètres mathématiques, de même qu'elles le sont dans le sous-espace des trois premiers formants. L'analogie est flagrante entre ces deux types de paramètres, mais les paramètres mathématiques présentent le grand avantage d'être beaucoup plus facilement accessibles que les fréquences formantiques.

Nous prévoyons la réalisation prochaine d'un dispositif électronique ou microprogrammé, qui permettra d'effectuer en temps réel le passage des informations spectrales à un sous-espace de dimension variable au choix. Cette "matrice de transformation" câblée sera obtenue après apprentissage sur d'importants corpus de données. Nous verrons, en partie C, tout l'intérêt que peut présenter un tel gain d'information pour les applications en programmation dynamique.

PARTIE C

RECONNAISSANCE DE LA PAROLE

AU NIVEAU ACOUSTIQUE

INTRODUCTION A LA PARTIE C.

Les notions développées dans les parties A et B nous permettent d'aborder maintenant la reconnaissance automatique de la parole. Nous nous limiterons, dans cette partie, au seul niveau acoustique ; l'utilisation de contraintes plus évoluées sera traitée en partie D. Nous avons vu (partie A chapitre III) que la majeure partie des expériences de reconnaissance vocale se sont limitées jusqu'à présent au niveau acoustique. Cela ne constitue à nos yeux qu'une première étape, certes nécessaire, mais insuffisante.

Après avoir décrit l'organisation générale du système et les conditions expérimentales, nous parlerons des expériences de reconnaissance globale de mots, et analytique de segments, effectuées, autant que possible, en temps réel.

L'examen des deux types d'approches du problème permettra enfin d'en faire une étude comparative.

CHAPITRE I

GENERALITES.

I-1 Introduction.

Ce chapitre est destiné à situer nos expériences (implantation du système, conditions expérimentales) et à préciser un certain nombre d'options prises : programmes conversationnels, temps réel, niveaux de reconnaissance. Auparavant, nous esquisserons rapidement l'évolution des idées mises en pratique depuis nos premières expériences de reconnaissance, vers 1968.

I-2 Evolution des travaux.

Nos recherches ont débuté par la simulation d'un appareil de reconnaissance de phonèmes, à matrice d'apprentissage $[B-29]/[C-1]$. Décrite, dans sa forme initiale, par K. STEINBUCH $[C-2]$, la matrice que nous avons utilisée possède des connexions à poids variables, que l'on ajuste au cours d'un apprentissage préalable.

La simulation sur ordinateurs IBM 1130 puis CH 10070 a permis d'effectuer une étude systématique des différents paramètres de réglage, et d'obtenir les poids optimaux de la matrice $[C-3]$. La réalisation électronique de ce système $[B-31]$ a ensuite vérifié le bien-fondé des hypothèses de départ, mais a aussi montré les limites de cette méthode de reconnaissance.

La trop grande rigidité de la matrice, permettant difficilement de traiter les variations qui apparaissent, même pour un locuteur unique, nous amenait alors à rechercher des méthodes de reconnaissance moins déterministes et plus souples, donc mieux adaptées au signal vocal. C'est ainsi que les méthodes de comparaison dynamique furent mises au point et appliquées à la reconnaissance de mots isolés (cf. chapitre II), puis à la reconnaissance de segments (cf. chapitre III).

Simultanément, l'utilisation de contraintes évoluées (linguistiques, syntaxiques) était envisagée (cf. partie D).

I-3 Description générale du système.

I-3-1 Implantation :

Nos systèmes de reconnaissance, de mots ou de segments, sont implantés sur l'ordinateur T 2000 (Télé-mécanique Electrique) du laboratoire. Ce minicalculateur possède une mémoire centrale de 8 K mots de 19 bits (qui va être portée à 12 K mots), un lecteur-perforateur rapide de ruban et un télétype. L'analyseur-codeur spectral, décrit en B, est relié directement par coupleur universel à l'ordinateur. De plus, les sonagrammes fournis par l'analyse sont visualisables instantanément sur oscilloscope cathodique, en attendant l'installation prochaine d'une console de visualisation graphique. Le schéma général de l'installation est celui de la figure I-1. Tous les programmes considérés sont écrits en langage d'assembleur ASMAT.

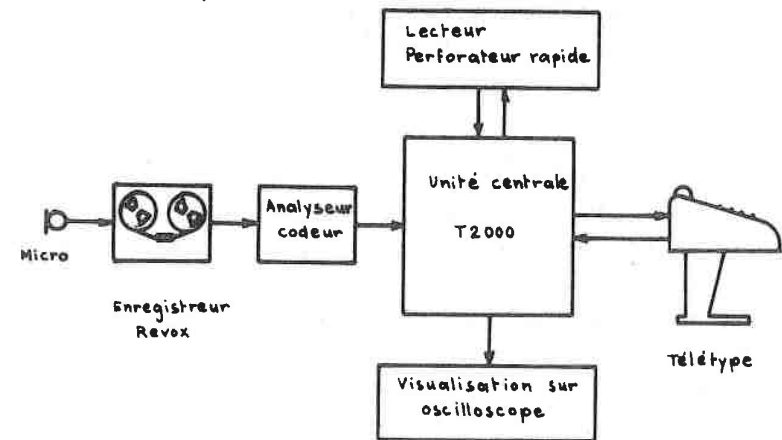


Schéma général de l'installation

Figure I-1.

I-3-2 Le choix du temps réel :

Comme nous l'avons déjà dit, toutes nos expériences de reconnaissance sont effectuées en ligne et, autant que possible, en temps réel. Cette façon de procéder possède de nombreux avantages :

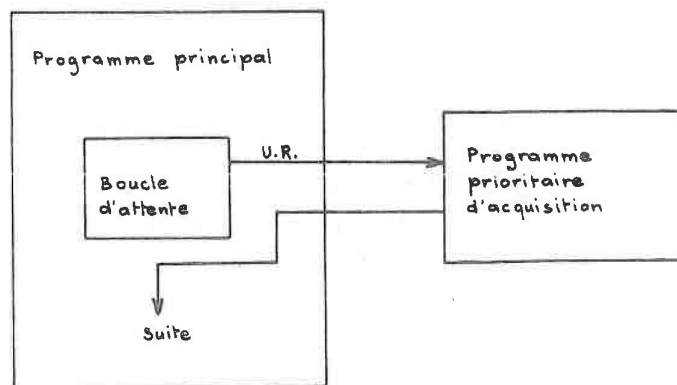
- c'est le type de fonctionnement qui, en pratique, conviendra à la plupart des applications industrielles (cf. partie A chapitre III), même si peu de systèmes ont opéré jusqu'à présent dans ces conditions,

- cela permet une très grande souplesse d'expérimentation (différents types de voix, différents locuteurs, différents réglages de paramètres) et une multiplicité des essais qui est une garantie de la validité des pourcentages de reconnaissance obtenus,

- les mots, ou phrases, prononcés peuvent être visualisés instantanément, sous forme de sonagrammes binaires, soit sur télétype, soit sur écran cathodique. Il est ainsi très facile de suivre le travail du système. De plus, cette visualisation facilite grandement la phase d'apprentissage, en conjonction avec un programme de service conversationnel (cf. chapitre II).

Mais le fonctionnement en temps réel pose aussi un certain nombre de problèmes concernant en particulier :

- l'acquisition des données spectrales d'analyse ; ces données, arrivant pas "bouffées" chaque 10 millisecondes, doivent être prises en compte en priorité, ce qui nécessite une interruption des calculs en unité centrale. Cette acquisition se fait au niveau de la programmation suivant l'organigramme schématique de la figure I-2.



Acquisition de données en interruption prioritaire.
Figure I-2

- l'optimisation des programmes ; l'écriture en langage symbolique facilite ce travail, indispensable si l'on veut vraiment travailler en temps réel.

- l'interaction locuteur-machine ; cet aspect important a déjà été noté par d'autres utilisateurs de systèmes temps réel [A-111]. Pour de tels systèmes, l'apprentissage apparaît comme une adaptation mutuelle de l'homme et de la machine. En effet, l'apprentissage de la machine a lieu au cours d'une phase préalable pendant laquelle le locuteur prononce tout ou partie du vocabulaire utilisé, tandis que "l'apprentissage", ou plutôt l'adaptation du locuteur à la machine, se fait progressivement, voire inconsciemment, selon les résultats que celui-ci obtient. Ce phénomène peut éventuellement fausser les performances d'un système, même s'il est normal et banal dans la conversation entre êtres humains, comme nous l'avons déjà fait remarquer (partie A chapitre III), et il est en fait très difficile de chiffrer objectivement un fait aussi... subjectif.

- la dérive de la voix humaine ; si le locuteur a tendance à améliorer ses performances lorsqu'il tente de se faire comprendre par un système temps réel, il est au contraire parfois trahi par la dérive de sa voix. Ce phénomène progressif consiste en une variation des caractéristiques de la voix - due à la fatigue, à l'accoutumance, etc... - provoquant une dégradation des performances de reconnaissance. Nous le mettrons en évidence dans le cadre de notre système de reconnaissance globale de mots (chapitre II). Cela est plus ou moins marqué selon les individus, et il faut alors procéder à un nouvel apprentissage de la machine. Cette solution n'est bien sûr qu'un pis-aller et nous avons déjà dit que l'idéal est un autoapprentissage du système au cours de la conversation

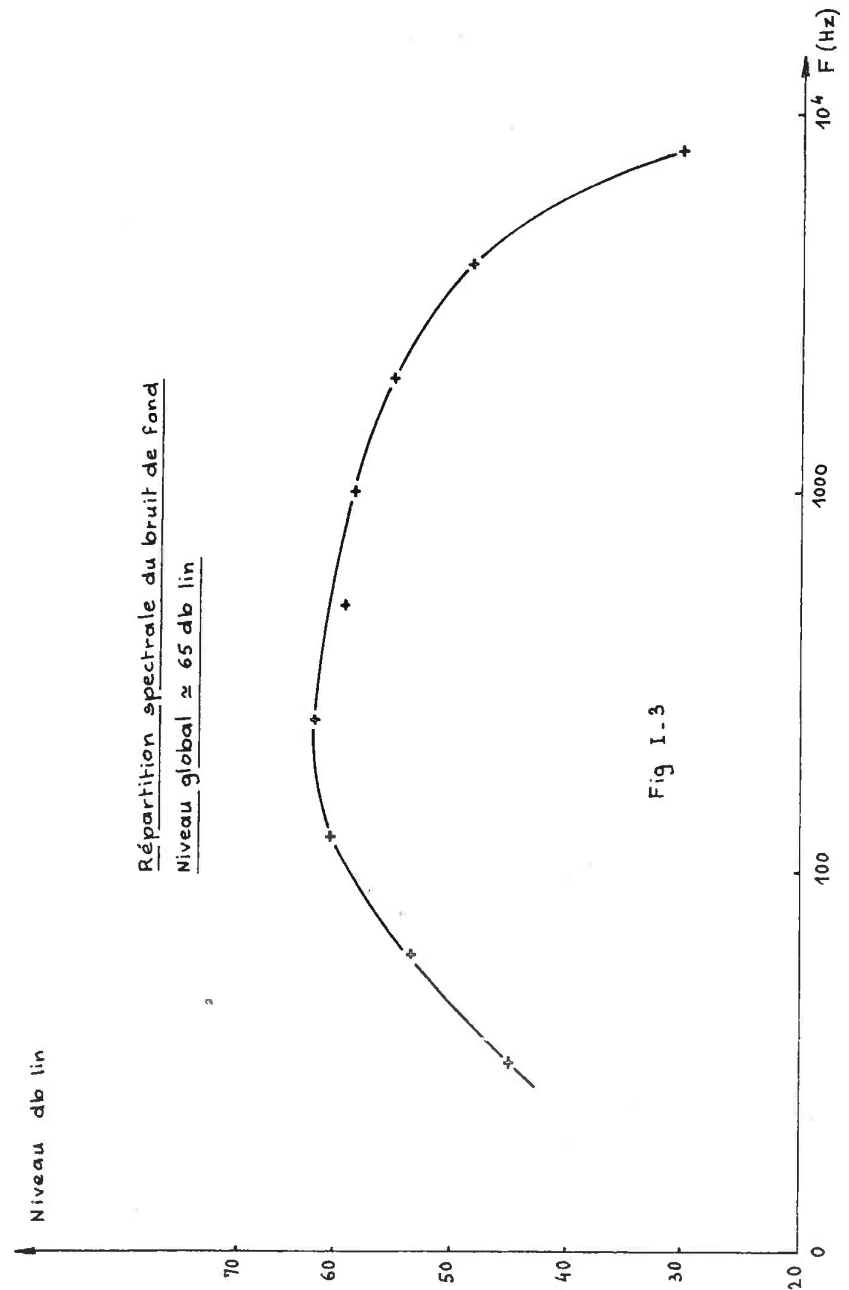
avec un locuteur. Cela suppose une certaine initiative de la machine, capable de s'autoévaluer et d'en tirer une conclusion. Le passage aux niveaux hiérarchiques supérieurs, peut-être jusqu'à la sémantique de la phrase, semble nécessaire. Ceci constitue un des aspects sur lesquels repose l'avenir des systèmes de compréhension de la parole, et il fait partie de nos préoccupations à long-terme.

I-3-3 Les conditions expérimentales :

Les expériences de reconnaissance ont lieu dans un local non insonorisé où se trouve l'ordinateur et ses périphériques. Le niveau global de bruit y est d'environ 65 décibels ; la répartition spectrale de ce bruit, par octave, est donnée sur la figure I-3. On voit que le maximum d'intensité du bruit se situe dans la zone du premier formant de la parole (500 - 1 000 Hz environ).

Le locuteur est assis devant le télécype, de façon à pouvoir utiliser le programme conversationnel. Aucun conseil de prononciation n'est donné, si ce n'est celui de parler distinctement. Les mots ou phrases prononcés sont enregistrés, à travers un microphone dynamique directionnel SHURE, sur un magnétophone REVOX. Cet appareil permet d' "archiver" éventuellement les expériences effectuées et il sert simultanément de préamplificateur pour l'analyseur-codeur (cf. figure I-4). Le rapport parole/bruit est de l'ordre de 20 db, et il existe une certaine réverbération du signal vocal.

Précisons qu'aucune régulation du niveau de parole n'est effectuée. Cela ne s'impose pas, étant donné le caractère relatif de notre analyse spectrale (cf. partie B chapitre II).



I-4 Choix du niveau de reconnaissance.

I-4-1 Système de reconnaissance et modèle de perception :

Un des aspects de la reconnaissance vocale est l'approfondissement des mécanismes humains de perception. En ce sens, un système de reconnaissance, aussi rudimentaire soit-il, est un modèle partiel de perception, permettant peut-être de faire évoluer les connaissances dans ce domaine. Fournir des pourcentages de reconnaissance apparaît alors comme un moyen de tester la validité du modèle. Un autre moyen consiste à diversifier les conditions d'utilisation du modèle, de façon à augmenter le champ d'investigation. C'est une des raisons qui ont motivé le choix de nos niveaux de reconnaissance.

I-4-2 Reconnaissance globale et reconnaissance analytique :

Nos expériences porteront ainsi à la fois sur la reconnaissance globale de mots isolés et sur la reconnaissance analytique de segments élémentaires, à partir desquels on peut déduire des mots. Notre démarche a été inverse de celle généralement adoptée, puisque nous avons étudié la reconnaissance analytique (expériences par matrice d'apprentissage) avant la reconnaissance globale.

Il a paru nécessaire d'aborder ces deux aspects du problème et cela permettra en particulier d'en effectuer une comparaison valable.

CHAPITRE II

EXPERIENCES DE RECONNAISSANCE

GLOBALE DE MOTS

II-1 Introduction.

Les problèmes de normalisations, en fréquence et en temps, sont essentiels en reconnaissance de mots et nous avons pour notre part développé une méthode de comparaison-normalisation temporelle par programmation dynamique dont les performances sont intéressantes.

Après avoir rappelé les problèmes de normalisations en général et décrit notre méthode de normalisation temporelle, nous donnerons quelques résultats obtenus par le système de reconnaissance de mots. L'influence de divers facteurs (nombre de locuteurs, taille du vocabulaire, apprentissage, dérive de la voix, ...) sera aussi étudiée ; nous terminerons sur quelques perspectives d'avenir des méthodes de reconnaissance globale, appelées rapidement à des applications commerciales, dans le cadre de machines à reconnaître autonomes, dont la structure reste à définir.

II-2 Les problèmes de normalisation de la parole.

II-2-1 Normalisation fréquentielle :

L'étude des sonagrammes de divers locuteurs fait apparaître des différences importantes dans la répartition spectrale de l'énergie émise. Ces différences influent très

peu sur l'intelligibilité du message transmis, mais elles dégradent au contraire les performances d'un système de reconnaissance. La normalisation fréquentielle conditionne ainsi la possibilité d'adaptation d'un système à un grand nombre de locuteurs, mais cela est loin d'être fait. En effet, le problème est très complexe car les différences que l'on peut constater entre le spectre d'un mot prononcé par un homme, dont la fréquence fondamentale est de l'ordre de 150 Hz par exemple, et le spectre du même mot prononcé par une femme, dont la fréquence fondamentale vaut environ 300 Hz, ne se réduisent pas à une simple translation du spectre en bloc. D'autres différences apparaissent; par exemple, les zones formantiques du locuteur féminin sont plus floues que celles du locuteur masculin.

Seules des solutions partielles ont été proposées, qui reviennent à effectuer une normalisation au niveau de paramètres ou directement au niveau du spectre, par calcul d'un coefficient de translation du spectre (anamorphose fréquentielle régulière [C-4]). Mais le problème n'est pas résolu, et c'est pourquoi aucun système actuel de reconnaissance vocal ne fonctionne pour n'importe quel locuteur.

II-2-2 Normalisation temporelle :

Les variations qui affectent le signal de parole sont aussi importantes dans le domaine temporel que dans le domaine fréquentiel. Il est ainsi très difficile de prononcer plusieurs fois le même mot avec la même durée, et sur le même rythme. Ces différences sont dues :

- à des variations de la vitesse d'élocution qui produisent une simple transformation linéaire de l'échelle des temps, et

- à des variations de rythme, qui produisent des distorsions non linéaires de cette même échelle des temps.

La reconnaissance globale de mots doit forcément être précédée d'une réduction de ces variations, c'est-à-dire d'une normalisation temporelle des mots prononcés, permettant de comparer entre elles des formes de tailles différentes.

Divers types de normalisations linéaires ont été proposés [A-39], consistant par exemple à raccourcir ou à allonger un mot prononcé, pour lui donner une durée standardisée. Cela peut être obtenu en considérant les zones stables des mots (voyelles par exemple), mais la méthode ne permet pas de traiter les distorsions dues aux changements de rythme.

Une normalisation temporelle idéale doit permettre de faire correspondre les parties semblables des mots à comparer, en compensant automatiquement les variations de durée, de façon à obtenir le taux de similitude optimal entre les deux mots. C'est précisément ce qu'effectue une méthode introduite récemment, dérivée des principes de la programmation dynamique, et que nous allons exposer maintenant.

II-3 Normalisation temporelle par comparaison dynamique.

II-3-1 Principe :

La programmation dynamique est une technique mathématique bien connue qui a été utilisée dans des domaines scientifiques très variés. Fondée sur le principe d'optimalité [C-5],

elle permet, dans le cas présent, de trouver la similitude optimale entre deux mots, avec le minimum de calcul, en assurant une normalisation temporelle non linéaire au cours de la comparaison [C-6].

Le principe a été proposé pour la première fois par Slucker [C-7] et appliqué depuis à la reconnaissance vocale, avec différents algorithmes [C-8] [C-9] [A-86]. Des méthodes voisines ont également été utilisées [C-10].

II-3-2 Algorithme :

a) Formulation du problème.

Appelons

$$X = \{x_1, x_2, \dots, x_N\}$$

une forme sonore échantillonnée dans le temps (correspondant par exemple à la prononciation d'un mot isolé), composée de N échantillons qui sont des vecteurs à 24 composantes binaires données par l'analyse spectrale :

$$x_i = \{x_i^\ell\}, \ell = 1, 2, \dots, 24$$

On dispose par ailleurs de R formes de référence qui sont stockées en mémoire :

$$Y_r = \{y_1, y_2, \dots, y_{M_r}\}, r = 1, 2, \dots, R$$

dont les longueurs, c'est-à-dire les nombres d'échantillons M_r qui les composent, sont variables.

Les vecteurs y sont également à 24 composantes binaires :

$$y_j = \{y_j^\ell\}, \ell = 1, 2, \dots, 24$$

Il s'agit de classer la forme X dans une des R classes de référence, en calculant un taux de similitude entre X et les différentes formes Y_r , en s'affranchissant des différences de longueur entre ces formes. Cette similitude s'entend au sens d'une certaine métrique, qui est dans notre cas la distance $L1$, ou distance de Hamming :

$$d(x_i, y_j) = d(i, j) = \sum_{\ell=1}^{24} |x_i^\ell - y_j^\ell|$$

En fait, l'algorithme calcule un "taux de dissemblance" entre deux formes, de sorte que l'opération de comparaison apparait comme un processus dynamique d'allocation bidimensionnel, au cours duquel il s'agit de minimiser le taux de dissemblance T , ce qui revient strictement au même.

b) Relation de récurrence.

Soit $s(i, j)$ le taux de dissemblance intermédiaire, correspondant à la comparaison de i premiers vecteurs de X et des j premiers vecteurs de Y_r .

On a :

$$s(1, 1) = d(1, 1) \quad (1)$$

condition initiale qui signifie que l'on commence par comparer le premier échantillon de chacune des formes.

Avec cette condition initiale (1) et la relation de récurrence f :

$$s(i, j) = f(i, j, i-1, j-1, \dots) \quad (2)$$

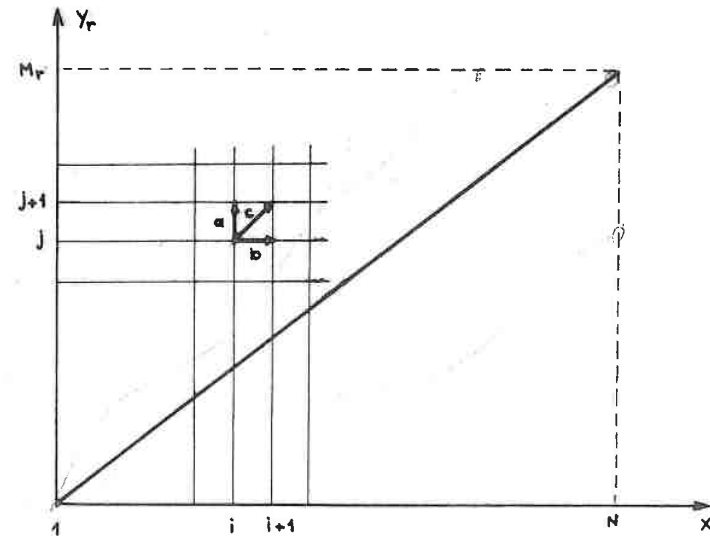
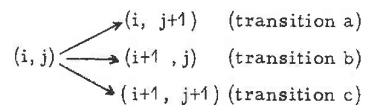
on obtient la dissemblance optimale T entre les formes X et Y_r lorsque l'on aboutit au dernier échantillon d'une des formes, ou des deux simultanément, soit :

$$T = \begin{cases} \frac{1}{K} s(N, j) \\ \text{ou} \\ \frac{1}{K} s(i, M_r) \\ \text{ou} \\ \frac{1}{K} s(N, M_r) \end{cases} \quad (3)$$

K représente, pour chaque cas, le nombre total de processus d'allocation successifs au cours de la comparaison.

L'efficacité de la méthode dépend beaucoup de la relation de récurrence f. Nous en avons testé plusieurs et il semble que l'on atteigne des performances du même ordre, indépendamment de l'algorithme, si celui-ci est valable. Il s'agit alors simplement d'une question de vitesse de calcul ; la relation de récurrence que nous allons exposer est particulièrement intéressante à ce point de vue.

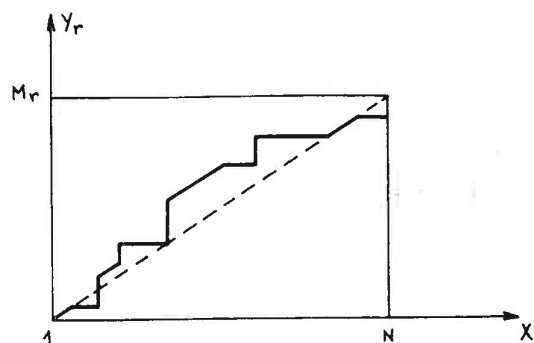
Lorsque l'on se trouve au point (i, j) dans le "plan de comparaison" $X - Y_r$, trois transitions sont autorisées (figure II-1) :



Transitions permises par l'algorithme (CEC)

Figure II-1

Ces transitions déterminent un chemin optimal de comparaison qui s'écarte en général assez peu de la diagonale du rectangle $X - Y_r$, comme sur l'exemple typique de la figure II-2.



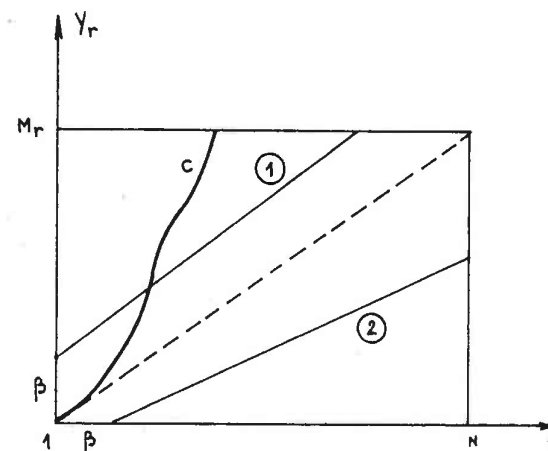
Exemple typique de chemin de comparaison

Figure II-2

Cependant ce chemin peut parfois diverger et fournir un taux de dissemblance T inexact (un tel cas est symbolisé sur la figure II-3). Pour éviter de telles erreurs, nous introduisons une "fenêtre temporelle" F , qui délimite une zone de part et d'autre de la diagonale, de laquelle le chemin ne peut pas sortir (figure II-3). Pour plus de souplesse, la fenêtre F augmente linéairement avec le rang i (ou j) :

$$F = \alpha i + \beta$$

Cette fenêtre permet par ailleurs de limiter le nombre de comparaisons effectuées, donc le temps de calcul (cf. II-5-5).



C est un exemple schématique de chemin divergent. Les droites 1 et 2 délimitent le domaine de comparaison (effet de fenêtre).

