Reconnaissance automatique du locuteur : présentation générale, méthodologies et expérimentation, perspectives d'application

Patrick Corsi

THESE

présentée à

Institut National Polytechnique de Grenoble

pour obtenir le grade de

DOCTEUR INGENIEUR

(Génie Informatique)

par

Patrick CORSI

RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR :

PRESENTATION GENERALE, METHODOLOGIES ET EXPERIMENTATION, PERSPECTIVES D'APPLICATION.

Thèse soutenue le 30 octobre 1979 devant la Commission d'Examen :

Rapporteur	:	C. BELLISSANT
Président	:	G. VEILLON
Examinateurs	:	G. BENBASSAT
		L.J. BOË
		R. CARRE
		J.P. HATON
		G. PERENNOU

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Année universitaire 1979-1980

Président : M. Philippe TRAYNARD

Vice-Présidents : M. Georges LESPINARD
M. René PAUTHENET

PROFESSEURS DES UNIVERSITES

MM.	ANCEAU François	Informatique fondamentale et appliquée
	BENOIT Jean	Radioéléctricité
	BESSON Jean	Chimie Minérale
	BLIMAN Samuel	Electronique
	BLOCH Daniel	Physique du Solide - Cristallographie
	BOIS Philippe	Mécanique
	BONNETAIN Lucien	Génie Chimique
	BONNIER Etienne	Métallurgie
	BOUVARD Maurice	Génie Mécanique
	BRISSONNEAU Pierre	Physique des Matériaux
	BUYLE-BODIN Maurice	Electronique
	CHARTIER Germain	Electronique
	CHERADAME Hervé	Chimie Physique Macromoléculaires
Mme	CHERUY Arlette	Automatique
MM.	CHIAVERINA Jean	Biologie, Biochimie, Agronomie
	COHEN Joseph	Electronique
	COUMES André	Electronique
	DURAND Francis	Métallurgie
	DURAND Jean-Louis	Physique Nucléaire et Corpusculaire
	FELICI Noël	Electrotechnique
	FOULARD Claude	Automatique
	GUYOT Pierre	Métallurgie Physique
	IVANES Marcel	Electrotechnique
	JOUBERT Jean-Claude	Physique du Solide - Cristallographie
	LACOUME Jean-Louis	Géographie - Traitement du Signal
	LANCIA Roland	Electronique - Automatique
	LESIEUR Marcel	Mécanique
	LESPINARD Georges	Mécanique
	LONGEQUEUE Jean-Pierre	Physique Nucléaire Corpusculaire
	MOREAU René	Mécanique
	MORET Roger	Physique Nucléaire Corpusculaire
	PARIAUD Jean-Charles	Chimie - Physique
	PAUTHENET René	Physique du Solide - Cristallographie
	PERRET René	Automatique

.../...

MM.	PERRET Robert	Electrotechnique
	PIAU Jean-Michel	Mécanique
	PIERRARD Jean-Marie	Mécanique
	POLOUJADOFF Michel	Electrotechnique
	POUPOT Christian	Electronique - Automatique
	RAMEAU Jean-Jacques	Chimie
	ROBERT André	Chimie Appliquée et des matériaux
	ROBERT François	Analyse numérique
	SABONNADIERE Jean-Claude	Electrotechnique
Mme	SAUCIER Gabrielle	Informatique fondamentale et appliquée
M.	SOHM Jean-Claude	Chimie - Physique
Mme	SCHLENKER Claire	Physique du Solide - Cristallographie
MM.	TRAYNARD Philippe	Chimie - Physique
	VEILLON Gérard	Informatique fondamentale et appliquée
	ZADWORNY François	Electronique

CHERCHEURS DU C.N.R.S. (Directeur et Maître de Recherche)

M.	FRUCHART Robert	Directeur de Recherche
MM.	ANSARA Ibrahïm	Maître de Recherche
	BRONOEL Guy	Maître de Recherche
	CARRE René	Maître de Recherche
	DAVID René	Maître de Recherche
	DRIOLE Jean	Maître de Recherche
	KAMARINOS Georges	Maître de Recherche
	KLEITZ Michel	Maître de Recherche
	LANDAU Ioan-Doré	Maître de Recherche
	MERMET Jean	Maître de Recherche
	MUNIER Jacques	Maître de Recherche

Personnalités habilitées à diriger des travaux de recherche (décision du Conseil Scientifique)

E.N.S.E.E.G.

MM.	ALLIBERT Michel	
	BERNARD Claude	
	CAILLET Marcel	
Mme	CHATILLON Catherine	
MM.	COULON Michel	
	HAMMOU Abdelkader	
	JOUD Jean-Charles	
	RAVAINE Denis	
	SAINFORT	C.E.N.G.

.../...

MM. SARRAZIN Pierre
SOUQUET Jean-Louis
TOUZAIN Philippe
URBAIN Georges — Laboratoire des Ultra-Réfractaires ODEILLO

E.N.S.M.E.E.

MM. BISCONDI Michel
BOOS Jean-Yves
GUILHOT Bernard
KOBILANSKI André
LALAUZE René
LANCELOT Françis
LE COZE Jean
LESBATS Pierre
SOUSTELLE Michel
THEVENOT François
THOMAS Gérard
TRAN MINH Canh
DRIVER Julian
RIEU Jean

E.N.S.E.R.G.

MM. BOREL Joseph
CHEHIKIAN Alain
VIKTOROVITCH Pierre

E.N.S.I.E.G.

MM. BORNARD Guy
DESCHIZEAUX Pierre
GLANGEAUD François
JAUSSAUD Pierre
Mme JOURDAIN Geneviève
MM. LEJEUNE Gérard
PERARD Jacques

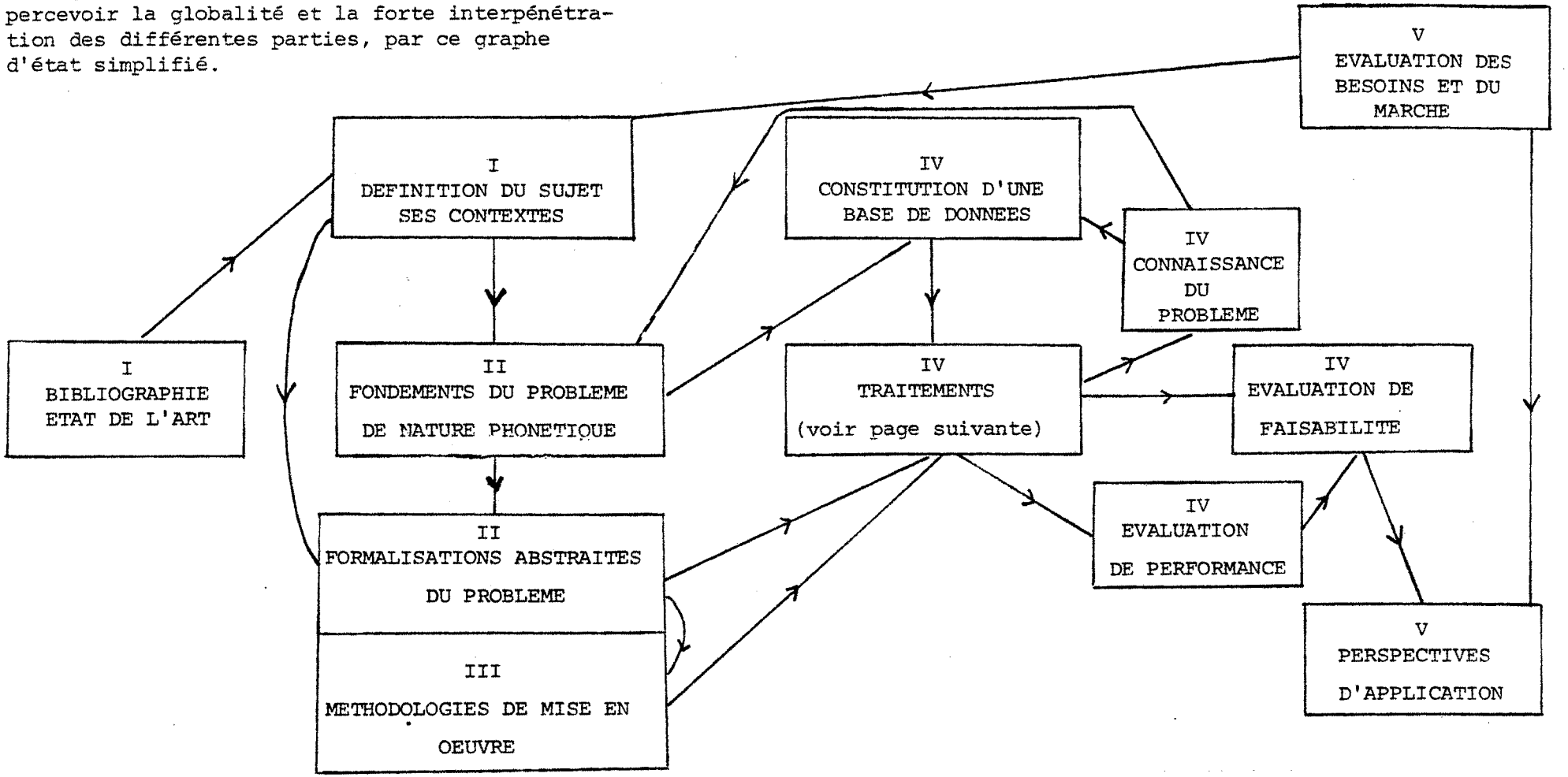
E.N.S.H.G.

M. DELHAYE Jean-Marc

E.N.S.I.M.A.G.

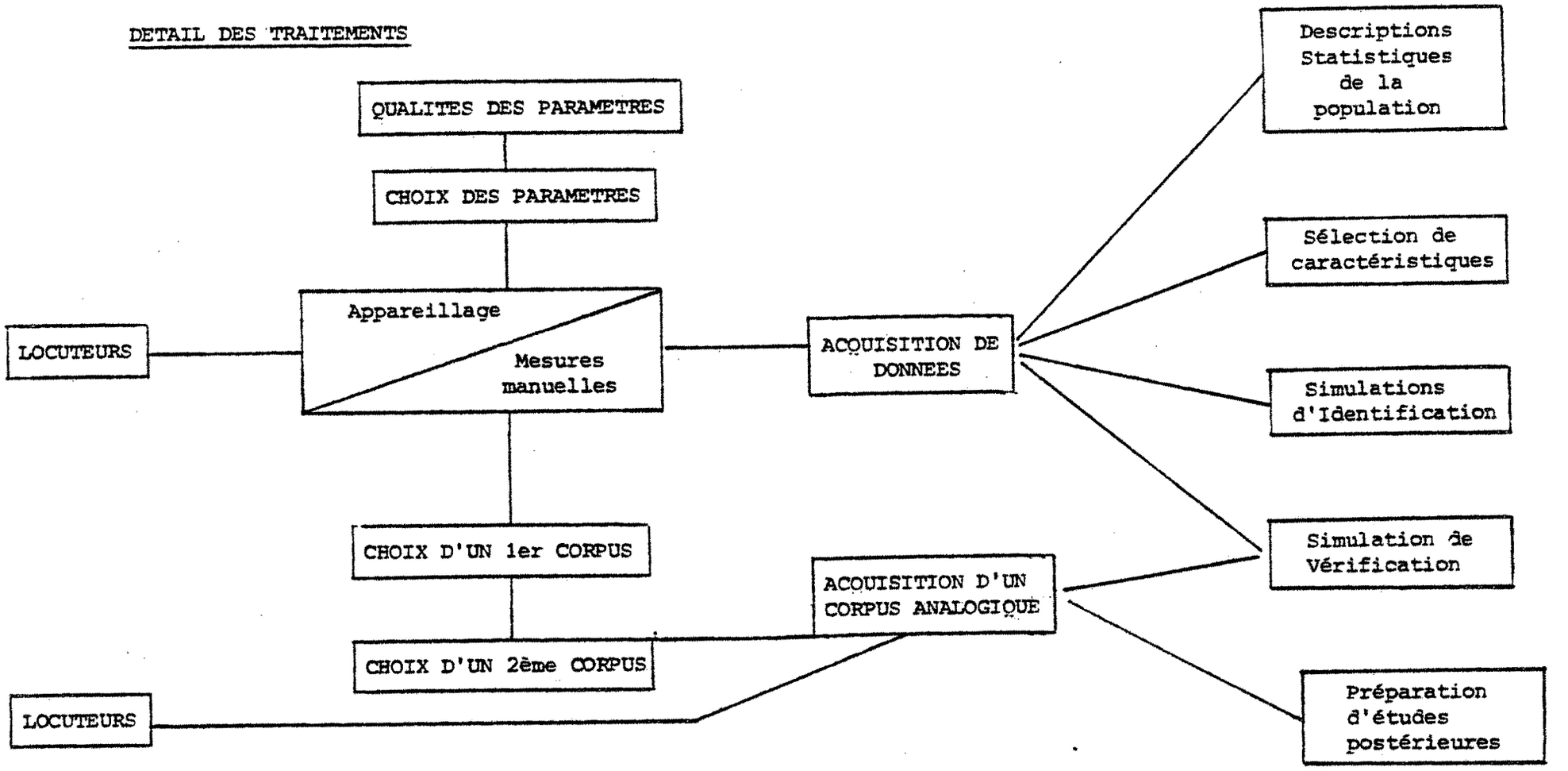
MM. COURTIN Jacques
LATOMBE Jean-Claude
LUCAS Michel
VERDILLON André

GRAPHE D'INTERRELATION DES DIFFERENTES SECTIONS
Bien que rédigées de façon indépendante, on peut percevoir la globalité et la forte interpénétration des différentes parties, par ce graphe d'état simplifié.



ONS NEES	ETAT DE L'ART I - II	EXPLORATION DU SUJET I - II - III	EXPERIMENTATIONS IV	LA FAISABILITE DE LA RECONNAISSANCE DU LOCUTEUR V

DETAIL DES TRAITEMENTS



PARAGRAPHES CONCERNES	PREPARATION DES EXPERIENCES	COLLECTION DE DONNEES	EXPLORATION

Quand la stérilité de la conversation me forçait d'y suppléer par d'innocentes fictions, j'avais tort, parce qu'il ne faut pas, pour amuser autrui, s'avilir soi-même.

ROUSSEAU, 4ème Promenade,
"Rêveries"

AVANT-PROPOS

Mes remerciements vont d'abord aux personnes qui ont assuré la réalisation des divers documents écrits : à Mesdames BOULESTEIX, DUBOIS, SOUILLARD et TREVISAN, et au Service de Reprographie de l'I.M.A.G., pour une assistance chaque fois amicale et compétente et un travail de qualité.

J'exprime :

ma gratitude et mes respects au Professeur G. VEILLON pour la présidence du jury,

ma sincère reconnaissance aux Professeurs J.P. HATON et G. PERENNOU, en qui j'ai pu trouver en plusieurs occasions, une écoute attentionnée, et qui m'ont assuré de leur conseil particulièrement précieux et pertinent ; je les remercie pour l'honneur de leur présence,

ma respectueuse considération pour R. CARRE, Maître de Recherches au C.N.R.S., pour l'honneur qu'il me fait en participant au jury ; je n'oublie pas que R. CARRE a été un précurseur français dans le domaine : c'est avec un profond respect que je lui soumets ces pages,

ma chaleureuse reconnaissance à l'équipe de l'Institut de Phonétique de Grenoble qui m'a permis d'élaborer et de réaliser toute la partie phonétique de ce travail. C'est tout particulièrement à L.J. BOE que je dois de m'avoir accepté et aidé d'une façon décisive, en ce qui concerne notamment la définition d'objectifs, et le démarrage des expériences. A L.J. BOE donc, un grand merci pour m'avoir consacré beaucoup de son temps et montré que la coopération interdisciplines est possible. Puisse-t-elle avoir été féconde !

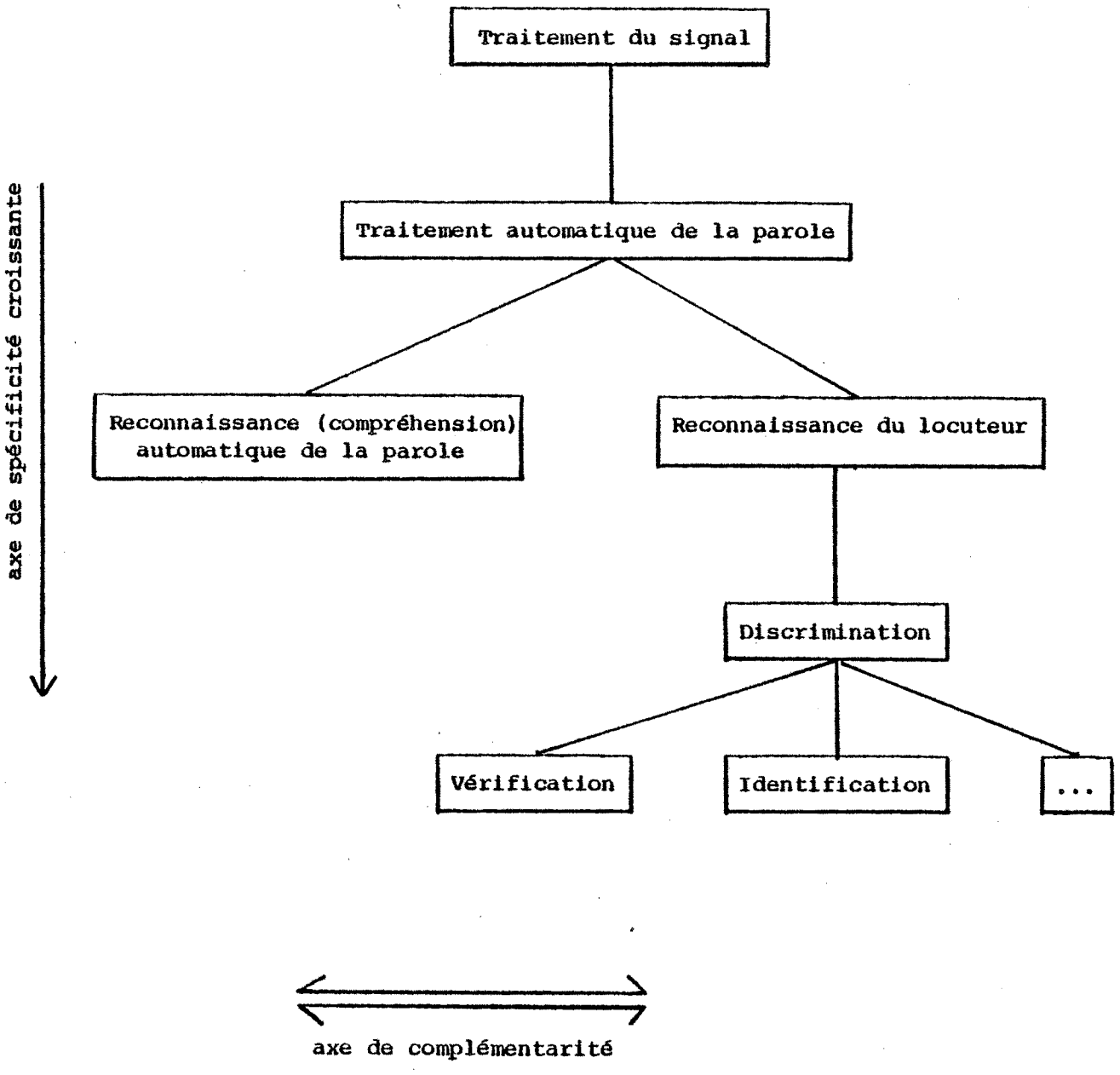
mes remerciements à G. BENBASSAT, Ingénieur à la CIT-ALCATEL, pour sa présence et la démonstration d'un intérêt soutenu,

mes remerciements amicaux à mes collègues de l'I.M.A.G., pour les discussions fructueuses inter-équipes ; je leur adresse toute ma sympathie,

enfin, une pensée reconnaissante pour toutes les personnes ayant participé aux expériences, et qui ont ainsi rendu les études possibles.

Je garde toutefois pour C. BELLISSANT la certitude de ma très profonde estime. Je lui dois, outre l'acceptation dans son équipe de recherche et la proposition précise du sujet, un indéfectible appui, matériel et moral, tout au long de mes hésitations, et un encouragement à nombre d'initiatives. Pour son attitude désintéressée et ses qualités humaines qui m'ont permis d'évoluer avec une relative liberté, je me sens redevable - envers C. BELLISSANT d'une réelle dette morale.

LES CONTEXTES D'ETUDE DU SUJET



ORGANISATION DU TEXTE

Le texte est présenté en cinq parties totalement indépendantes (livres I à V). On pourra s'orienter dans son parcours de la façon suivante :

Plan d'ensemble :

Une perspective d'ensemble : synthèses d'introduction

LIVRE I : approche du sujet
LIVRE II : fondements du problème
LIVRE V : applications

Un guide de lecture :

Dictionnaire-index (reporté en fin de texte)
Table des matières, liste des illustrations
Plan d'ensemble, organigrammes

Le corps principal :

LIVRE III : méthodologie de la mise en oeuvre
LIVRE IV : Conception de Système.

Il arrive que quelques points aient été brièvement repris ailleurs que dans leur section privilégiée. La bibliographie rappelle en tête son mode de classement.

I - APPROCHE DU SUJET

II - LES FONDEMENTS DU PROBLEME

I - LA VARIABILITE DE LA VOIX

1 - Quelques rappels sur la production de la parole
1.1. Les mécanismes de production
1.2. La production des formants
2 - Ebauche d'une idiolectologie
3 - Les sources de variabilité intra et inter-locuteurs
4 - Modes d'utilisation des variabilités intra et inter-locuteurs

II - FORMALISATIONS DU PROBLEME

1 - Introduction
2 - Première formalisation
2.1. Machine à discriminer les locuteurs
2.2. Machine à identifier un locuteur
2.3. Machine à vérifier un locuteur
3 - Description détaillée des tâches
3.1. Description de la machine à discriminer les locuteurs
3.2. Description de la machine à identifier un locuteur
3.3. Description de la machine à vérifier un locuteur
4 - Conclusion

III - CONSTITUTION D'UNE BASE DE DONNEES

1 - Considérations générales
2 - La constitution de la base
3 - Présentation de SYSIPHE

III - METHODOLOGIE DE LA MISE EN OEUVRE

I - Reconnaissance du locuteur par l'audition

II - Reconnaissance du locuteur par l'analyse visuelle de spectrogrammes

III - Reconnaissance du locuteur par l'analyse automatique de spectrogrammes

1. Introduction
2. Techniques utilisant un spectrogramme discret

IV - Les modes opératoires

1. Tests contemporain et non contemporain
2. Systèmes en ligne
3. Systèmes hors ligne
4. Résumé synoptique

V - Reconnaissance du locuteur par ordinateur

1. La chaîne de traitement
2. Les problèmes fondamentaux

IV - POUR LA CONCEPTION D'UN SYSTEME OPERATIONNEL

INTRODUCTION : Les facteurs de l'étude et les hypothèses implicites

PREMIERE PARTIE : ETUDES PREPARATOIRES

I - METHODE DE CHOIX A PRIORI DES PARAMETRES

1 - Définition d'un paramètre

2 - Les qualités idéales d'un paramètre

- 2.1. La pertinence
- 2.2. La disponibilité
- 2.3. La résistance
- 2.4. L'interprétabilité
- 2.5. La polyvalence

3 - Choix de paramètres temporels et acquisition des données

- 3.1. Quelques motivations guidant les choix
- 3.2. Liste des paramètres retenus a priori
- 3.3. Appareillage utilisé et conditions d'enregistrement
- 3.4. Acquisition des données

II - MODELISATION DES LOCUTEURS

. Les facteurs en présence

IV

DEUXIEME PARTIE : ETUDES EXPLORATOIRES PILOTES ET SIMULATIONS D'EXPERIENCES

III - MISE EN OEUVRE DE PARAMETRES PERTINENTS

1 - Sélection de caractéristiques

2 - Description statistique de la population et discrimination descriptive

3 - Expériences d'identification

4 - Synthèse des résultats

5 - La fréquence fondamentale moyenne : obtention, discriminance et stabilité

5.1. La modélisation de la production de la fréquence laryngienne

5.2. Un programme de traitement statistique de la fréquence fondamentale

1) Implémentation

2) Algorithme de traitement statistique

5.3. Quelques résultats

1) Discriminance

2) Stabilité

6 - Exploitation de l'intensité du Signal

IV - CONCLUSIONS D'ENSEMBLE

Limites de l'étude

V - LES PERSPECTIVES D'APPLICATION

I - LES TENDANCES SOCIO-ECONOMIQUES ET LES BESOINS ACTUELS

1 - Le contexte de la communication homme-machine

2 - Une analyse du marché potentiel

2.1. Quels avantages attendre d'une communication orale avec une machine ?

2.2. Que peut permettre un système de vérification d'identité ?

2.3. Quelles performances envisager en reconnaissance du locuteur ?

2.4. Quelles applications envisager en reconnaissance du locuteur ?

2.5. Etude économique du terminal vocal et du terminal financier

II - PREMIERE EVALUATION DES COUTS D'EXPLOITATION

III - ESTIMATION DE L'IMPACT SUR L'INDIVIDU ET LA SOCIETE

VI - PARTIE CONCLUSIVE

I - Conclusions d'ensemble

II - Glossaire de termes phonétiques

III - Bibliographie

1. Organisation des références
2. Références en Reconnaissance de la Parole et Divers
3. Références en Reconnaissance du Locuteur (1938-1979)
4. Références en Reconnaissance des Formes

RESUME

RESUME GENERAL :

Cette monographie est consacrée à l'étude de la possibilité de reconnaissance de l'identité des personnes à partir de leur voix.
On élabore progressivement un cadre conceptuel adapté à cette reconnaissance, et effectue diverses expériences de reconnaissance.
Les aspects de la communication homme-machine incluent le cas où un ordinateur est piloté par la voix humaine. Cela suppose une certaine capacité à s'adapter à la spécificité de la voix du locuteur présent. Le problème complémentaire consiste à tirer parti des différences inter-individuelles de la voix, afin d'établir des typologies et jusqu'à pouvoir reconnaître l'identité des locuteurs. Cette approche est relativement récente, notamment chez les phonéticiens qui ont traditionnellement recherché des éléments universaux pour chaque groupe linguistique. On essaie de justifier l'existence de ces différences individuelles.
Du point de vue de la Reconnaissance des Formes, il s'agit de sélectionner un ensemble de paramètres possédant une forte capacité de discrimination dans la population de locuteurs de référence et de réaliser un appren-

ABSTRACT

OVERVIEW :

This study is directed towards the improvement of speaker recognition feasability. We built a conceptual, practical and experimental frame which is adapted to this context and we run various experiments.

Man-machine communication yet involve the situation in which a computer is controlled by a human voice. That implies a special skill of the machine to adapt itself to the actual talker. Conversely, one can try to discriminate among the voices, in order to establish typologies with the particular case of identifying the persons.
This is a somewhat original point of view since the phoneticians traditionnally so far considered universal elements in a given language. The reality of the personal differences is herebu assessed.

From the Pattern Recognition stand point, we have to select a set of parameters on the basis of their good discrimination power upon the given reference speaker set. Than we must make a supervised learning with the available references of each speaker. We get some insight in the many basic problems - both theoretical and practical - that have to be solved first, especially concer-

tissage supervisé fondé sur le traitement de plusieurs références associées à chaque locuteur.
L'exposé évalue les caractéristiques d'une base de données adaptée à ce problème et traite des problèmes généraux - théoriques et expérimentaux - rencontrés,

On évalue les limites de l'étude actuelle et on distingue les deux contextes fondamentaux de :
- la vérification de l'identité de personnes avec ses applications potentielles spécifiques (banques, verrous d'accès à des systèmes,...)
- la recherche d'une personne à partir d'éléments de sa voix (identification de locuteurs) ; ce contexte plus difficile et moins étudié que le précédent, possède des applications propres.

ning the acquisition of a good data-base.

Limits to the present study are pointed out.
We naturally distinguish the two contexts :
- speaker Verification which may lead to specific applications (banking, system accesses,...)
- speaker Identification : a more difficult and less studied task which leads to proper kinds of applications.

RESUMES PAR SECTIONS

I

Ce chapitre tend à définir le cadre contextuel actuel de la reconnaissance de l'identité de la personne à travers sa voix, en précisant avec autant de rigueur qu'il est possible, les termes consacrés par l'usage dont la définition reste souvent imprécise, ainsi que d'autres termes dont l'usage semble commode. On fait un tour d'horizon des problèmes fondamentaux posés, qui sont de nature phonétique et statistique ; on esquisse les concepts de base méthodologiques et de nature informatique, ainsi que le schéma d'un système automatique fictif. Cette étude cherche donc à faire le point sur la question et se place dans une perspective bibliographique et historique.

II

Après avoir mis en évidence au chapitre I, les variations de la voix, on tente ici une investigation des sources de ces variabilités, en liaison avec les mécanismes de production de la voix.
Dans une deuxième partie, on propose diverses formalisations abstraites de la Vérification et de l'Identification, susceptibles d'être exploitées dans les parties suivantes.

I

This chapter presents the contextual frame for the identification of persons by means of their voices : common definitions and other ones which are less common are described in a rather rigorous way. We make an overview of fundamental problems, both theoretical and statistical and present some methodology for solving them adding a rough description of a possible automatic system. We then make a review of literature.

II

After having presented speech variation, we now get more insight into the sources of these variabilities mainly at the level of speech production.
In a second part, a mathematical formulation Speaker Discrimination, Identification and Verification is presented.

III

On détaille chacune des diverses méthodologies existantes (reconnaissance par l'audition, par spectrogrammes, par machine), et propose leur comparaison ; on introduit la reconnaissance par ordinateur, dont les principes font appel à la reconnaissance des formes, et qui fait l'objet du chapitre IV.

IV

On s'intéresse ici à l'aspect temporel du signal, contrairement à la plupart des travaux antérieurs. Après avoir présenté les qualités à rechercher dans les paramètres, on en fait une première sélection, fondée entre autres sur la relation entre les segments acoustiques mesurés et des considérations physiologiques.
Les données portent sur 12 hommes lisant 10 fois un texte d'une minute environ, ainsi que 6 phrases courtes. A partir de ces paramètres choisis a priori, on opère une sélection sur la base des corrélations observées, d'un rapport de dispersion des classes, et d'un taux de dispersion empirique moyen. On observe une décorrélation entre les groupes naturels de définition des caractéristiques, ce qui conduit à prendre pour solution simple sous-optimale du problème de sélection, le n-uplet des meilleures caractéristiques de chaque groupe. Parmi les paramètres ayant été observés comme les plus pertinents, on note : la fréquence fondamentale, la durée des occlusives sourdes, et le taux de voisement.

III

Each existing methodology : recognition by audition, by spectrograms and by machine is detailed with comparative remarks. Recognition by computer is then emphasized with an introduction to pattern recognition problems.

IV

We are here mainly interested in the temporal aspect of the speech signal.
After having proposed some important a priori qualities of a parameter, we define a set of 37 parameters ; many of them resulting from close definitions. We consider 9 definition groups, based upon their acoustical and physiological significance.
Data are made from 10 repetitions taking place during a single session and spoken by 12 adult male reading a long text (about one minute long and 6 short sentences).
Descriptive studies include a selection of the parameters, mostly based upon their empirical correlations, F-ratio and relative group dispersion.
Natural groups of definition of the parameters are observed to be practically non-correlated.
A non-optimal solution to the feature selection phase is given by the feature vector made with the best single features in each group. Some of them are : the fundamental frequency, non-voiced stops durations, voicing rate.

On effectue une analyse descriptive de la population, à l'aide des mêmes paramètres, qui tend à mettre en valeur leur capacité discriminante.
On décrit diverses expériences d'identification à l'aide des mêmes paramètres.
La fréquence fondamentale fait l'objet d'une étude séparée.
Enfin, on exploite individuellement l'intensité du signal de parole sur une même phrase pour 10 locuteurs, en mesurant l'écart entre deux répétitions quelconques.

We make a descriptive analysis of the population which reveals us the discriminant power of all the parameters.

Various identification experiments are described, which use the same parameters.
Fundamental frequency is alse studied in a separate paragraph.

Intensity of speech signal is processed by a method which evaluates the proximity between two intensity curves resulting from two repetitions of the same sentence by 10 talkers.

L'un des aspects essentiels de la communication homme-machine, est d'introduire des données en direction d'un ordinateur, à partir d'un poste de saisie des données et souvent encore, par l'intermédiaire d'un opérateur.
Les caractéristiques externes du poste de saisie n'ont guère évolué, en ce qui concerne l'usage permanent des mains et la direction fixe du regard, qui doit vérifier la justesse des informations frappées au clavier, ce qui empêche toute disponibilité pour une tâche parallèle.
En outre :
- le langage de commande reste compliqué et nécessite une formation spéciale de l'opérateur,
- la vitesse de l'acquisition des données est limitée par la vitesse de frappe.
Parallèlement, les recherches fondamentales entreprises depuis quelques dizaines d'années, attestent un progrès certain dans la compréhension de la parole humaine par ordinateur, même si le problème n'est pas encore résolu dans son ensemble. Elles montrent aussi qu'il est possible de vérifier l'identité des personnes par leur voix avec une bonne confiance.
La reconnaissance de la parole offre une simplification de l'entrée des données. La situation de connexion à un réseau d'ordinateurs (y compris les mini-machines), se généralise et multiplie les postes de saisie, libéralisant

V

One non-minor aspect of man-machine communication is the input of data from a terminal, this terminal beeing operated by hand-typing.
So far, data input caracteristics have not yet been very strongly improved, as far as hand and eye coordinated usage is maintained while entering and verifying the inputs. As we know, this is an obstacle for the operator to process a parallel task.
Morever :
- command language still remains rather complex and requires a particular training,
- data acquisition speed is naturally limited by typing speed.
On the contrary, fundamental research in speech recognition by computers has reached in the past decade an interesting level of feasability even if this problem is not solved in a general sense. The results show that it is possible to verify the identity of talkers with a good level of confidence.
Generally speaking, speech recognition authorizes a more simplified data-entry. On the other hand, computer network connexions - including the mini-machines - tend to become more and more common. We are then facing the forever problems of checking the identity when accessing to the system, and the problem of the protection of data from impostors.
This is an attempt to analyse the potential market and to describ some situations in which speaker recognition would result in some

leur accès. Il se pose alors le problème qui n'est pas nouveau, de la protection et de l'accès réservé à des données : personne habilitée à opérer sur le système. On essaie, dans cette section, de faire une première analyse du marché potentiel et présentant quelques situations d'application en référence aux avantages apportés, ou aux besoins actuellement rencontrés dans chaque contexte.
Il est fait une évaluation qualitative sommaire des coûts d'exploitation. On termine en évoquant quelques problèmes sociaux et juridiques, posés par l'introduction de tels systèmes sur le marché.

MOTS-CLES : Applications de la reconnaissance - Vérification, Authentification - Identification du locuteur - Signature vocale - Protection - Sécurité - Confidentialité des données - Communication homme-machine.

improvement of the given task. We evaluate the supposed qualitative running costs, and, as a consequence of all the preceding analysis, point out some social and legal problems that would be supposed to appear.

ACKNOWLEDGMENTS :
The author is indebted to M. Louis-Jean BOË from Institut de Phonétique de Grenoble for assistance and guidance through this study and expresses his thanks to other members of this Institute and of IMAG for constructive discussion.

INTRODUCTION : LA PORTEE DU SUJET

Ce travail voudrait offrir un examen relativement large et général des divers aspects impliqués dans la reconnaissance du locuteur, en mettant l'accent sur les principes ou idées qui semblent fondamentaux pour chacun d'entre eux. Ceci afin de fixer ce qui peut l'être, comme autant de points de repère pour la suite. Le problème de la reconnaissance des personnes d'après leur voix, n'est pas seulement un thème scientifique abstrait auquel se greffent des méthodes pratiques de mise en oeuvre, c'est aussi un sujet qui a focalisé l'intérêt de beaucoup d'organismes divers, intéressés par les perspectives d'application.

Comme le font remarquer FASOLO & MIAN (ICASSP, 1978), les résultats obtenus par beaucoup de chercheurs durant les années passées suggèrent que le signal de parole contient tant d'information concernant l'identité du locuteur que, malgré la complexité très problable de son codage, la reconnaissance de cette identité peut être menée de bien des manières, par l'usage de différentes techniques d'extraction de paramètres, et de différents algorithmes de reconnaissance des formes.
Ainsi, si la question de reconnaître les personnes d'après leur voix n'est pas globalement résolue, personne n'a pu montrer qu'elle était irréalisable.
Au contraire, les nombreuses études qui ont été menées notamment depuis deux décades, indiquent une certaine "faisabilité". On s'accorde à reconnaître que les méthodes automatiques peuvent dépasser dans ce domaine l'habileté humaine, ne serait-ce que par le stockage d'un important volume de données, la rapidité des calculs et de l'accès mémoire et la finesse possible de l'analyse.
Pourtant, les auteurs (par exemple, ROSENBERG, 1976), estiment généralement qu'il s'agit d'un sujet difficile (1) et ouvert dans une certaine mesure. Il est indispensable, affirme cet auteur, de bien circonscrire ce problème si l'on veut y réussir.