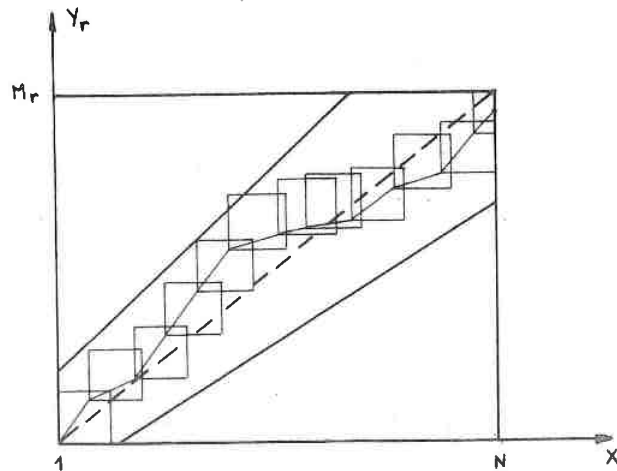
Figure II-3.

c) Généralisation de l'algorithme.

Les trois transitions qui constituent la relation de récurrence, forment ce que nous appelons une cellule élémentaire de comparaison (CEC). Guidés par le souci d'assouplir l'algorithme de comparaison dynamique et de minimiser le temps de calcul, nous avons été amenés à associer plusieurs cellules CEC. Des expériences préliminaires ont permis de fixer à 4 le nombre de CEC à associer.

En dessous de ce chiffre, les performances obtenues sont moins bonnes. Au-delà, elles restent stationnaires, mais le temps de calcul augmente très vite car on effectue de plus en plus de comparaisons inutiles.

On obtient alors un chemin de comparaison plus global, mais qui donne finalement un taux de dissemblance T optimal (figure II-4).



Exemple de chemin global avec association de 4 CEC.

Figure II-4

II-4 Organisation du système de reconnaissance.

II-4-1 Programme de service conversationnel :

Le fonctionnement en temps réel nécessite un dialogue avec l'ordinateur, de façon à pouvoir utiliser au maximum les possibilités offertes, et à contrôler à tout moment le déroulement des programmes. Un programme de service, sorte de superviseur, est utilisé pour cela (un programme analogue sert également en reconnaissance analytique, cf. chapitre III).

Ce programme comprend :

- des sous-programmes de service :

- * d'acquisition des données binaires correspondant à la prononciation d'un mot ;
- * d'écriture des résultats de la reconnaissance ;
- * de visualisation sur télétype du sonagramme binaire du mot prononcé ;
- * de visualisation sur oscilloscope cathodique de ce même sonagramme. Cette visualisation peut avoir une durée réglable à volonté et elle peut concerner un mot tout entier, ou une partie de mot à préciser. Elle possède l'avantage important, par rapport à la visualisation sur télétype, d'être instantanée. C'est pourquoi elle est très utile, en particulier en phase d'apprentissage d'un vocabulaire, car elle permet de ne conserver en mémoire que des mots correctement prononcés.

- une fonction conversationnelle, qui permet de sélectionner, à partir du clavier du télétype, une des fonctions suivantes :

- * N: demande de la taille du vocabulaire en mémoire (nombre de mots de référence), avec réponse sur télétype ;
- * W: écriture sur télétype des noms des mots de référence, avec leur durée en nombre d'échantillons;
- * I: demande d'entrée d'information. Après une temporisation d'environ 1s, l'ordinateur se met en attente pour acquérir les données correspondant à un mot qui sera placé en référence (rappelons que la fin d'un mot est détectée automatiquement après 0,3s de "silence") ;
- * L: demande de la longueur (en nombre d'échantillons, c'est-à-dire en 0,01s) du mot qui vient d'être acquis, avec réponse sur télétype ;
- * D: visualisation sur télétype du sonagramme du mot qui vient d'être acquis (en entrée d'information I, ou en reconnaissance R) ;
- * V: visualisation sur oscilloscope de ce même sonagramme, avec indication de la durée de visualisation désirée et de la partie de mot à visualiser ;
- * B: validation du mot qui vient d'être acquis (après, par exemple, une visualisation de contrôle) et indication, au clavier, du nom de ce mot ;
- * F: rejet du mot qui vient d'être prononcé. L'ordinateur se met à nouveau en attente d'information, après une temporisation ;
- * F xx: visualisation du mot de référence n° xx qui avait été précédemment stocké en mémoire ;
- * C xx: même fonction que la précédente mais sur oscilloscope cathodique ;

- * R: appel du programme de reconnaissance.

L'ordinateur se met en attente d'information, après temporisation, puis donne le résultat sur télétype dès la détection de la fin du mot prononcé. L'ordinateur se remet ensuite en attente d'information, sauf spécification contraire.

Ces diverses fonctions rendent le système extrêmement souple et facilitent beaucoup son utilisation en temps réel.

II-4-2 Le programme de reconnaissance :

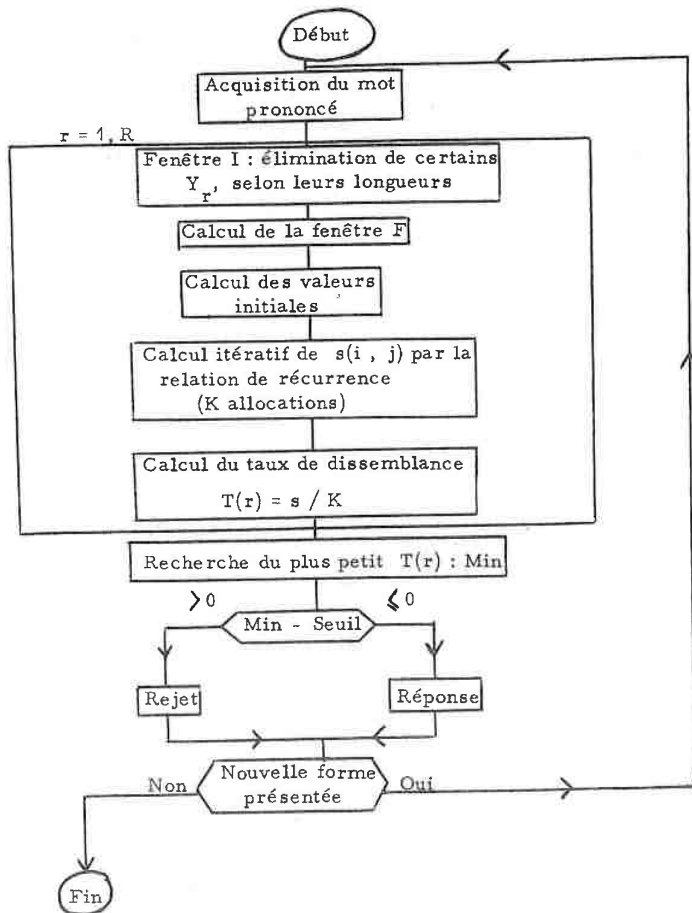
Fondé sur l'algorithme de comparaison précédemment décrit en II-3, ce programme fonctionne schématiquement comme l'indique la figure II-5.

Il est écrit en langage d'assembleur ASMAT et comporte 200 instructions environ.

Une fenêtre I permet de ne retenir que les mots de référence ne différant pas trop en longueur du mot présenté, ce qui constitue un nouveau gain de temps. Nous ne conservons en général que les mots de référence Y_r tels que :

$$\frac{N}{2} \leq M_r \leq 2N$$

De plus, le système ne donne une réponse que si le taux de dissemblance correspondant au mot reconnu (M_{in}) est inférieur à un certain seuil. Dans le cas contraire (correspondant à l'existence d'un doute), le mot prononcé est rejeté et le système se met en attente du mot suivant.



Fonctionnement schématique du programme de reconnaissance globale

Figure II-5.

II-4-3 Organisation du lexique :

Théoriquement, la reconnaissance d'un mot prononcé nécessite la comparaison de ce dernier à tous les mots de référence qui constituent le vocabulaire et qui sont stockés dans un lexique.

En pratique, il est intéressant de restreindre le nombre de comparaisons, pour augmenter la vitesse du processus de reconnaissance et diminuer les erreurs. La fenêtre I, mentionnée au paragraphe précédent, joue déjà ce rôle. Nous avons, de plus, cherché à classer les mots en grands types facilement différenciables :

- mots contenant une, plusieurs occlusions,
- mots contenant un, plusieurs segments fricatifs.

Un mot prononcé est classé dans un type, qui est déterminé au cours de l'acquisition de ce mot, à partir de l'analyse spectrale. Le lexique est organisé de telle façon que l'on n'examine que les mots de référence du même type. Les performances du système sont ainsi nettement améliorées, encore que des erreurs peuvent s'introduire au niveau de la détermination du type. Ce cas, cependant très rare, est dû au principe de l'analyseur-codeur utilisé, dont nous avons déjà mentionné les faiblesses (cf. partie B chapitre II).

II-5 Résultats expérimentaux.

Les résultats concernant les performances d'un système de reconnaissance vocale doivent toujours être considérés avec circonspection. Nous avons déjà fait remarquer que le fait de travailler en temps réel et donc de multiplier les expériences était un garant de la validité statistique des résultats obtenus, malgré le phénomène, également signalé, de l'adaptation du locuteur.

Il n'en reste pas moins que ces résultats sont "instantanés", en ce sens qu'ils caractérisent l'état actuel du système de reconnaissance et qu'ils ne doivent donc pas être considérés comme définitifs.

II-5-1 Performances pour différents vocabulaires :

Ces résultats dépendent également beaucoup du vocabulaire utilisé (mots de longueur voisine, phonétiquement semblables, etc...) ; c'est pourquoi nous avons effectué nos expériences de reconnaissance globale avec trois vocabulaires différents. Ces vocabulaires ont été choisis en vue des applications possibles de la reconnaissance vocale et non pas avec le souci de faciliter la tâche du système. Nous les désignerons par :

- V1 : vocabulaire des 26 mots-clés utilisés dans les Télécommunications pour identifier les 26 lettres de l'alphabet ;
- V2 : vocabulaire de 25 noms de villes françaises ;
- V3 : vocabulaire de 20 mots anglais, empruntés à l'expérience de reconnaissance de Vicens/B-15]. Il nous a paru intéressant de tester notre système avec des mots prononcés avec un accent étranger, pour augmenter la variabilité au cours de la prononciation.

Le tableau II-1 donne la liste de ces trois vocabulaires.

<u>Vocabulaire V1</u>	<u>Vocabulaire V2</u>	<u>Vocabulaire V3</u>
ANATOLE	PARIS	SOMEONE
BERTHE	NANCY	EXPRESSION
CELESTIN	NICE	ANOTHER
DESIRE	TOULOUSE	APPEAR
EUGENE	AMIENS	BESIDE
FRANCOIS	CHARTRES	USUALLY
GASTON	AUXERRE	STRIKE
HENRI	ANGERS	PERFORM
IRMA	MARSEILLE	REALIZE
JOSEPH	AUTUN	SLIDE
KLEBER	BEAUVAIS	CHRISTMAS
LOUIS	EPINAL	BROKEN
MARCEL	MENTON	SUNLIGHT
NICOLAS	BEZIERS	EVENING
OSCAR	LYON	RESULT
PIERRE	BESANCON	PAPER
QUINTAL	AGEN	SPACE
RAOUL	SENS	TEACHER
SUZANNE	LEN	BOTTOM
THERESE	ARLES	JUNGLE
URSULE	METZ	
VICTOR	MONTLUCON	
WILLIAM	QUIMPER	
XAVIER	ROUEN	
YVONNE	LILLE	
ZOE		

Les vocabulaires utilisés.

Tableau II-1

Dans chaque expérience, tous les mots du vocabulaire sont prononcés une première fois et stockés en mémoire : ils constituent le lexique de référence. Ces mots sont ensuite prononcés par le même locuteur dix fois chacun, dans un ordre quelconque avec des arrêts, et à cadence raisonnable.

Les pourcentages de reconnaissance sont calculés à partir des résultats obtenus pour les dix prononciations.

Pour un locuteur masculin, les résultats obtenus sont les suivants :

- vocabulaire V1 : 99,7 % de reconnaissance
- vocabulaire V2 : 99 % de reconnaissance
- vocabulaire V3 : 95 % de reconnaissance

On constate une diminution de 0,7 % quand on passe de V1 à V2, dans les mêmes conditions : V1 est un vocabulaire bien adapté à la communication orale, choisi sur des critères d'intelligibilité. V2, au contraire, est composé de mots pris au hasard en parcourant une carte de France. Il s'agit là d'un exemple rudimentaire d'optimisation d'un vocabulaire, problème important en communication homme-machine.

Les résultats sont moins bons pour V3, ce qui traduit la variabilité de la prononciation lorsque l'on parle une langue étrangère.

Pour un locuteur féminin, les résultats obtenus, par exemple :

- vocabulaire V1 : 99 % de reconnaissance,
- sont tout à fait comparables.

II-5-2 Influence de la taille du vocabulaire :

Pour tester l'influence du nombre de mots contenus dans le vocabulaire sur les performances du système, nous avons regroupé les vocabulaires V1 + V2.

La figure II-6 donne la variation du pourcentage de reconnaissance (pour un locuteur), en fonction du nombre de mots N du vocabulaire. Ces pourcentages ont été obtenus de la façon exposée en II-5-1.

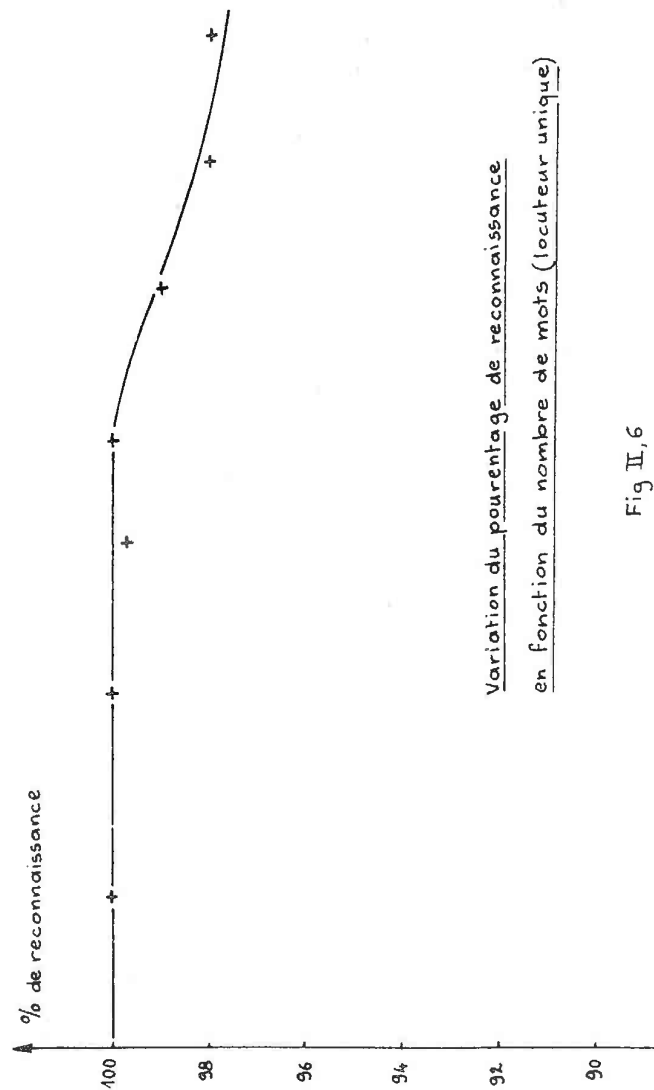
La courbe est limitée à $N = 40$ à cause de la saturation de la mémoire centrale actuelle du T 2000. Il sera intéressant de poursuivre cette expérience dans le cadre de l'extension prochaine du T 2000.

On remarque que la dégradation des performances, lorsque N croît, est relativement lente. L'approche globale de la reconnaissance vocale semble ainsi intéressante jusqu'à une valeur limite de N, sans doute de l'ordre de quelques centaines, au-delà de laquelle l'approche analytique prévaut.

II-5-3 Influence du nombre de locuteurs :

Ce point très important est lié à la notion de généralisation d'un système, telle qu'elle a déjà été définie.

Un locuteur masculin (1) a prononcé une fois un vocabulaire de 20 mots, sous-ensemble de V1, pour former le lexique de référence. Ensuite, ce même vocabulaire a été prononcé trois fois de suite par cinq locuteurs, sans aucune préparation. De plus, les cinq locuteurs possèdent des voix nettement différentes (accent, hauteur, ...). Les pourcentages



de reconnaissance cumulés sont donnés dans le tableau II-2 :

Locuteurs	Pourcentage de reconnaissance
(1)	100 %
(1) + (2)	88 %
(1) + (2) + (3)	79 %
(1) + (2) + (3) + (4)	75 %
(1) + (2) + (3) + (4) + (5)	71 %

Tableau II-2

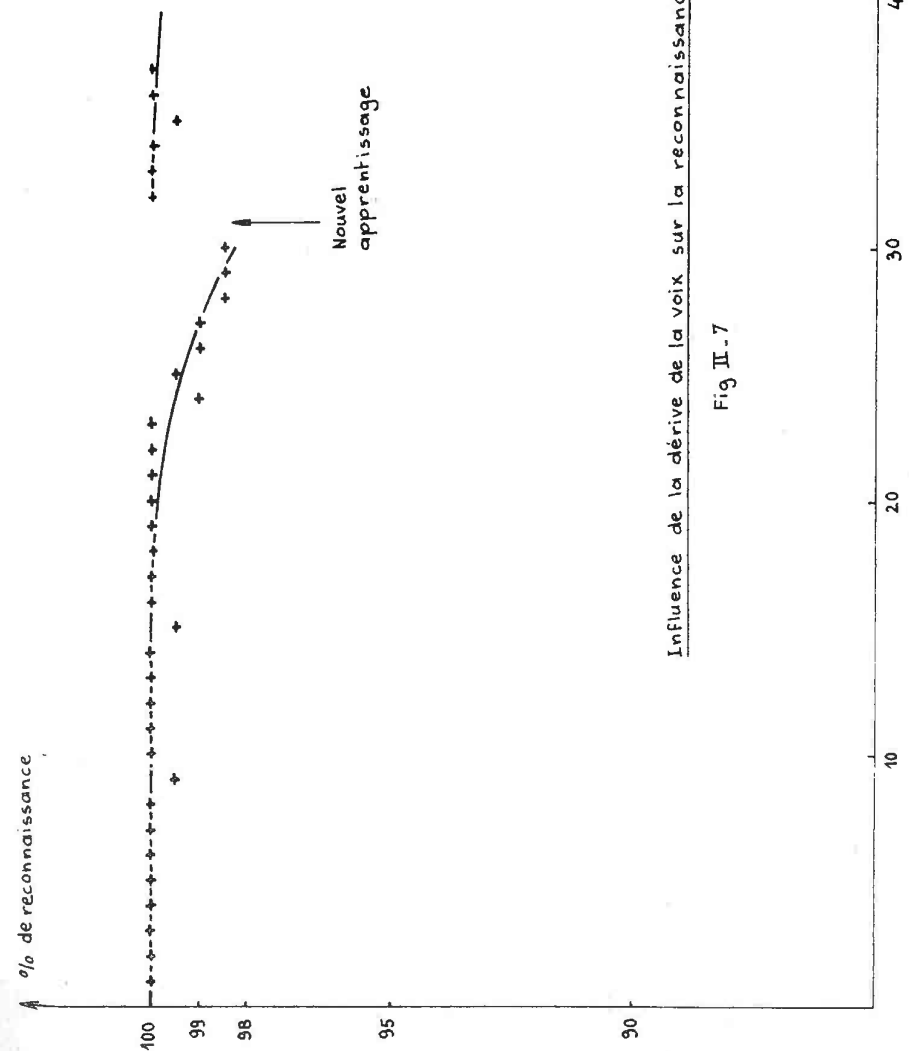
On constate une rapide détérioration des résultats lorsque le nombre de locuteurs augmente. Ces pourcentages correspondent au cas où aucune adaptation au locuteur n'est envisagée. Dans une autre expérience, nous avons procédé à une adaptation individuelle, en ajoutant au lexique les mots les plus mal reconnus pour chacun des locuteurs. La diminution du pourcentage de reconnaissance est alors beaucoup moins rapide. On voit tout l'intérêt d'une procédure plus complexe d'adaptation, incluse par exemple dans l'apprentissage du système.

Les expériences précédentes peuvent également servir à optimiser le vocabulaire utilisé, les mots souvent mal reconnus étant éliminés. Ceci constitue un aspect de la mise au point d'un langage oral de communication homme-machine.

Elles permettent enfin d'étudier le comportement individuel des locuteurs face à la machine : chaque locuteur a tendance à modifier sa prononciation pour améliorer ses performances. Ce phénomène est surtout net chez les personnes habituées aux systèmes de reconnaissance, et il est en fait bien difficile de préciser à quoi tient cette faculté d'adaptation.

II-5-4 Influence de la dérive de la voix :

Nous avons déjà parlé de ce phénomène, spécifique de la reconnaissance vocale en temps réel (cf. I-3-2). Il est mis en évidence ici avec un locuteur, et un vocabulaire de 20 mots, sous-ensemble de V2. La courbe de la figure II-7 donne la variation du pourcentage de reconnaissance sur ce vocabulaire, prononcé en entier à chaque fois, en fonction des prononciations successives effectuées à intervalle régulier. Après des fluctuations statistiques, ce pourcentage décroît incontestablement. A la suite de la 30^{ème} prononciation du vocabulaire (soit 2 heures plus tard), nous avons procédé à un nouvel apprentissage ; le pourcentage de reconnaissance correspondant à la 31^{ème} prononciation du vocabulaire a alors repris sa valeur initiale.



Influence de la dérive de la voix sur la reconnaissance

Fig II-7

II-5-5 Temps moyen de calcul :

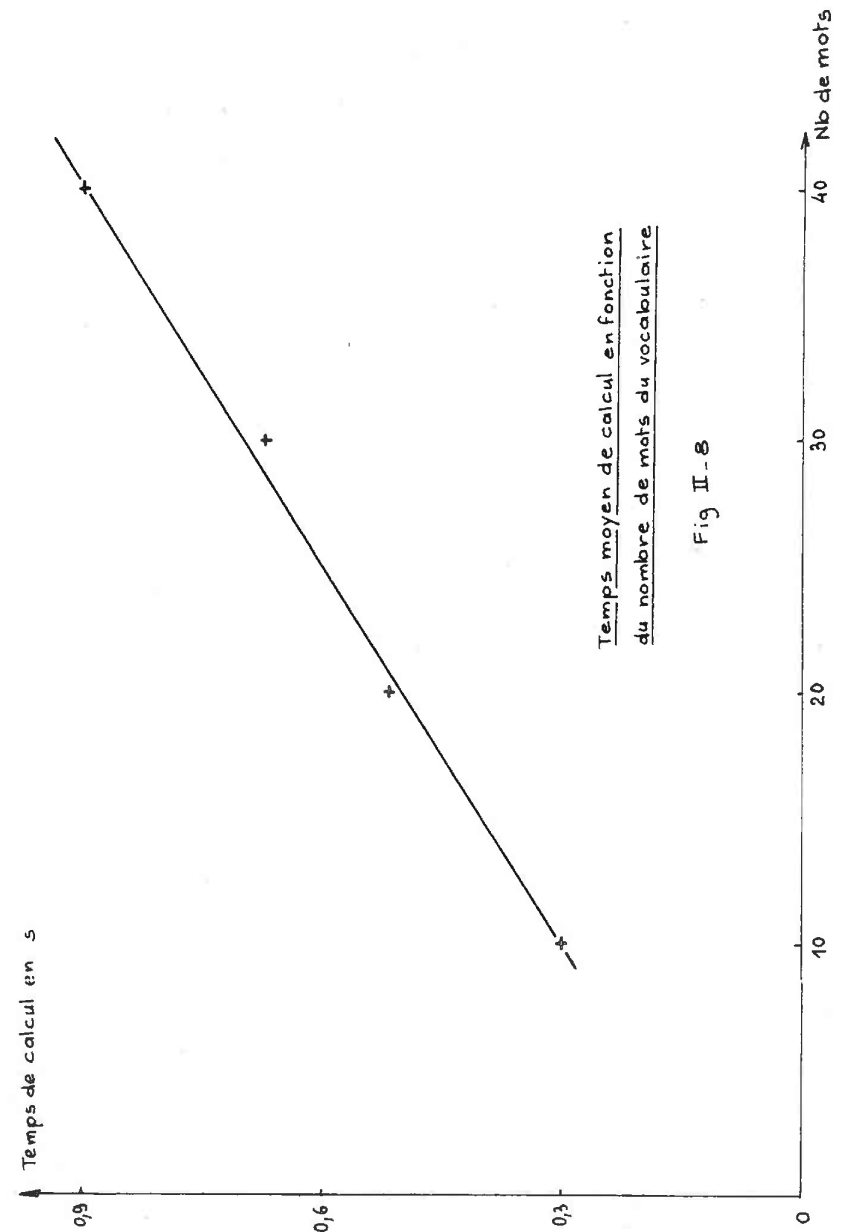
Il est très difficile de chiffrer l'intervalle de temps séparant la fin de l'énoncé d'un mot, du début de l'impression de la réponse sur télétype, que l'on peut considérer comme le temps de réponse du système.

De plus, ce temps de réponse n'est pas seulement le temps de calcul de reconnaissance car il comprend d'une part la détection de fin de mot (attente de 0,3 s après le dernier échantillon non vide) et d'autre part le temps de validation du télétype. En valeur moyenne, le temps de calcul est de l'ordre de 0,5 s pour un vocabulaire de 20 mots, et il augmente sensiblement avec le nombre de mots, comme l'indique la figure II-8.

Ces performances sont intéressantes, car le T 2000 n'est pas un calculateur très rapide (cycle de base : $1,3 \mu s$ en moyenne) et, de plus, l'algorithme de programmation dynamique nécessite énormément de bouclages. L'optimisation du temps de calcul a porté sur :

- l'optimisation des programmes,
- l'emploi des "fenêtres de comparaison" F et I,
- l'optimisation du calcul de la distance entre deux échantillons.

La distance entre deux échantillons est, rappelons-le, la distance de Hamming, puisque l'analyseur fournit des données binaires. Calculer la distance entre deux échantillons revient ainsi à effectuer leur disjonction et à compter les bits à 1 du mot binaire résultant. Nous avons accéléré ce calcul par un hardware spécialisé placé dans le bac coupleur et permettant de compter le nombre de bits à 1 de l'accumulateur,



à l'aide d'une simple instruction de décalage, ce qui divise par 15 environ le temps de calcul.