---

(1) mais - probablement - d'un degré de difficulté inférieur à celui de la reconnaissance de la parole (il n'est pas nécessaire d'extraire le message sémantique. Voir la section I - II - 3).

Quatre points sont à développer dans ce sujet :

1 - Etablir que la voix peut à elle seule servir de support pour identifier une personne.
2 - Valider des méthodes d'identification par la voix.
3 - Evaluer la possibilité de l'application de ces méthodes à divers contextes d'application.
4 - Mettre au point un système pour une application particulière.

Ce programme en quatre points figure dans la préface du rapport du Department of Michigan State Police (cf. Bibliographie).
Comme dans la plupart des recherches expérimentales, on se sert principalement de la méthode inductive, c'est-à-dire qu'on procède du particulier au général.
Comme toujours, trois éléments sont indispensables pour mener à bien l'étude :

- un corpus de données,
- une méthode,
- un outil (l'ordinateur) capable de faire agir la méthode sur les données.

Comme dans d'autres problèmes de reconnaissance des formes, il faut essayer de développer trois aspects :

- aspects théoriques : formalisation du sujet et des méthodes,
- aspects méthodologiques : mise en oeuvre des outils valables,
- aspects pratiques : études expérimentales.

Sur le choix des méthodes, on accorde généralement la préférence aux statistiques actuelles, celles qui opèrent l'examen le plus neutre, le plus "innocent" possible des données en laissant le soin aux techniques descriptives de découvrir la répartition des individus ou des variables, en divers groupes.(1)
Comme on le sait, les hypothèses de normalité deviennent très contraignantes dès que l'on passe à une étude multidimensionnelle : on peut proposer une multitude de variables offrant chacune quelque intérêt.

Le domaine qui nous concerne (vérification et identification de locuteurs) exige une décision après l'analyse : il faut prendre une option sur la vraisemblance

(1) Cet engouement pour les études statistiques est fort général et si répandu qu'il mérite le nom de "statistite".

de l'hypothèse relative au locuteur courant. Tout outil d'aide à la décision est a priori à examiner.

Pour ce qui est de l'évaluation de performances, la plupart des auteurs se précipitent sur les "taux de performance", et limitent cette évaluation à la seule donnée d'un nombre dont la portée est faible parce que de très nombreux facteurs influencent les résultats. En reconnaissance des formes, on sait que l'évaluation de performances est une branche toute récente, bien peu évoluée parce que difficile, et dépendant de nombreux facteurs.

Au niveau des applications potentielles dans le secteur socio-économique, la reconnaissance automatique du locuteur serait-elle susceptible d'introduire de nouveaux types de services tels que le remplacement de la carte magnétique, l'accès contrôlé par la voix à des zones réservées ou à des informations privilégiées ? Depuis 10 ans, les auteurs posent ces mêmes questions. Pour certains, la Vérification du locuteur est un problème résolu ; reste celui de l'Identification que l'on assure être de nature beaucoup plus fondamentale. Pourtant, les articles spécifiques sur la vérification continuent à être de mode (il y a tant de paramètres pouvant être utilisés, et en plus, on est relativement assuré d'avoir un effet positif !), alors qu'on trouve encore peu d'études profondes générales et théoriques sur cette reconnaissance en général, et peu de systèmes opérationnels. Il n'y a à ma connaissance que les monographies suivantes qui offrent une étude générale de la question :(1)

- le livre de HECKER (ASHA monographs, 1971),
- le rapport du Michigan State Police, (U.S. Dept. of Justice, 1972),
- l'article de ATAL (IEEE, ICASSP, 1976),
- l'article de ROSENBERG (IEEE ICASSP, 1976),
- l'étude de JESORSKY (1978),
- un récent rapport collectif de la National Science Foundation.

Mais, au terme de ceci, on s'aperçoit qu'on a souvent oublié d'étudier l'essentiel : les différences ou caractéristiques individuelles d'un point de vue phonétique. Les phonéticiens eux-mêmes semblent avoir jusqu'ici négligé ce point de vue au profit du point de vue complémentaire : la recherche des traits, indices phonétiques d'un système linguistique.
Le choix d'une approche pluridisciplinaire semble justifié. C'est, comme le fait remarquer humoristiquement un auteur (2), au prix d'une "approche intelligente

---

(1) Voir la bibliographie. et une revue de la littérature par le Sensory Research Center (Stanford Research Institute)
(2) A.E. ROSENBERG, Proc. IEEE, 1977, 64, 4, 475.

du sujet, que les chances de succès en reconnaissance de la parole semblent meilleures que celles des alchimistes".

I

APPROCHE DU SUJET

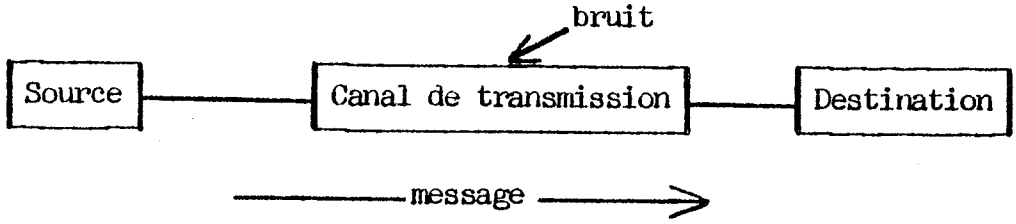
Les voix des personnes ne sont inouies qu'une fois.

I - DEFINITION DU SUJET

1 - Quelques remarques d'ensemble sur le signal de parole

1.1. Les aspects informatifs du signal de parole :

a) Un système de communication peut être représenté par le schéma général suivant :



On dit que le canal transporte de l' "information" entre l'émetteur (ici , le locuteur) et le récepteur (la machine à analyser la parole, l'auditeur). Le canal est, dans la pratique, perturbé par du bruit : la parole atteint son destinataire après distorsion ou atténuation. Cette transmission suppose un codage du message, adapté à la nature du canal (les propriétés physiques de l'air ambiant ou de tout autre milieu de propagation du son : hélium, ...).

Mais qu'est-ce que cette "information" ? C'est une quantité qui est directement fonction de la disparité (différence, rapport) entre le nombre de réponses possibles concernant l'état de la source, avant et après la transmission : elle est d'autant plus grande qu'on lève l'ambiguïté entre les réponses possibles (un apport d'information se traduit par une réduction d'incertitude).

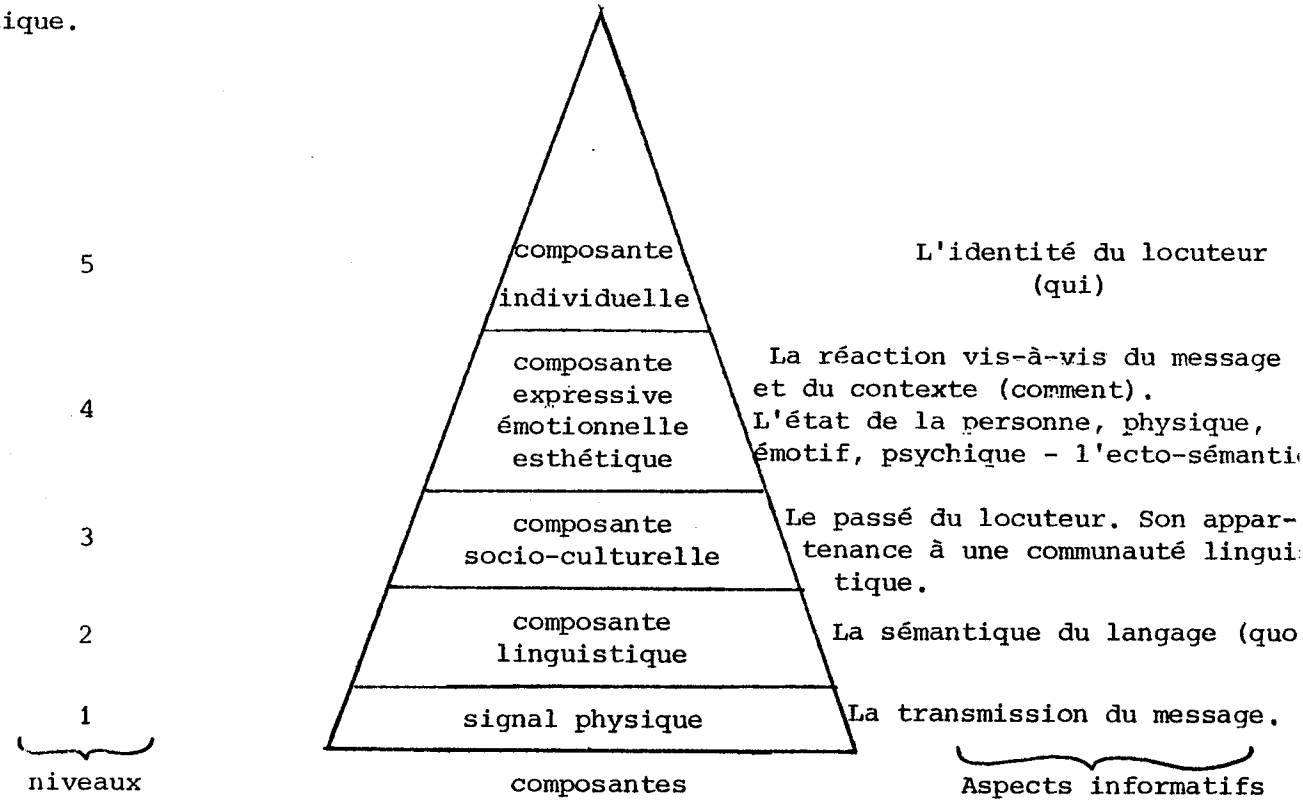
b) La Communication Parlée consiste en un échange d'informations bidirectionnel entre au moins deux locuteurs. Sachant qu'elle a lieu dans un contexte où la rétroaction a une action permanente sur le dialogue oral, et qu'elle est à considérer comme un processus hautement hiérarchisé de cerveau à cerveau, on est en présence de trois phénomènes qui s'intègrent dans le processus :

- la superposition (ou mieux : le recouvrement) des caractéristiques acoustiques des sons successifs (par phénomène d'anticipation ou aussi de coarticulation),
- des modifications dues à la vitesse d'élocution, à l'accentuation,
- enfin, toute la gamme des variations d'une élocution à une autre, ou d'un locuteur à l'autre.

c) Revenons à la notion d'information et de réduction de l'incertitude attachée à une communication. Quelles sont les incertitudes attachées au signal de parole ?

Tout d'abord, le contenu du message linguistique. Ensuite il est reconnu que ce signal est quelque peu révélateur de l'anatomie et la physiologie du locuteur. Egalement, qu'il contient une composante émotionnelle... La figure 1 décompose les différentes "couches d'information" codées - chacune différemment - dans le signal de parole et qui s'appuient toutes sur une onde porteuse : la pression acoustique.

Figure 1 : l'information contenue dans la parole humaine. Le but avoué d'un locuteur est de transmettre des idées (niveau 2) par des mots et des phrases (niveau 1). En fait, il exprime plusieurs aspects de son patrimoine linguistique.



La détérioration due au bruit n'est pas indiquée sur le schéma. On voit que de nombreux facteurs du signal ne sont pas critiques du point de vue de la communication.

Figure 2 - Les informations contenues dans le signal de parole :

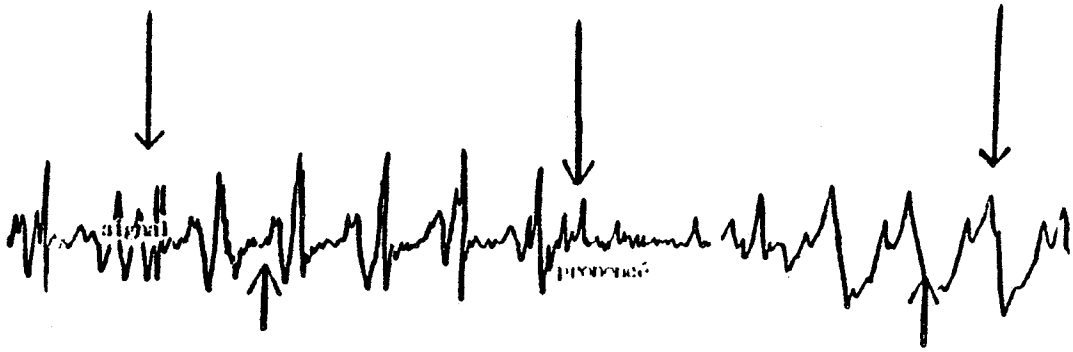
Elles appartiennent à des types qui sont autant de facteurs influant sur le signal. C'est la dynamique des relations entre chaque type qui crée la spécificité du signal de parole provenant d'un locuteur donné.

PATRIMOINES

Bilan socio-culturel

état émotionnel actuel propre au locuteur

état de l'environnement



origine géographique
(linguistique, dialectale)

disposition du locuteur vis-à-vis du message

(bilan émotionnel situationnel)

facteurs lointains
(un bilan)

facteurs immédiats
(un état)

La conjecture fondamentale est alors la suivante : y a-t-il des éléments (sans que l'on cherche à définir davantage ce mot) linguistiquement pertinents, et/ou pertinents du point de vue de l'identité du locuteur, etc... ?

Une question très importante est la suivante : est-ce que la totalité de l'information relative au locuteur est située dans certaines classes de sens (1), comme les nasales ? KASHYAP (1976) a répondu négativement à cette question. Cependant, tous les sons semblent contenir une information utile. La question

(1) Un son est la réalisation physique d'un phonème.

se transforme alors en : "quels sont les sons les plus pertinents", et elle nécessite l'investigation de séquences plus longues que des sons.
L'extraction de l'information pertinente vis-à-vis d'une composante donnée, sous-entend le choix (sélection) de ces éléments.

1.2. La nécessité de l'information temporelle :

En reconnaissance du locuteur, le domaine fréquentiel a été fortement étudié, plus que le domaine temporel du signal(1). On sait pourtant que la force de l'analyse de la parole dans le domaine temporel réside, entres autres, dans la détection et la caractérisation des transitions acoustiques. En outre, la résolution temporelle peut être aussi fine qu'on le souhaite. (Mais, meilleure est la résolution en fréquence, plus mauvaise est la résolution temporelle, et inversement). En outre, les phénomènes temporels ont une base perceptuelle et physiologique.

Les études en fréquence, qui bien sûr ne révèlent correctement l'information qu'en présence de signaux au moins presque périodiques, ont été dominées par des études spectrographiques, et par la détermination de composantes fréquentielles sur un intervalle de temps fixé. Cet examen de la tessiture formantique explique le grand usage qui a été fait des voyelles : on peut compter plusieurs périodes assez stables durant la tenue de la voyelle, surtout si celle-ci est accentuée. On sait que l'intelligibilité de la parole est due pour une bonne part aux transitions acoustiques (principalement les éléments consonantiques), et non aux voyelles. Qu'en est-il pour l'intelligibilité du locuteur ?
Les régions caractérisées par des transitions acoustiques semblent seulement observables par une résolution temporelle précise et fine. Cela pourrait établir la nécessité de l'étude temporelle du signal.
Les analyses fréquentielles et temporelles semblent être autant complémentaires qu'en reconnaissance de la parole. Il y a peut-être redondance entre les deux, comme cela est le cas pour la reconnaissance de la parole (cf. les études parallèles de BELLISSANT, 1978).

Au cours de la phase d'exploration des méthodes appartenant au domaine temporel, nous avons extrait des paramètres à partir du signal pour plusieurs

(1) BAKER (1975) invoque une raison à cela : la variabilité du signal temporel est telle qu'elle n'a pu être valablement réduite jusqu'à en tirer des paramètres robustes, c'est-à-dire peu sensibles aux variations pour un même locuteur.

locuteurs, à la main ou directement suivant le cas. De nombreux segments phonétiques ont été étudiés, incluant : des phrases complètes, un texte long, des groupes de phonation, des pauses et des sons, le tout avec une énonciation normale et naturelle. Ces études ont permis d'envisager la définition de procédure discrimination adaptées aux paramètres et au matériau phonétique en présence, et leur utilisation.

2 - Approche intuitive et vocabulaire

Reprenant HECKER (1971), nous définissons la locution : "reconnaissance du locuteur", par : tout processus de prise de décision fondé sur les variations inter-personnes du signal de parole(1). Cette définition au sens large, nous permet d'inclure différents moyens aboutissant à la dite reconnaissance : par l'audition, par l'analyse visuelle de spectrogrammes, par ordinateur. Toute méthode de reconnaissance du locuteur doit se fonder sur la diversité constatée de la réalisation phonique de la voix des personnes. Nous faisons tous, et quotidiennement, de la reconnaissance du locuteur. Au téléphone, bien sûr, à l'écoute de la radio, mais aussi dans les rassemblements où des "voix connues" nous parviennent. Nous bénéficions d'un nombre d'années égal à notre âge pour l'apprentissage de cette tâche, contrairement aux systèmes plus ou moins automatiques. Il y a d'ailleurs là un champ d'investigation qui semble présenter de l'intérêt sur l'étude du mécanisme de la mémoire, de la recherche des souvenirs, et leur comparaison avec les messages auditifs perçus, étude dont les résultats aideraient vraisemblablement pour la conception d'un système.

Etant donnée une certaine population de locuteurs, on dispose au moins d'un échantillon de voix de chacun d'entre eux, portant ou non sur le même contenu phonétique , et dont l'identité est connue. Il y a alors deux tâches fondamentales, que l'on distingue habituellement à cause de leur représentation de deux contextes d'application distincts, bien que, nous le verrons, elles soient assez fortement indépendantes des techniques de base employées.

---

(1) Les définitions plus récentes ne changent pas : ainsi, pour EL CHAFEI (1978) : "processus de décision effectué par un système employant certaines caractéristiques du signal vocal d'un certain message pour reconnaître le locuteur".

2.1. La Vérification :

Elle consiste simplement à décider l'acceptation ou le refus de l'admission à un bien quelconque d'une personne sur la base de :

a) une déclaration, par elle-même, de son identité[1] - supposée et annoncée comme telle - parallèlement à :

b) l'analyse de son signal vocal, émis au même moment et confronté à un modèle disponible de la voix du locuteur annonceur.

Cette définition éclaire déjà sur la méthodologie de la Vérification exposée au paragraphe 3 ci-dessous. Ainsi, la Vérification met en jeu une et une seule paire d'échantillons des voix : celui qui a été prononcé est comparé avec l'échantillon de référence. La décision donne lieu à une réponse du type oui/non : les deux échantillons proviennent/ne proviennent pas de la même personne (c'est-à-dire : sont suffisamment "semblables" ou non). La réponse n'est pas certaine mais comporte une erreur possible.

Il est à remarquer que l'on ne considère qu'un nombre fixé M[2] de locuteurs. La probabilité de décision incorrecte est indépendante de la taille M de la population.

Rappelons pour terminer que la Vérification nécessite deux entrées : l'échantillon de parole auquel est jointe une déclaration d'identité. On se reportera à la planche I pour une présentation visuelle, ainsi qu'à la figure 3.

2.2. L'Identification :

Disposant d'un échantillon de parole dont l'auteur est inconnu, il s'agit de retrouver l'identité de ce locuteur par comparaison avec une population de $N \geq 1$ locuteurs connus. A la différence de la tâche précédente, on effectue N comparaisons au lieu d'une seule, on classe les N échantillons par "ressemblance décroissante" avec l'échantillon témoin et on décide d'accepter/refuser l'identification avec le (ou éventuellement les) plus proche(s) échantillon(s), tout ceci sur la base d'une entrée unique.

Ainsi, la probabilité de décision incorrecte est ici une fonction croissante de N.[3] Paradoxalement, il se peut que l'échantillon inconnu appartienne à un individu ne faisant pas partie de la population de référence (Identification

(1) Par un lecteur de badge, de carte magnétique, la frappe au clavier d'un code alphanumérique, ou tout autre moyen.

(2) Qui peut être égal à 1.

(3) Bien qu'on ait d'autant plus de chance de faire figurer le locuteur cherché dans la population, que N est grand.

de type ouvert). L'hypothèse de travail fréquente → la personne recherchée est représentée par au moins un échantillon dans la population de référence (Identification de type fermé)→ permettra de mieux cerner la tâche de l'identification.

2.3. Discussion :

Evidemment, dans le cas d'Identification de type ouvert, et si N = 1, nous sommes dans le cas de la Vérification. D'une façon générale, les différences entre Identification et Vérification ont été discutées notamment par LI & al. (1966), DODDINGTON (1970). HECKER (1971) appelle Discrimination (et d'autres auteurs Authentification), la tâche de Vérification. Nous utiliserons exclusivement le terme discrimination lorsqu'il s'agit d'évaluer la dissemblance entre deux locuteurs, distincts ou non, quelle que soit la tâche en vue. EL CHAFEI (1978), utilise la notion de personne adhérente à la population de référence. Dans le cas de la Vérification, le locuteur entrant est généralement supposé être adhérent.

En résumé, la Vérification apporte une preuve, une attestation (c.a.d. une signature), tandis que l'Identification conduit à une indexation. Si l'on recherche des synonymes de ces deux termes, on s'aperçoit qu'ils se reflètent en majorité par leur signification spécifique, à un contexte d'application particulier. Pour nous, le mot reconnaissance s'emploie dans un sens général et recouvre tous les cas possibles. Le tableau 1 donne une liste d'acceptions relativement moins courantes qui sont susceptibles de jouer le rôle de synonymes.

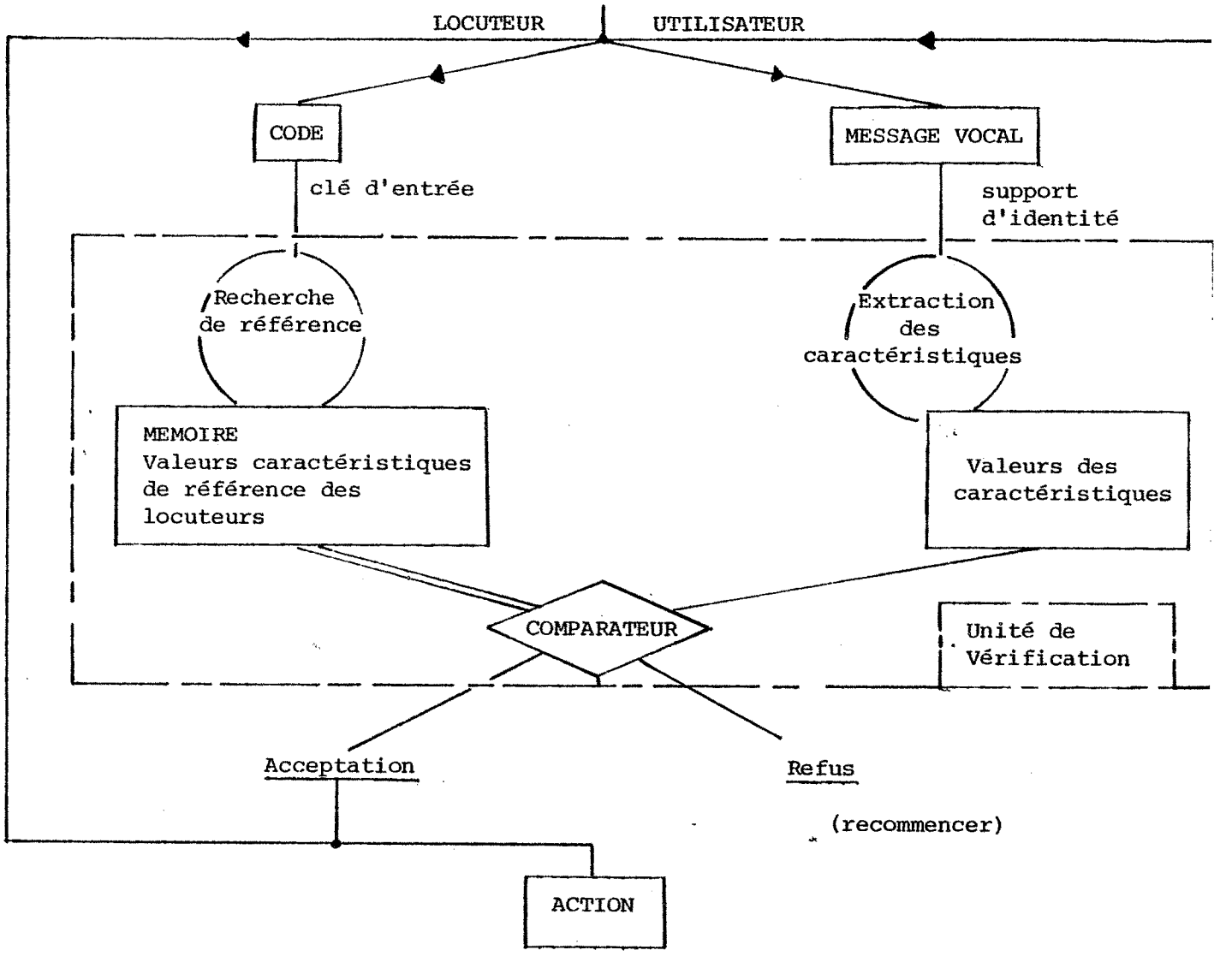
1 - Vérification	authentification justification légalisation	certification confirmation réclamation
2 - Identification	détection notification repérage	détermination recherche

Tableau 1

Formes paradigmatiques de la Vérification et de l'Identification.

PLANCHE I - Schéma directeur de la Vérification du locuteur

On compare le résultat de l'analyse de la voix du locuteur à un modèle en ordinateur de la voix de la personne dont l'identité est celle proclamée.



LEGENDE

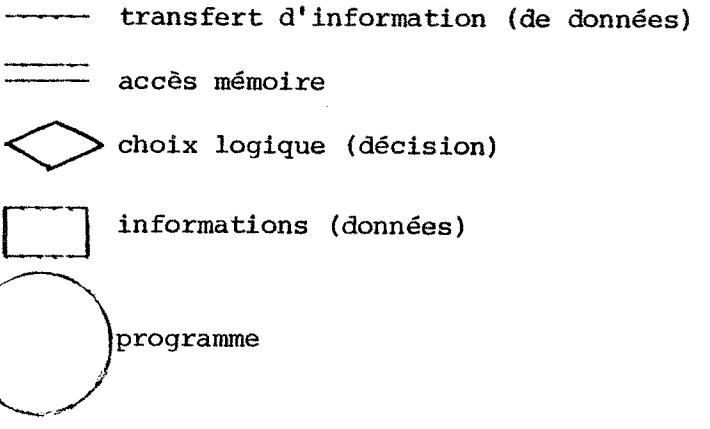
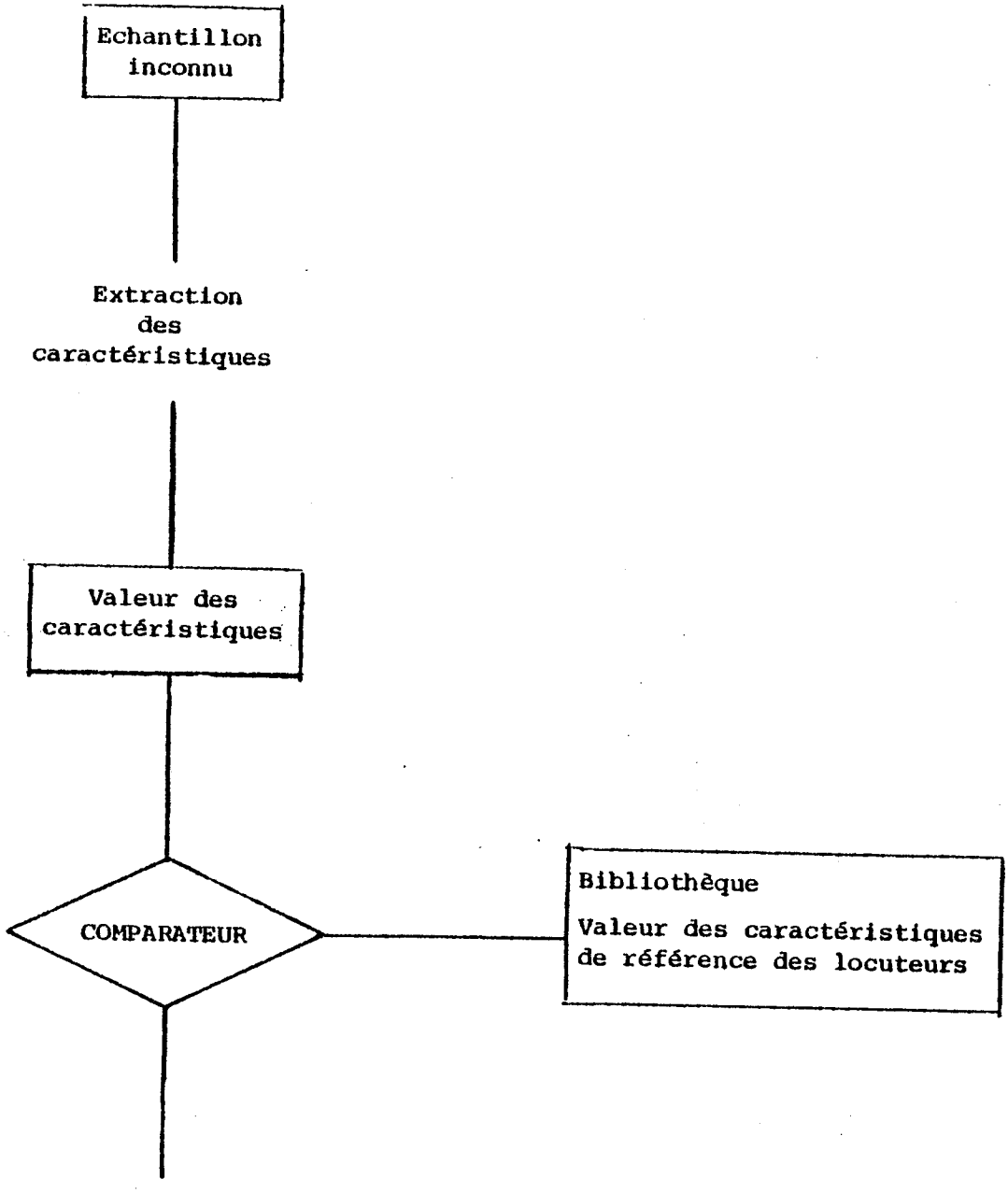


PLANCHE II - Schéma directeur de l'Identification du locuteur



2.4. Autres tâches

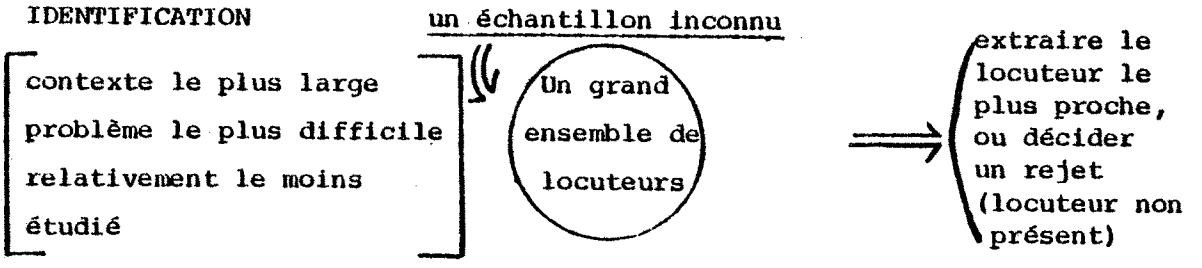
Le mot compréhension, si utilisé en traitement automatique de la parole, ne l'a pas été dans le présent contexte. Il suppose, on le sait, une perception globale de la sémantique d'une phrase (la compréhension consiste à saisir le sens de la phrase, même si la reconnaissance de cette phrase est imparfaite, bien que suffisamment correcte, de façon à exécuter l'action demandée ; la reconnaissance consiste à retrouver dans une phrase l'ensemble des éléments (mots, syllabes, phonèmes) la composant. La compréhension s'inscrit dans le cadre de l'intelligence artificielle, dont elle utilise les techniques : définitions comparées de HATON, 1977).

D'un point de vue de recherche fondamentale, l'étude des différences inter-individuelles à l'intérieur des groupes linguistiques, peut et devrait ultérieurement conduire à une certaine compréhension du système linguistique (en donnant au mot système un sens suffisamment étroit pour dépendre précisément de ces différences - ce qui n'a jamais été fait jusqu'ici[1]) dans lequel évolue chaque locuteur. La tâche de compréhension du système propre au locuteur prend alors tout son sens.

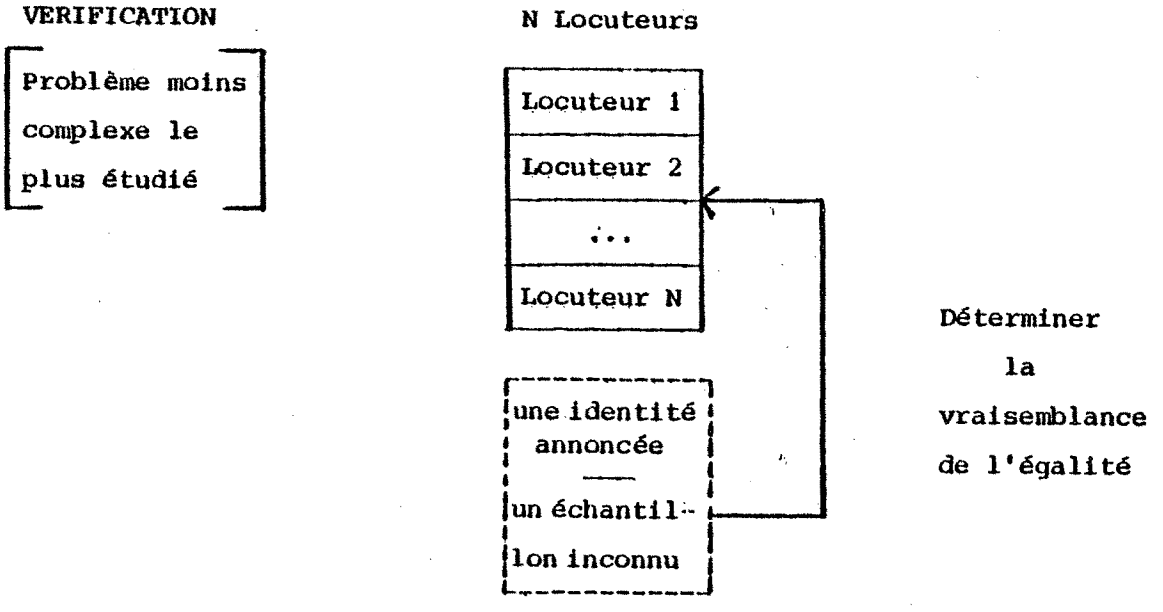
La planche III liste conjointement l'ensemble des tâches, en incluant la Discrimination qui est en fait une sous-tâche, sous-tendue à la fois par la Vérification, l'Identification et la Compréhension.

(1) Il est indispensable de se reporter à l'article de ABRY & BOE (1979), qui sera discuté quelque peu à la section II-I.

Planche III - Présentation conjointe permettant la comparaison des différentes tâches relatives au traitement du locuteur.



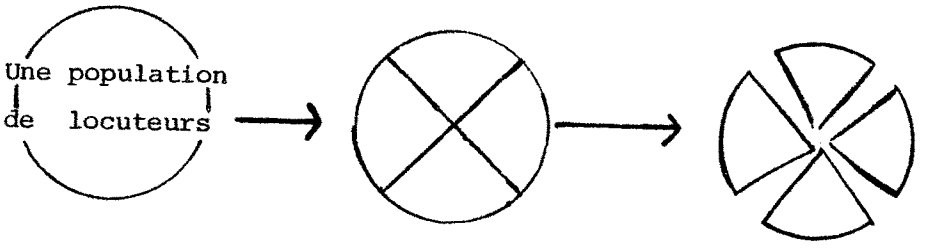
La probabilité de décision incorrecte (non reconnaissance du vrai locuteur), croît avec le nombre de locuteurs potentiels. Les modalités de cette croissance sont inconnues.



La probabilité de décision incorrecte est indépendante de la taille N.

Planche III - Suite

DISCRIMINATION



détermination de frontières séparatrices entre les locuteurs

minimisation du recouvrement (ou maximisation de l'éclatement).

La croissance de la difficulté de la discrimination avec la taille de la population n'est ni stricte, ni linéaire (par exemple, on sépare généralement mieux deux voix provenant de sexes différents que deux voix du même sexe).

TYPOLOGIE DE LOCUTEURS, COMPREHENSION DE SYSTEMES LINGUISTIQUES dépendant du locuteur.

C'est le problème de la discrimination pour une population non restreinte. En outre, les paramètres permettant la séparation des locuteurs doivent posséder une interprétation claire à l'intérieur d'un système de classification organisé, lui aussi à définir.