II-5-6 Le problème de l'apprentissage :

Dans les expériences précédentes, la phase d'apprentissage a été réduite au minimum, puisque les mots de référence composant le lexique correspondent simplement à une seule prononciation de chacun d'eux. Ceci est bien entendu insuffisant, même pour un seul locuteur, et un apprentissage plus élaboré améliorerait les performances du système. Une simple adaptation partielle au locuteur entraîne ainsi une nette amélioration des performances, comme on l'a constaté en II-5-3. Les mots de référence peuvent par exemple être obtenus en faisant la moyenne de plusieurs prononciations, et dans certains cas, on peut envisager de stocker plusieurs références du même mot. Cette dernière solution résoud partiellement le problème de l'adaptation à divers locuteurs, dans la mesure où chacune des références correspond à un type de voix.

II-6 Perspectives d'avenir et extensions possibles.

L'apprentissage est une des clés de la généralisation de notre système de reconnaissance de mots par comparaison dynamique.

Cette généralisation nécessite aussi une diminution du temps de réponse, si l'on veut continuer à travailler en temps réel. Nos travaux sont déjà orientés dans ce sens :

- par l'utilisation des méthodes de compression d'information, développées en partie B, qui permettent d'effectuer une

diminution notable de la dimension des formes à comparer, donc du temps de calcul,

- par la conception d'un processeur câblé de programmation dynamique. Ce processeur sera, dans un premier temps, contrôlé par l'ordinateur et effectuera, avec un gain de temps considérable, la comparaison dynamique entre deux formes.

Par la suite, on peut concevoir des machines à reconnaissance vocale autonomes, utilisant ces processeurs et dont la structure devra faire appel à des techniques nouvelles de traitement de l'information, telles que le calcul associatif, ou le calcul parallèle [C-19].

Enfin, la méthode décrite ici peut être appliquée à la reconnaissance des phonèmes (comme nous le verrons dans le chapitre suivant) et aussi au problème délicat de la reconnaissance de parole continue, plus particulièrement à la recherche de mots dans une phrase prononcée continûment.

II-7 Conclusion.

Le problème de la normalisation temporelle de la parole a trouvé une solution efficace par la méthode de comparaison dynamique. La reconnaissance globale de mots, en temps réel, fondée sur ce principe, donne de bons résultats surtout pour un locuteur unique, car le problème de l'adaptation à divers locuteurs reste à résoudre. Le taux de reconnaissance apparaît ainsi comme une fonction lentement décroissante du nombre de mots à reconnaître, et rapidement décroissante du nombre de locuteurs.

Les résultats obtenus permettent d'envisager une réalisation câblée d'un processeur de programmation dynamique et, par ailleurs, nous utiliserons cette méthode de comparaison pour la reconnaissance des segments, dans l'approche analytique exposée au chapitre suivant.

CHAPITRE III

SEGMENTATION ET RECONNAISSANCE ANALYTIQUE DE LA PAROLE

[C-11]

III-1 Introduction.

Ce chapitre est consacré à l'étude de l'approche analytique de la parole. Cette approche, plus générale, permet de traiter des vocabulaires pratiquement illimités ; de plus, c'est la seule utilisable en reconnaissance de parole continue. Mais les résultats obtenus sont moins spectaculaires que pour la reconnaissance globale, comme on le verra, car d'importants problèmes restent à résoudre, tels que la segmentation du signal vocal.

Après avoir rappelé brièvement les diverses possibilités que l'on a de segmenter la parole et décrit notre propre méthode de segmentation acoustique, nous parlerons du système de reconnaissance analytique que nous avons mis au point. Ce système est encore évolutif, l'objectif à moyen terme étant de le faire fonctionner en reconnaissance de phrases prononcées continûment.

Quelques exemples de résultats seront donnés et nous verrons enfin comment il est possible de passer de la chaîne phonémique obtenue par cette méthode, au mot effectivement prononcé (problème de la recherche lexicale d'un mot).

III-2 Segmentation du signal vocal.

III-2-1 Position du problème - Les segments :

La segmentation a pour but de fragmenter le signal vocal en segments, dont la nature dépend du principe de la segmentation et des critères utilisés. Ces segments peuvent être des sons élémentaires, constitutifs des phonèmes [C-13], des pseudo-phonèmes, des diphonèmes [A-127] ou des syllabes [A-46]. La méthode de segmentation la plus courante consiste à étudier la stabilité du spectre de la parole et d'en déduire la position des zones stables et des zones transitoires. Diverses méthodes ont aussi été employées : segmentation sur critères articulatoires [C-14], par transformée de Haar [B-17], par analyse temporelle zero-crossings [B-15], ou en utilisant une cochlée artificielle [A-114].

On peut noter aussi qu'un certain nombre de dispositifs ont été réalisés électroniquement, permettant une segmentation en temps réel [B-15], [C-18], [C-15].

En réalité, le problème, très récent, n'est encore qu'effleuré, peu de chercheurs l'ayant abordé dans les années passées [A-49], [C-16], [A-46], [C-17].

De toute façon, il est difficile de parler dans l'absolu de segmentation de la parole, car celle-ci ne peut être jugée que dans le cadre du système de reconnaissance où elle s'insère et qui, seul, lui donne un sens.

III-2-2 Principe de la segmentation utilisée :

La segmentation que nous utiliserons se fonde sur des critères purement acoustiques ; elle permet de distinguer des zones de stabilité relative de la parole (corps de voyelles, de fricatives, occlusions, etc...) et des zones de transition (explosion d'une consonne plosive, fin ou début de voyelle, etc...). Les segments obtenus sont parfois assimilables à des phonèmes,

mais cette règle n'est pas générale : une consonne plosive sera généralement constituée de deux segments (une occlusion et une explosion), tandis qu'une voyelle tenue pourra donner lieu à plusieurs segmentations.

Les critères de segmentation permettent en outre de donner aux segments une étiquette qui les range dans l'un des quatre grands groupes suivants :

- F : segment "type fricatif",
- P : segment "type occlusion de plosive",
- V : segment "type voyelle",
- N : segment "type nasal".

ce qui facilite la reconnaissance ultérieure.

En fait, cette préclassification doit être entendue au sens large : elle est suffisamment floue pour ne pas introduire d'erreurs qui seraient ensuite difficilement réparables. Par exemple, le type V contient aussi bien les voyelles que les semi-voyelles, les liquides, les nasales.

Pour assurer cette double fonction de segmentation et de préclassification, nous utilisons deux types de paramètres, déterminés chaque centième de seconde, à partir des résultats de l'analyse-codage binaire :

- trois paramètres principaux, reliés au nombre de "1" binaires dans les trois zones de fréquence 0 - 700 Hz, 700 Hz - 1 800 Hz, 1 800 Hz - 7 000 Hz,

- des paramètres secondaires propres à chacune de ces trois zones de fréquence, et permettant de rendre compte de la stabilité du segment.

III-2-3 Description du programme :

Le programme de segmentation comporte 280 instructions ASMAT. Conçu dans l'optique du fonctionnement en temps réel, il fonctionne pendant l'acquisition d'un mot, comme le programme de reconnaissance analytique (cf. paragraphe IV-3). La durée moyenne d'une segmentation est de l'ordre de 10^{-3} s. Chaque fois qu'un segment est détecté, il est rangé en mémoire, avec l'indication de son type et sa longueur.

La figure III-2 donne le principe de fonctionnement de ce programme.

III-2-4 Résultats obtenus :

Il est difficile et de peu d'intérêt d'estimer dans l'absolu les performances d'un tel programme de segmentation. En effet, ce programme est conçu en vue d'un traitement ultérieur de reconnaissance analytique, et c'est l'adaptation mutuelle de ces deux programmes qui importe surtout. La validité de la segmentation est ainsi révélée a posteriori, selon les résultats de reconnaissance.

De plus, comme les segments que nous avons définis ne correspondent pas exactement aux phonèmes, la comparaison de la segmentation obtenue à une segmentation "idéale" est très subjective.

Ceci dit, l'évaluation du programme doit porter sur deux points :

- la segmentation proprement dite, et la validité des segments obtenus. La multiplication des expériences nous a montré

que le nombre de segments est en moyenne de 1,5 à 2 fois le nombre de phonèmes déterminés a priori (les expériences concernent des mots bien articulés). Ce nombre important correspond à des segmentations multiples sur les phonèmes soutenus (voyelles, fricatives, ...), que le traitement ultérieur permettra d'éliminer. On constate aussi quelques omissions, assez rares, qui se situent presque exclusivement au niveau d'associations vocaliques voyelle-voyelle ou voyelle-liquide, etc... Ces erreurs sont bien plus graves que les précédentes, encore qu'elles seront atténuées en reconnaissance par l'utilisation d'un lexique donnant une description phonétique adéquate des mots du vocabulaire (cf. III-4). On touche ici à la limite des possibilités de l'analyse binaire utilisée, bien plus qu'à celle du programme de segmentation,

- la préclassification en types V, F, P, N. Cette opération s'effectue bien la plupart du temps, car la classification désirée est suffisamment large, comme nous l'avons déjà fait remarquer. Les seules erreurs constatées sont également dues à l'analyse : par exemple certains phonèmes /i/ seront classés dans le type F (segment fricatif) car leur premier formant n'apparaît pas, à cause de l'effet de seuil introduit par le codage binaire. De même, une information, même rudimentaire, sur l'état vibratoire des cordes vocales (pitch) permettrait d'affiner cette préclassification, avec la distinction segments sourds-segments sonores.

Ce sont de telles insuffisances qui nous ont, entre autres, amenées à concevoir un nouvel analyseur vocal, actuellement en cours de réalisation.

La figure III-2 donne un exemple de segmentation du mot "SAPIN", obtenu en temps réel et listé sur télétype.

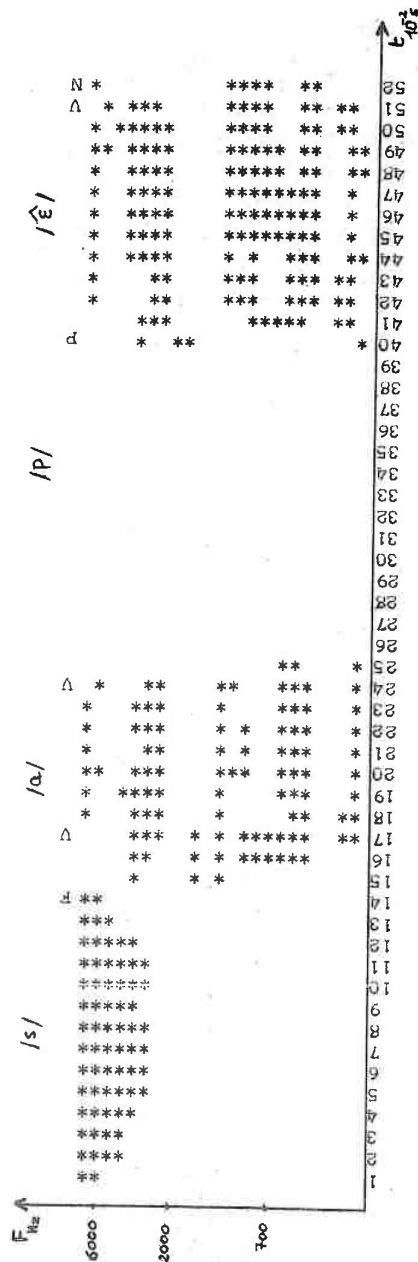


Figure III-2

Segmentation du mot "SAPIN"

III-3 Reconnaissance analytique de la parole.

III-3-1 Présentation :

La reconnaissance analytique est une méthode suffisamment générale pour ne pas introduire de restrictions importantes sur le vocabulaire utilisé. Dans notre cas, la seule nécessité est de prononcer distinctement les mots, de façon à ce que le programme de segmentation fonctionne correctement. Le système de reconnaissance analytique décrit ci-après fonctionne sur T 2000 en temps réel, pour des mots isolés. Il ne constitue en fait qu'une étape vers la conception d'un système plus évolué, permettant finalement le traitement de la parole continue par utilisation des niveaux linguistique et syntaxique (cf. partie D).

III-3-2 Le système de reconnaissance :

a) Organisation générale.

Conçu également pour le temps réel, ce système est organisé de façon analogue au système de reconnaissance globale précédemment décrit. Un programme de service conversationnel supervise l'ensemble et possède un certain nombre d'options permettant de sélectionner diverses fonctions : segmentation seule (S), segmentation-reconnaissance (I), visualisation sur télécype ou sur écran du mot prononcé avec indication de la segmentation obtenue (V), acquisition des segments de référence (L).

Pour permettre d'effectuer un traitement ultérieur de la chaîne phonémique (linguistique, syntaxique, etc...), la reconnaissance a lieu au cours de prononciation du mot, comme la segmentation. De cette façon, la chaîne phonémique représentant

un mot est disponible dès la fin de prononciation de ce mot.

Le schéma de principe de fonctionnement de ce système est donné figure III-3.

b) Le programme de reconnaissance.

Lorsqu'un segment a été détecté, il est immédiatement classé par le programme de reconnaissance, par comparaison à des segments-types, avant l'acquisition de l'échantillon de données suivant. L'intervalle de temps séparant deux échantillons étant de 10 ms, il reste donc environ 8 ms pour cette opération (en comptant 2 ms au plus pour la segmentation). Le programme de reconnaissance doit donc être à la fois rapide et suffisamment performant pour ne pas introduire d'erreurs graves que les traitements ultérieurs ne retrouveraient que difficilement.

Nous avons vu que le problème de normalisation temporelle était important en reconnaissance globale de mots ; il l'est encore plus en reconnaissance analytique car les variations de longueur constatées au niveau des segments peuvent être très importantes. Il est résolu ici par une méthode de comparaison dynamique semblable à celle que nous avons utilisée en reconnaissance globale de mots.

Il est possible d'accélérer le processus de reconnaissance en envisageant une comparaison de quelques critères, plutôt que la comparaison globale des spectres effectuée ici. Nous étudions ainsi actuellement un processus de reconnaissance fondé sur la détermination, à partir du spectre, de quelques critères caractéristiques des segments. Ces critères peuvent

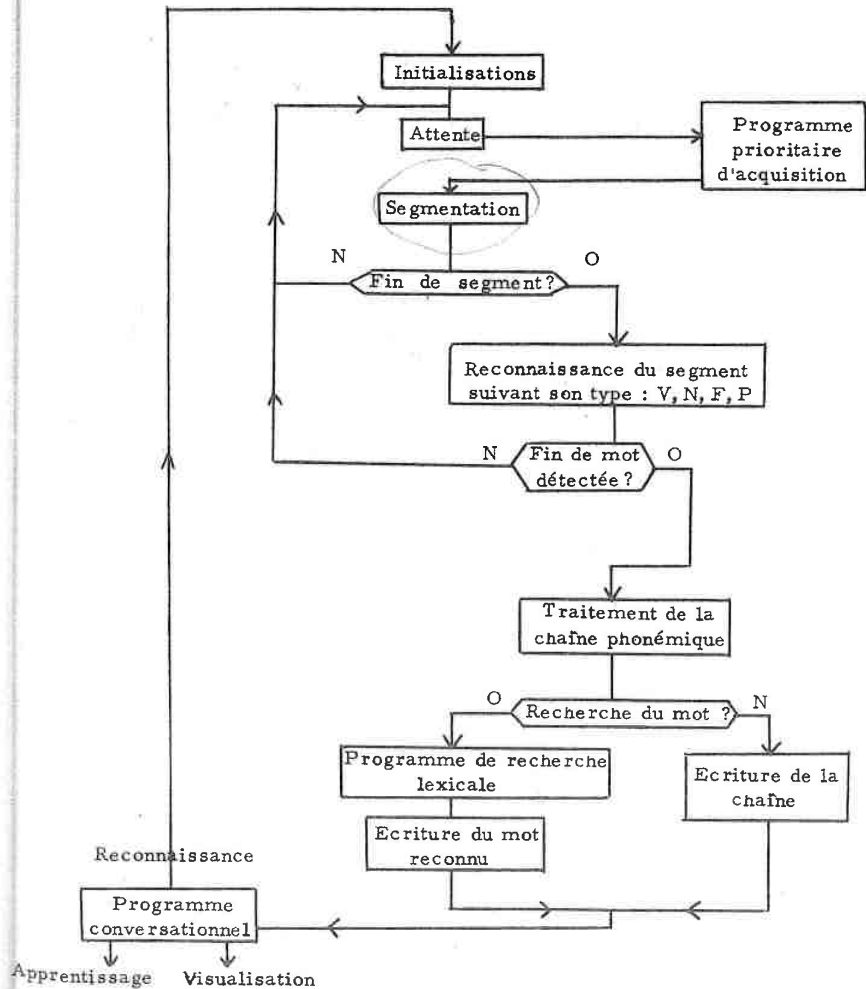


Figure III-3
Fonctionnement du système de reconnaissance analytique.

être aussi bien obtenus par des considérations physiques (voisement, fréquences des formants pour les segments vocaliques, nature du bruit pour les segments fricatifs, etc...), que mathématiques (sous-espace optimal (cf. partie B chapitre III)).

Pour l'instant, un segment inconnu est rangé dans la classe du segment de référence dont le spectre est le plus semblable. Ces segments de référence ont été obtenus dans une phase d'apprentissage préalable, au cours de laquelle on peut, par exemple, prendre la moyenne de plusieurs segments de chaque type. Cela résout en partie le problème de l'adaptation au locuteur, mais il est aussi possible d'envisager ici une normalisation fréquentielle par anamorphose.

c) Traitement de la chaîne phonémique.

Il n'est pas question ici de l'optimisation ultérieure de la chaîne aux niveaux linguistique ou syntaxique qui sera abordée en partie D, mais d'un simple "lissage", effectué à partir de critères très simples. Ce traitement se situe dès la fin du mot prononcé, et permet d'éliminer les erreurs les plus flagrantes de segmentation, ou de diminuer la redondance de la chaîne phonémique (cas des segmentations multiples de voyelles).

Les critères utilisés peuvent être purement physiques (ou "de bon sens"), par exemple :

- suppression des segments de longueur 0,01 s. Cette opération a lieu en fait en cours d'acquisition et non à la fin du mot ;

- suppression des segments de type P (occlusion) en début ou en fin de mot. La fin de mot est, de plus, détectée comme un segment P durant plus de 0,3 s (le test parole/non parole est donc analogue à celui utilisé en reconnaissance globale de mots),

- suppression des segments P de longueur inférieure à 0,03 s. Ces segments correspondent en général à une fluctuation du bruit dans une fricative, ou plus généralement à une période non significative, pendant laquelle rien n'est détecté.

Ces critères peuvent être aussi morphologiques, en ce sens qu'ils traduisent les règles régissant une succession de phonèmes. Par exemple :

- concaténation de segments "fricatifs" successifs, s'ils sont identiques ou si un type domine :

$$\begin{array}{l} \text{SS} \\ \text{SZS} \end{array} \} \rightarrow \text{S}$$

par contre $\text{SZ} \rightarrow \text{SZ}$

- concaténation des segments "voyelles" successifs identiques, mais avec mise en mémoire pour tenir compte des mots contenant éventuellement des voyelles redoublées,
- suppression des segments de type N ou F en fin de mot si leur durée est inférieure à 0,05 s. Ces segments correspondent à des "queues" de phonèmes et ne sont pas significatifs.

III-3-3 Résultats obtenus :

Le tableau III-1 donne une liste de 24 mots employés pour tester le système avec leur transcription phonétique, ainsi

Mot	Transcription phonétique	Réponse du système
1 SABOT	/sab o/	sapâo
2 AVOCAT	/avoka/	atoakə
3 CHATEAU	/fato/	/apo
4 ETE	/ete/	elte
5 ASSEZ	/ase/	asi
6 JETER	/zate/	zəate
7 EPOUSE	/epuz/	eituzɜ
8 CHANTER	/ʃâte/	ʃɛâte
9 ACHETER	/aʃate/	asʃatei
10 CERISE	/səriz/	səlsɜ
11 SILENCE	/silâs/	sieîâ ʃ
12 APPAT	/apa/	abə
13 CHOSE	/foz/	/oɜs
14 CHAT	/ʃa/	/aə
15 SOUPER	/supe/	supie
16 BATEAU	/bato/	apo
17 PECHER	/peʃe/	eʃse
18 EPEE	/epe/	ete
19 APACHE	/apaʃ/	abəasʃ
20 HARASSE	/arase/	aətase
21 SAPIN	/sapê/	sateê

Tableau III-1

Exemples de résultats obtenus

que la chaîne phonémique obtenue par reconnaissance, après traitement physique et morphologique.

On peut constater un certain nombre d'erreurs dans les chaînes reconnues. Cependant, il est possible de retrouver sans difficulté le mot prononcé, à l'aide d'une recherche lexicale du type exposé en III-4.

Le tableau III-2 donne par ailleurs une matrice de confusion (en pourcentages) pour les 19 phonèmes du vocabulaire considéré. Cette matrice correspond à plusieurs prononciations de ce vocabulaire par un seul locuteur masculin. Il faut préciser que le choix de ces 19 phonèmes résulte du hasard et qu'il n'a pas du tout été fait pour faciliter la reconnaissance.

D'autre part, il nous a paru plus intéressant de fournir les confusions les plus courantes, plutôt que des pourcentages globaux de reconnaissance de phonèmes, de peu d'intérêt.

Les voyelles sont, le plus souvent, bien reconnues et les erreurs les plus courantes mettent surtout en évidence l'insuffisance des informations fournies par l'analyse binaire. Voici quelques exemples :

- mauvaise distinction entre consonnes plosives : ces consonnes sont caractérisées par une brusque variation d'énergie à l'explosion, qui n'apparaît pas ici ; de plus l'absence d'information sur le fondamental ne permet pas de distinguer les plosives sourdes de leurs homologues sonores (par exemple /p/ reconnu à la place de /b/ dans SABOT) ;

- confusion entre fricatives sonores et sourdes qui pourraient être évitées, également par l'utilisation du fondamental (par exemple dans CERISE) ;

- "oubli" des consonnes plosives en début de mot, à cause du manque d'information sur les variations brusques d'énergie (par exemple dans BATEAU, PECHER) ;

Phonème → Reconnu
↓
Prononcé

	/s/	/ʃ/	/v/	/ʒ/	/z/	/p/	/t/	/k/	/b/	/a/	/o/	/e/	/ə/	/u/	/i/	/â/	/ê/	/ɛ/	/l/
/s/	90	6		2	2														
/ʃ/	3	90		6	1														
/v/	10		32			15	19	18	6										
/ʒ/	5	40		55															
/z/	29		10	10	51														
/p/						40	19	18	23										
/t/						9	41	43	7										
/k/						5	42	45	8										
/b/						36	17	15	32										
/a/										85		5	10						
/o/											92			8					
/e/												90	2		8				
/ə/										8		1	89			2			
/u/										1	11			81		7			
/i/	5											7			88				
/â/											2		4			83	11		
/ê/												6				4	90		
/ɛ/							2	3	2	4						6	8	40	3
/l/										3		16	7					32	4

Tableau III-2

Matrice de confusion de phonèmes

- confusions introduites par l'effet de seuil en tension de l'analyseur (cf. partie B chapitre II) : par exemple le /i/ de CERISE est assimilé à la fricative /s/, tandis que le /r/ de HARASSE, trop peu intense, est considéré comme une occlusion de plosive et donne /t/.

La méthode de reconnaissance analytique qui vient d'être exposée sera donc reprise et améliorée lorsqu'un nouvel analyseur, plus informant, sera réalisé.

III-4 Recherche lexicale d'un mot à partir de la chaîne phonémique

III-4-1 Position du problème :

La chaîne de phonèmes donnés par le système de reconnaissance analytique peut être considérée comme la réponse définitive de ce système, par exemple pour des problèmes de transmission de la parole. Cependant, pour la plupart des applications possibles, il est nécessaire de rechercher dans un lexique le mot, ou la phrase, qui correspond à la transcription phonétique obtenue. Cette transcription est entachée d'erreurs, que l'on peut classer en deux grands types, comme on l'a constaté sur le tableau III-1 :

- les répétitions d'un même phonème, ou de phonèmes voisins (phénomène de segmentation multiple) ;
- les omissions ou insertions de phonèmes, qui sont les erreurs les plus délicates à traiter.

La recherche lexicale d'un mot à partir d'une suite de phonèmes n'est donc pas un problème trivial. En particulier, il est intéressant d'avoir une chaîne phonémique aussi bonne que possible, les performances des méthodes de recherche lexicale

se dégradant rapidement lorsque le pourcentage d'erreurs dans la chaîne augmente. Dans cette optique, une optimisation préalable au niveau linguistique de la chaîne phonémique est très utile. Le rôle des contraintes linguistiques sera abordé en détail en partie D, chapitre II. Pour l'instant, les méthodes de recherches lexicales que nous utilisons ne bénéficient pas de cette optimisation car le système de reconnaissance linguistique décrit en partie D ne peut pas être actuellement implanté sur T 2000.

En réalité, les méthodes de recherche lexicale pure qui vont être développées ici -consistant à identifier un mot d'un vocabulaire à partir d'une succession de phonèmes- sont surtout intéressantes lorsqu'elles sont associées à une analyse syntaxique, donc en reconnaissance de phrases et non de mots isolés. Elles préparent alors directement à la reconnaissance du discours continu.

Dans le cas de mots isolés, la seule application intéressante concerne la recherche d'un mot dans de grands vocabulaires (jusqu'à mille mots au plus), pour lesquels les méthodes de reconnaissance globale semblent moins adaptées (cf. chapitre II). Pour ces grands vocabulaires, on ne peut d'ailleurs pas concevoir une recherche lexicale qui prenne en compte tous les mots possibles. Il est nécessaire de prévoir une sélection des mots intéressants, par exemple à l'aide de mémoires associatives.

III-4-2 Les méthodes de recherche lexicale :

Les travaux dans ce domaine sont encore en nombre relativement restreints, dans le contexte général de la reconnaissance vocale.

Une méthode de "comparaison élastique" qui tient compte à la fois des éléments composant la chaîne d'éléments et de leur ordre a été proposée [C-20]. Utilisée pour corriger les erreurs dans un texte dactylographié, elle semble pouvoir être utilisée également dans des cas plus complexes tels que la reconnaissance de mots écrits ou la reconnaissance vocale.

Une seconde méthode est fondée sur une procédure d'extraction de sous-chaînes d'éléments ordonnés de longueur maximale, communes à deux chaînes d'éléments. Par extension, la méthode permet le calcul d'un "indice de ressemblance" entre deux chaînes de phonèmes, avec possibilités d'avoir des réponses multiples pour chaque élément de la chaîne [C-21]. Cette méthode paraît puissante et suffisamment rapide pour un traitement en temps réel.

Une autre technique de décodage qui ouvre des perspectives intéressantes dans le domaine de la parole, moyennant un certain nombre de transformations, est la méthode fondée sur l'algorithme de Fano [C-22]. Il s'agit essentiellement d'une procédure de recherche de chemin dans un graphe, avec une métrique liée à la ressemblance entre une chaîne d'éléments donnée et la chaîne correspondant à ce chemin. Une application intéressante dans le cas de la parole a été proposée [C-23], mais elle est en fait utilisée avec une analyse syntaxique pour la reconnaissance de phrases [C-24], ce qui rejoint le travail d'Alter sur le Fortran/A-125].

Deux méthodes sont proposées ici.

- La première revient à chercher un cheminement dans une sorte de graphe obtenu en considérant des réponses multiples dans la chaîne phonémique. Le but est identique à celui de la méthode de Fano mais le principe en est essentiellement différent.

- La seconde utilise à nouveau les principes de la programmation dynamique pour s'affranchir des variations de longueur des chaînes phonémiques à comparer.

Ces méthodes ont été testées sur les données du système de reconnaissance analytique précédent. Les restrictions, déjà formulées, concernant l'analyse binaire, nous amèneront à ne donner que quelques résultats. Il s'agit là, en fait, de modèles qui seront testés systématiquement avec d'autres analyses du signal vocal.

III-4-3 Méthode de cheminement dans un tableau :

Pour chaque élément de la chaîne phonémique représentant un mot inconnu, on ne considère pas un seul phonème mais les trois phonèmes les plus probables ; ceci augmente la complexité de la recherche lexicale, mais aussi les chances de succès car la probabilité d'avoir le phonème exact est alors bien plus importante. Le nombre de phonèmes pris en compte a été limité à trois pour que l'algorithme puisse opérer en temps réel. A chacun de ces phonèmes est affecté un "poids", acoustique ou acoustico-linguistique, déterminé par la reconnaissance acoustique et, éventuellement, par une optimisation au niveau linguistique (partie D chapitre II) ; on s'intéresse donc plutôt à des présomptions de présence de phonèmes.

Dans ces conditions, une chaîne phonémique composée de L éléments est alors représentée par un tableau GRAPH de dimensions $(L \times 3)$.

D'autre part, chaque mot de référence du lexique est décrit sous forme d'un arbre donnant sa transcription phonétique avec, éventuellement, les variantes possibles (élisions, insertions, substitutions de phonèmes). La figure III-4 symbolise l'arbre

correspondant au mot "PETITE". Ces représentations sont obtenues par apprentissage et tiennent compte en même temps des résultats fournis par la segmentation acoustique.

On cherche à parcourir tout le tableau GRAPH par un "chemin" continu constitué de transitions d'un phonème à un autre ; chaque transition possède un poids déterminé par la forme de la représentation lexicale du mot considéré, par les poids propres des phonèmes et par un certain nombre de règles de transitions.

La comparaison d'une chaîne phonémique avec un mot du lexique revient ainsi à rechercher le chemin optimal -de plus grand poids- permettant de décrire GRAPH. Plusieurs chemins sont en général possibles et, de plus, certains chemins sont abandonnés en cours de route, grâce à des tests d'impossibilité (par exemple si toutes les transitions possibles à un moment donné ont un poids inférieur à un seuil). Ces abandons nécessitent un retour en arrière et augmentent beaucoup le temps de calcul. Il serait intéressant de trouver une procédure permettant d'abandonner le plus tôt possible un chemin erroné, mais ce problème n'est pas simple.

Le système donne des résultats intéressants et il faudra le coupler à l'optimisation linguistique de la chaîne phonémique pour en connaître toutes les possibilités.

La seconde méthode qui va être exposée maintenant nécessite en moyenne moins de temps de calcul et elle est particulièrement bien adaptée au traitement des segmentations multiples.

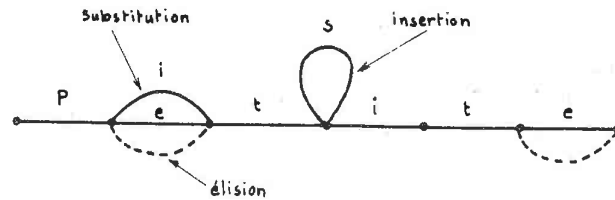


Figure III-5

Représentation lexicale du mot
"PETITE"

III-4-4 Méthode de comparaison dynamique :

La recherche lexicale telle qu'elle a été définie précédemment consiste à comparer deux chaînes de phonèmes en général de longueurs différentes. Il se pose donc un problème de normalisation en longueur de ces chaînes, analogue au problème de normalisation temporelle rencontré en reconnaissance acoustique. Utilisant l'expérience acquise, nous avons adopté une méthode de comparaison dynamique du type de celle décrite au chapitre II. Il s'agit de minimaliser le taux de dissemblance entre deux chaînes, au sens d'une métrique inter-phonèmes. Cette métrique utilise les données d'une matrice de confusion de phonèmes, obtenue par apprentissage avec le système de reconnaissance analytique et semblable à celle donnée en III-3-3. Cela permet de tenir compte à la fois du caractère propre de chaque phonème et des performances de la reconnaissance acoustique.

Il faut noter aussi que cette méthode permet de s'affranchir du traitement préalable de la chaîne sur critères simples, tel qu'il a été décrit en III-3-2-c, alors que la méthode précédente le nécessite.

Il est possible de travailler en temps réel pour des vocabulaires allant jusqu'à quelques centaines de mots, sans doute. Au-delà, une présélection de type associatif doit être envisagée.

L'exemple schématisé sur la figure III-6 montre bien les possibilités de la méthode. On a représenté le résultat de la comparaison de la chaîne phonémique correspondant à la prononciation du mot "SAPIN", avec la représentation lexicale de ce même mot (/ sapin /), qui a été finalement reconnu.

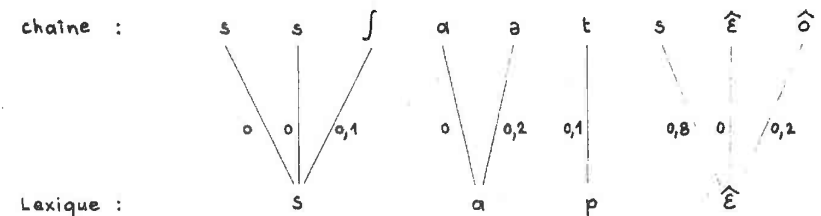


Figure III-6

Comparaison d'une chaîne inconnue avec
le mot "SAPIN"

Bien entendu, la méthode serait inopérante, comme n'importe quelle autre, si le mot SATIN figurait au lexique. Le problème se situe alors au niveau sémantique.

On peut envisager, au prix d'une augmentation de temps de calcul, la généralisation au cas où l'on considère des présomptions de présence de phonèmes. Enfin, si le traitement sur critères simples de la chaîne phonémique est inutile, une optimisation au niveau linguistique améliorerait par contre les performances.

III-5 Conclusion - Reconnaissance globale et reconnaissance analytique.

Un système de reconnaissance analytique de segments, associé à une segmentation acoustique du signal vocal, a été décrit dans ce chapitre. Ce système utilise une méthode de comparaison dynamique analogue à celle décrite au chapitre II et permet d'obtenir en temps réel la transcription phonémique d'un mot. Deux procédures de recherche lexicale ont aussi été développées ; elles permettent de retrouver le mot prononcé à partir de la chaîne de phonèmes. En pratique, cette recherche lexicale est surtout intéressante en conjonction avec des analyses de niveaux supérieurs : linguistique, syntaxique ; on obtient alors une ébauche de système de reconnaissance du discours continu.

Ces expériences de reconnaissance analytique ont mis en évidence quelques insuffisances du système d'analyse spectrale binaire que nous utilisons actuellement. En particulier, on peut constater un manque d'informations pour distinguer entre eux certains phonèmes : consonnes plosives, quelques voyelles...

Le système proposé apparaît en ce sens comme un modèle opérationnel qu'il sera intéressant d'utiliser avec une méthode d'analyse plus fine.

Quoiqu'il en soit, cette étude, associée à celle du chapitre II, permet d'effectuer une comparaison valable des deux approches -globale et analytique- de la reconnaissance vocale.

Tout d'abord, l'approche analytique apparaît beaucoup plus générale, car elle permet de traiter des vocabulaires théoriquement illimités. On restreint un peu cette généralité en utilisant un lexique, mais le gain de place reste important par rapport à l'approche globale puisqu'il suffit de mettre en mémoire la transcription phonétique des mots et non pas toute leur représentation, spectrale ou autre. De plus, l'approche analytique est la seule qui permette d'aborder le problème du discours continu, en association avec des analyses aux niveaux lexical, linguistique, syntaxique et sémantique. Enfin, une reconnaissance analytique peut être très rapide, si le processus de reconnaissance se déroule au cours de la prononciation du mot, comme c'est le cas dans le système décrit ici.

Cependant d'importants problèmes restent à résoudre avant qu'un système analytique devienne vraiment performant, en particulier dans le domaine de la segmentation.

L'approche globale, la plus utilisée jusqu'à présent, pose par contre moins de difficultés et un certain nombre de systèmes de reconnaissance globale sont déjà opérationnels. Les études rapportées au chapitre II ont montré la validité de cette approche pour des vocabulaires de taille moyenne, dont la limite doit se situer vers quelques centaines de mots. Elles ont mis en évidence, également, que de tels systèmes pouvaient se satisfaire d'une analyse assez rudimentaire du signal vocal car la reconnaissance se fonde plus sur l'aspect global d'un mot que sur sa structure fine.

PARTIE D

UTILISATION DE CONTRAINTES LINGUISTIQUES,

SYNTAXIQUES ET SEMANTIQUES

EN RECONNAISSANCE VOCALE.

INTRODUCTION A LA PARTIE D.

Les deux approches, globale et analytique, de la reconnaissance vocale au niveau acoustique ont été abordées en partie C. Les résultats obtenus, pour intéressants qu'ils soient, sont loin d'être parfaits. Il subsiste encore beaucoup d'ambiguïtés et d'erreurs, ce qui est normal car l'homme lui-même commet des erreurs lorsqu'il se place au niveau acoustique (reconnaissance de logatomes sans signification, par exemple).

Il est très important d'étudier dans quelle mesure l'utilisation de contraintes de niveaux hiérarchiques supérieurs : linguistique, syntaxique, sémantique, peuvent améliorer les performances d'une reconnaissance purement acoustique. C'est un fait maintenant admis que le passage à ces niveaux est très utile en reconnaissance vocale et indispensable, en particulier dans le cas de la parole continue (cf. partie A chapitre III). Cette façon de procéder est d'ailleurs en accord, comme nous le verrons, avec le processus de perception humaine. Cependant, les réalisations de systèmes à ces niveaux sont encore rares.

Dans une première étape, nous reviendrons sur le problème de l'optimisation d'une chaîne phonémique en reconnaissance analytique, cette fois avec l'aide de contraintes linguistiques. Ces contraintes sont exprimées en résultats statistiques, obtenus par une étude préalable du français parlé. L'amélioration obtenue, forcément limitée par le caractère statistique des données utilisées, est cependant intéressante car peu coûteuse en moyens et en temps de calcul.

Ensuite, l'utilisation de contraintes syntaxico-sémantiques sera envisagée, tant en reconnaissance globale de mots qu'en reconnaissance analytique. Indispensable pour reconnaître la parole continue, l'utilisation d'une grammaire se révélera également très intéressante dans le cas de phrases de mots prononcés isolément. Pour illustrer cela, nous appliquerons le système de reconnaissance de mots, décrit en partie C chapitre II, à un langage très simple utilisé en commande numérique de machines-outils.

La présence, au laboratoire, d'un groupe actif de commande numérique a été déterminante pour le choix de ce langage, dont l'intérêt pratique est certain et qui représente pour nous une bonne introduction à l'analyse syntaxique.

Ce travail est actuellement poursuivi, dans le cadre de langages plus évolués (programmation orale), pour déboucher, dans notre esprit, sur la conception de langages adaptés à la communication orale homme-machine.

CHAPITRE I

STATISTIQUE PHONETIQUE DU FRANCAIS PARLE [B-3]

I-1 Introduction.

Après avoir introduit le phonème en acoustique, on s'intéresse maintenant à son comportement sur le plan phonétique et linguistique. Il ne s'agit bien entendu nullement d'effectuer une étude linguistique complète mais simplement de dégager des résultats qui permettront une amélioration de la reconnaissance au niveau du phonème. Nous nous sommes placés à un point de vue statistique, permettant d'obtenir, par comptage, les fréquences d'apparition de différentes entités. Cette approche nécessite l'utilisation de corpus importants et suffisamment représentatifs de la langue parlée (50 033 phonèmes répartis en 11 443 mots). La stabilité des résultats que nous avons constatée pour différents corpus permet à la rigueur de parler de probabilités et non de fréquences d'occurrence, avec des intervalles de confiance que nous calculerons à la fin du chapitre. Cependant, la comparaison des différentes études de ce type concernant le français parlé laisse apparaître quelques divergences dans les résultats [D-12]. Il faut donc être prudent dans l'utilisation de ces résultats statistiques.

I-2 Principe de l'étude.

La validité des résultats obtenus repose sur la validité du corpus choisi. Ce dernier se compose d'échantillons suffisamment longs (au moins 2 000 phonèmes par extrait), tirés de la langue parlée, conversationnelle et narrative de milieux sociaux divers, dans des domaines variés. Certaines de ces conditions de validité seulement étaient réalisées dans des études précédentes concernant la langue anglaise $[D-2] \rightarrow [D-5]$. Les textes retenus sont ensuite transcrits en écriture phonétique en utilisant 34 phonèmes, y compris la pause que nous appelons "silence" (en fait, ce nombre a été réduit à 30 pour les expériences de reconnaissance). Le tableau I-1 donne la liste de ces phonèmes avec leur mot-clé ainsi que le symbole machine choisi.

n°	phonème	mot clé	symbole machine
1	Silence		1
2	α + a	plat	A
3	R	rue	R
4	l	lent	L
5	e	blé	J
6	s	sou	S
7	i	il	I
8	z	lait	C
9	ø (eu + e)	le peu	E
10	k	cou	K
11	t	tas	T
12	p	pas	P
13	d	dans	D
14	m	ma	M
15	n	an	3
16	ø	nous	N
17	U	ou	U
18	V	vie	V
19	Y	nu	Y
20	5	cn	2
21	o	eau	ø
22	z	je	J
23	o	bol	O
24	z	lin	4
25	f	feu	F
26	b	bon	B
27	w	voir	W
28	Y	huit	+
29	œ	heure	—
30	z	zéro	Z
31	J	bailler	.
32	f	chat	X
33	g	gare	G
34	œ	un	5

Tableau I-1 : Liste des phonèmes.

La transcription phonétique a été effectuée manuellement, de préférence à l'utilisation d'un programme. Nous avons choisi cette solution, beaucoup plus fastidieuse, de façon à pouvoir trancher dans le cas où une ambiguïté se présente. De toute façon, le caractère parfois subjectif de cette transcription apparaît ici : la précision des résultats obtenus sera ainsi moins grande que ne l'indique la théorie (cf. paragraphe I-7).

L'étude statistique comprend deux thèmes principaux :

a) comptage des phonèmes, diphonèmes et triphonèmes du corpus.

Nous appelons diphonème tout élément formé par l'association de deux phonèmes. Ainsi le mot "maison" se décompose en trois diphonèmes : mE - E z - zS.

Un diphonème peut aussi résulter de l'association d'un phonème X et du "silence" : (silence - X) et (X - silence).

Un triphonème sera de même un élément formé de l'association de trois phonèmes. Parmi les associations comportant le "silence", seules celles du type (X - Y - silence) et (silence - X - Y) présentent un intérêt. L'étude des triphonèmes nécessite déjà l'introduction d'un tableau comportant $34^3 = 39\ 304$ éléments, ce qui exclut, pour l'instant, l'étude systématique des éléments phonémiques d'ordre supérieur ;

b) études liées aux mots.

L'introduction du "silence" permet de tirer quelques conclusions concernant les mots. Nous avons simplement étudié la distribution des mots suivant leur longueur, ainsi que les commencements et terminaisons de mots.

I-3 Etude des phonèmes.

Le tableau I-2 donne, dans l'ordre décroissant, les pourcentages d'occurrence pour tous les phonèmes réunis (I-2-a), ainsi que pour voyelles et consonnes séparées (I-2-b).

Phonème	% d'occurrence	Phonème	% d'occurrence
R	8,37	u	2,28
a	7,50	o	2,16
l	6,79	ô	2,13
l	6,71	f	1,44
s	5,90	z	1,23
e	5,82	b	1,18
æ	5,42	q	1,07
t	4,93	ê	0,99
e	4,81	w	0,94
d	4,51	z	0,75
p	3,94	œ	0,60
k	3,90	ʃ	0,56
â	3,73	g	0,54
m	2,98	œ	0,49
n	2,63	J	0,43
y	2,46	Y	0,20
v	2,46		

I-2-a

Phonème	% d'occurrence	Phonème	% d'occurrence
R	8,37	a	7,50
l	6,71	i	6,79
s	5,90	e	5,82
t	4,93	æ	4,42
d	4,51	z	4,81
p	3,94	â	3,73
k	3,90	Y	2,46
m	2,98	u	2,28
n	2,63	o	2,16
V	2,46	ô	2,13
f	1,44	o	1,07
z	1,23	ê	0,99
b	1,18	œ	0,60
w	0,94	œ	0,49
z	0,75	Y	0,20
ʃ	0,56		
g	0,54		
J	0,43		
Total	51,96	Total	48,04

Consonnes

I-2-b

Voyelles

Tableau I-2

Fréquences d'occurrence des phonèmes.

La répartition globale s'établit à 48 % de voyelles et 52 % de consonnes, c'est-à-dire pratiquement à égalité alors que, par exemple, les Anglo-Saxons utilisent deux fois plus de consonnes que de voyelles [D-5].

Les cinq voyelles /a, i, e, æ, ε / représentent 65 % des occurrences de voyelles, tandis que les quatre consonnes /r, l, s, t/ représentent 50 % des occurrences totales.

Enfin, huit phonèmes seulement (/r, a, i, l, s, e, æ, t/) sont responsables de 50 % des occurrences totales.

On peut remarquer que les liquides et semi-voyelles (/r, l, z, w/) représentent à elles seules près de 32 % des consonnes.

Enfin, les consonnes sourdes sont plus fréquentes que leurs homologues sonores (tableau I-3).

Sourdes	Pourcentage	Sonores	Pourcentage
p, t, k	24,6	b, d, g	12,0
f, s, ʃ	15,2	v, z, ʒ	8,5

(pourcentages donnés sur la totalité des consonnes)

Tableau I-3.

La comparaison de ces résultats avec ceux du français littéraire [D-6] fait apparaître quelques différences, provenant de la richesse plus grande du vocabulaire de la langue écrite et de phénomènes propres à la langue parlée, tels que l'élision du /ə / en fin de mot, ou l'usage plus fréquent du /a/.

La figure I-1 montre la courbe donnant les pourcentages cumulés d'apparition des phonèmes en fonction du nombre de phonèmes. Sur ce même graphique sont également portées les courbes analogues relatives aux mots et aux syllabes.

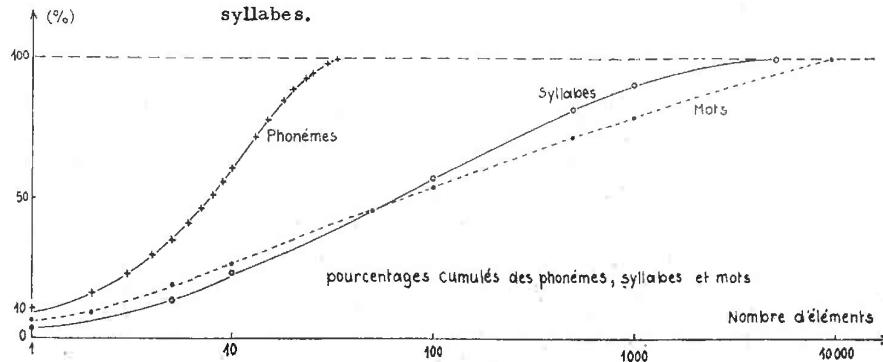


Figure I-1

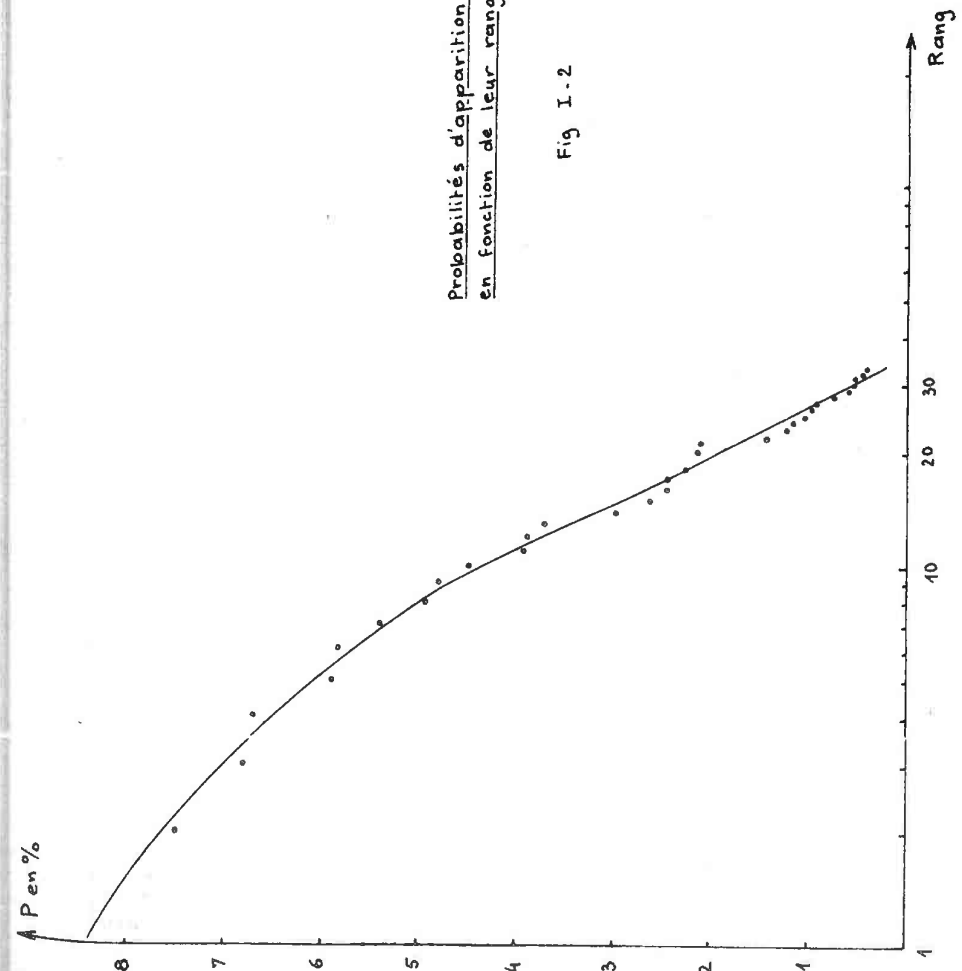
La distribution des phonèmes peut être étudiée en traçant la courbe donnant leur fréquence d'apparition en fonction de leur rang. Cette courbe est donnée sur la figure II-2, en coordonnées logarithmiques. Elle ne présente pas de points d'inflexion à ses extrémités : les phonèmes souvent employés sont nombreux, tandis que les phonèmes peu employés sont peu nombreux.

I-4 Etude des diphonèmes.

Les fréquences d'apparition des diphonèmes, comptées sur l'ensemble du corpus, sont données dans le tableau I-4. Dans ce tableau, la fréquence f_{ij} du diphonème (I-J) se trouve à l'intersection de la ligne n° i et de la colonne n° j. Ainsi, les sommes par ligne $\sum_j f_{ij}$ et par colonne $\sum_i f_{ij}$ représentent

Probabilités d'apparition des phonèmes
en fonction de leur rang R

Fig I-2



respectivement la fréquence d'apparition du phonème I en première position d'un diphonème et celle du phonème J en deuxième position d'un diphonème. On constate que ce classement diffère du classement des phonèmes obtenu en I-3, puisque dans ce tableau les phonèmes sont rangés par ordre d'occurrence décroissante.

Sur les 1089 (33²) diphonèmes possibles, 615 seulement (soit 56 %) apparaissent dans le corpus et parmi ces derniers, 61 n'apparaissent qu'une seule fois. On peut considérer ce nombre comme caractéristique du français puisque des études analogues, portant sur d'autres données, ont fourni 613 diphonèmes sur 900 possibles [D-7] et 630 diphonèmes sur 1225 possibles [D-8].

Dans le tableau I-5 sont rassemblés les 27 premiers diphonèmes, correspondant à 25 % des occurrences totales. On remarque, parmi ces 27 diphonèmes, plusieurs "mots-chevilles" (articles, prépositions, ...) très courants dans la langue parlée.

Avec les voyelles V et les consonnes C, on constate que les associations

C-V représentent 52,2 % du total
V-C représentent 33,5 % du total
C-C représentent 11,2 % du total
V-V représentent 3,1 % du total.