DETERMINATION DE L'ETAT EMOTIONNEL, ou d'autres particularités expressives.

Cette tâche relève d'abord de la composante expressive et émotionnelle de la voix. Elle est cependant parfois pratiquée dans le contexte de la reconnaissance d'un locuteur.

ADAPTATION AU LOCUTEUR d'une machine de reconnaissance automatique de la parole.

Cette tâche est très liée à toutes celles qui précèdent et exprime en fait une complémentarité vis-à-vis de l'ensemble des tâches particularisant les locuteurs (cf. Section I-II-3).

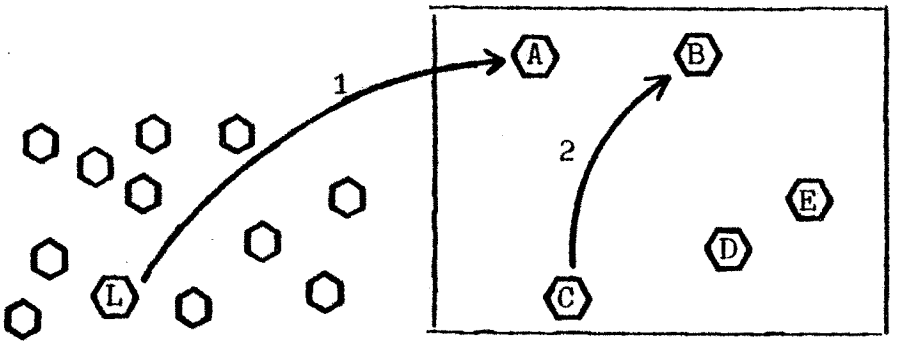
2.5. Les différents types de locuteurs et d'erreurs :

On ne peut exclure en pratique les deux cas irréguliers suivants :
- le numéro d'identification (ou code) est utilisé par une personne autre que son propriétaire (cas de la Vérification),
- le locuteur altère volontairement sa diction.

Mais il faut être sûr de l'exactitude de l'identité de la personne lors de l'élaboration de ses références, puisque l'initialisation des codes locuteurs conditionne la suite du traitement, lors de chaque remise à jour des références. Les deux cas d'irrégularité précédents conduisent à une situation d'imposture chez les locuteurs en cause. Nous définissons deux types de locuteurs imposteurs (voir figure 3) :

- Figure 3-Cas d'imposture :

Le locuteur L (resp. C) prend l'identité du locuteur A (resp. B) : imposture de type 1 (resp. 2).



a) type 1 : locuteur non adhérent, mais prétendant l'être,
b) type 2 : locuteur adhérent se faisant passer pour une autre personne.
Tout locuteur non imposteur sera dit locuteur vrai.

Nous ne pouvons pas avoir la connaissance de tous les imposteurs possibles. En d'autres termes, on ne peut modéliser l'imposteur.

Lorsque l'imposteur réussit son opération, il y a une erreur du système. Nous distinguons 2 types d'erreurs suivant la tâche traitée :

1. Vérification :

a) Fausse acceptation (ou fausse reconnaissance) : elle correspond à l'imposture de type 1 ou 2.

b) Faux rejet (ou fausse alarme) : rejet d'un locuteur vrai par le système.

Mentionnons que l'opération de Vérification réussit si elle accepte un locuteur vrai, ou aussi si elle rejette un imposteur de type 1 ou 2.

2. Identification : appelons locuteur réel (ou authentique), le locuteur inconnu à mettre en cause.

a) Fausse identification : acceptation d'un locuteur adhérent autre que le locuteur réel recherché (qui peut ou non être adhérent).

b) Non détection : rejet de tous les locuteurs adhérents, y compris le locuteur réel recherché (qui ici est adhérent, sinon il n'y a pas d'erreur).

2.6. Contrôle des divers types d'erreur :

Un point important consiste à déterminer les seuils de décision. Ils sont fonction de l'importance relative que l'on attribue à chaque erreur, Dans le contexte de la Vérification, il semble qu'il soit plus logique de pénaliser l'erreur a) (fausse acceptation) plutôt que b) (faux rejet) : en effet, la vérification automatique du locuteur doit avant tout réaliser une protection du système dont elle commande l'accès.
Ainsi, on cherche souvent à résoudre le problème :

minimiser (Prob(fausse acceptation)),

sous la contrainte :

Prob(faux rejet) ≤ constante.

Nous appellerons taux d'erreur la probabilité d'erreur de décision de tout type. Le but est de réduire le taux d'erreur du système. Un rejet est la décision de rejeter (abandonner) l'échantillon courant. Il peut correspondre à une bonne

décision aussi bien qu'à une mauvaise.

On connaît la correspondance entre taux d'erreur et taux de rejet dans le cadre bayésien (pour une règle de décision de risque minimum).

Il importe donc de concevoir une stratégie qui autorise un certain contrôle des erreurs et, le cas échéant, des rejets. Elle passe par une évaluation des coûts des erreurs respectives.

2.7. Les différents états des locuteurs :

Nous avons exposé les types de locuteurs en relation avec les erreurs possibles. Mais ces types sont des résultats d'un souhait de la part des locuteurs ou bien, peuvent être fortuits. Au niveau du locuteur, il y a donc la distinction suivante :

1. Locuteur imposteur :

a) Imposteur volontaire : a modifié volontairement sa voix, ou certains aspects de sa voix[1]. C'est le cas notamment de l'imitation, où la voix est sensée se rapprocher de celle d'une autre personne : Vérification souhaitée avec une autre personne. Sinon, c'est le cas du souhait de l'échec de l'Identification avec soi-même.

b) Imposteur fortuit : il y a eu modification accidentelle de la voix, par exemple à cause de la dérive temporelle de la voix, d'un état pathologique, etc.

c) Imposteur naïf[2] : imposteur non volontaire et non fortuit.

2. Avant que le locuteur ne soit reconnu d'un type particulier, il faut faire l'importante distinction recouvrant tous les cas du :

a) Locuteur coopératif (ou "dédié") : s'efforce d'être naturel parce que c'est son intérêt (cas de la vérification avec un locuteur vrai).

b) ocuteur ordinaire : élocution la plus naturelle possible (ce cas ne se prête guère à une définition).

---

(1) Comme pour d'autres plans expressifs de la personnalité, ces altérations artificielles sont probablement trahies par des phénomènes compensatoires.

(2) ROSENBERG (1972), utilise l'expression " casual impostor".

c) locuteur intrus : cherche à modifier sa voix parce que c'est son intérêt.

Ces états doivent être posés clairement lors des expériences puisqu'ils ont valeur d'hypothèses de travail.
En particulier, le contexte de la Vérification bénéficie des conditions suivantes (qui s'avèrent favorables du point de vue de la fiabilité des techniques) :

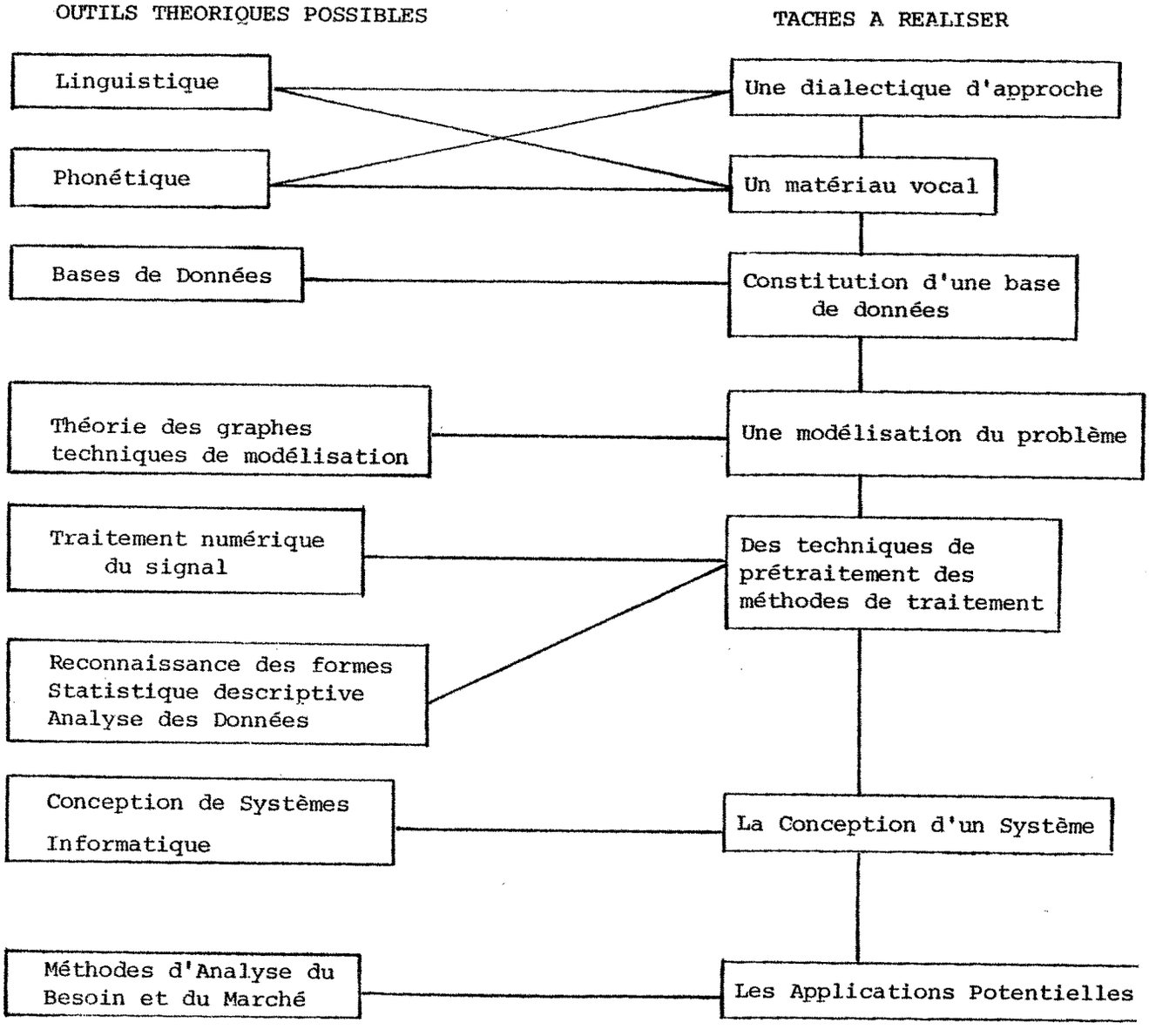
- les locuteurs sont coopératifs lors de l'apprentissage (ce n'est pas le cas de l'identification !) : ils voudront être authentifiés à chaque fois,
- le contenu phonétique de l'apprentissage et du test est généralement le même, et il peut être commandé par la machine,
- il est possible de faire répéter sur place le locuteur en cas de rejet,
- on peut contrôler l'environnement acoustique avec obtention d'un bon rapport signal sur bruit,
- on peut adapter le choix des paramètres au locuteur.

II - LA SITUATION GENERALE DU SUJET DANS SES DIVERS CONTEXTES

1 - La pluridisciplinarité du sujet :

Dans la mesure où l'on propose une approche quelque peu systématique de la reconnaissance du locuteur, c'est-à-dire qui se veuille générale, il faut admettre d'emblée - me semble t-il - qu'une telle étude ne peut être décemment entreprise par une approche purement informatique, car il manquerait une connaissance spécifique du signal vocal. Comme on le verra ultérieurement, la connaissance des différences individuelles dans la voix est encore faible.
Il faut donc conduire une investigation majeure dans le domaine de la phonétique afin de savoir ce que l'on veut chercher qui soit pertinent, et quelle peut en être la raison a priori. Les études aveugles à grand renfort de techniques mathématiques à la mode, ne sont pas toujours d'un grand secours à la longue. On aime pouvoir interpréter les phénomènes mis en jeu, ainsi que comparer les performances obtenues.
Pour cela, il faut partir de la spécificité (bien réelle) du signal de parole et ménager à chaque niveau des moyens d'interprétation de ce qui est fait.
La chaîne complète de traitement exposée au paragraphe 3 ci-dessous préfigure le schéma ci-contre.
La science phonétique comprend au sens large, trois branches : l'articulatoire, qui est l'étude des commandes mécaniques dont le larynx est un organe important, l'acoustique qui étudie le signal de parole et la phonétique perceptive qui concerne le système oreille-cerveau.

Figure - Mise en évidence graphique de la pluridisciplinarité du sujet. Au fur-et-à-mesure de sa "résolution", la reconnaissance du locuteur par ordinateur fait appel à des disciplines très variées, dont la liste ci-dessous n'est pas limitative.



2 - Le contexte de la communication orale homme-machine :

Par communication homme-machine, on entend habituellement un type de communication qui a des points communs avec la communication humaine inter-personnes. On est alors amené à se poser des questions en retour, telles que :

q1 - Comment reconnaissons-nous les locuteurs lors de l'audition ?
Quels sont les processus impliqués ?

q2 - Pouvons-nous réellement imiter le comportement vocal d'autres personnes ?
Quelle est la vraisemblance de cette imitation ?

q3 - Peut-on classer les voix d'une façon relativement intrinsèque ?

Les réponses à ces questions, et surtout aux deux premières, ne nous aideraient pas obligatoirement : par exemple, dans q1, les paramètres humains ne sont pas nécessairement les bons paramètres pour une machine, et dans q2, l'imitation humaine semble ne porter que sur des modifications macroscopiques.

Bien qu'il y ait d'assez nombreuses formulations mathématiques pour décrire les processus de la production de la parole (FANT(1960), FLANAGAN (1972)), , il n'existe pas de théorie générale pouvant résoudre le problème ultime de la sélection de la partie informative dans le signal de parole, c'est-à-dire de la partie utile du point de vue de la reconnaissance. D'ailleurs, la communication orale homme-machine a été souvent définie jusqu'ici en termes d'antagonisme : message-locuteur ; soit on cherche à faire comprendre à une machine le contenu du discours sans se préoccuper du locuteur en tant que tel, soit on fait de la Vérification ou de l'Identification de locuteur sans regard sur la reconnaissance du message pouvant être émis, ce qui est signal dans un cas, devient bruit dans l'autre, et inversement. Pourtant, remarque COLLINS (1979), "reconnaissance du locuteur et reconnaissance de la parole sont intimement reliées à cause du fait que l'on ne sait pas séparer l'information sémantique de celle due au locuteur".

L'attitude précédente semble due à ce que la plupart des théories linguistiques classiques cherchent à rendre compte uniquement du message. GUIBERT (1979), l'affirme : "bien peu considèrent la caractérisation de l'identité du locuteur

comme du domaine d'étude de la linguistique. DE SAUSSURE a signalé ce problème dans son cours de linguistique générale (1916), mais ce n'est qu'assez récemment que l'interrelation entre la personnalité et le langage a été l'objet d'études approfondies".

Ainsi, ce qui limite la reconnaissance au niveau phonétique, est ce qui justifie la caractérisation de l'identité du locuteur. La variabilité du signal de parole entre locuteurs distincts sera donc à considérer; nous l'appellerons variabilité extrinsèque ou variabilité inter-locuteur.

3 - Le contexte idiolectal et les aspects de complémentarité message/locute

Le vrai contexte d'étude de la reconnaissance du locuteur est l'étude des caractéristiques individuelles ou traits idiolectaux(1). Ce qui est donc en cause, c'est l'approfondissement des mécanismes de perception. Cependant, l'étude de ces descripteurs de différences individuelles entre patrimoines vocaux des locuteurs, n'est pas très avancée. Nous reportons le lecteur à la section II-I, sur l'étude de la variabilité de la voix. Nous nous bornerons ici à mettre en évidence les variabilités de la voix.

Il existe des paramètres perceptifs caractérisant un son (intensité, hauteur, spectre, durée,...), mais la question est de savoir si les paramètres sont, ou bien linguistiquement pertinents et/ou bien pertinents du point de vue du locuteur. On distingue les paramètres :

- micro-structuraux, véhiculant une information micro-structurale : sons,...
- macro-structuraux, se rapportant à l'intonation,...

Le fait que la voix soit, dans une très relative mesure - et pour certains locuteurs - imitable, devrait renseigner sur :

- l'existence de la variabilité intra-locuteurs et son amplitude,
- la variation de l'étendue de la variabilité intra-locuteurs suivant les sujets.

3.1. Evidence de la variabilité inter-locuteur :

Il est habituel de constater l'augmentation des erreurs en reconnaissance, lorsqu'on augmente le nombre de locuteurs distincts. C'est que les difficultés supplémentaires de reconnaissance de mots dits par une autre personne que celle

(1) Un idiolecte est un ensemble de réalisations propres à un locuteur.

qui en a fait l'apprentissage, sous-tendent une variabilité d'un locuteur à l'autre, c'est-à-dire liée à leur identité. Ainsi - afin de préserver un niveau de performance donné - les machines à reconnaître la parole nécessitent généralement des ajustements pour chaque nouveau locuteur. Ceux-ci ont d'abord été accomplis manuellement, pas essais et erreurs. Leur automatisation passe par une modélisation des locuteurs.

3.2. Evidence de la variabilité intra-locuteur :

On sait que l'on améliore les performances en reconnaissance, si l'on accroît l'étendue temporelle des données enregistrées de référence : enregistrements couvrant mieux une même journée, s'étalant sur plusieurs jours et semaines consécutifs, etc.

En outre, l'absolescence de la machine peut être très rapide en cas d'utilisation de références trop vieilles, datant de plusieurs mois par exemple.

3.3. Le point de vue idiolectal en phonétique :

Puisque, lorsque l'on communique par la parole, ces deux facteurs sont prései
- différences individuelles caractérisant le locuteur et l'auditeur,
- connaissance commune de la langue (champ sémantique commun) et expérience d'une même pragmatique,

il devient nécessaire - pour décrire cet acte - de comprendre à la fois les caractéristiques individuelles de la parole et les règles générales du langage. Schématiquement on peut dire que l'intersection de ces réalisations est appelée code linguistique et est évidemment directement utilisable en reconnaissance automatique, et probablement dans le processus du décodage par l'auditeur.
Les différences individuelles ont de multiples causes, nous renvoyons à STEVENS (1971) pour une analyse systématique de leurs origines physiologiques. ABRY & BOË (1979) ont insisté sur le fait que jusqu'ici la variabilité a été traitée hors système phonétique et remarqué que par exemple : "la plupart des procédures de normalisation des espaces vocaliques tendent à se débarasser de la variance interlocuteur". Pourtant ils ont avancé l'existence à l'intérieur d'un même système phonologique, de réalisations différentes, c'est-à-dire qu'il y a une part d'individuation du système lui-même.

Il devient nécessaire de trouver une description linguistique de la performance des locuteurs qui s'applique à la reconnaissance automatique.