Diphonème	Fréquence en %
DE	19,9
LA	16,1
AR	15,2
SI	15
TR	13,9
ÈR	13,6
IL	11,8
PA	11,5
Lé	12,6
Dé	11,8
LE	11,7
WA	11,4
SE	10,5
ouR	9,9
RI	9,5
Rè	
Sè	9,4
PR	
OR	9,3
AV	8,5
Men	8,1
KI	7,8
AL	7,7
Té	
KE	7,5
TI	
DI	7,2
Total	252,1
	soit 25 % environ

Tableau I-5

DICTIONNAIRE																														Σ				
R	a	l	e	e	t	e'	d	p	k	u	y	v	u	z	h	f	z	b	o	e	w	u	a	f	g	u	j	q						
R	0	136	297	10	75	145	132	97	205	46	5	25	124	63	32	17	25	44	63	18	7	14	10	17	19	28	0	5	17	4	0	3	158	
a	329	1	2	167	148	1	0	103	1	20	69	51	0	48	54	2	184	0	1	0	15	60	67	0	0	0	21	0	71	8	0	44	141	
l	113	20	0	254	137	85	36	145	44	50	20	80	7	48	54	0	58	0	20	139	30	11	32	2	75	0	63	15	6	9	0	28	162	
e	0	349	142	0	0	274	253	14	85	1	0	21	37	18	0	110	2	8	60	27	2	2	1	18	19	23	0	55	0	3	4	2	154	
s	1	46	328	0	0	131	227	105	205	0	54	48	101	4	0	96	0	26	62	93	4	0	0	6	18	24	0	22	0	0	0	19	166	
e	69	4	16	29	29	2	0	116	2	14	41	54	4	24	16	8	24	0	3	0	30	14	18	1	0	0	29	0	10	25	0	12	58	
a	44	0	0	57	23	0	0	38	0	3	33	13	0	127	48	0	47	0	2	0	8	1	0	0	0	20	0	6	12	0	0	0	48	
e	304	73	156	2	0	162	43	0	130	0	1	1	93	4	0	92	0	118	11	13	0	0	0	12	16	10	0	18	0	0	0	0	126	
e	294	0	1	140	104	0	1	146	0	3	3	83	0	53	57	0	6	0	1	11	2	3	0	0	0	21	0	3	17	1	20	0	97	
d	51	55	156	0	0	255	431	0	39	0	1	0	102	3	0	122	0	14	26	21	0	1	0	12	6	12	0	3	0	0	22	2	1331	
p	203	792	26	141	2	44	60	4	64	0	0	0	42	1	0	25	0	155	79	13	0	0	0	14	2	6	0	12	0	0	0	0	149	
g	49	70	169	28	50	15	162	36	60	0	0	0	44	0	5	47	0	65	118	122	0	0	15	2	24	0	8	0	0	8	0	0	109	
e	6	0	0	9	126	0	0	123	0	97	25	32	0	0	1	0	13	0	0	15	22	23	1	0	0	1	0	9	6	0	0	0	50	
m	0	105	97	0	1	46	46	0	146	1	1	0	177	0	1	13	0	4	38	28	0	0	1	10	24	36	0	3	0	0	0	1	77	
o	0	58	132	0	0	85	111	1	49	0	0	0	27	3	0	32	0	71	26	19	0	0	0	9	2	3	1	1	0	0	0	12	6	591
y	110	10	59	39	31	7	0	19	5	18	13	9	2	18	139	0	0	0	0	1	5	7	0	0	0	14	1	0	0	0	0	0	50	
v	46	42	119	0	0	66	37	0	128	0	0	0	43	1	0	13	0	46	25	18	0	0	0	15	15	68	0	2	0	0	0	0	72	
u	214	0	1	23	24	9	0	42	2	4	12	1	1	2	0	0	78	0	0	0	4	10	5	0	1	2	7	0	0	0	0	0	450	
z	201	0	2	44	44	1	0	43	0	11	22	26	0	106	77	0	8	0	0	13	10	19	0	0	0	27	0	2	2	0	0	0	661	
h	0	0	0	0	39	0	0	40	0	32	22	14	0	0	0	0	6	0	0	11	2	18	0	0	1	0	1	0	2	0	0	0	0	188
f	36	51	57	11	0	24	10	0	81	0	0	1	14	0	0	29	0	7	43	18	0	0	0	25	10	16	0	1	0	0	0	0	434	
o	0	24	23	1	0	68	79	0	13	0	0	0	43	0	1	15	0	46	1	1	0	0	0	1	9	0	7	0	0	0	0	0	352	
b	43	37	61	46	6	4	1	14	0	0	11	0	1	18	0	13	24	12	0	1	14	0	6	0	1	0	0	0	0	0	0	0	359	
o	14	0	0	6	16	0	0	52	1	6	2	19	0	4	2	5	0	0	0	2	5	1	0	0	0	11	0	4	0	0	0	0	150	
e	0	0	0	0	24	0	0	38	0	14	23	5	0	0	0	0	2	0	0	0	8	3	0	0	0	0	2	0	1	1	1	0	122	
w	0	248	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	291	
z	0	18	61	0	0	28	4	0	7	0	0	0	24	1	1	6	0	0	5	14	0	0	0	2	3	3	0	0	0	0	0	1	178	
u	136	0	0	20	2	0	0	0	0	5	0	0	0	6	0	10	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	187	
f	0	25	4	0	0	34	14	0	19	0	0	0	26	0	0	0	10	2	0	0	0	9	0	6	0	2	0	0	0	0	0	0	151	
g	45	55	2	10	0	1	0	0	12	0	0	0	1	1	0	3	0	7	4	3	0	0	0	3	0	1	15	1	0	0	0	0	164	
u	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
j	0	7	9	0	0	23	14	0	7	0	0	0	10	0	0	0	0	0	0	0	3	0	0	1	7	0	0	13	0	0	0	0	94	
y	0	0	55	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	
2308 1869 881 4668 1313 352 933 495 468 622 147 207 253 231 86 37 57																																		
1761 1117 1492 1166 321 486 523 648 664 564 167 177 255 172 89 132																																		

Tableau I-4

Fréquence d'apparition des diphonèmes.

Les associations mixtes C-V et V-C prédominent largement avec 85,7 % du total.

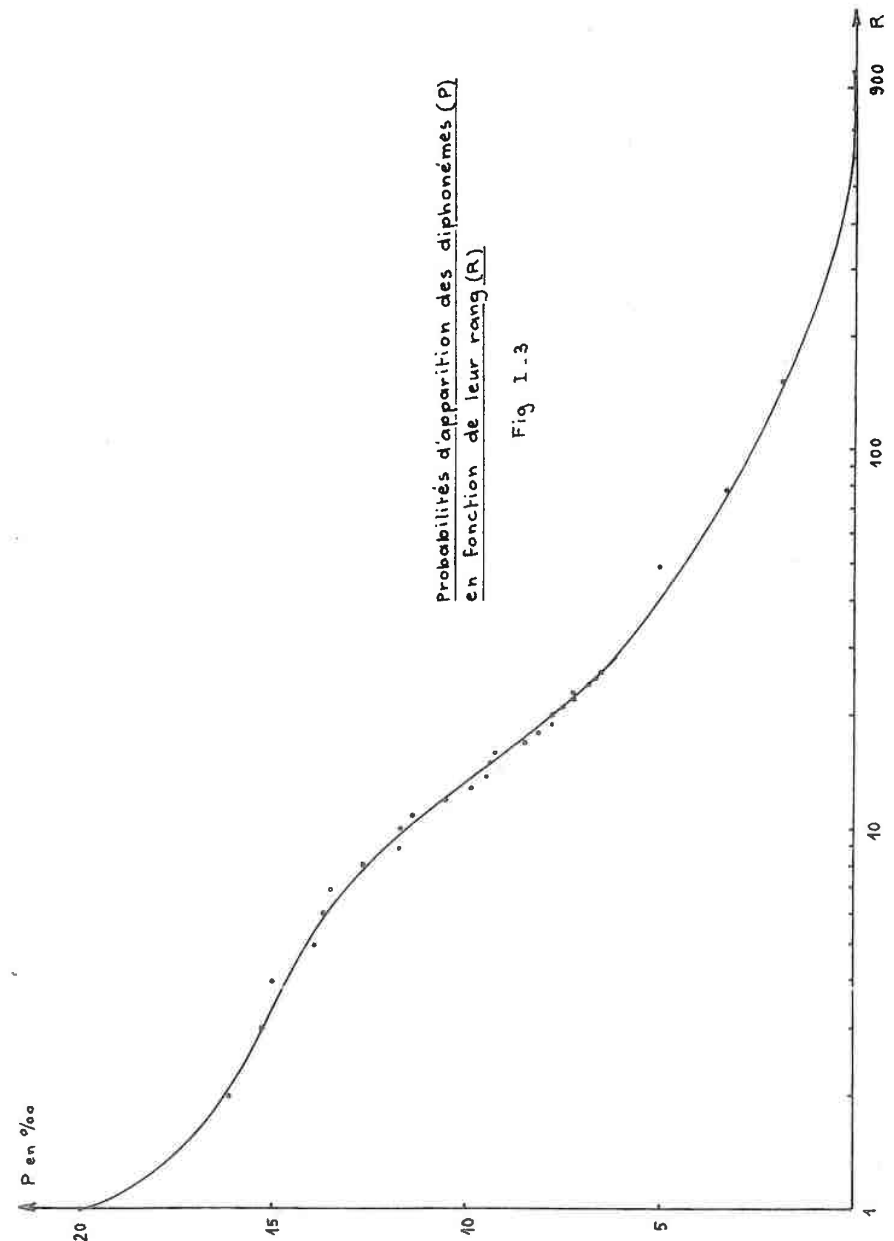
Afin d'étudier la distribution des diphonèmes, la courbe donnant, en coordonnées logarithmiques, la fréquence des diphonèmes en fonction de leur rang, a aussi été tracée sur la figure I-3. Outre la régularité de cette distribution, on constate que cette courbe s'infléchit aux deux extrémités : les diphonèmes très employés sont peu nombreux, ceux peu employés très nombreux. Ce comportement est donc opposé à celui que nous avons constaté pour les phonèmes (cf. figure I-2).

L'existence de contraintes phonétiques au niveau des diphonèmes apparaît déjà dans le nombre limité de diphonèmes effectivement permis. Elle se retrouve également dans les probabilités d'apparition de ces diphonèmes, comme on l'a noté au chapitre I partie B, à propos de la redondance de la parole.

Par exemple :

$$\frac{p(\ell - a)}{p(\ell) \cdot p(a)} = 3,21 \quad \text{et} \quad \frac{p(a - \ell)}{p(a) \cdot p(\ell)} = 1,42$$

Nous rencontrerons à nouveau cette propriété avec les triphonèmes. Ces contraintes sont également responsables de la dissymétrie qui existe, pour un phonème donné, dans les distributions de P_{ij} et de P_{ji} , où j désigne un phonème quelconque. A titre d'exemple les diagrammes circulaires des phonèmes /d/ et /k/ sont portés sur la figure I-4. On constate aussi que les distributions de P_{ij} et $P_{i'j}$, pour deux phonèmes i et i' , peuvent varier considérablement. Ainsi la distribution correspondant au /d/ est très dissymétrique -les diphonèmes (d-ə) et



Probabilités d'apparition des diphonèmes (P)
en fonction de leur rang (R)

Fig I-3

(d-e) émergeant très nettement -, tandis que la distribution correspondant au /k/ est plus uniforme.

I-5 Etude des triphonèmes.

Il n'est pas possible d'écrire en entier le tableau des fréquences d'occurrence de tous les triphonèmes et nous nous contenterons de donner, dans le tableau I-6, les 200 premiers triphonèmes classés par ordre décroissant, avec leurs fréquences. Ici encore, les mots de trois phonèmes (par, pour, etc...) comptent parmi les associations les plus courantes.

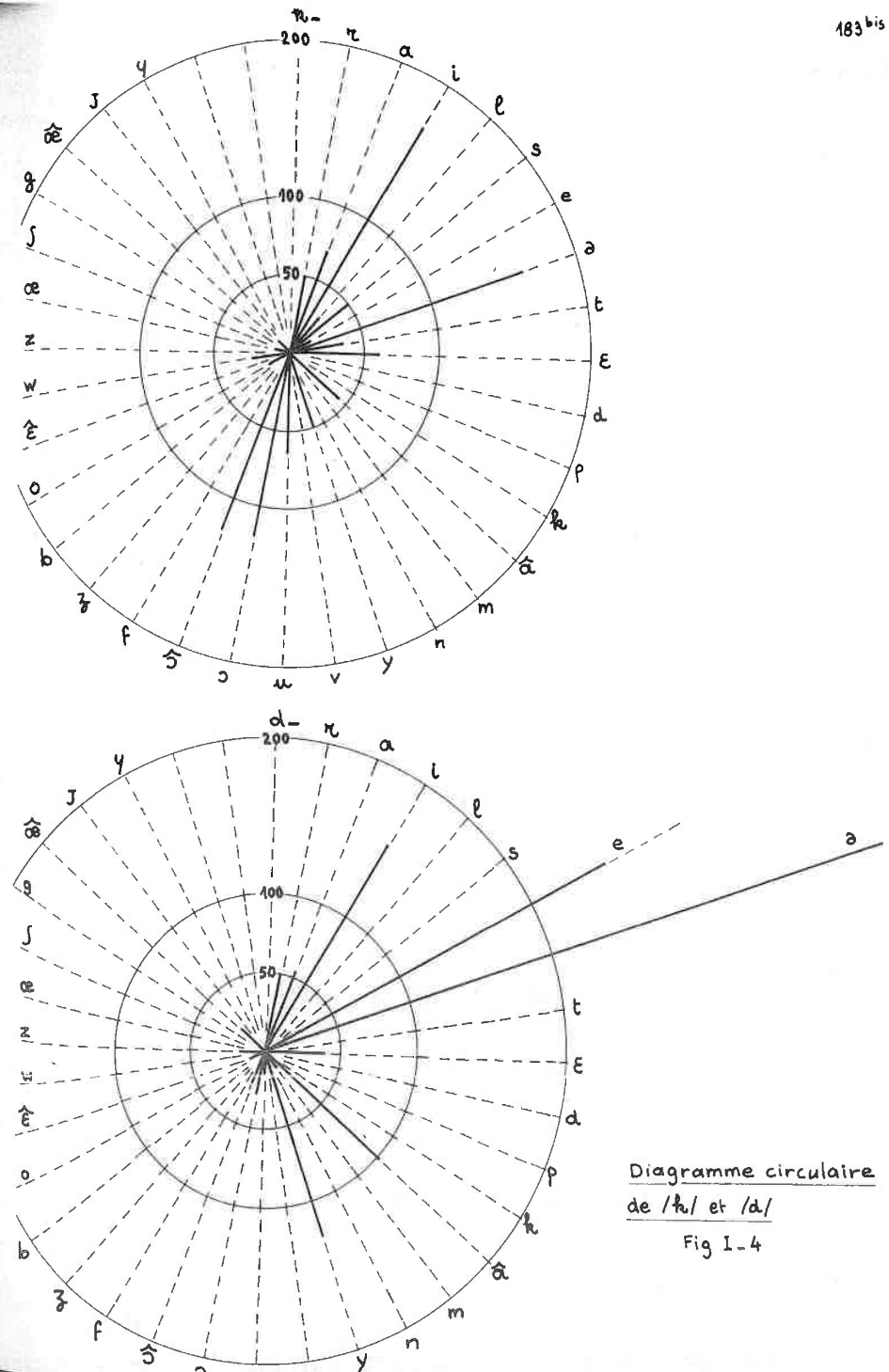
La redondance constatée au niveau des diphonèmes est encore plus importante pour les triphonèmes.

En effet, sur les 35 937 (33^3) triphonèmes possibles, seuls 2 509 sont effectivement présents dans le corpus, soit 7 %.

Les associations mixtes CVC et VCV sont encore les plus courantes, tandis que les associations VVV, totalement absentes de notre corpus, doivent être considérées comme très rares dans la langue.

Comme on pouvait s'y attendre, de nouvelles contraintes apparaissent au niveau des triphonèmes. Il est possible d'en rendre compte par un raisonnement analogue à celui que nous avons tenu pour les diphonèmes. En se limitant aux diphonèmes, supposés indépendants, on obtient pour la probabilité du triphonème (I-J-K) :

$$P_{ijk} = \frac{P_{ij} \cdot P_{jk}}{\sum_k P_{jk}}$$



P A R	953	L I T	206	K R I	140	W A L L	110
P O U R	946	L E R	206	K U L	140	L E U R	110
S I R N	761	T R A	206	A N K S	140	R I I N	103
E U M A N	643	A I K S	206	S N A I	140	R E U V	103
A V A I	517	T U R	199	M A I T	140	A R S	103
V V A	502	K S I	199	A R D	133	A T U	103
A S I	473	A N D R	199	A L I	133	A V A N	103
P L U	458	M A I M	199	A V W	133	T E U R	103
S E U E	436	I A I R	192	S I Y	133	I D E	103
L E U R	413	L A S	192	S E U R	133	T Z E	103
S E R	406	L U I	192	S P B	133	E U R A I	103
K P M	399	P U I	192	T E R	133	E U S T	103
W A R	399	M W A	192	P A S	133	E U V E U	103
P R H	369	N I E	192	K A N P	133	E K M	103
K I L	369	R O U V	184	A N S I	133	T R E U	103
A I T P	362	S A N T	184	R E Z	125	T A L	103
I T E	340	T R O U	184	A V A N	125	T I T	103
S T R	332	V E U M	184	I T U	125	T E U H	103
T A I R	317	L E U M	177	E U M I	125	T A I L	103
R Q T	310	P A N S	177	S I L	125	A I G Z	103
S S I	310	A R T	170	T R I	125	P R E U	103
P R I	303	A T I	170	T I K	125	P A N O	103
K A I L	303	S A A	170	T O U R	125	K A N	103
A N T P	303	E R I	170	T O U T	125	A R T E	103
I S I	295	A T R N	170	D E U M	125	A N D U	103
E T A I	295	P S Z	170	D E P	125	A N J E	103
S A I T	288	R R S	170	P E U T	125	R R A I	103
S I S	288	M A R	170	K A R	125	V I Z	103
P R A I	288	G A P	170	O U V E	125	R L I	103
S A I R	280	R T I	162	S R A N	125		
P A I R	280	S A V	162	R A L	118		
F R R	280	D U H	162	A R A I	118		
P R R	273	S M A N	162	A M I	118		
F A I S	266	A S E	155	A V A	118		
V A I R	266	A P R	155	A W I	118		
V A I K	266	E T E	155	L A I R	118		
F A I R	258	A I L K	155	L B T	118		
I S T	251	K O U R	155	E U T I	118		
E U L A	251	K W A	155	S B M	118		
T R A I	251	F A S	155	A I K T	118		
A I S T	243	R A V	147	D E B	118		
R E U M	236	R T V	147	D S N	118		
R A N D	236	R E U P	147	P O U V	118		
S S I	236	I L I	147	M A T	118		
J O U R	236	S T R	147	F W A	118		
R I J N	236	T R A N	147	B L A I	118		
A I S S	229	D E U V	147	R A P	110		
A I J	221	M A E	147	A L B	110		
A B L	221	R A N S	140	I S K	110		
P R F	221	A V E	140	L W A	110		
K R R	221	L I K	140	E U N U	110		
K O N S	221	S T J	140	T E U M	110		
A R I	214	T E U R	140	P E A	110		
D I R	214	A J R T	140	P E R	110		
K O N T	214	A I R V	140	A N T I	110		
R R M	214	A I S P	140	A N D A N	110		
R V A	206	P R A N	140	V R A I	110		

Tableau I-6

Compte-tenu des incertitudes sur ces grandeurs, les écarts constatés entre cette valeur théorique et les valeurs expérimentales sont souvent significatifs.

Ainsi :

- pour le triphonème EULA (/ə l a/) : $\frac{\text{fréquence mesurée}}{\text{fréquence "théorique"}} \approx 2,8$

- pour le triphonème DEUV (/d ə v/) : $\frac{\text{fréquence mesurée}}{\text{fréquence "théorique"}} \approx 0,4$

I-6 Etude des mots.

I-6-1 Longueur des mots :

Le tableau I-7 donne les fréquences d'occurrence des mots suivant le nombre de phonèmes qu'ils contiennent.

Mots de	Fréquence en %
1 phonème	107,3
2 phonèmes	361,3
3 phonèmes	168,7
4 phonèmes	116,1
5 phonèmes	87,4
6 phonèmes	67,7
7 phonèmes	42,1
8 phonèmes	24,0
9 phonèmes	13,4
10 phonèmes	7,4
11 phonèmes	2,5
12 phonèmes	0,7
13 phonèmes	0,5
14 phonèmes	0,1

Tableau I-7

Fréquence d'occurrence des mots suivant leur longueur

On constate que les mots de deux phonèmes sont de loin les plus courants, puisqu'ils constituent plus de 36 % du total. Parmi ceux-ci, les "mots-chevilles", dont nous avons

déjà parlé, prédominant largement.

Les autres pourcentages sont régulièrement répartis, comme l'indique la figure I-5.

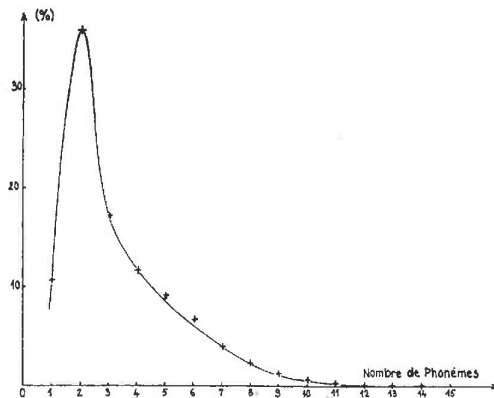


Figure I-5

Répartition des mots en fonction du nombre de phonèmes

I-6-2 Commencements et terminaisons :

Les fréquences d'occurrence des phonèmes en début et en fin de mot sont données dans le tableau I-8. Les voyelles apparaissent peu en début de mot (25 %) et beaucoup en fin de mot (74 %), leur pourcentage général dans la langue étant de 48 %. Les résultats sont bien entendu symétriques pour les consonnes, 75 % en début de mot, 26 % en fin de mot, pour un pourcentage général de 52 %.

A l'exception du /ə/, très fréquent en fin de mot et très rare en début de mot, la distribution des voyelles en début et en fin de mot diffère peu de la distribution générale des voyelles, comme l'indique le tableau I-9-a, dans lequel les phonèmes sont classés par ordre décroissant. Le tableau

Répartition des phonèmes en début et en fin de mot
Les fréquences sont en dix-millièmes

DEBUT		FIN	
Phonème	Fréquence	Phonème	Fréquence
D	1.172	e	1.306
L	1.044	ə	1.300
S	1.028	R	1.090
P	945	a	983
K	786	ɛ	703
a	607	L	584
M	427	ɛ	562
T	388	i	523
N	350	ô	515
e	329	T	279
V	321	u	277
F	312	y	271
R	300	N	238
i	252	ɛ	203
ɔ	240	O	196
z	234	S	178
ɛ	181	œ	166
B	172	M	154
o	157	K	119
œ	126	D	63
Y	120	Z	60
θ	104	ʃ	54
f	97	J	44
G	87	V	41
ɛ	60	f	26
u	49	P	17
w	43	F	14
œ	16	ɔ	9
ə	12	G	5
J	3	B	4
Y	2	ɣ	1
z	2	œ	0
		w	0

Tableau I-8

I-9-b montre qu'il existe des variations plus importantes en ce qui concerne les consonnes.

Dans l'optique de la reconnaissance, nous donnons aussi, dans le tableau I-10, les fréquences d'apparition des 50 premiers diphonèmes en début et en fin de mot.

Général	Début de mot	Fin de mot
a	a	e
i	e	ə
e	i	a
ə	â	â
ɛ	ɛ	ɛ
â	o	i
y	œ	ø
u	y	u
ɔ	ô	y
ø	ê	ê
o	ɔ	o
ê	u	œ
œ	œ	ɔ
œ	ə	ɥ
ɥ	ɥ	œ

Tableau I-9-a
Voyelles

général	Début de mot	Fin de mot
r	d	r
l	l	l
s	s	t
t	p	n
d	k	s
p	m	m
k	t	k
m	n	d
n	v	z
v	f	z
f	r	J
z	z	v
b	b	ʃ
w	ʃ	p
z	g	b
ʃ	w	g
g	J	b
J	z	w

Tableau I-9
Classement des phonèmes

DÉBUT DE MOT

1 DEU	3089
1 LEU	1951
1 LA	1921
1 PA	1899
1 AI	1855
1 SEU	1722
1 LE	1604
1 DE	1596
1 IL	1249
1 EL	1241
1 SA	1145
1 KI	1138
1 POU	1123
1 PR	1057
1 KEU	990
1 AN	894
1 IN	864
1 KBN	820
1 AV	776
1 MAI	768
1 TR	746
1 TOU	717
1 NEU	717
1 SU	709
1 DI	709
1 KB	694
1 BN	694
1 PL	665
1 DU	657
1 UN	635
1 BI	628
1 FEU	620
1 VMO	598
1 JEU	591
1 SN	569
1 DAN	561
1 NU	517
1 MA	510
1 SA	487
1 SAN	480
1 SI	473
1 PEU	473
1 AI	458
1 SE	450
1 VI	421
1 RE	413
1 PAI	406
1 ET	399
1 KA	384
1 SB	362

FIN DE MOT

DEU	2744
IL	1899
LA	1842
LE	1737
LEU	1618
OU	1153
MAN	1116
SEU	1094
KEU	1094
AI	1086
AR	1080
DE	1020
BN	1005
EU	997
FE	960
PA	938
TR	931
WA	931
UN	916
AIL	805
PAI	790
KI	790
NEU	694
SE	657
BN	657
DAN	657
IR	635
SAT	620
DU	620
UI	576
SI	569
IE	561
AT	554
JEU	547
IN	539
MAI	495
UR	495
VOU	495
BR	480
TOU	458
AL	443
ANS	428
JE	428
NOU	421
LI	413
BM	406
VE	406
RE	399
TAN	399
BL	340

Tableau I-10

Fréquences des 50 premiers diphonèmes
de début et de fin de mot (en 100 000^{ème}).

I-7 Validité des résultats.

L'étude statistique rapportée ici a été effectuée à partir d'un n-échantillon de phonèmes extrait d'une population de moyenne α_1 , d'écart-type σ .

Afin d'avoir une estimation des résultats obtenus, on définit pour chaque phonème une variable aléatoire X, pouvant prendre les valeurs 1 (présence du phonème) ou 0 (absence) avec les probabilités respectives p et q = 1-p.

On montre que, pour n suffisamment grand, X tend en loi vers une loi normale $\mathcal{N}(\alpha_1, \frac{\sigma}{\sqrt{n}})$.

Soit encore, en introduisant une variable centrée, réduite

$$u = \frac{X - \alpha_1}{\sigma / \sqrt{n}},$$

$$u \xrightarrow{\text{loi}} \mathcal{N}(0, 1)$$

Cela signifie par exemple que :

$$\Pr \{ |u| > k \} = \frac{2}{\sqrt{2\pi}} \int_k^{+\infty} e^{-\frac{u^2}{2}} du = p_0$$

Pour $p_0 = 1\%$, les tables donnent

$$k = 2,5768$$

soit, en revenant en X :

$$\Pr \left\{ |X - \alpha_1| > 2,57 \frac{\sigma}{\sqrt{n}} \right\} = 0,01$$

Ou encore, en remarquant que $\sigma = \sqrt{p - p^2} = \sqrt{p(1-p)}$

$$\Pr \left\{ |X - \alpha_1| > 2,57 \sqrt{\frac{p(1-p)}{n}} \right\} = 0,01$$

En d'autres termes, cela signifie que, au seuil de 99 %, l'intervalle de confiance des probabilités p calculées est :

$$p \pm 2,57 \sqrt{\frac{p(1-p)}{n}}$$

On peut remarquer que la taille n du corpus intervient par sa racine carrée. L'augmentation de la précision des résultats ne peut donc être obtenue que par une augmentation importante du corpus.

Dans notre cas, la formule précédente donne les intervalles suivants :

p = 8 %	$\longleftrightarrow$	7,7 % $\leq p \leq$ 8,3 %
p = 1 %	$\longleftrightarrow$	0,9 % $\leq p \leq$ 1,1 %
p = 0,5 %	$\longleftrightarrow$	0,42 % $\leq p \leq$ 0,58 %
p = 0,5 %	$\longleftrightarrow$	0,25 % $\leq p \leq$ 0,75 %

On s'aperçoit que les fréquences des phonèmes, ainsi que celles des diphonèmes et triphonèmes les plus courants, sont connues avec une précision suffisante.

CHAPITRE II

UTILISATION DE CONTRAINTES LINGUISTIQUES EN RECONNAISSANCE ANALYTIQUE [D-17].

II-1 Introduction.

Nous avons vu (partie C chapitre III) qu'en reconnaissance analytique de la parole, il s'agissait de retrouver un mot ou une phrase prononcés, à partir d'une chaîne phonémique plus ou moins entachée d'erreurs. Un algorithme simple nous a permis d'éliminer les erreurs les plus flagrantes dans cette chaîne. Mais cela est insuffisant et on peut se demander dans quelle mesure une optimisation, au niveau linguistique, de la chaîne de phonèmes peut être intéressante avant d'appliquer une méthode de recherche lexicale, telle qu'elle a été définie en C chapitre III, ou une analyse syntaxique, telle que nous l'étudierons au chapitre suivant.

L'étude statistique qui vient d'être exposée a montré l'existence de contraintes linguistiques assez strictes au niveau des associations de phonèmes. Nous allons voir dans ce chapitre que l'utilisation de ces contraintes permet, en partie, d'optimiser la chaîne phonémique. Il semble d'ailleurs que l'homme possède en mémoire plus ou moins consciemment ces statistiques, acquises par apprentissage [D-16].

Après avoir précisé le principe de l'utilisation de ces contraintes, nous décrirons le système de reconnaissance au niveau linguistique et nous donnerons quelques exemples de ré-

sultats obtenus en temps différé. Pour terminer, nous préciserons les suites possibles de ces expériences.

Mais on peut déjà remarquer que l'utilisation de données statistiques doit être effectuée très prudemment et que l'amélioration escomptée ne peut être également que statistique.

II-2 Principe d'utilisation des contraintes linguistiques.

II-2-1 Position du problème :

Les données linguistiques qui seront utilisées pour améliorer la reconnaissance acoustique sont les résultats statistiques précédents. L'un des problèmes qui se posent est le stockage de ces données ; nous verrons que cette difficulté a été en partie tournée par l'utilisation de "triphonèmes binaires".

Dans l'historique du chapitre III, partie A, nous avons noté que Denes/A-48/ avait montré l'utilité des probabilités d'occurrence des diphonèmes en reconnaissance acoustique, dans un cas très simplifié. Plus récemment, ces probabilités ont à nouveau été utilisées avec succès, dans le cas de la parole continue/A-107/.

Par ailleurs, des méthodes analogues, utilisant les probabilités concernant les associations de graphèmes, ont été mises au point avec plus ou moins de succès/D-13//D-14//D-15/.

Avant de passer aux expériences proprement dites, nous allons examiner en détail sous quelle forme ces données linguistiques seront utilisées.

II-2-2 Utilisation des diphonèmes :

Les fréquences d'occurrence des diphonèmes correspondant aux 30 phonèmes utilisés dans nos expériences ont été normalisées et stockées dans un tableau à deux entrées ; elles seront utilisées telles quelles pour calculer un "score combiné" pour chaque phonème, en les associant au "score acoustique". Ce score combiné permet d'effectuer, à tout instant dans la chaîne phonémique, un nouveau classement des candidats, qui peut être différent du classement acoustique. Le nombre optimal de candidats considérés à tout instant, ainsi que le pourcentage optimal à accorder aux informations linguistiques dans l'élaboration du score combiné, sont des paramètres du système qui ont été étudiés au cours d'expériences préliminaires (cf. paragraphe II-3).

II-2-3 Utilisation des triphonèmes :

Il serait beaucoup trop onéreux en place mémoire de stocker les fréquences d'apparition des triphonèmes sous la même forme que celles des diphonèmes. Pour réduire les données à stocker, les fréquences d'apparition ont été codées sous forme binaire :

$$\begin{cases} 0 & \text{pour un triphonème qui n'existe pas} \\ 1 & \text{pour un triphonème qui existe.} \end{cases}$$

Ceci permet de coder un triphonème sur un bit et non plus sur un mot mémoire. Le gain de place est donc considérable. La prise en considération des triphonèmes permet donc d'éliminer les candidats phonèmes qui donneraient des triphonèmes inexistantes ; par contre, les triphonèmes n'interviennent pas dans l'élaboration des scores combinés.

II-2-4 Retour sur le système de reconnaissance acoustique :

Les expériences de reconnaissance linguistique ont été effectuées en temps différé sur ordinateur 10070, car la configuration actuelle du T 2000 ne permet pas de mettre simultanément en mémoire le système de reconnaissance acoustique avec le dictionnaire de segments et le système de reconnaissance linguistique avec les données statistiques.

Dans ces conditions, nous avons utilisé le système de reconnaissance acoustique par matrice d'apprentissage, dont il a été fait mention. Ses performances sont nettement inférieures à celles du système fonctionnant en temps réel décrit au chapitre III, partie C, et les résultats obtenus au niveau linguistique s'en ressentent forcément. Cependant, le but est plutôt de tester la validité de la méthode qui sera implantée sur le système temps réel dès que cela sera possible.

II-3 Expériences préliminaires.

Le système de reconnaissance acoustique fournit, pour chacun des 30 phonèmes possibles, un "score acoustique" et il donne comme réponse la classe de phonèmes dont le score est maximal, lorsqu'il fonctionne seul. Dans ces expériences, le niveau acoustique ne donne pas une réponse unique, mais fournit simplement au niveau linguistique une liste des trente phonèmes, classés par score acoustique décroissant. Après élimination des candidats qui, compte-tenu des deux phonèmes précédemment reconnus, donneraient des triphonèmes n'existant pas, un "score combiné" est élaboré, par addition au score acoustique d'un score linguistique, proportionnel à la fréquence d'occurrence du diphonème correspondant. La réponse finale est alors le phonème dont le score combiné est maximal.

Des expériences préliminaires ont été nécessaires pour déterminer deux paramètres fondamentaux du système :

- i) l'importance à accorder aux informations linguistiques par rapport aux informations acoustiques dans l'élaboration du score combiné,
- ii) le nombre de phonèmes candidats qu'il faut conserver au niveau acoustique. En effet, une erreur de reconnaissance sur un phonème peut être extrêmement préjudiciable pour la suite du message à reconnaître et peut aller en s'amplifiant. Pour réduire ce risque de contre-réaction positive en erreur, on restreint le nombre de candidats avant d'utiliser le niveau linguistique, en ne conservant que ceux dont le score acoustique est suffisant.

Une étude systématique a donc été menée en faisant varier ces deux paramètres et en calculant à chaque fois un pourcentage de reconnaissance. Ce calcul n'est d'ailleurs destiné qu'à trouver les valeurs optimales des deux paramètres étudiés. Les données concernant les triphonèmes n'ont pas été utilisées dans ces expériences. Le tableau II-1 donne les pourcentages obtenus, selon les différentes valeurs du nombre de candidats retenus (N) et du coefficient multiplicatif des informations linguistiques (K), ces dernières étant prises, au départ, de l'ordre de grandeur des scores acoustiques.

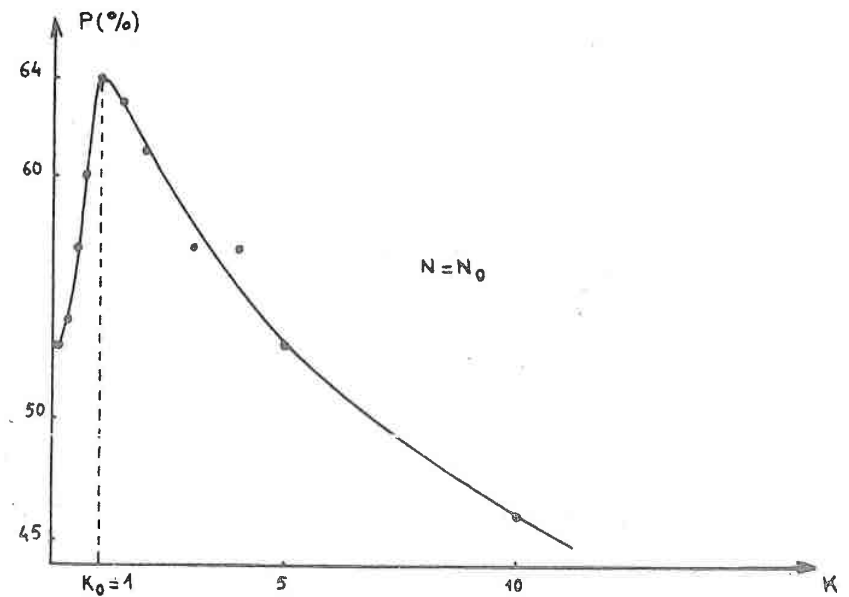
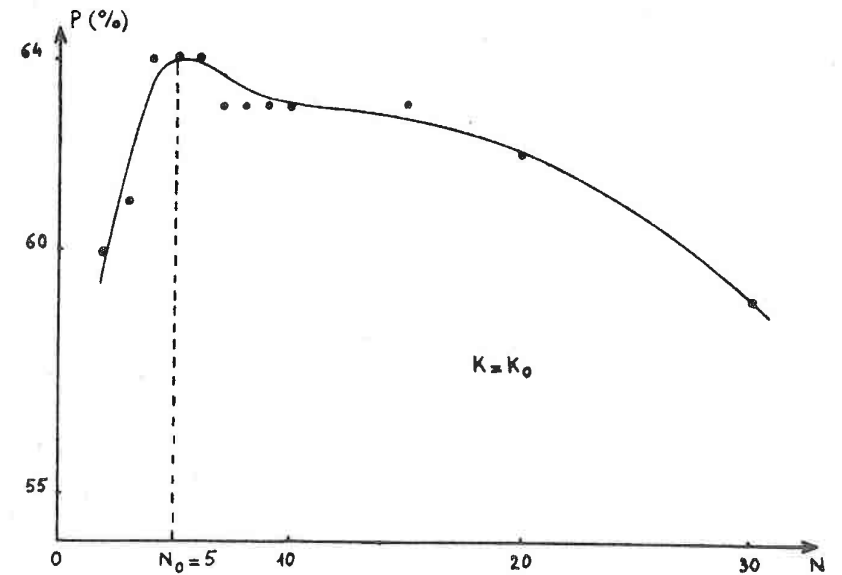
Ces expériences montrent qu'un nombre de candidats égal à 5 et un coefficient multiplicatif de l'ordre de 1 donnent les meilleurs résultats. Ce sont ces valeurs qui ont été adoptées par la suite.

Il est remarquable que le choix du coefficient multiplicatif K soit plus critique que celui du nombre N de candidats à

$K \backslash N$	2	3	4	5	6	7	8	9	10	15	20	30
0,1	53	53	53	53	53	53	53	53	53	53	53	53
0,3	53	54	54	54	54	54	54	54	54	54	54	54
0,5	57	57	57	57	57	56	56	56	56	56	56	53
0,7	58	58	60	60	59	59	59	59	59	59	58	56
1	60	61	64	64	64	63	63	63	63	63	62	59
1,5	58	60	63	63	62	61	61	61	61	61	61	56
2	56	59	61	61	59	59	59	59	59	59	58	54
3	53	56	57	57	56	56	56	56	56	56	54	51
4	53	55	56	57	55	55	56	55	55	55	53	49
5	51	53	53	53	52	52	52	52	52	52	51	48
10	46	46	47	46	45	45	44	44	44	43	43	43
0	53	53	53	53	53	53	53	53	53	53	53	53

Tableau II-1

retenir, comme le montre l'examen des courbes de variation du pourcentage de reconnaissance en fonction du nombre de candidats N (pour la valeur optimale de K) d'une part, et en fonction du coefficient K (pour la valeur optimale de N) d'autre part (figure II-1).



Variation du pourcentage de reconnaissance

Fig II.1

II-4 Résultats expérimentaux pour le système complet.

II-4-1 Schéma général du système :

Le fonctionnement du système à deux niveaux qui vient d'être décrit est schématisé sur la figure II-2. Les retours en arrière du niveau linguistique vers le niveau acoustique qui y sont symbolisés vont être explicités.

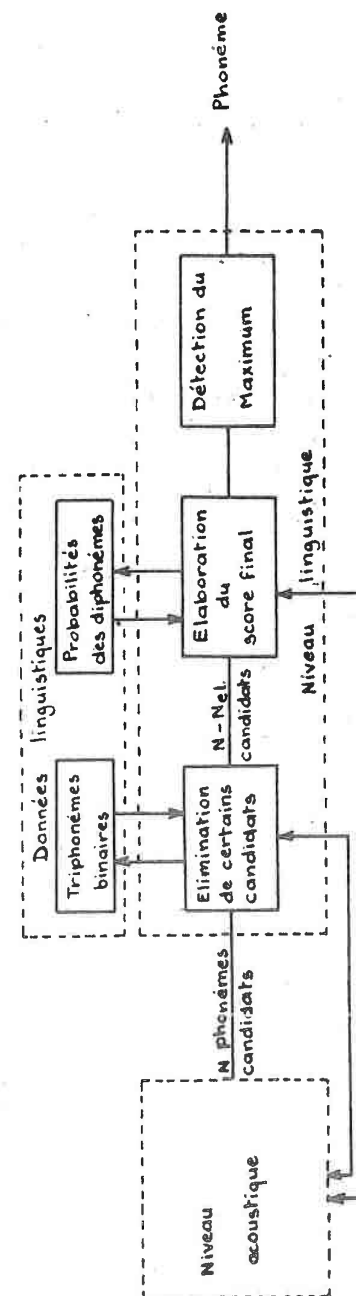
II-4-2 Interaction niveau acoustique-niveau linguistique :

Pour avoir le maximum d'efficacité, l'action du système ne doit pas se réduire à une simple correction au niveau linguistique des résultats acoustiques. Le niveau linguistique doit au contraire pouvoir réagir sur le niveau acoustique grâce à l'existence de retours en arrière, comme indiqué sur la figure II-2. De tels retours sont prévus dans deux cas :

- 1) lorsque tous les phonèmes candidats à un instant donné sont éliminés par considération des triphonèmes binaires,
- 2) lorsque le phonème considéré comme réponse du niveau linguistique possède un score combiné inférieur à un seuil déterminé expérimentalement.

Les retours reviennent à remplacer le phonème précédemment reconnu par le phonème qui avait été classé deuxième; le processus est, le cas échéant, limité à trois cycles, au-delà desquels on peut, par exemple, revenir plus en arrière ou sauter simplement le phonème considéré.

Le premier retour en arrière est évidemment nécessaire sous peine de perdre le phonème considéré. Quant au second,



Fonctionnement du niveau linguistique

Fig II.2

il semble désavantager les associations phonémiques rares qui, au niveau de la chaîne de phonèmes, transportent le plus d'information; mais cela est beaucoup moins vrai au niveau du mot, surtout lorsque la reconnaissance des mots est dirigée par une analyse syntaxique (cf. chapitre III). Cependant, on constate que l'efficacité de ce second "backtrip" est plus grande dans le cas où les données statistiques concernent en particulier le vocabulaire utilisé dans l'expérience et non la langue parlée en général; l'utilisation de la procédure s'avère de toute façon, en moyenne, intéressante.

Ces retours en arrière, sans commune mesure avec ceux rencontrés en perception humaine où ils sont fondamentaux [A-17], constituent cependant une première approche qu'il faudra généraliser dans un système complet de reconnaissance.

II-4-3 Exemples de résultats :

Comme il a déjà été précisé, le système a été testé en temps différé avec un vocabulaire, préanalysé et enregistré sur bande magnétique, de 54 mots seulement prononcés par un seul locuteur (3 séries de 18 mots, dont une pour l'apprentissage du système). Il n'est donc pas question de parler de taux de reconnaissance au sens de ceux fournis pour les systèmes temps réel. Nous donnerons malgré tout quelques chiffres, pour fixer les performances, avec les restrictions qui s'imposent.

Un exemple typique de reconnaissance est représenté sous forme de listing sur le tableau II-2. Il s'agit de la phase de reconnaissance du phonème /i/ dans le mot "étiquette". La réponse initiale du niveau acoustique (/e/) a été corrigée au niveau linguistique, pour donner finalement le phonème /i/ et la reconnaissance du mot s'est ensuite poursuivie.

REPONSES PRECEDENTES : RPAS1 = AI
RPAS2 = I

SCORES ACOUSTIQUES

NUMERO	1	2	3	4	5
SCORE	35463	32902	29424	22013	19786
PHONEME	AI	S	I	A	L

REPONSE ACOUSTIQUE: AI

CANDIDATS ELIMINES PAR TRIPHONEMES

PHONEME	S	L
---------	---	---

SCORES FINAUX

NUMERO	1	2	3
SCORE	63514	52391	39138
PHONEME	I	AI	A

REPONSE FINALE: I

Reconnaissance du phonème /i/
(dans le mot "étiquette")

TABLEAU II.2

Au total, sur le nombre limité de mots utilisés en reconnaissance (36, les 18 premiers mots ayant servi à l'apprentissage du système acoustique), le pourcentage de reconnaissance passe de 53 % à 65 % par utilisation des données statistiques concernant la langue parlée, déterminées au chapitre I. Avec une statistique réduite ne portant que sur le vocabulaire utilisé, le pourcentage passe de 53 % à 70 %, ce qui était prévisible, puisque l'on augmente ainsi les contraintes, mais au détriment de la généralité du système.

Il faut d'autre part remarquer que l'utilisation de telles données linguistiques à caractère statistique est d'autant plus intéressante que les performances au niveau acoustique sont meilleures. En effet, les risques de contre-réaction positive en erreur déjà mentionnés sont alors plus réduits ; c'est pourquoi il sera intéressant de tester le système dans le cadre de la reconnaissance analytique décrite au chapitre III, partie C.

II-5 Poursuite de ces travaux.

Le niveau de reconnaissance linguistique sera implanté en temps réel dès que possible. Cela permettra par ailleurs de chiffrer de façon plus précise l'amélioration obtenue en reconnaissance phonémique.

Dans la reconnaissance de la parole continue, l'utilisation des contraintes linguistiques ne constitue qu'un élément parmi d'autres (la prosodie en est un important [D-19]) qui permettent d'accéder finalement au message prononcé, par une analyse syntaxico-sémantique. En particulier, les données linguistiques peuvent être très utiles dans la détermination des frontières de mots, problème majeur pour l'analyse syntaxique : en effet, l'étude statistique du chapitre précédent nous a fourni des renseignements intéressants concernant les commencements et les terminaisons de mots.

II-6 Conclusion.

Dans le cadre de la conception d'un système hiérarchisé de reconnaissance, ce chapitre nous a permis d'étudier plus précisément le fonctionnement du niveau linguistique. L'utilisation de contraintes statistiques concernant les diphonèmes et les triphonèmes permet, moyennant quelques précautions, une amélioration notable des performances. Cependant, il faudra attendre l'implantation en temps réel du système pour chiffrer exactement cette amélioration.

D'autre part, le système étudié a constitué un bon modèle pour tester les interactions entre niveaux et, en particulier, les retours en arrière.

Enfin, la suite à donner à ces travaux a été rapidement esquissée, les contraintes linguistiques constituant un des indices dont il faut tenir compte pour la reconnaissance du discours continu.

C H A P I T R E I I I

RECONNAISSANCE AU NIVEAU SYNTAXIQUE

III-1 Introduction.

Le passage au niveau syntaxique permet de lever des ambiguïtés et même de corriger des erreurs apparues au niveau acoustique. Cela est tout à fait en accord avec ce que l'on connaît du processus de perception chez l'homme.

Jusqu'à présent, les réalisations dans ce domaine complexe et mal connu, sont restées rares, comme nous l'avons déjà vu (partie A chapitre III). Après quelques considérations sur l'utilisation d'une grammaire, nous verrons une application pratique du système de reconnaissance globale de mots par comparaison dynamique, décrit en partie C, dans le cadre d'un langage de commande numérique de machines-outils. Ce langage, à syntaxe simple, montrera tout l'avantage que l'on peut tirer de l'utilisation des contraintes syntaxiques, indispensable à la reconnaissance du discours continu.

Enfin, nous reviendrons sur la notion de langages adaptés à la communication orale avec la machine, qui est un des points-clés du développement futur de la reconnaissance vocale. Il faut remarquer que le développement des recherches en traitement syntaxique, en vue de la reconnaissance du discours continu, a permis de mettre en évidence la distinction entre deux types de systèmes :

- a) les systèmes à reconnaître la parole (speech recognition systems¹), dans lesquels on cherche à reconstituer élément

par élément (segments, phonèmes, mots, groupes de mots) le message qui a été prononcé ;

b) les systèmes à comprendre la parole ("speech understanding systems"), dans lesquels on s'efforce de retrouver le sens général du message, en vue de l'exécution de l'action demandée, même si des erreurs subsistent au niveau acoustique.

La plupart des systèmes existants sont du type a). La distinction est apparue nettement dans le rapport déjà cité de Newell et al. [A-124], et les Etats-Unis sont les plus avancés dans la mise au point des systèmes du type b) [A-69] [D-9] [D-10] [D-11].

On peut s'attendre à ce que la plupart des projets à venir s'orientent vers ce second type de systèmes qui correspondent à presque toutes les applications déjà citées (partie A chapitre III), hormis la "machine à écrire phonétique", encore utopique.

III-2 Utilisation d'une grammaire en reconnaissance automatique.

Pour nous, une grammaire sera l'ensemble des règles syntaxiques d'une langue, permettant de générer une phrase de cette langue.

Le passage au niveau syntaxique peut grandement améliorer les performances d'un système de reconnaissance : ce fait se retrouve dans le cas de la perception humaine, comme l'ont montré diverses expériences [D-18].

La principale utilisation de l'analyse syntaxique consiste à faire diriger la reconnaissance par la syntaxe. Celle-ci prédit à tout instant les mots, ou du moins le type de mots

qui sont susceptibles d'apparaître, ce qui restreint la reconnaissance à des sous-vocabulaires de taille raisonnable. Dans le cas de phrases à mots prononcés isolément, une analyse "de gauche à droite", sans retours en arrière, suffit. En parole continue, au contraire, il faut prévoir à la fois une analyse "de gauche à droite" et une analyse "de droite à gauche" car les frontières de mots ne sont pas a priori localisées. Nous reviendrons en III-4 sur le rôle particulier joué alors par les mots-clés.

Le cas de la parole continue est d'ailleurs encore plus complexe car, à cette analyse que l'on peut qualifier de montante (permettant d'accéder finalement à la phrase), il est nécessaire d'associer une analyse descendante, partant de la phrase et revenant en arrière, dans le cas où une erreur de syntaxe est détectée.

C'est en particulier ce type d'analyse qui est effectuée le plus souvent par le cerveau humain qui ne descend aux niveaux inférieurs d'analyse que lorsqu'il y a ambiguïté ou que le contexte ne permet pas de prévoir ce qui va être dit.

III-3 Un exemple pratique en commande numérique de machines-outils.

III-3-1 Présentation :

En commande numérique, une machine-outil effectue une succession d'opérations d'usinage sous le contrôle d'une "armoire de commande", sans autre intervention extérieure. La succession d'opérations, conçue par l'homme, est traduite sous forme d'une "phrase" qui est codée puis entrée dans l'armoire de commande, en général sous forme d'un ruban perforé.

Ce ruban perforé, servant à la commande de la machine-outil, doit être préparé au préalable, par transcription de la phrase en clair dans un certain code et perforation. Nous nous sommes proposé de générer automatiquement ce ruban, à partir du perforateur du T 2000, par le simple énoncé oral de la phrase en clair. Il en résulte pour l'utilisateur un gain de temps (la reconnaissance ayant lieu en temps réel) et une plus grande commodité, puisqu'il suffit de prononcer la phrase en séparant simplement les mots, sans pause prolongée entre ces mots.

Le vocabulaire utilisé comporte 36 mots, donnés dans le tableau III-1.

On peut constater plusieurs ambiguïtés au niveau acoustique parmi ces mots : par exemple entre "Z" et "sept", "arrosage" et "alésage", etc... Grâce à l'utilisation de la syntaxe, aucune de ces ambiguïtés ne subsistera en cours d'utilisation.

Une "phrase" de commande numérique fait appel à une grammaire à syntaxe simple, décrite en III-3-2.

1 DEBUT	19 ROTATION
2 STOP	20 ARROSAGE
3 NUMERO	21 CHANGEMENT D'OUTIL
4 PERÇAGE	22 FIN DE PROGRAMME
5 TARAUDAGE	23 FIN DE BLOC
6 ALESAGE	24 ERREUR
7 FIN DE CYCLE	25 BON
8 VITESSE RAPIDE	26 0
9 VITESSE TRAVAIL	27 1
10 RELATIF	28 2
11 ABSOLU	29 3
12 COORDONNEES	30 4
13 X	31 5
14 Y	32 6
15 Z	33 7
16 AVANCE	34 8
17 VITESSE DE BROCHE	35 9
18 FIN DE COORDONNEES	36 EFFACER

Tableau III-1

Le vocabulaire de commande numérique.

III-3-2 Syntaxe et sémantique du langage étudié :

a) Syntaxe.

Le langage utilisé comporte des pseudo-phrases à structure très rigide, dont la syntaxe est donnée par une grammaire du type context-free. Cette grammaire est représentée dans le formalisme de Backus-Naur sur le tableau III-2. On voit que le vocabulaire de 36 mots se subdivise en quelques grands groupes : <chiffres>, <validation>, <indication de coordonnées>, <spécification 1>, <spécification 2>, ... Le nombre total de phrases permises par cette grammaire est de l'ordre de plusieurs millions.

b) Sémantique.