3.4. Définition et comparaison entre reconnaissance du message, et du locuteur :

a) En reconnaissance et en compréhension multi-locuteurs, on cherche fondamentalement à normaliser le message d'un locuteur donné, pour l'adapter à une référence. C'est en quelque sorte une reconnaissance de mots parlés (ou phrases, ou autres réalisations d'unités phonologiques.
La procédure duale consiste à tirer parti des différences inter-individuelles, jusqu'à pouvoir séparer (au sens d'une métrique et d'un algorithme de classification), deux locuteurs quelconques : c'est l'identification du locuteur par discrimination.
Ce premier aspect majeur de la complémentarité message/locuteur, est symbolisé sur la figure 1.

b) Il existe une méthode de reconnaissance qui résout trivialement la question précédente : c'est le recours à un apprentissage distinct par locuteur (voir la partie supérieure de la figure 2).

Nous allons voir qu'il y a là un aspect de complémentarité, relativement faible : en effet, étant donnée une référence nouvelle (un échantillon de test), la question réciproque consiste à savoir s'il provient de tel locuteur donné (vérification du locuteur).

Figure 1 - Représentation symbolique du premier aspect de dualité message/locuteur

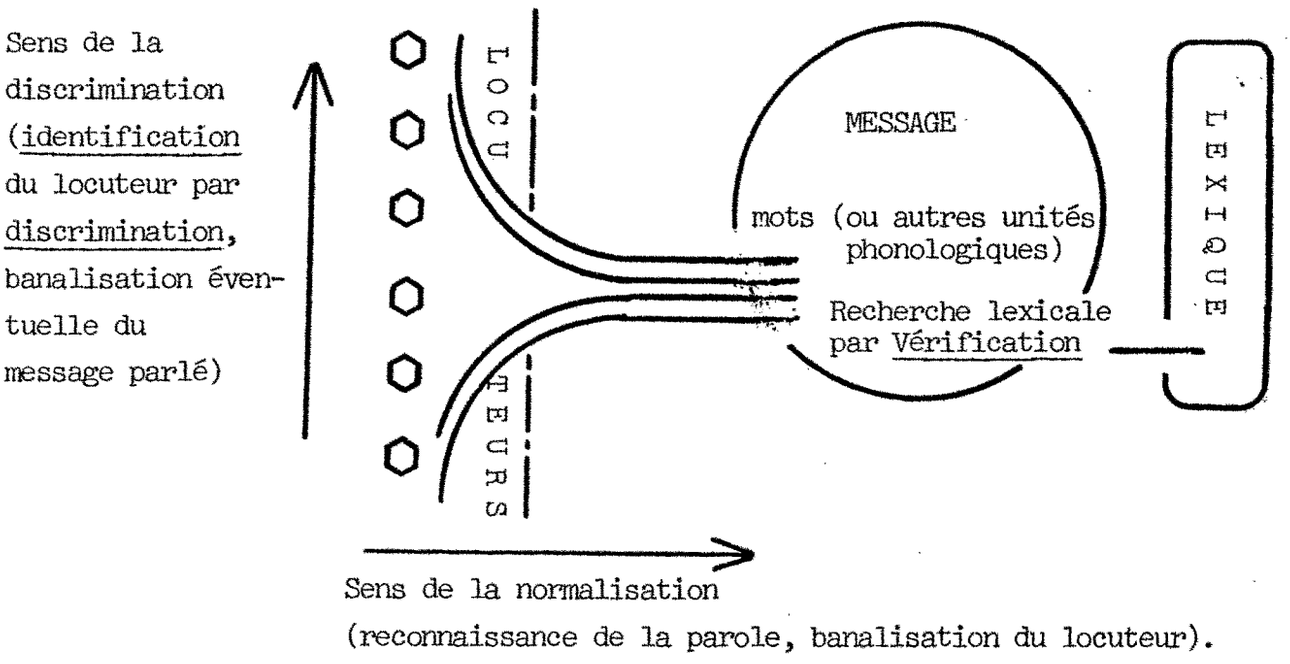
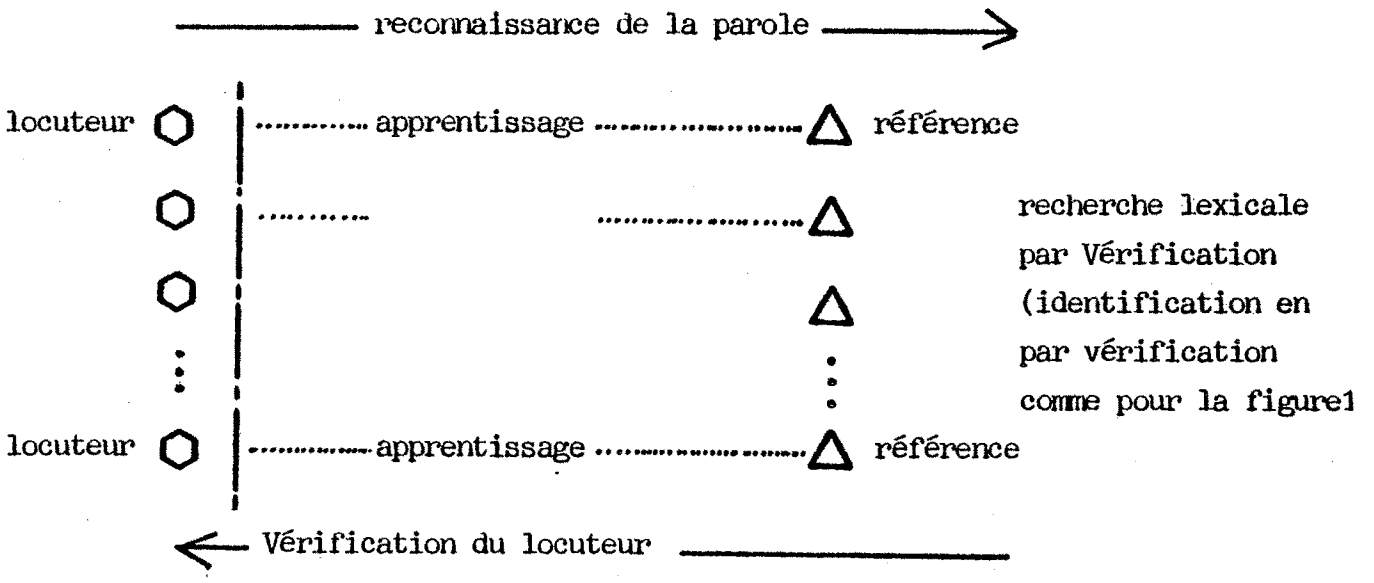


Figure 2 - Schématisation du second aspect de dualité message/locuteur.



Adaptation : (la spécificité décroît, l'adaptabilité des références croît)

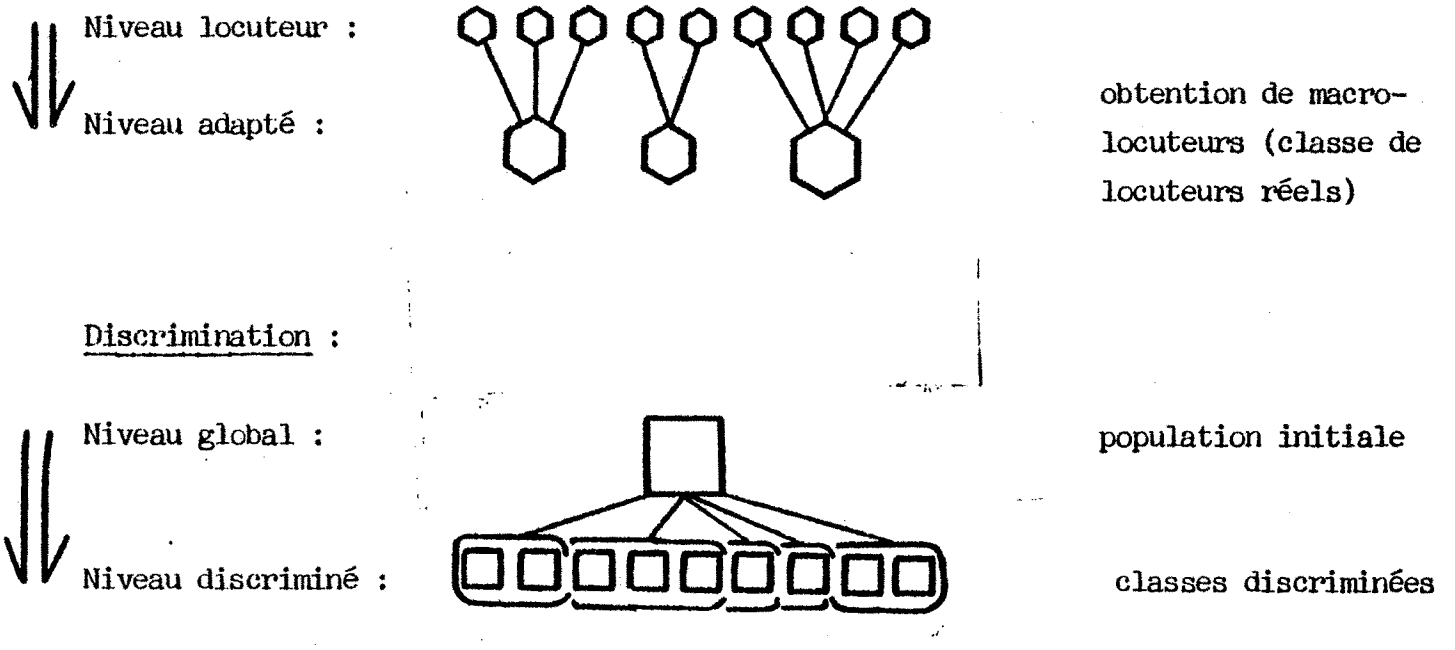


Figure 3 - Mise en évidence des natures respectives de l'adaptation et de la discrimination sur une population de 9 locuteurs. La discrimination est une opération d'analyse, l'adaptation de synthèse.

La figure 2 schématise cet aspect de complémentarité.
Il faut enfin mentionner le "phénomène" d'adaptation volontaire ou non du locuteur à la machine, conséquence de l'acquisition d'une certaine expérience de l'expérimentateur de la machine à laquelle il parle.
Comme conséquence de ce qui précède, on vérifie bien que deux obstacles fondamentaux à une bonne reconnaissance de la parole sont le nombre de locuteurs, et leur mode d'élocution (plus le nombre de voix est élevé, et moins elles sont coopératives, plus il est difficile en moyenne de reconnaître le mot prononcé).
La figure 3 situe les natures respectives de l'adaptation et de la discrimination, pour une taille de la population fixée.
Un troisième obstacle est bien sûr constitué par la taille du vocabulaire.
Enfin, on se sera aperçu que la reconnaissance de la parole est une opération d'identification et non de vérification.

3.5. La complémentarité au niveau méthodologique :

Les travaux précédents indiquent que, pour être économiquement rentable, un système de reconnaissance de parole doit être souple, c'est-à-dire pouvoir s'adapter, avec un coût minimum, à chaque locuteur potentiel. Cette exigence implique l'emploi de méthodes quelque peu générales, tant au niveau des capteurs et des prétraitements, qu'à celui de la stratégie de normalisation.

Prenons un exemple : la reconnaissance des voyelles est entravée par les différences inter-locuteurs. Si ces différences sont connues, la reconnaissance pourra s'adapter au locuteur courant, sinon un échantillon court de texte (éventuellement une partie antérieure du message déjà émis par le locuteur), peut être utilisée à des fins d'ajustement.
Citons ici WAKITA (1975) et KASHYAP (1976), qui effectuent une reconnaissance simultanée de voyelles et du locuteur, en utilisant les coefficients de prédiction linéaire.
Différentes méthodes d'adaptation ont été proposées (récemment, GRENIER (1977), id. MAURIN (1979) : analyse des corrélations canoniques).

On trouvera dans le tableau suivant quelques points de complémentarité, entre :
a) la discrimination de locuteurs (séparation en classes plus ou moins fines),
b) la banalisation de locuteurs (c'est-à-dire l'adaptation de la machine au locuteur). N représente la taille de la population considérée, et i est l'indice du paramètre (ou caractéristique) considéré.

Côté Adaptation (banalisation)	Côté Discrimination
1) - recherche de paramètres universels (valable pour tous les locuteurs) ou de grandes classes de locuteurs exemples : fricatives, plosives distinction voisé/non-voisé indice d'universalité : U_i	- paramètres offrant un F-ratio élevé (ou tout autre critère d séparation). Mais il n'y a pas d'heuristique permettant d'aug menter le F-ratio. Mais résistance aux déformatio de la voix d'une même personne indice de spécificité : S_i
(un paramètre bon (mauvais) pour la reconnaissance, est mauvais (bon) pour la discrimination). Les deux indices sont duaux	
U_i = taux de locuteurs "bien" adaptés $\frac{1}{p} \sum_{1}^{p}$ ui est la capacité d'adaptation moyenne du système $\frac{1}{N} \left\| \bigcup_{i} \{\text{locuteurs adaptés par } U_i\} \right\|$ est la capacité d'adaptation totale du système	C_i = taux de "bonne" discrimination $= \frac{\text{nombre d'arcs manquants}}{N(N-1)/2}$ (il existe un arc entre 2 locuteurs si et seulement si ils sont non assez discriminés).

2) Recherche d'un algorithme d'adaptation au locuteur Exemple : modification linéaire de l'ensemble des valeurs d'un paramètre. indice d'adaptabilité : a_i a_i est la facilité, simplicité (évaluée selon des critères de temps de calcul) avec laquelle le paramètre i est mesuré.	Algorithme de cheminement dans arbre de décision : les sommets sont les paramètres, le problème est de trouver la meilleure suite ordonnée de paramètres pour séparer la population en classes jusqu'à éventuellement identifier un locuteur. Même situation, mais simplifiée pour la Vérification.
3) Utilisation prioritaire de paramètres simples. La simplicité est évaluée selon le coût et le temps du prétraitement, des calculs, de l'interprétabilité des paramètres. Exemple : fondamental formants	Ces mêmes paramètres simples. Exemple : on peut distinguer le sexe par la longueur du pharynx

Le fait que les deux tâches discutées ici soient complémentaires, ne signifie évidemment pas qu'elles puissent être accomplies par un algorithme unique. Par exemple, la séparation de la tâche de reconnaissance du message parlé et de l'identification de l'auteur du message, élimine une étape de décision, réduit le volume des calculs, et aussi le taux d'erreur global final.

Cependant, et GUIBERT (1979) l'affirme, les prochaines machines à reconnaître la parole, procèderont à la caractérisation du locuteur, ainsi qu'à celle du message. On assiste déjà à la conception de tels appareils aux Etats-Unis. Bien sûr, la question reste de savoir si ces deux opérations seront

simultanées (processeurs parallèles ; la méthode de cette simultanéité ne semble pas encore claire) ou séquentielles. Pour certaines applications, il peut suffire de catégoriser le locuteur sans l'identifier, ceci dans le but de fournir la référence adéquate ou le modèle adéquat pour la reconnaissance du contenu de son message vocal courant.

3.6. Conclusion :

On a mis en évidence que la reconnaissance de la parole et la reconnaissance de locuteurs sont deux tâches fortement complémentaires. Cette complémentarité se situe au niveau de la dualité message/locuteur, et s'appuie en outre sur des méthodologies relativement duales. Un système de reconnaissance de la parole fonctionne de manière optimale lorsqu'il connaît le locuteur. Comme conséquence de ceci, il est ressenti le besoin d'étudier les différences individuelles au niveau phonétique, dans le but de permettre une fine modélisation des locuteurs, utilisable à la fois en reconnaissance des messages parlés et de l'identité des locuteurs. Cette modélisation fondée sur les réalisations phonétiques des individus, devrait permettre de définir des typologies intrinsèques et prenant place à l'intérieur d'un même système linguistique. Elle offre en outre pour les systèmes de reconnaissance de la parole un niveau de traitement supplémentaire se superposant aux niveaux habituels acoustique, phonologique, morphologique, syntaxique et sémantique.

Cependant, dans le cas de compression de parole, comme c'est le cas pour les vocodeurs , la transmission économique du signal qui s'accompagne d'une faible réduction de l'intelligibilité du message, devra ne pas détruire l'information relative au locuteur.

4 - Le contexte de la reconnaissance des Formes :

La méthode d'analyse, en Vérification, fait appel au principe de reconnaissance par comparaison d'une forme à une classe de formes : comparaison du signal vocal courant au modèle témoin correspondant à la personne dont l'identité est celle proclamée. Ainsi, le résultat - c'est un résultat d'une décision - est fondé sur un apprentissage supervisé du système, préalable à toute opération de

reconnaissance d'un certain nombre de formes prototypiques.
L'Identification, si elle se fonde sur le même principe très général, fait appel à un apprentissage supervisé, préalable dans la mesure où l'on veut distinguer un locuteur de tous les autres (ces autres locuteurs doivent être connus).

Il y a enfin apprentissage non supervisé, lors de l'étude d'une population de locuteurs - en nombre connu ou non - où l'on cherche à percevoir les voix en présence et à les séparer. Leur perception plus leur séparation, constituent ce que l'on a appelé la Discrimination.
Mais c'est essentiellement au niveau des applications que les trois tâches diffèrent.

Nous renvoyons le lecteur à la section III-V, pour une discussion détaillée de la méthodologie de la reconnaissance des formes, appliquée à ce sujet, et à KANAL (1974) pour une présentation générale des tendances récentes de ces recherches.

III - LE POINT SUR LE SUJET :

Panorama sommaire des recherches en reconnaissance du locuteur

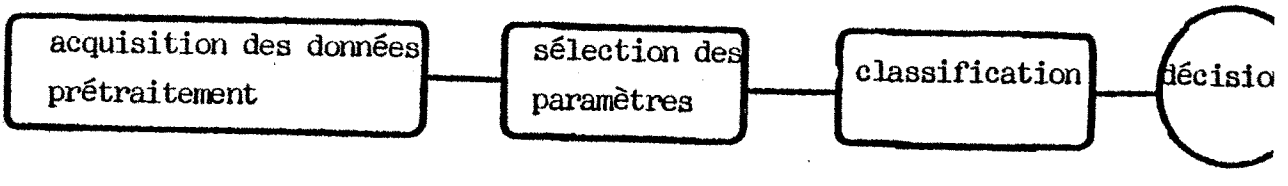
1 - Vue d'ensemble sur la période 1938-1979.