La signification des mots du vocabulaire est le plus souvent non-ambigue. Seuls, quelques termes ont un sens particulier que nous allons expliciter :

"STOP" : cette spécification permet d'interrompre au départ l'énoncé d'une phrase et de se renvoyer au programme conversationnel. Cela évite d'avoir à prononcer une phrase entière lorsque, pour une raison ou une autre, on désire annuler la phrase en cours.

"FIN PROGRAMME" / "FIN BLOC" : ces deux mots terminent indifféremment une phrase et commandent l'écriture sur télétype de la phrase entière reconnue. "FIN BLOC" signifie que la phrase qui vient d'être prononcée sera suivie directement d'une autre ; "FIN PROGRAMME" signifie qu'il n'y aura pas d'autres phrases et on sera donc ramené au programme conversationnel.

"ERREUR" / "BON" : ces deux mots permettent de valider, ou non, la phrase qui vient d'être reconnue et imprimée et commandent la perforation (BON) ou le rejet (ERREUR) de cette phrase.

"COORDONNEES" / "FIN DE COORDONNEES" : ces "mots-clés" ne sont pas insérés dans la phrase ; ils permettent de passer directement au groupe syntaxique qui suit immédiatement (un groupe peut ainsi être vide, par exemple < Spécification 1 >).

<Phrase> ← STOP | <Ordre>
 <Ordre> ← DEBUT | <Libellé> <Spécification 1> <Spécifications>
 <Libellé> ← NUMERO <Nombre 1>
 <Nombre 1> ← <Chiffre> <Chiffre> <Chiffre>
 <Chiffre> ← 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
 <Spécification 1> ← COORDONNEES | <Spécification 2>
 <Spécification 2> ← PERCAGE | TARAUDAGE | ALESAGE | FIN DE CYCLE |
 VITESSE RAPIDE | VITESSE TRAVAIL | RELATIF |
 ABSOLU
 <Spécifications> ← <Coordonnées> <Spécification 3> | <Spécification 3>
 <Coordonnées> ← <Spatiales> FIN | <Spatiales> <Autres> FIN |
 <Autres> FIN
 <Spatiales> ← <x> <y> <z> | <x> <y> | <y> <z> | <x> <z>
 <x> ← X <Nombre 2>
 <y> ← Y <Nombre 2>
 <z> ← Z <Nombre 2>
 <Nombre 2> ← <Chiffre> <Chiffre> <Chiffre> <Chiffre> <Chiffre>
 <Chiffre> <Chiffre>
 <Autres> ← <Avance> | <Avance> <Broche> | <Broche>
 <Avance> ← AVANCE <Nombre 3>
 <Nombre 3> ← <Chiffre> <Chiffre> <Chiffre> <Chiffre> <Chiffre>
 <Broche> ← VITESSE DE BROCHE <Nombre 4>
 <Nombre 4> ← <Chiffre> <Chiffre>
 <Spécification 3> ← ROTATION | ARROSAGE | CHANGEMENT D'OUTIL |
 FIN DE PROGRAMME | FIN DE BLOC

III-3-3 Le programme de reconnaissance :

a) Reconnaissance acoustique.

Nous avons adopté, pour cette expérience, une approche globale de la reconnaissance acoustique. Cela signifie que la phrase à reconnaître doit être prononcée mot par mot, avec un intervalle minimum de 0,3 sec. entre deux mots (attente correspondant à la détection d'une fin de mot). Grâce à l'analyse syntaxique, le nombre de mots "candidats" possibles est limité et le temps de reconnaissance négligeable ; de cette façon, une brève pause, presque naturelle, entre chaque mot, est suffisante ; le choix d'une reconnaissance acoustique globale n'introduit ainsi pas de contraintes gênantes au niveau de l'utilisation du système.

Chaque mot reconnu est stocké en mémoire, à moins que le mot suivant ne soit "EFFACER". Cette spécification permet en effet de supprimer systématiquement le mot précédemment reconnu et de le remplacer par un autre, de la même catégorie syntaxique.

Lorsque l'énoncé d'une phrase est terminé, la reconnaissance (aussi bien acoustique que syntaxique) est également terminée : la phrase reconnue est immédiatement imprimée sur télétype. Nous avons vu qu'elle pouvait alors être validée oralement pour la perforation, ou répétée ("ERREUR").

La perforation, si elle est demandée, s'effectue avec le code symbolique de l'armoire de commande utilisée.

La méthode de reconnaissance utilisée est strictement celle de comparaison dynamique globale qui a été décrite en détail en partie C chapitre II. Le système de reconnaissance comporte simplement en plus l'analyseur syntaxique, tandis que le

programme conversationnel possède la spécification supplémentaire "M" orientant vers la reconnaissance syntaxique. En particulier le vocabulaire de commande numérique peut être aussi bien étudié en reconnaissance avec ou sans analyse syntaxique, comme nous le verrons en III-3-4.

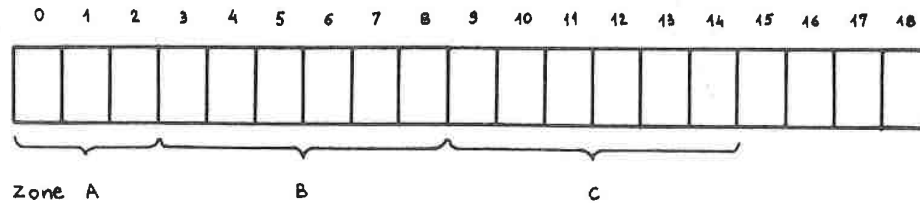
b) Reconnaissance syntaxique.

La reconnaissance syntaxique a pour but ici de dresser la liste des mots susceptibles de suivre le mot reconnu, en accord avec les règles de la grammaire utilisée. En pratique, la forme de l'analyse syntaxique dépend beaucoup de la reconnaissance acoustique qui la précède : type de reconnaissance (globale ou analytique), pourcentage de reconnaissance, etc... L'algorithme développé ici est adapté au cas de "phrases" composées de mots isolés et nous avons admis que la réponse donnée par la reconnaissance acoustique est toujours bonne. Ce dernier point constitue plus qu'une hypothèse commode, les résultats obtenus en partie C chapitre II montrant que le système utilisé est parfaitement fiable pour des vocabulaires d'une dizaine de mots, taille des sous-vocabulaires fournis par l'analyseur syntaxique.

L'analyse syntaxique est par conséquent dirigée de gauche à droite en vue de la prédiction des mots à venir, sans retours en arrière.

La grammaire exposée sur le tableau III-2 a été formalisée en faisant correspondre à chaque mot du vocabulaire, hormis les chiffres, un mot d'un tableau grammaire organisé comme l'indique la figure III-1.

bit n°



Zone A : nombre de chiffres (de 0 à 7) suivant le mot reconnu (CHIF)

Zone B : adresse du premier mot du sous-vocabulaire susceptible de suivre le mot reconnu (N1)

Zone C : nombre de mots composant le sous-vocabulaire (N2)

Figure III-1

Organisation d'un mot du tableau grammair

Toutes les informations nécessaires à la reconnaissance sont rassemblées dans ce tableau et accessibles directement par de simples décalages logiques.

Les 10 chiffres ne comportent pas d'indication syntaxique, ils sont simplement considérés comme des auxiliaires susceptibles de suivre, ou non, un mot et le qualifiant.

Dans ces conditions, le temps de traitement syntaxique est de l'ordre de 0,5 milliseconde, donc négligeable par rapport au temps de traitement acoustique.

Un organigramme schématique du système est donné figure III-2.

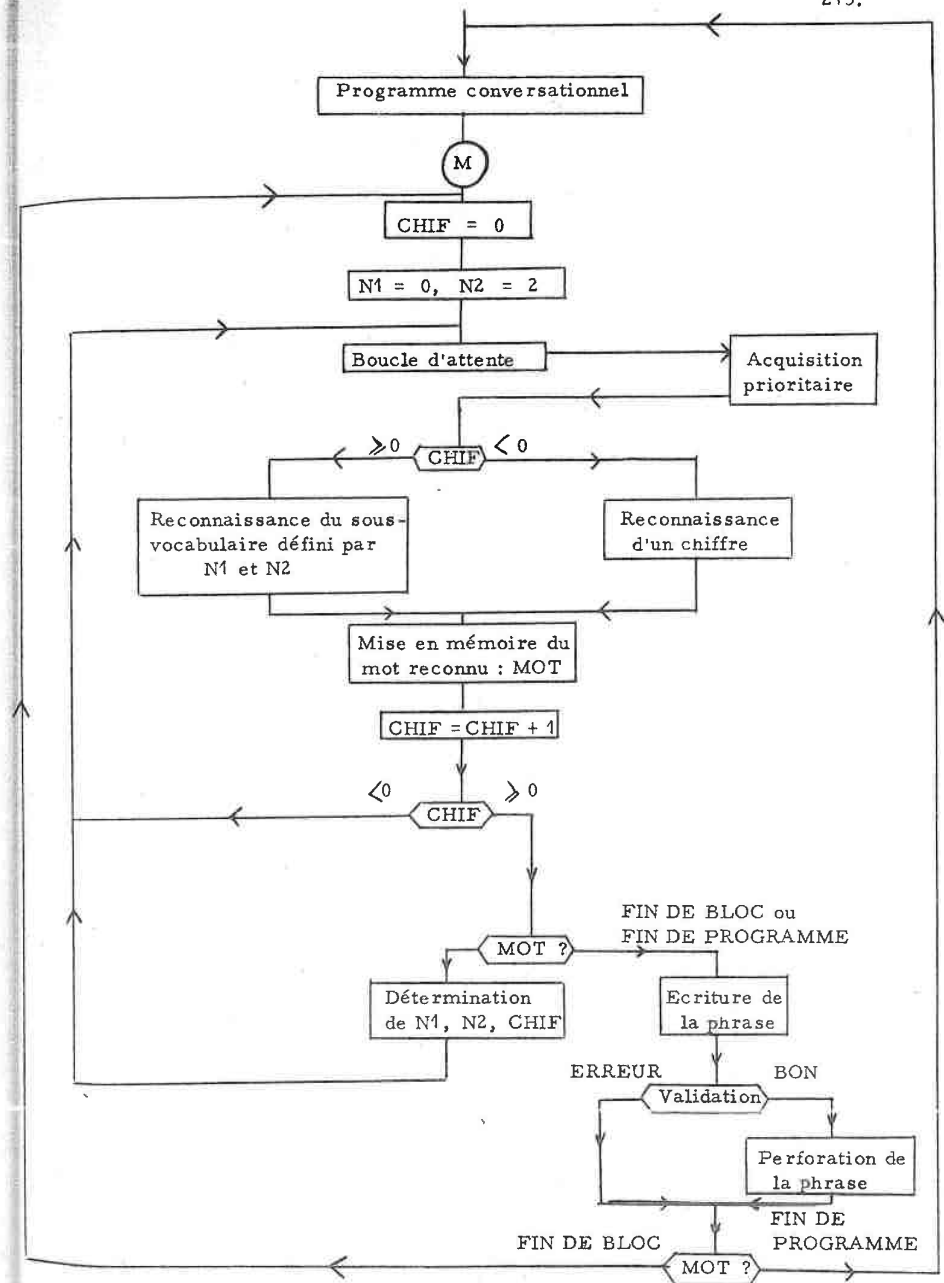


Figure III-2

Schéma de fonctionnement du système de reconnaissance syntaxique.

III-3-4 Résultats obtenus :

Le système a été testé en temps réel sur de nombreuses "phrases" de commande numérique. Le taux de reconnaissance obtenu est de l'ordre de 100 % pour un locuteur masculin. Les très rares erreurs constatées proviennent de confusions entre des chiffres (5 et 7 par exemple) contre lesquelles l'analyse syntaxique ne peut rien. On peut en grande partie supprimer ces confusions en s'orientant, dans le cas des chiffres, vers un algorithme de reconnaissance plus sophistiqué, ce qui ne poserait aucun problème.

Le tableau III-3 reproduit un listing de télétype donnant quelques exemples de reconnaissance. Les différentes phases de l'expérience sont les suivantes :

a) demande d'écriture du nombre de mots constituant le vocabulaire (36), puis appel du programme de reconnaissance syntaxique (M) ;

b) reconnaissance de la phrase :

DEBUT, NUMERO 001, PERÇAGE, X = 0137405,

Y = 4490021, Z = 6103397, ARROSAGE, FIN DE BLOC.

La phrase a été bien reconnue, elle a alors été perforée avec le code numérique de l'armoire de commande (la perforation est obligatoirement couplée à l'impression sur la télétype utilisée) ;

c) reconnaissance de la phrase :

DEBUT, NUMERO 002, TARAUDAGE, AVANCE = 53804,

ROTATION, CHANGEMENT D'OUTIL, FIN DE BLOC.

Comme dans l'exemple b, cette phrase a été bien reconnue et perforée. Le signe "?" a été tapé par l'ordinateur en cours de prononciation de la phrase, correspondant au rejet d'un mot dont l'acquisition a été mauvaise et qui a dû être répété.

d) reconnaissance de la phrase :

DEBUT, NUMERO 003, VITESSE TRAVAIL, VITESSE

DE BROCHE = 05, ROTATION, FIN DE PROGRAMME

On constate une erreur dans la phrase reconnue (la vitesse de broche a été reconnue égale à 07). La phrase n'a pas été validée, grâce à l'ordre "ERREUR".

e) reconnaissance de la même phrase qu'en d :

Cette fois, la reconnaissance est bonne et la perforation est effectuée.

Ensuite, le système se ramène au programme conversationnel (#) pour une autre expérience, car la dernière phrase reconnue se terminait par "FIN PROGRAMME".

f) expérience de reconnaissance acoustique sans syntaxe (R) :

Sur le vocabulaire complet de 36 mots (la réponse du système est suivie de la longueur du mot prononcé, en nombre d'échantillons, donc en centisecondes).

Les mots "DEBUT", "PERÇAGE", "AVANCE" ont été bien reconnus. "ALESAGE" a été reconnu à la place de "ARROSAGE". C'est un cas typique d'erreur évitée par l'analyse syntaxique.

Pour un deuxième locuteur masculin, le taux de reconnaissance de phrases demeure très bon, mis à part des erreurs sur les chiffres, plus nombreuses que pour un seul locuteur.

III-4 Retour sur les langages oraux de communication homme-machine.

L'extension de l'utilisation d'analyseurs syntaxiques en reconnaissance vocale, c'est-à-dire finalement le cheminement

l'entrée orale en ordinateur d'un programme complet ne présente pas d'intérêt. La programmation orale doit plutôt être envisagée comme un "outil" de travail dans certains cas particuliers : consultation de banques de données, de fichiers, calcul bien délimité (les machines à calculer orales mentionnées en partie A chapitre III en constituant une première approche).

Cependant, l'étude de l'entrée orale d'un programme en langage de programmation écrite est actuellement une bonne introduction à ces problèmes, dans la mesure où elle peut aboutir à la définition de normes concernant les langages oraux. C'est dans cette optique que nous avons abordé ce problème depuis quelques mois déjà.

III-5 Conclusion.

Nous n'avons fait, dans ce chapitre, qu'aborder le problème de l'utilisation d'une syntaxe en reconnaissance vocale, car l'aboutissement de ces recherches doit être la reconnaissance du discours continu.

L'exemple d'un langage de commande numérique de machine-outils, que nous avons traité, nous a cependant permis de montrer que cette utilisation permettait d'augmenter le taux de reconnaissance au niveau de la phrase, même dans le cas d'une prononciation mot par mot. La reconnaissance est, en quelque sorte, dirigée par la syntaxe qui prédit les mots qui peuvent apparaître et qui, éventuellement, détecte les erreurs commises.

La généralisation de ces études doit conduire à la conception de langages spécifiques à la communication orale entre l'homme et la machine, à la fois adaptés au mode oral et suffisamment rigides pour que les performances restent bonnes.

CONCLUSION

Le thème principal de nos recherches a été la reconnaissance automatique, en temps réel, de la parole et les problèmes annexes de traitement du signal, de paramétrisation, etc...

Ce travail a été entrepris dans l'optique de la conception d'un système hiérarchisé de reconnaissance, dont les différents niveaux -acoustique, linguistique, syntaxique- ont été étudiés, ainsi que leurs connexions.

Dresser un bilan des sujets abordés c'est aussi donner les orientations actuelles de nos recherches, car ce mémoire ne constitue qu'une étape dans un travail de longue haleine, pour aboutir à la reconnaissance automatique du discours continu.

L'exposé des méthodes d'analyse et de paramétrisation du signal vocal a permis de rappeler le principe de l'analyseur spectral utilisé dans nos expériences, qui consiste à effectuer un codage binaire du spectre à court-terme de la parole. Une méthode mathématique de paramétrisation par compression d'information a été exposée. Ce travail est actuellement poursuivi en temps réel et doit permettre la réalisation d'un système électronique de paramétrisation.

Le niveau acoustique de reconnaissance a été abordé par les deux approches possibles, globale et analytique :

- en reconnaissance globale, nous avons mis au point une méthode puissante de normalisation temporelle par comparaison dynamique, d'efficacité supérieure à celles des méthodes non dynamiques utilisées jusqu'à présent. Ces recherches doivent conduire maintenant à la réalisation d'un système autonome de reconnaissance de mots, dans le cadre de la conception de nouvelles structures de machines. En même temps, nous recherchons des méthodes d'apprentissage plus performantes, en vue de la généralisation du système ;

- la reconnaissance analytique, qui utilise les résultats d'une segmentation sur critères acoustiques du signal vocal, est également fondée sur une méthode de comparaison dynamique. Le problème de la recherche lexicale d'un mot à partir d'une chaîne phonémique a également été abordé. Cette seconde approche a montré l'insuffisance de l'analyseur spectral actuel, mal adapté à distinguer entre eux les sons élémentaires. Un nouvel analyseur, fournissant plus d'informations, est actuellement en cours de réalisation.

Enfin une étude comparée des deux approches a été menée. L'approche globale, plus simple, est adaptée à la reconnaissance de vocabulaires limités, -quelques centaines de mots au plus-, tandis que l'approche analytique, plus générale, est préférable dans le cas de vocabulaires plus importants ; cette dernière approche est, de plus, la seule utilisable dans le cas de la parole continue.

L'utilisation des niveaux hiérarchiques supérieurs de reconnaissance a été envisagée sous deux points de vue :

- l'optimisation de chaînes de phonèmes par contraintes linguistiques (statistiques sur les associations phonémiques) ;
- la reconnaissance de phrases d'un langage à syntaxe simple à l'aide d'une reconnaissance acoustique globale dirigée par une analyse syntaxique. Le processus donne des performances

presque parfaites. Ces recherches portent maintenant sur des langages plus évolués, en vue de l'étude des langages oraux de communication homme-machine.

Les différents systèmes de reconnaissance mis au point fonctionnant en temps réel dans des conditions normales de travail se placent dans le cadre de l'amélioration du dialogue homme-machine. Nous pensons avoir contribué à cette amélioration et nous espérons surtout pouvoir continuer à le faire.

BIBLIOGRAPHIE

Abréviation.

JASA : Journal of the Acoustical Society of America.

Introduction.

- [I-1] R. FATEHCHAND "Machine recognition of spoken words" in "Advances in Computers" Academic Press - New-York 1960.
- [I-2] T. MARILL "Automatic recognition of speech" I R E Tr. on H. F. E., 2, pp. 34-38, 1961.
- [I-3] N. LINDGREN "Machine recognition of human language" part 1. IEEE Spectrum, March, pp. 114-36, 1965.
- [I-4] S. R. HYDE "Speech recognition : State of the art in 1972" Int. Conf. on Machine Perception of patterns and pictures - Teddington - April 1972.

PARTIE A.

- [A-1] A. MOLES "Physique et technique du bruit", Dunod, Paris, 1952
- [A-2] A. MOLES "Théorie de l'information et perception esthétique" Flammarion - Paris - 1958.
- [A-3] L. PIMONOW "Vibrations en régime transitoire" Dunod - Paris - 1962.
- [A-4] J. I. FLANAGAN "Speech : Analysis, Synthesis and perception" 2^d édit. Springer Verlag - Berlin - 1972.
- [A-5] E. GARDE "La Voix" P. U. F. 4^{ème} édit. 1970.
- [A-6] G. FANT "Acoustic theory of speech production" Mouton - the Hague - 1970.

- [A-7] B. MALMBERG "La Phonétique" P. U. F. 9ème édit. 1971.
- [A-8] R. JAKOBSON et al. "Preliminaries to speech analysis" Tech. Rept. Acoustic Labs. MIT. Cambridge - USA - n° 13 - 1952.
- [A-9] G. E. PETERSON, H. L. BARNEY "Control methods used in the study of vowels" JASA, 24, n° 2, pp. 175-84, 1952.
- [A-10] P. DELATTRE "Le jeu des transitions de formants et la perception des consonnes" Proc. of 4th Int. Congress. of Phonetic Sciences - Helsinki - 1961 - pp. 407-17.
- [A-11] G. Von BEKESY "Experiments in hearing" Mac Graw Hill, New-York, 1960.
- [A-12] A. MOLES, B. VALLANCIEN "Communications et langages" Gauthier-Villars - Paris - 1963.
- [A-13] E. LEIPP "Le problème de l'intelligibilité de la parole" Bulletin du GAM, n° 37 - 1968.
- [A-14] A. A. MOLLER "Acoustic reflex in man" JASA, 34, pp. 1524-34, 1962.
- [A-15] J. C. R. LICKLIDER "On the process of speech perception" JASA, 24, n° 5, pp. 590-4, 1952.
- [A-16] H. FLETCHER "Speech and hearing in communications" Van Nostrand Co. Princeton - 1965.
- [A-17] J. C. RISSET "Sur l'analyse, la synthèse et la perception des sons, étudiés à l'aide de calculateurs électroniques" Thèse d'Etat - Orsay - 1967.
- [A-18] D. GABOR "Theory of communication" Jour. IEEE, 93, pp. 429-57, 1946.
- [A-19] E. A. FEIGENBAUM, J. FELDMAN "Computers and thought" Mac Graw Hill - New-York - 1963.
- [A-20] G. S. SEBESTYEN "Decision-making processes in pattern recognition" Mac Millan - New-York - 1962.

- [A-21] N. NILSSON "Learning machines. Foundations of trainable pattern classifying systems" Mac Graw Hill - New-York - 1965.
- [A-22] Y. C. HO, A. K. AGRAWALA "On pattern recognition algorithms - Introduction and survey" IEEE Tr. A. C., 13, n° 6, pp. 676-90, 1968.
- [A-23] F. ROSENBLATT "Principles of Neurodynamics" Spartan Books, Washington, 1962.
- [A-24] T. M. COVER, P. E. HART "Nearest Neighbor pattern classification" IEEE Tr. Inf. Th., IT-13, n° 1, pp. 21-6, 1967.
- [A-25] "Special Issue on Syntactic Pattern Recognition" Pattern Recognition, 4, n° 1, 1972.
- [A-26] R. S. FU "Stochastic automata, stochastic languages and pattern recognition" J. of Cybernetics - 1, n° 3, pp. 31-49, 1971.
- [A-27] N. CHOMSKY "Syntactic Structures" Mouton. The Hague. 1957
- [A-28] J. C. SIMON, C. ROCHE, G. SABAH "On automatic generation of pattern recognition operators" IEEE Int. Conf. on Cybernetics - Washington USA - Oct. 1972.
- [A-29] J. P. HATON "Reconnaissance de la parole : bilan de vingt années de recherches et tendances actuelles" Ann. Telecom. 27, n° 3-4, pp. 77-88, 1972.
- [A-30] H. F. OLSON "Speech processing systems" IEEE Spectrum, Feb., pp. 90-102, 1964.
- [A-31] S. R. HYDE "Recognition of speech by machine" Post. off. El. Eng. J., 62, n° 2, pp. 100-4, 1969.
- [A-32] E. E. DAVID, O. G. SELFRIDGE "Eyes and ears for computers" IRE Tr., 41, pp. 1093-101, 1962.
- [A-33] J. PFEIFFER "The Computer Age" Harper and Row. New-York - 1965.
- [A-34] C. WEITZMAN "Voice recognition and response systems" Datamation, 15, n° 12, pp. 165-70, 1969.

- [A-35] W. A. LEA "Evaluating speech recognition work" JASA, 47, n° 6, pp. 1612-14, 1970.
- [A-36] A. DEWEZE "Techniques de reconnaissance automatique" Automatisme, IX, n° 3, pp. 95-102, 1964.
- [A-37] W. A. LEA "Establishing the value of speech communication with computers" IEEE Tr. AU, 16, pp. 184-97, 1968.
- [A-38] BELL Laboratories "Experimental voice-controlled device gives 'command' performances" JASA, 50, n° 6, p. 1450, 1971.
- [A-39] J. W. GLENN, M. H. HITCHCOCK "with a speech pattern classifier, computer listens its master's voice" Electronics, 44, n° 10, pp. 84-9, 1971.
- [A-40] J. DREYFUS-GRAF "Sonograph and sound Mechanics" JASA, 22, pp. 731-39, 1950.
- [A-41] J. DREYFUS-GRAF "Actuateur phonétique (Phonacteur) et calculateur" Revue d'Acoustique, 9, pp. 10-6, 1970.
- [A-42] C. P. SMITH "A phoneme detector" JASA, 23, pp. 446-51, 1951.
- [A-43] K. H. DAVIS et al. "Automatic recognition of spoken digits" JASA, 24, p. 637, 1952.
- [A-44] H. DUDLEY, S. BALASHEK "Automatic recognition of phonetic patterns in speech" JASA, 30, pp. 721-32, 1958.
- [A-45] H. F. OLSON, H. BELAR "Phonetic typewriter" JASA, 28, pp. 1072-81, 1956.
- [A-46] H. F. OLSON, H. BELAR "Phonetic typewriter III" JASA, 33, pp. 1610-15, 1961.
- [A-47] J. WIREN, H. L. STUBBS "Electronic binary system for phoneme classification" JASA, 28, pp. 1082-91, 1956.
- [A-48] P. DENES "The design and operation of the mechanical speech recognizer at University College, London" J. Brit. Inst. Radio. Eng., 19, pp. 219-29, 1959.