Si l'on cherchait à classer les travaux effectués, on pourrait le faire en fonction de chacune des caractéristiques suivantes :

- le contexte de l'étude (Vérification, Identification, Discrimination),
- paramètres utilisés,
- le nombre de locuteurs pris en compte dans l'expérience,
- les supports linguistiques et phonétiques utilisés (sons, logatomes, phrases, textes, ...),
- les méthodes de classification,
- la performance obtenue.

D'autres types de classifications (par matériel informatique utilisé, ...), semblent mineures. Ce travail de classification a été quelque peu entrepris dans les deux panoramas du sujet, présentés par ATAL (1976) et ROSENBERG (1976). On se bornera ici à quelques commentaires locaux et à la mise en évidence des motivations et des tendances. Quant aux fondements du problème, ils peuvent être recouverts par les sections I, II et III.

Une première remarque est que les diverses étapes de la chaîne de traitement :



n'ont pas été étudiées en même temps et les études ont porté essentiellement sur l'un ou l'autre de ces aspects du traitement.
Il ne manque pas de précurseurs de l'étude de la parole en liaison avec la personnalité et l'individualité (SAPIR, 1927 ; TAYLOR, 1933 ; HERZOG, 1933 ; ALLPORT & CANTRIL, 1934 ; BONAVENTURA, 1935). Mais il faut surtout citer l'oeuvre de KAISER (1938,...)

Une première étude prometteuse sur les "empreintes vocales"est réalisée par GRAY & KOFF (1944) (cit. par EL CHAFEI).

Les premières liaisons avec les questions criminelles apparaissent avec Mc GEHEE (1937, 1944), qui a eu à résoudre un problème d'origine légale (The Lindbergh Case, 1938).
Parmi les méthodes de reconnaissance, la reconnaissance par l'audition est la plus ancienne (Mc GEHEE, 1944), et a suscité un maximum d'intérêt, également à cause de la comparaison avec des méthodes plus récentes, vers 1965-1970 : VOIERS, 1964 ; BRICKER & PRUZANSKY, 1966 ; HOLMGREN, 1967 ; STEVENS & al., 1968 ; CLARKE & BECKER, 1969 ; ROSENBERG, 1973. PTACEK & SANDER (1966) ont tenté la reconnaissance auditive de l'âge.

La reconnaissance par utilisation de spectrogrammes a rapidement fait l'objet de commentaires passionnés, lesquels ont oscillé entre deux tendances :
- soit un éloge de la méthode, prétendue très sûre lors des applications pratiques (KERSTA, 1962 ; cet auteur a développé une méthode d'entraînement et un laboratoire spécialisé),
- soit des mises en garde sévères quant à son utilisation, sous prétexte qu'il est nécessaire de faire progresser les recherches permettant d'établir clairement les limites de la méthode (BOLT & al., 1969), ou tout au moins, des critiques quant au manque de fiabilité de la méthode (HOLLIEN, 1974).

TOSI & al. (1970-1972), ont essayé de clarifier la situation : sous des conditions idéales (1), un locuteur peut selon lui être reconnu parmi une population de 10 à 40 individus avec presque certitude (99%).

Il est à remarquer que les grands laboratoires (Bell Laboratories, ...), ne se sont jamais impliqués dans l'un ou l'autre des points de vue. Pour BOLT & al. (1973), aucune explication ne semble pouvoir être donnée quant à savoir comment (sous quel prétexte), peut-on rejeter, dans le cadre de tests non contemporains, le spectrogramme inconnu, dès que l'on a un doute.
Divers auteurs ont essayé de décrire les voix, par exemple en considérant les paramètres perceptifs (BORDONE-SACERDOTE & SACERDOTE, 1969). D'autres auteurs ont utilisé les paires d'attributs opposés, par exemple : intense-faible ; celles-ci sont peu utilisables pour une procédure automatique en raison de leur subjectivité.

---

(1) Spectrogramme de mots isolés, test contemporain (fait à la même époque que l'apprentissage des observateurs entraînés). Une expérience avec des tests postérieurs de un mois à l'apprentissage, ont donné 18% d'erreur.

Enfin, pour des études comparatives entre reconnaissance par audition et spectrogramme, il faut citer STEVENS & al. (1968).

PRUZANSKY (1963), a ouvert la voie aux recherches modernes, en traitant les spectrogrammes par 17 bancs de filtres, et en utilisant la corrélation entre deux spectrogrammes digitaux pour mesure de similarité. PRUZANSKY & MATHEWS (1964), ont amélioré cette technique, ainsi que LI & al. (1966), par l'emploi de discriminations linéaires.

DODDINGTON (1971) a remplacé les bancs de filtres par l'analyse des formants, et l'idée suivant laquelle les formants élevés sont plus représentatifs du locuteur s'est imposée : ceux-ci correspondent à des différentiations plus fines, qui ne sont donc pas caractéristiques de l'espèce humaine mais plutôt des individus.

Jusqu'ici, l'étude des spectrogrammes a donné une importante égale aux dimensions fréquentielle, temporelle et d'amplitude du signal. On assiste à cette époque, à une ramification des méthodes, lesquelles s'établissent sur le choix de l'une de ces dimensions :

- la méthode fréquentielle, qui cherche à éliminer la variable temps, soit par un calcul de moyenne sur une durée suffisante, ou bien l'extraction de paramètres statistiques, l'utilisation du spectre à long terme (BORDONE-SACERDOTE & SACERDOTE, 1969), de la matrice de covariance des spectres instantanés (LI & HUGUES, 1974), de la moyenne de l'auto-corrélation (YEGNANARAYANA & al., 1977 ; BRICKER & al., 1971), des histogrammes sur le fondamental et le spectre (HUNT & al., 1977 ; BEEK & al., 1971), des coefficients de prédiction linéaire (SAMBUR, 1976). La méthode fréquentielle tend à être indépendante du contexte phonétique, surtout si les paramètres sont extraits sur des durées suffisamment longues.

D'une façon générale, le domaine fréquentiel a précédé et a été considérablement plus étudié que les autres. (AL PERT & al., 1963 ; BORDONE-SACERDOTE & SACERDOTE, 1969 ; BRICKER & al., 1971 ; FURUI & al., 1972 ; FURUI, 1974 ; LI & HUGUES, 1974 ; YEGNANARAYANA & al., 1977 ; MAJEWSKI & HOLLIEN, 1974 ; ZALEWSKI & al., 1975). Ces études s'appuient sur des recherches antérieures sur le filtrage (POLLACK & al., 1954 ; PETERS, 1954, 1956 ; STARKWEATHER, 1956 ; CLARKE & al., 1966, ...), et suscitent toujours beaucoup d'intérêt (par ex GRENIER, 1977).

Les méthodes temporelles sont plus récentes et semblent rester l'apanage de quelques chercheurs (ATAL, 1972, 1974 ; ROSENBERG & SAMBUR, 1975 ; SARMA & YEGNANARAYANA, 1976 ; VIDALON & al., 1977).
IBBA & al. (1979) ont mesuré la durée d'établissement du voisement pour les consonnes [k], [t] et [p] en position intervocalique, et constaté, par un test statistique sous des hypothèses de normalité, la présence d'un "effet locuteur" et également de fortes corrélations entre groupes de paramètres, définis sur une même consonne (par exemple : [pa], [pe], [po]).

- la méthode temporelle est par contre fortement dépendante du contexte et se heurte au problème de la segmentation. Elle a traité le taux de passage par zéro du signal dans différentes bandes de fréquences (LIN & PILLAY, 1976), le spectre de certaines voyelles et nasales (WOLF, 1972 ; SAMBUR, 1975), les formants calculés par prédiction linéaire (GOLDSTEIN, 1976).

Il est vrai que la durée des voyelles dépend de l'accentuation, mais aussi et surtout, des consonnes qui les suivent, et, dans une moindre mesure, les précèdent (PETERSON & LEHISTE, 1960) ; (elles sont plus longues: 0,1 s. pour les consonnes sonores, que pour les consonnes sourdes : 0,08 s).
D'autre part, les voyelles sont très sensibles à la coarticulation (par ex. : DEMICHELIS & DE MORI, 1978). Selon CHEN (1970), ces propriétés sont valables pour toutes les langues et dépendent de mouvements articulatoires V → C.
Des auteurs conseillent le contexte <occlusive> <voyelle> .

Le codage prédictif a été appliqué avec succès (SAMBUR, 1972 ; ATAL, 1974 ; ROSENBERG & SAMBUR, 1975 ; GRENIER, 1977, 1978).

Les méthodes temporelles et fréquentielles apparaissent comme complémentaires : par exemple, importance des premières pour l'étude des phases transitoires, et des secondes pour les segments stables. Ce point de vue rejoint celui de la reconnaissance de la parole (J.J. BAKER, 1975).

Les variations inter et intra-locuteurs ont été étudiées assez systématiquement par examen de leurs causes, par STEVENS & HOUSE (1961), STEVENS (1971, 1972), qui ont effectué plusieurs expériences de démonstration de la variabilité, notamment en ce qui concerne la fréquence fondamentale et la vibration glottique, le débit glottal, la fréquence et la largeur de bande des formants, les sons produits par turbulence de l'air et les consonnes nasales. Ils remarquent que les

différences dans la dynamique articulatoire produisent des attributs acoustiques différents (même si ce sont plutôt des habitudes articulatoires apprises,plutôt que des différences anatomiques ou physiologiques), lesquels attributs permettent de discriminer les locuteurs.
Avec le téléphone, se pose bien sûr le problème de la reconnaissance du voisement ou de la substitution de la fréquence fondamentale.
Les dégradations les plus importantes dues à la voix téléphonique concernent évidemment les paramètres spectraux.
KERSTA (1965), a étudié les variations à l'intérieur d'une même famille, et ALLPERT & al. (1963), les jumeaux.
Il n'y a pas eu beaucoup d'études ayant ce caractère fondamental; les facteurs accroissant la variabilité inter-locuteurs n'ont guère été étudiés, au contraire des caractéristiques acoustiques (ENDRES & al., 1971 ; WOLF, 1972 ; HAZEN, 1973, a, b, qui a étudié les effets du contexte ; SAMBUR, 1973). Quant à la dérive temporelle de la voix, elle l'a été d'une façon très locale : surtout par la variation du fondamental, avec l'âge (MYSAK, 1959 ; PTACEK & al., 1966 ; PTACEK & SANDER, 1966 ; SHIPP & HOLLIEN, 1969 ; ENDRES & al., 1971 ; RYAN & BURL, 1972 ; HARTMAN & DAN HAUER, 1976).

La très grande majorité des études a porté sur la constitution d'un matériau sonore de haute qualité (une chambre sourde souvent, éventuellement des bandes de bonne qualité, et des conditions d'enregistrement optimales). Or, c'est un point important de la reconnaissance du locuteur : d'un côté, les diverses et très nombreuses dégradations du signal vocal sont invariablement des causes de diminution de la performance d'ensemble du système, d'autre part, la nécessité de pouvoir appliquer les méthodes au contexte du monde réel, rend nécessaire la prise en compte de ces dégradations. La table ronde de Padoue (1978) a eu le mérite de faire ressortir ce point de vue. On trouvera dans ROSENBERG (1971), une évaluation des causes de dégradation dues à l'utilisation du signal téléphonique pour le présent problème. KRAUSE (1976) recense les modifications correspondantes intervenant dans le spectre de fréquence, ainsi que les causes suivantes :

- le rétrécissement de la bande de fréquence à la largeur de la bande téléphonique (300-3300 Hz pour une ligne ordinaire),
- les distorsions linéaires et non linéaires de la fréquence deus aux voies de trans mission,
- l'acoustique du local utilisé et les interférences de toutes sortes.

Des travaux importants d'acoustique ont permis de mettre en évidence la variabilité inter-locuteur (par les propriétés acoustiques des consonnes nasales : FANT (1960), FUJIMURA (1962), DICKSON (1962)). Ces résultats sur les nasales ont été appliqués par GLENN & KLEINER (1968) à l'Identification (les consonnes nasales sont isolées manuellement, en diverses positions à l'intérieur des mots : 7% d'erreur (3%) pour 30 locuteurs (10 locuteurs), dont 2/3 d'hommes avec une décision fondée sur le maximum de corrélations entre échantillo de test et d'apprentissage. La variabilité des paramètres acoustiques a été aussi étudiée par HOLLIEN, ELLIOT et CHILDERS (importance des transitions).
Les nasales ont été utilisées très tôt par GLEN & KEINER (1968). FUJIMURA (1962) leur attribue des propriétés acoustiques sujettes à beaucoup de variation suivant les individus. Elles sont très sensibles à l'état de santé ORL du locuteur (SAMBUR, 1975), ont une variabilité intra-locuteur faible mais possèdent un spectre complexe.

Dans le but d'étudier la dérive temporelle à long terme de la voix, SAMBUR (1972) a sélectionné un ensemble de paramètres acoustiques parmi une liste de 92 : structure, transitions et bandes de fréquence des formants des voyelles et des consonnes nasales, divers aspects du fondamental et diverses duré Ces paramètres sont extraits à la main et analysés par des techniques de prédiction linéaire. La collection des données s'étend sur 3,5 années et se répartit en quelques séances. Les paramètres retenus pour leur haut pouvoir de discrimination, sont le 3è formant de [u], le 2è de [n], les 3è et 4è de [m], le 2è et 4è des voyelles antérieures et la fréquence fondamentale.
Les deux premiers formants sont généralement suffisant pour définir une voyelle du point de vue phonétique. Les deux suivants ainsi qu'au delà semblent par contre être révélateurs de l'identité du locuteur.
(l'emplacement des formants est toujours assorti d'une mesure difficile ; en outre, la fréquence des différents formants est liée à la qualité de la voix, ce qui complique la tâche en cas d'enregistrements de mauvaise qualité).

La question du choix et de la sélection des paramètres pertinents, a été envisagée par WOLF (1972), qui a défini leurs qualités "idéales", PRUZANSKY & MATHEWS (1964), DAS & MOHN (1971), MOHN (1971), DAS & al. (1972). Aussi : HAIR & REKIETA (1972), ROSENBERG & SAMBUR (1975), SAMBUR (1975), ATAL (1976), ROSENBERG (1976) . DAS & al. ont sélectionné 200 paramètres parmi 1600.

On a parfois considéré des paramètres faciles à extraire (les sorties des bancs de filtres tout d'abord), ou bien des paramètres plus sophistiqués (fréquences des formants[1]), résultats d'une théorie quelconque. Très souvent, les mesures ont été faites manuellement, afin de conserver un maximum de précision. Puis, au fur et à mesure que les idées de la Reconnaissance des Formes se précisaient, on a cherché à élaborer des prétraitements antérieurs à la phase d'apprentissage (par ex. : SMITH, 1962). Des auteurs ont utilisé des paramètres très simples (RAMISHVILI, 1966 : la distribution des points extrêmaux adjacents qui ne nécessitent pas de normalisation d'amplitude ; HOLMGREN (1963, 1967) et CLARCKE & BECKER (1969) : le débit ; RAMISHVILI & TUSHVILI (1976) : les pauses).

Des études ont porté sur la totalité des paramètres (PRUZANSKY & MATHEWS (1964), MOHN (1971), WOLF (1972), BROWN & al. (1973), SAMBUR (1975).

WOLF (1969) a testé systématiquement des paramètres mesurés à la main et a sélectionné : la fréquence du fondamental, les caractéristiques spectrales des voyelles et des consonnes nasales, les fréquences des deux premiers formants des voyelles [a], [i] et [u], des durées de segments phonétiques, et la durée d'établissement du voisement - pour autant qu'elle ait une existence bien définie indépendamment du contexte. Ces paramètres ont été utilisés avec des résultats parfaits lors de la Vérification.

L'analyse des formants a retenu considérablement l'attention : citons SCHAFFER & RABINER (1970), qui ont utilisé les trois premiers formants, associés au fondamental et au cepstre du signal, et GOLDSTEIN (1976), pour l'étude de la structure formantique des diphtongues. KRAUSE & ENDRES (1976) ont traité les transitions formantiques.

C'est l'étude du fondamental qui a suscité le plus de recherches, en connexion avec divers aspects de la production et de la reconnaissance de la parole, et de l'extraction du fondamental sur signal téléphonique. Citons : VOIERS (1965), HOLMGREN (1967), CLARKE & BECKER (1969), STEFFEN-BATÓG (1970), WOLF (1971), LUMMIS & ROSENBERG (1971), LUMMIS (1973), JASSEM & al. (1973), BOULOGNE & al. (1973), ATAL (1974, 1976), EL CHAFEI (1978).

---

(1) Sujettes à des variations rapides et importantes, surtout pour F2 et F3. Quant au fondamental, sa valeur moyenne garde de l'intérêt.

Le paramètre le plus évident et le plus facile à imiter semble être le fondamental.

L'école polonaise (CZAJKA, GROSZKA, GUBRINOWICZ, J ASSEM, KOŚCIELAK, KRZYSKO, FRACKOWIAK-RICHTER, STEFFEN-BATÓG, etc.) a publié de 1970 à 1973 une série d'articles traitant notamment de la distribution à court terme de la fréquence fondamentale et des fonctions discriminantes.

MATHEWS & al. (1961), ont mis au point une procédure d'extraction par filtrage inverse de l'onde glottique, à partir du signal de parole.

ATAL (1968) a montré l'utilisation de la variation de la fréquence fondamentale sur une phrase (contour du fondamental) et SAMBUR (1975) a mis en évidence certains de ses avantages.

LEVIN & LORD (1975) ont fait de cette fréquence un indicateur de l'état émotionel et physiologique, après WILLIAMS & al. (1970) lesquels avaient inclus sa moyenne, sa dispersion et la forme de son enveloppe.

Il a été démontré la variabilité à long-terme du spectre moyen et de la fréquence fondamentale (FURUI 1973-78).

On a aussi considéré l'importance du fondamental quant à sa résistance à l'imitation (LUMMIS , 1971), sa variabilité (CLARKE & BECKER, 1969), sa variabilité en fonction du "stress" (HECKER & al., 1968 ; HAYRE, 1976), son importance par rapport à d'autres paramètres . Le fondamental apparaît toujours comme un paramètre solide, mais non suffisant : on lui reproche d'être sujet à des variations à long terme. Certains l'utilisent pour une pré-classification des locuteurs.

EDIE & SEBESTYEN (1962), puis FLOYD (1964), ont effectué une étude systématiq sur 13 paramètres extraits à la main (fréquences des quatre premiers formants fréquence fondamentale, enveloppe de l'amplitude, les dérivées de ces 6 paramètres, et un paramètre fondé sur les intervalles de voisement), qui les a conduits à un sous-ensemble efficace : 90 à 93% de bonne Vérification pour 10 locuteurs et 4 imposteurs en utilisant les formants et le fondamental, 57% de bonne Identification avec les 4 formants pour 11 locuteurs.