- [A-49] T. SAKAI, S. DOSHITA "The Phonetic Typewriter", Inform. Proc. Proc. of IFIP Congress 1962, pp. 445-49, 1963.
- [A-50] J. W. and C. D. FORGIE "Results obtained from a vowel recognition computer program" JASA, 31, pp. 1480-89, 1959.
- [A-51] J. W. and C. D. FORGIE "A computer program for recognizing the English fricative consonants /f/ and /θ/" 4th Int. Cong. on Acoustics - Aug. 1962.
- [A-52] J. E. K. SMITH, L. KLEM "Vowel recognition using a multiple discriminant function" JASA, 33, p. 358, 1961.
- [A-53] P. DENES, M. V. MATHEWS "Spoken digit recognition using time-frequency pattern matching" JASA, 32, pp. 1450-55, 1960.
- [A-54] G. L. SHULTZ "Investigation procedures for speech recognition" Proc. Seminar. on Speech Comp. and Proc. Cambridge - Mass - Sept. 1959.
- [A-55] P. N. SHOLTZ, R. BAKIS "Spoken digit recognition using vowel-consonant segmentation" JASA, 34, pp. 1-5, 1962.
- [A-56] J. SUZUKI, K. NAKATA "Phonemic classification and recognition of Japanese monosyllables" J. Radio Res. Lab., 10, n° 49, pp. 269-83, 1963.
- [A-57] J. SUZUKI et al. "Formant frequency extraction by the method of moment calculations" J. Radio. Res. Lab., 9, n° 46, pp. 469-85, 1962.
- [A-58] B. WIDROW et al. "Practical applications for adaptive data processing systems" Revue A, X, n° 1, pp. 27-39, 1968.
- [A-59] L. R. TALBERT et al. "A real-time adaptive speech recognition system" Tech. Rept. n° 6760-1, Stanford Elec. Lab. May 1963.
- [A-60] J. F. HEMDAL, G. W. HUGHES "A feature based computer recognition program for the modeling of vowel perception" Proc. of Symp. on Models for the perception of speech and visual forms - AFCRL - Nov. 1964.

- [A-61] W. BEZDEL, H. J. CHANDLER "Results of an analysis of recognition of vowels by computer using zero-crossings data" Proc. IEE, 112, nov. 1965.
- [A-62] J. H. KING, C. J. TUNIS "Some experiments in spoken word recognition" IBM Journal, 10, n° 1, pp. 65-79, 1966.
- [A-63] B. GOLD "Word recognition computer program" Tech. Rept n° 452, Lincoln Lab. MIT, 1966.
- [A-64] S. PRUZANSKY "Pattern matching procedure for automatic talker recognition" JASA, 35, pp. 354-58, 1963.
- [A-65] B. S. ATAL "Automatic speaker recognition based on pitch contours" JASA, 45, pp. 309, 1969.
- [A-66] J. J. WOLF "Acoustic measurements for speaker recognition" Prog. Rept n° 94, R. L. E., MIT, pp. 216-22, 1969.
- [A-67] Dr W. S. "Identification des personnes par la spectrographie vocale" Automatisme, VIII, n° 7-8, pp. 280-81, 1963.
- [A-68] J. L. FLANAGAN, R. C. LUMMIS "Automatic identity verification by voice characteristics" JASA, 53, n° 4, p. 1188, 1973.
- [A-69] D. R. REDDY et al. "The C-MU Speech Recognition Project" Proc. IEEE Syst. Sci. Cyb. Conf. 1971.
- [A-70] D. G. BOBROW, D. H. KLATT "A limited speech recognition system" Proc. AFIPS Fall Joint Comp. Conf. Thompson, t; 33, pp. 305-18, 1968.
- [A-71] M. MEDRESS "Computer recognition of single-syllable english words" 77th Meeting ASA, 1969.
- [A-72] R. C. OHLENDORF, C. L. COATES "Recognition of spoken digits utilizing sequential patterns" Rept. CFSTI, N.69-30170, Texas Univ. 1968.

- [A-73] J. I. MAKHOUL "Speaker-machine interaction in a limited speech recognition system" MIT, RLE quart. Prog. Rep., n° 96, pp. 195-202, 1970.
- [A-74] G. D. EWING, J. F. TAYLOR "Computer recognition of speech using zero-crossing information" IEEE Tr. AU, 17, n° 1, pp. 37-40, 1969.
- [A-75] J. C. MILLER et al. "An adaptive speech recognition system operating in a remote time-shared computer environment" IEEE Tr. AU, 18, n° 1, pp. 26-32, 1970.
- [A-76] C. F. TEACHER et al. "Experimental, limited vocabulary, speech recognizer" IEEE Tr. AU, 15, n° 3, pp. 127-30, 1967.
- [A-77] L. R. FOCHT "The single equivalent formant" IEEE Int. Conf. Commun. Conf. Digest. p. 108, 1966.
- [A-78] M. R. ITO, R. W. DONALDSON "Zero-Crossings measurements for analysis and recognition of speech sounds" IEEE Tr. AU, 19, n° 3, pp. 235-42, 1971.
- [A-79] R. S. VONUSA, L. R. FOCHT "A spoken word recognizer incorporating semiautomatic speaker adaptation" IEEE Conf. on feature extraction and selection in pattern recognition - Argonne - Ill. oct. 1970.
- [A-80] Y. KATO et al. "Speech processing equipment ; advanced studies" NEC Res. and devel., 20, pp. 1-7, Janv. 1971.
- [A-81] S. CHIBA "Spoken word recognition by multiple linear separation" 6th Int. Cong. on Acoustics. Tokyo, Aug. 1968.
- [A-82] A. KUREMATSU, S. INOUE "Speech recognition with time-normalized frequency pattern" 6th Int. Cong. on Acoustics, Tokyo, Aug. 1968.
- [A-83] S. ITAHASHI et al. "Discrete-Word recognition utilizing a word dictionary and phonological rules" IEEE Tr. AU, 21, n° 3, pp. 239-49, 1973.

- [A-84] H. SAKOE, S. CHIBA "A dynamic programming approach to continuous speech recognition" 7th Int. Cong. on Acoustics - Budapest - 1971.
- [A-85] M. CARTIER, R. CARRE, J. Y. GRESSER "Compte-rendu d'une mission d'étude en URSS : analyse, synthèse, reconnaissance automatique et perception de la parole - déc. 1971 - CEI/7 - TMA/5.
- [A-86] V. M. VELICHKO, N. G. ZAGORUJKO "Avtomaticheskoe raspoznavanie ogranichenovo nabora oustnich kommand" Vychisl. Sist. SSR, 36, pp. 101-10, 1969.
- [A-87] G. Y. VYSOTSKIY et al. "An experiment in oral control of a computer" Eng. Cybern., n° 2, pp. 320-27, 1970.
- [A-88] E. T. SEMENOVA "Analysis of speech recognition criteria" For. Tech. Div. Wright - Patterson AFB, n° AD 680-571, april 1968.
- [A-89] V. N. TRUNIN-DONSKOI "Recognition of spoken word ensembles with the aid of a digital computer" in "Works on technical Cybernetics" Moscou - URSS - 1969.
- [A-90] L. L. MYASNIKOV et al. "Infrasonic cues for the automatic recognition of speech sounds" Soviet Phys-Acoust., 14, pp. 522-24, 1969.
- [A-91] W. KLEIN et al. "Vowel spectra, vowel spaces and vowel identification" JASA, 48, pp. 999-1009, Oct. 1970.
- [A-92] L. C. W. POLS "Real-time recognition of spoken words" IEEE Tr. Computers, 20, n° 9, pp. 972-78, 1971.
- [A-93] N. I. M. DE WIT "Sprakherkenning middels foneemspektra" Technische Hogeschool - Delft - Juin 1973.
- [A-94] J. P. HATON "Compte-rendu d'une mission aux Pays-Bas : Etudes sur la Parole et Reconnaissance de formes" Mai 1973 - GALF.
- [A-95] E. ZWICKER et al. "Erkennung gesprochener Zahlworte mit Funktionsmodell und elektronischer Rechenanlage" Kybernetik, 3, n° 6, pp. 267-72, 1967.

- [A-96] D. ANKE, P. HOESCHELE "Un système simple de reconnaissance des chiffres 0 à 9" (en allemand) Kybernetik, 4, pp. 228-34, 1968.
- [A-97] H. G. TILLMANN et al. "DAWID 1 - Ein Beitrag zur Automatischen Spracherkennung" 5th Int. Cong. on Acoustics - Liège - Sept. 1965.
- [A-98] L. FAVELLA et al. "The probabilistic classification of time sequences and its application to the recognition of human voice" Il Nuovo Cimento, X, n° 36, pp. 1023-34, 1965.
- [A-99] L. GILLI, A. R. MEO "Sequential system for recognizing spoken digits in real time" Acustica - 1969.
- [A-100] A. R. MEO, G. RIGHINI "Riconoscitore istantaneo di suoni vocalici" Alta Frequenza, XXXIV, n° 4, pp. 256-63, 1965.
- [A-101] R. DE MORI et al. "A flexible real-time recognizer of spoken words for man-machine communication" Int. I. on Man-Machine Studies, 2, pp. 317-26, 1970.
- [A-102] J. N. SHEARME, P. F. LEACH "some experiments with a simple word recognition system" IEEE Tr. AU, 16, n° 2, pp. 256-61, 1968.
- [A-103] W. F. GILMOUR "Electronic speech recognition" Wireless World, 75, pp. 76-80, Feb. 1969.
- [A-104] J. P. HATON, G. MERCIER "Rapport de mission effectuée en Angleterre en Avril 1972" CNET-LEA.
- [A-105] M. J. UNDERWOOD et al. "The evaluation of certains parameters for the automatic recognition of spoken words" Proc. Int. Conf. on Mach. Perc. of patterns and pictures - Teddington - G. B. pp. 117-25, April 1972.
- [A-106] J. H. WARREN "A dynamic pattern matching algorithm with applications to automatic speech recognition" Proc. Int. Conf. on Mach. Perc. of patterns and pictures - Teddington - G. B. pp. 141-50 - April 1972.

- [A-107] P. D. GREEN, W. A. AINSWORTH "Towards the automatic recognition of spoken Basic English" Proc. Int. Conf. on Mach. Perc. of patterns and pictures - Teddington - G. B. pp. 161-68 - April 1972.
- [A-108] G. FERRIEU et al. "Synthèse et reconnaissance de la parole par ordinateur" L'écho des recherches, pp. 30-42, Juin 1968.
- [A-109] J. P. TUBACH "Reconnaissance de chiffres parlés" Colloque Int. sur la Télé-informatique - Paris - Mars 1969.
- [A-110] S. CASTAN, G. PERENNOU "Séparation linéaire des formes par apprentissage" C.R. Ac. Sci. 262, pp. 52-4, 1966.
- [A-111] J. Y. GRESSER, G. MERCIER "Exemple de reconnaissance automatique de la parole" Commutation et électronique, 32, pp. 48-63, 1971.
- [A-112] J. P. TUBACH "Reconnaissance automatique de la parole" Thèse d'Etat - Université de Grenoble - Juillet 70.
- [A-113] G. MERCIER "Approximation stochastique et reconnaissance d'un vocabulaire limité" Ann. Telecom., 25, n° 5-6, pp. 207-216, 1970.
- [A-114] P. ALINAT "Reconnaissance de phonèmes au moyen d'une cochlée artificielle" Thèse - Université de Nice - Juin 1973.
- [A-115] M. HALLE, K. STEVENS "Speech recognition : a model and a program for research" IRE Tr. I. T. 8, n° 2, pp. 155-59, 1962.
- [A-116] H. J. ZADELL et al. "Acoustic recognition by analog-feature abstraction techniques" Proc. Symp. On models for perc. speech and visual forms - Boston - Nov. 1964.
- [A-117] J. HERAULT "Représentation électronique du neurone et modèles de réseaux nerveux" Thèse - Université de Grenoble - 1970.
- [A-118] J. J. SPARKES "Pattern recognition and a model of the brain" Int. J. Man-Mach. Studies, 1, n° 3, pp. 263-78, 1969.

- [A-119] D. R. HILL "An Esoteric approach to some problems in automatic speech recognition" Int. J. Man-Mach Stud., 1, n° 1, pp. 101-21, 1969.
- [A-120] J. DANSAC, M. CAPUANO "Quelques techniques simples de traitement de l'information par procédés optiques" Onde Electrique, 48, n° 492, pp. 205-8, 1968.
- [A-121] D. LEVERINGTON "Speech recognition using optical filtering" IERE Conf. Proc., n° 14, pp. 298-9, 1969.
- [A-122] Q. C. WILSON "Optical processing in speech recognizer development" IEEE Tr. AU, 16, n° 2, pp. 280-7, 1968.
- [A-123] M. B. HERSCHER, R. B. COX "An adaptive isolated word recognition system" Proc. 1972 Conf. On Speech Comm. and Proc. Boston - pp. 89-92.
- [A-124] A. NEWELL et al. "Speech understanding systems : final report of a study group" May 1971.
- [A-125] R. ALTER "Utilization of contextual constraints in automatic speech recognition" IEEE Tr. AU, 16, n° 1, pp. 6-11, 1968.
- [A-126] D. R. REDDY, A. E. ROBINSON "Phoneme-to-Grapheme translation of English" IEEE Tr. AU, 16, n° 2, pp. 240-6, 1968.
- [A-127] J. S. LIENARD "Analyse, synthèse et reconnaissance automatique de la parole" Thèse d'Etat - Université de Paris VI - Avril 1972.

PARTIE B.

- [B-1] J. P. HATON "A propos de la redondance du signal de parole". L. E. A. Note interne - mars 1973.
- [B-2] W. ENDRES "Probleme der Sprachübertragung" N. T. Z. pp. 74-80 - 1957.
- [B-3] J. P. HATON, M. LAMOTTE "Etude statistique des phonèmes et des diphonèmes dans le français parlé" Revue d'Acoustique, 4, n° 16, pp. 258-62, 1971.
- [B-4] C. SHANNON "A mathematical theory of communication" Bell Syst. Tech. J. pp. 379-423 - 1948.
- [B-5] J. W. COOLEY, J. W. TUKEY "An algorithm for the machine calculation of complex Fourier Series" Math. Comput., 19, pp. 297-301, Avril 1965.
- [B-6] Special Issue on Fast Fourier Transform - I3E Tr. AU, 17, n° 2, 1969.
- [B-7] H. DUDLEY "Remaking Speech" JASA, 11, pp. 169-77, 1939.
- [B-8] M. R. SCHROEDER "Vocoders : analysis and synthesis of speech" Proc. I3E, 54, n° 5, pp. 720-34, 1966.
- [B-9] M. LECOURS, J. J. SPARKS "Adaptive spectral analysis for speech sound recognition" IE3 Tr. AU, 16, n° 4, pp. 523-5, 1968.
- [B-10] M. J. DUGRAVOT "Prétraitement numérique du signal vocal dans le domaine spectral" Thèse de Spécialité Univer. de Nancy - Octobre 1969.
- [B-11] M. J. DUGRAVOT, M. LAMOTTE, J. P. HATON, J. BREMONT "Recherches actuelles sur l'extraction de caractéristiques et la reconnaissance de la voix parlée" Automatisme, - XV, n° 12, pp. 646-649, 1969.

- [B-12] Proc. 1972 Symp. Appl. Walsh Functions - Washington D. C.
- [B-13] J. MAKHOUL "Spectral Analysis of Speech by linear prediction" Int. Conf. on Speech Com. and Proc. Newton Mass. April 1972.
- [B-14] E. PETERSON "Frequency detection and speech formants" JASA, 23, pp. 668-74, 1951.
- [B-15] P. VICENS "Aspects of speech recognition by computer" Ph. D. Thesis - Stanford University - April 1969 - Memo A. I. 85.
- [B-16] D. R. REDDY "Segmentation of Speech Sounds" JASA, 40, pp. 307-12, 1966.
- [B-17] G. A. SITTON "Acoustic segmentation of speech" Int. J. Man-Mach. Stud., 2, n° 1, pp. 61-102, 1970.
- [B-18] J. T. TOU "Feature extraction in pattern recognition" Pattern Recognition, 1, 1, pp. 3-11, 1968.
- [B-19] J. N. MAKSYM "Real-time pitch period extraction by adaptive prediction of the speech waveform" Int. Conf. on Speech Com. and Proc. Newton Mass April 1972.
- [B-20] A. H. MAISSIS "Méthode d'extraction du pitch" Journées d'étude sur la parole - GALF - Lannion Juin 1972.
- [B-21] A. N. MUCCIARDI, E. E. GOSE "A comparison of seven techniques for choosing subsets of pattern recognition properties" I3E Tr. Comp., 20, n° 9, pp. 1023-31, 1971.
- [B-22] J. P. HATON, M. LAMOTTE "Extraction de paramètres et compression de l'information ; application à la reconnaissance vocale" C. R. Ac. Sci., 273, pp. 415-8, 1971.
- [B-23] S. WATANABE "Karhunen-Loeve expansion and factor analysis : Theoretical remarks and applications" Proc. 4th Cong. on Inform. Theory - Prague - pp. 635-60 - 1965.

- [B-24] T. Y. YOUNG "The reliability of linear feature extractors" I3E Tr. Comput., 20, n° 9, pp. 967-71, 1971.
- [B-25] Y. T. CHIEN, K. S. FU "Selection and ordering of feature observations in a pattern recognition system" Information and Control, 12, pp. 395-414, 1968.
- [B-26] T. KAMINUMA et al. "Reduction of clustering problem to pattern recognition" Pattern Recognition, 1, n° 3, pp. 195-205, 1969.
- [B-27] T. T. KADOTA, L. A. SHEPP "On the best set of linear observables for discriminating two Gaussian signals" I3E Tr. Inf. Theory, 13, n° 4, pp. 278-84, 1967.
- [B-28] K. FUKUNAGA, W. KOONTZ "Application of the Karhunen-Loeve expansion to feature selection and ordering" I3E Tr. Comput. 19, n° 4, pp. 311-18, 1970.
- [B-29] M. LAMOTTE, J. BREMONT, J. P. HATON "Simulation de la reconnaissance des formes vocales par apprentissage" C. R. Ac. Sci., 269, pp. 286-8, 1969.
- [B-30] J. P. HATON, M. LAMOTTE "Prétraitement et reconnaissance de la parole. Simulation et réalisations pratiques" Journées d'étude sur la Parole - GALF - Aix-en-Provence 1971 - également dans Automatisme, XVII, n° 3, pp. 63-4, 1972.
- [B-31] R. HORNUNG "Dispositif de reconnaissance de la voix parlée" Thèse de Docteur-Ingénieur - Université de Nancy I - Février 1973.
- [B-32] J. P. HATON, M. LAMOTTE "Selection of features by information compression" Proc. Int. Conf. on Machine Perception of Patterns and Pictures - Teddington G. B. - avril 1972 - pp. 134-40.
- [B-33] L. F. PAU "Méthodes statistiques de réduction et de reconnaissance des formes. Normalisations des paramètres phonémiques - Application à la reconnaissance" E. N. S. T. Thèse 1972.

- [B-34] J. P. HATON, M. C. HATON "Some statistical feature preselection methods and their properties" proposé à Acustica.
- [B-35] J. MAX "Méthodes et techniques de traitement du signal et applications aux mesures physiques" t. I Masson - Paris 1972.
- [B-36] A. BLANC - LAPIERRE, R. FORTET "Théorie des fonctions aléatoires" Masson - Paris - 1960.
- [B-37] B. GOLD, C. M. RADER "Digital processing of signals" Mac Graw Hill - New-York - 1969.

PARTIE C.

- [C-1] M. LAMOTTE, J. BREMONT, J. P. HATON "Simulation d'un système à matrice d'apprentissage pour la reconnaissance des formes vocales" Cong. Int. d'Electron. Méd. et de Gén. Biol., Nancy, Juillet 1969, également dans Gén. Biol. Méd., 1, n° 3, pp. 138-43, 1970.
- [C-2] K. STEINBUCH "Die Lernmatrix" Kybernetik, n° 1, pp. 36-45, 1961.
- [C-3] A. FRUHLING et collaborateurs "Prétraitement numérique et reconnaissance de la voix parlée" Actes du 6^e Cong. Int. de Cybernétique - Namur - Sept. 1970 - pp. 657-66.
- [C-4] J. S. LIENARD et al "Normalisation fréquentielle de la parole" Journées d'étude sur la parole - GALF - Bruxelles - Mai 1973.
- [C-5] R. BELLMAN "Dynamic programming" Princeton Univ. Press. 1957.

- [C-6] J. P. HATON "The use of dynamic algorithms in automatic speech recognition" 7th Int. Cong. on Cybernetics - Namur - 7-11 Sept. 1973.
- [C-7] G. S. SLUCKER "Nelinejnyj method analiza recevych signalov" trudy N.I.I.R., n° 2, pp. 76-82, 1968.
- [C-8] H. SAKOE, S. CHIBA "A dynamic programming approach to continuons speech recognition" Proc. 7th Int. Cong. on Acoustics - Budapest - Aug. 1971.
- [C-9] J. F. BARS, J. Y. GRESSER, M. QUERRE "Application de la programmation dynamique à la reconnaissance des mots" Journées d'Etude sur la Parole - GALF - Bruxelles - Mai 1973.
- [C-10] J. S. LIENARD et al "Quelques idées directrices en reconnaissance automatique de la parole" Automatisme, XVIII, n° 3, pp. 89-98, 1973.
- [C-11] J. P. HATON "Segmentation et reconnaissance par comparaison dynamique de la voix parlée" Journées d'études sur la Parole - GALF - Bruxelles - Mai 1973.
- [C-12] J. P. HATON "Acoustic segmentation and real-time recognition of speech" Accepté pour le 8^e Cong. Int. d'Acoustique - Londres - Juillet 1974.
- [C-13] D. R. REDDY, P. VICENS "A Procedure for segmentation of connected speech" J. Audio Engr. Soc., 16, 4, 1968.
- [C-14] T. G. VON KELLER "An on-line recognition system for spoken digits" JASA, 49, n° 4, pp. 1288-96, 1971.
- [C-15] M. LAMOTTE, R. HORNUNG "Un dispositif simple de segmentation phonémique de la parole" A paraître dans la Revue d'Acoustique.
- [C-16] D. R. REDDY "Computer recognition of connected speech" JASA, 42, pp. 329-47, 1967.
- [C-17] N. R. DIXON, C. C. TAPPERT "Toward objective phonetic transcription. An on-line interactive technique for machine-processed speech data" IEEE Tr. Man-Mach. Syst., 11, n° 4, pp. 202-10, 1970.

- [C-18] J. GONSCHOREK, O. HINRICHS "Ein Spracherkennungsgerät mit selbsttätiger Anpassung an Sprechgeschwindigkeit und Lautstärke" N. T. Z. H. 4, April, 1971.
- [C-19] Special Section on parallel Computation - IEEE Tr. Comput, 22, n° 8, 1973.
- [C-20] A. J. SZANSER "Elastic matching in automatic pattern recognition" Proc. Int. Conf. on Mach. Perc. of patterns and pictures - Teddington - G. B. - Avril 1972.
- [C-21] R. VIVES "Elaboration d'un indice de ressemblance entre deux chaînes de longueur finie quelconque" Note Technique CEL/CSI/27 - CNET - Oct. 1972.
- [C-22] R. M. FANO "A heuristic discussion of probabilistic decoding" IEEE Tr. on Inf. Theory, 9, pp. 64-74, 1963.
- [C-23] A. S. RABINOWITZ "On the value of the Fano algorithm in establishing the graphemic form of machine-derived phonetic strings" Ph. D. Thesis Univ. of North-Carolina - Raleigh - March 1972.
- [C-24] C. C. TAPPERT et al "Application of sequential decoding for converting phonetic to Graphemic representation in automatic recognition of continuons speech (ARCS)" Int. Conf. on Speech Comm. and Proc. - Newton - Mass. April 1972.

PARTIE D.

- [D-1] N. R. FRENCH, C. W. CARTER, W. KOENIG "The words and sounds of telephone conversations" Bell System Tech. J., 9, pp. 290-324, 1930.
- [D-2] D. B. FRY "The frequency of occurrence of speech sounds in Southern English" Arch. Neerl. Phon. Exp., 20, pp. 103-6, 1947.
- [D-3] W. S. Y. WANG, J. CRAWFORD "Frequency studies of english consonants" Language and speech, 3, pp. 131-9, 1960.

- [D-4] P. B. DENES "On the statistics of spoken English" JASA, 35, pp. 892-904, 1963.
- [D-5] J. C. LAFON "Message et Phonétique. Introduction à l'étude acoustique et physiologique du phonème" P. U. F. 1961.
- [D-6] J. S. LIENARD "Le rôle des éléments phonétiques dans la synthèse de la parole et leur importance en linguistique quantitative" Revue d'Acoustique, n° 3-4, 1968.
- [D-7] J. P. TUBACH "Etude des contraintes statistiques des groupements phonématiques" C. E. T. A. Grenoble 1970.
- [D-8] W. A. WOODS, J. MAKHOUL "Mechanical Inference problems in continuous speech understanding" 3th Int. Joint. Conf. on Artificial Intelligence - Stanford - USA - August 1973.
- [D-9] D. E. WALKER "Speech understanding through syntactic and semantic analysis" Stanford Research Inst. Tech. Note 80, June 1973.
- [D-10] J. BARNETT "A vocal data management system" 1972 Conf. on Speech Comm. and Proc. - NEWTON - Mass, USA - April 1972.
- [D-11] F. WIOLAND "Constantes et variables dans l'estimation de la fréquence des phonèmes en français" Journées d'étude sur la Parole - GALF - Bruxelles - Mai 1973.
- [D-12] R. W. DONALDSON, G. T. TOUSSAINT "Use of contextual constraints in recognition of contour-traced handprinted characters" IEEE Tr. Comput., 19, n° 11, pp. 1096-99, 1970.
- [D-13] E. M. RISEMAN, R. W. EHRICH "Contextual word recognition using binary digrams" IEEE Tr. Comput., 20, n° 4, pp. 397-403, 1971.

- [D-14] R. NARASIMHAN, V. S. N. REDDY "A syntax-aided recognition scheme for handprinted English letters" Pattern Recognition, 3, n° 4, pp. 345-61, 1971.
- [D-15] H. B. SAVIN "Word frequency effect and errors in the perception of speech" JASA, 35, pp. 202-6, 1963.
- [D-16] J. P. HATON "Rôle des contraintes linguistiques en reconnaissance de la parole" Journées d'études sur la Parole - GALF - Lannion - Juin 1972.
- [D-17] G. A. MILLER "Decision units in the perception of speech" IRE Tr. Inf. Theory. 3, n° 2, pp. 81-83, 1962.
- [D-18] W. A. LEA "An approach to syntactic recognition without phonemics" Int. Conf. on Speech Comm. and Proc. - Newton - Mass. April 1972.

NOM DE L'ETUDIANT : HATON Jean-Paul

Nature de la Thèse : Thèse d'Etat

Vu, Approuvé
et permis d'imprimer

NANCY, le 19 /12/ 73

Le Président du Conseil de l'Université de
NANCY I



J. -R. HELLUY