FLOYD a rajouté d'autres paramètres, qu'il appelle rudimentaires : maximum, moyenne, minimum de la fréquence du fondamental, durée d'intervalles voisés, l'extraction étant cette fois automatique. Le fait étonnant est que ces mesures rudimentaires ont été assorties d'un taux de reconnaissance supérieur au précédent.

CLARKE & BECKER (1969) ont comparé des types de paramètres : fréquence fondamentale moyenne, variabilité du fondamental, spectre à long terme et en certaines portions du signal, durées d'ensemble. Le spectre à long terme est apparu comme la meilleure caractéristique, ce que d'autres auteurs ont constaté (PRUZANSKY, 1963 ; ...). Les fortes différences inter-locuteurs sur le spectre à long terme sont donc un résultat important ; en outre, les parties basse et haute dans ce spectre semblent y contribuer de façon égale.
La puissance dans le spectre a été étudiée par HALL (1975), et BORDONE-SACERDOTE & SACERDOTE (1968), qui remarquent une forte sensibilité de ce paramètre à l'émotivité.

FURUI (1974-1976) a utilisé des techniques de filtrage inverse (le spectre à long terme est utilisé ici pour construire un filtre du second ordre).

KASHYAP (1976) a proposé une formalisation de type statistique du problème de la Vérification et de l'Identification, fondée sur la réalité physique des phonèmes et sur des hypothèses statistiques. Cet auteur a en outre mis l'accent sur le choix controversé des niveaux de confiance.
Sur le choix des sons, divers auteurs anglo-saxons se réclament des donsonnes nasales et des voyelles, souvent prises isolément (Mc GEE, 1965 ; GLENN & KLEINER, 1968 ; REDDY, 1975). Le trait de nasalité a donné lieu à de nombreuses études, dont GLENN & KLEINER (1968), WOLF (1972), SU & al. (1974), SAMBUR (1975).

La voix modifiée a été très tôt à l'ordre du jour, que ce soit pour les caractéristiques propres à la voix chuchotée, (POLLACK & al., 1954 ; WILLIAMS, 1964 ; RAMISHVILI, 1966 ; SCHWARTZ & RINE, 1968), pour les modifications causées par le vocodeur (SCHEARME & HOLMES, 1969) ou récemment par les dégradations dues au téléphone (FLANAGAN, 1973 ; ROSENBERG, 1976 ; BROWNER & STENZEL, 1976 ou par un milieu ambiant inhabituel (hélium en acoustique sous-marine).

Le cas des imposteurs est couramment traité actuellement en raison de la nécessité d'évaluer la résistance des systèmes à l'imposture, et des études incluent un grand nombre d'imposteurs des divers types (ROSENBERG, 1972 ; FURUI, 1976). DAS & al. (1972), ont réalisé une expérience d'entraînement à l'imposture.

Les différences hommes/femmes/enfants, ont été mises en évidence sur le spectre (SOTTER & STEINBERG, 1950), et surtout du fondamental (HOLMGREN, 1963, 1967 ; MATSUMOTO & al. 1973 ; RAMISHVILI & TUSHVILI, 1976 ; CLARKE & BECKER, 1969, mais SCHWARTZ (1968) et INGEMAN (1968), ont utilisé les [s] et [ʃ] isolés). Deux autres études sont à signaler (SCHWARTZ & RINE, 1968 ; WEINBERG & BENNETT, 1971).

Etant donnée la diversification croissante des méthodes d'approche de la reconnaissance du locuteur, des auteurs (ATAL, 1976 ; ROSENBERG, 1976), se sont attachés à faire le point sur le sujet.

En entrée, une phrase entière est analysée pour obtenir l'amplitude du spectre (voir détails dans ROSENBERG, 1977). La stratégie de discrimination est de type séquentiel : le système traite autant de phrases qu'il faut pour atteindre un seuil de confiance donné.

Dans une expérience préliminaire, DODDINGTON annonce, pour une population de 10 femmes, 40 hommes et 100 sessions réparties sur 2 mois, et 70 imposteurs (dont un quart de femmes) intervenant dans 20 sessions, un taux d'erreur de 1,6 % pour une phrase 0,42 % sur 2 phrases et 0,23% sur 3.

Dans les versions récentes, le locuteur est instruit de la phrase qu'il doit prononcer : phrase de 4 mots pris au hasard par le système parmi un vocabulaire de 16 mots monosyllabiques. Il prononce ainsi 4 phrases. Au cours de l'apprentissage, qui dure 5 sessions, chaque mot est répété 5 fois à l'intérieur de phrases porteuses. Une adaptation au locuteur est automatiquement réalisée si elle est estimée nécessaire.

Le système est maintenant opérationel dans un contexte de contrôle d'accès : probablement 400 essais de Vérification en moyenne par jour sur quelques centaines de locuteurs des deux sexes. La durée de l'opération annoncée est de 5,8 s. en moyenne. Cette rapidité est obtenue par l'emploi de bancs de filtres lors du prétraitement.

Toutefois, l'ensemble de ces expériences ne nous renseigne guère sur les processus de décision humains. L'étude de l'application des techniques de décision date de très peu d'années (notamment : PRUZANSKY & MATHEWS (1964), DAS & MOHN (1971), BRICKER & al. (1971), (avec analyse discriminante), LUMMIS (1973), ROSENBERG (1973, 1976), ATAL (1974), DODDINGTON (1974), (stratégie séquentielle), BUNGE (1975), SAMBUR (1975), ROSENBERG & SAMBUR (1975)).

Un projet de Vérification automatique de grande envergure a débuté chez Texas Instrument vers 1970, le but essentiel semblant être la conception d'un système opérationnel très fiable à grande échelle. Selon un commentaire de EL CHAFEI (1978), ce système devrait répondre aux caractéristiques précises suivantes, permettant seules de le rendre opérationnel : empêcher les essais d'imposture (déguisement, imitation), ainsi qu'aux changements d'origine pathologique (c'est pourquoi le système offre une large place aux fréquences formantiques, et à l'énergie du signal, qui sont assez résistantes de ce point de vue), éviter les conséquences d'un excès d'émotivité de la part du locuteur, mettre à jour les références selon la dérive de la voix, et assurer un temps de calcul le plus faible possible, dans un souci de coût et de commercialisation.

L'équipe HOLLIEN, ELLIOT et CHILDERS a décrit le système SAUSSI (System for Automatic Speech and Speaker Identification).
ROSENBERG (1976) a fait une évaluation du système de reconnaissance automatique par téléphone des Laboratoires Bell, fondé sur l'évolution du fondamental, de l'intensité et des trois premiers formants, et la normalisation en durée des courbes obtenues, tout ceci sur des phrases assez longues (durée supérieure à 2s). L'analyse est fondée sur les contours temporels et une analyse prédictive. Pour 120 locuteurs et 5.000 essais, le taux de reconnaissance variait de 90% à 96%, suivant les paramètres utilisés. FURUI (1976) a rajouté des paramètres cepstraux (le cepstre est interpolé polynomialement) avec un taux d'erreur inférieur à 1%.

Actuellement, des agences militaires (ROME AIR DEVELOPMENT) et d'autres compagnies (International Telephone and Telegraph) s'intéressent à la Vérification et à l'Identification, parfois en complément d'autres techniques de reconnaissance (empreintes digitales, écriture manuscrite).

PLANCHE 1 : Travaux comparés en Reconnaissance du locuteur (extrait de SARMA, YEGNANARAYANA, 1975).

	Author	Measured speech characteristic	Number of Speakers	Number of Patterns	Performance and main contribution
A.	SPEAKER RECOGNITION BY HUMAN LISTENERS :				
1	Pollack Picket and Sumby (1954)	Monosyllabic words and sentences	2, 4, 6, 8 (All male)		Duration of speech is an important factor. Recognition scores increase from 20% for 0.5 sec. to 95% and above for a 2 sec utterance.
2	Ramishvili (1966)	Use of Russian phonemes in different contexts Use of several frequency bands			All phonemes carry speaker information. Voiced sounds (vowels and semi-vowels) carry more information than unvoiced sounds. 700-4000 Hz retains speaker information.
3	Rosenberg (1972)	Comparison of listeners and automatic verification	8 customers 33 imposters and professional mimics		Automatic system 1% miss and 0% false alarm for casual imposters. 14% miss for mimics (miss=acceptance of imposter, false alarm=rejection of customer). Automatic system is much better than average listener rates with 22% miss and 3-4% false alarm.
B.	SPEAKER RECOGNITION BY VISUAL EXAMINATION OF SPECTROGRAMS :				
4	Kersta (1962, 1965, 1966)	Voice print : visual comparison by trained observers	5, 9, 12 (All Male)	0-2% error	Responsible for raising the question of unique identification of a person by voice.
5	Young and Campbell (1967)	„	5	21.6% error for isolated words and 62.7% error for words in context.	Observed large discrepancies with Kersta's results.
6	Stevens *et al.* (1968)	Voice prints : (visual examination) and listening tests.	—	18-50% error	Aural recognition better than visual comparison of spectrogram.
7	Bolt *et al* (1970)	—	—		Critical survey of voice print work, comparison with finger prints and comments on legal use. Conclusion that more evidence is needed.
8	Tosi *et al* (1972)	Voice prints : and listening tests.	40 out of 250 speakers	34996 trials by 29 examiners	Exhaustive experiments on voice print identification for legal purposes. Confirmed Kersta's 1% error for closed trials and reported for open trials 6% miss and 13% false alarm.
C.	AUTOMATIC SPEAKER RECOGNITION :				
9	Keithsmith (1962)	Filterbank analysis (35 channel)	Not available		Used analysis of variance for maximization of interspeaker variation.
10	Hargreaves and Starkweather (1963)	Spectra from 18 channel filter bank	12, 192	10% error	Decision analysis.

	Author	Measured speech characteristic	Number of Speakers	Number of Patterns	Performance and main contribution
11	Pruzansky (1963)	Spectral data from filterbank	10, 393 (3M, 7F)	11% error	Cross correlation analysis.
	,,	Longterm spectra from voiced speech	10, 40	0% error	,,
12	Becker *et al.* (1964)	Spectral data from 17 channel filter bank.	10, 693	3% error	Comparison of several statistical distance measures.
13	Li *et al* (1966)	Quantized 3-dimensional spectrographic data from 15 channel filterbank	11	7% error	Use of adaptive techniques.
14	Glen and Kleiner (1968)	Nasal sounds spectral analysis	10 30 (20M, 10F)	3% error 7% error	
15	Atal (1968)	Pitch contours	10, 60 (All females)	2% error	
16	Luck (1969)	Cepstral analysis	4 speakers 30 imposters	8% error	System built around a mini-computer.
17	Das and Mohn (1969, 1971)	Filterbank analysis	118, 7,000	1% error 10% no decision	Much larger data base. Adaptive system sensitive to segmentation.
18	Doddington (1971)	Pitch Intensity and formant analysis	40 (Male) (8 customers, 32 imposters)	1.5% error	Nonlinear time-warping of utterances based on 2nd formant to correct for variable length of utterances.
19	Beek *et al* (1971)	Spectral data	—	—	Use of speech recognition circuitry.
20	Wolf (1972)	Several features in a set of five sentences			Comparison of features and selection of efficient features on the basis of F-ratio.
21	Lummis (1973)	Pitch intensity formants	40 (8 cutomers, 32 imposters)	1% error	Implemented on a large computer and a small computer.
22	Atal (1974)	Predictor coefficients and others derived from LPC.	6,60	2% error	Cepstral coefficients are the best in several sets of parameters derived via LPC analysis.
23	Su *et al* (1974)	Nasal coarticulation	4, 10	1.1% error	Coarticulation a better feature than nasal spectra.
24	Rosenberg and Sambur (1975)	LPC analysis	22 customers 55 casual imposters 4 professional mimics.		Improved version of Lummis' scheme
25	Wasson and Donaldson (1975)	Zero crossings and amplitude measurements	10 100 (7M, 3F)	3-6% error	Minimal computational complexity.

Concernant le matériel utilisé au cours de ces années, on note deux tendances :

- l'utilisation de calculateurs de taille moyenne pour la saisie et les expériences de Vérification,
- l'utilisation de gros calculateurs pour des études spécifiques (statistiques, extraction de paramètres, ...) (par exemple : Honeywell haut de gamme : DODDINGTON 1970, 1971).

ROSENBERG (1976) décrit un système de Vérification fonctionnant sur NOVA 500 auquel est couplé un système de réponse vocale par voix pré-enregistrée. Une centaine de locuteurs y accèdent par postes téléphoniques individuels (le code des locuteurs est envoyé par enfoncement d'un numéro d'identification au clavier : système Touc Tone ; les paramètres sont le fondamental, l'intensité des trois premiers formants et quelques coefficients de prédiction linéaire ; taux d'erreur : 5 % pour des loduteurs adaptés, 10 % pour les nouveaux locuteurs ; le temps de calcul est de 20 à 30 s.).
Les paramètres sont la valeur moyenne du fondamental, l'écart quadratique moyen du fondamental, et son étendue mélodique (environ 90 % de bonne identification et vérification pour 15 locuteurs).

Il semble que les systèmes sur mini-ordinateurs n'aient guère vu le jour, probablement à cause du manque de certitude quant au bon choix de paramètres et à l'automatisation de la segmentation. EL CHAFEI (1978, 1979), a étudié une telle application. Le temps de réponse est un autre obstacle à la construction de matériel spécialisé figé, le temps réel ne semble pas pouvoir être envisagé dans l'état actuel lorsqu'on utilise de nombreux paramètres.
On aura vu que les systèmes sont largement hybrides quant à la nature des paramètres qu'ils utilisent ; par exemple : coefficients de prédiction linéaire, sons caractéristiques, etc. Bien qu'aucun paramètre n'ait été reconnu comme suffisant ou universel, on cherche à en utiliser un certain nombre de façon à augmenter l'information relative au locuteur.

Pour classer les systèmes opérationnels, SARMA & YEGNANARAYANA (1975) se posent les questions suivantes :

1.	What sort of speech ?	Continuous (more than 1, 2 sec. duration). Single words or phrases specially chosen or arbitrary
2.	What language ?	English Native language of each speaker Nonsensional utterances
3.	How many speakers ?	Few, many 10, $\lhd$ 30, $\rhd$ 30
4.	What type of recognition ?	Verification Identification
5.	What sort of speakers ?	Cooperative, casual, trained, untrained, sex (M, F, mixed set)
6.	In what environment does the customer speak ?	Quiet Public place
7.	What sort of communication medium ?	Quality microphone Telephone (Analog, PCM, Vocoder)
8.	How much training of the system ?	No. of design samples per speaker No. of recording sessions No. of imposters to test for vulnerability.
9.	What kinds of errors can be tolerated ?	Mits, False alarm
10.	What is the order of computer processing capacity ?	Large computer or small one
11.	What is the computation time?	Near real-time or much greater than real-time.

2 - Quelques questions majeures :

2.2. La mystique du "voice print" :

a) Existe-t-il des "empreintes vocales" suivant l'expression propre KERSTA, qui pourraient jouer un rôle analogue aux empreintes digitales (1). Cet auteur a annoncé avoir la preuve expérimentale de l'Identification des locuteurs par des "empreintes vocales": "I think my publication indicate the possibilities of voice print identification always being correct". (2)
Décrivons brièvement la technique de KERSTA. La fourniture consiste en bandes analogiques (bande témoin, et bande référence). L'analyse commence par plusieurs écoutes attentives des enregistrements témoins, en repérant les mots qui sont présents dans les deux bandes et jugés comme importants ou adaptés à la reconnaissance.
Exemple de mots : "know, be, you, the, out, a, or, I, is, on, to, me, and".

---

(1) On connaît la phrase de KERSTA, utilisée lors de ses témoignages (cf. plus loin) : "It is what we call a pattern matching test and it works exactly the same as it does in fingerprints identification".

(2) The People VS. Edward Lee King, p. 866.

Les segments à large bande de ces mots échelle sur échelle logarithmique sont découpés en segments adjacents. La comparaison entre paires de segments correspondants pris dans les deux enregistrements est alors effectuée.
KERSTA raffine la procédure en ne fondant ses décisions que sur l'examen d'au moins 5 paires de mots et de 20 "points de similarité" (cette locution n'a pas été définie).
Selon cet auteur, il est facile d'enseigner de telles techniques à des experts de la police : "two weeks of comprehensive class-room lectures and laboratory work will prepare a student to operate the spectrograph and make positive voice print identification a work previously performed by KERSTA and his staff exclusively" (1).

On doit remarquer que le cas n'est pas rare, où la justice dans plusieurs pays a demandé à des "experts" de répondre à la question : "est-ce que ces deux échantillons proviennent de la même personne?"

b) Les réponses à KERSTA ont été très vives. LADEFOGED et VANDERSLICE ont essayé de mettre sur pied une preuve expérimentale à l'en contre des thèses de KERSTA, et lui reprochent d'avoir trompé "à la fois le public, la loi, et probablement lui-même", (2) et rien de moins que d'avoir fait preuve de dogmatisme. A l'appui de cette réplique, sont les preuves expérimentales suivantes :

- même en suivant la technique de KERSTA, il est possible de produire des paires de spectrogrammes pratiquement identiques, provenant de prononciations de deux locuteurs distincts (cf. aussi KRAUSE, 1971),

- même si les dires de KERSTA étaient vrais, sa technique ne ferait que prouver ceci : les mots effectivement utilisés dans la procédure de comparaison ont été prononcés par le même locuteur ; mais justement, ces mots ont pu être copiés ou insérés dans la bande témoin.

En outre, un locuteur peut, à différentes occasions, prononcer un mot de différentes manières, tandis que les variations d'un locuteur donné peuvent se superposer avec celles d'un autre. Selon les circonstances, deux locuteurs

---

(1) Voice print Laboratories, Sommerville, New Jersey, Press release, Nov. 1966.
(2) "There presently exists, as we hope to show, no such thing as a voice print technique. We prefer, therefore, tu use the term "mystique", meaning the esoteric skill or mysterious faculty essential in a calling or activity".

peuvent produire des sons dont les spectrogrammes sont très semblables. LADEFOGED & VANDERSLICE ont réalisé un contre-exemple de la manière suivante : 3 locuteurs hommes ont prononcé en chambre sourde des phrases durant environ 2s. Ces auteurs ont trouvé qu'il était tout à fait possible de produire une "preuve" conduisant à une fausse identification de l'un des trois locuteurs, avec l'un des deux restants, ou aussi les deux à la fois, indifféremment.

Leur conclusion est contenue dans les deux phrases suivantes : "No one is more eager than we are to see any bona fide advances in the technology of our field applied to the legitimate purposes of the police. It would be nice if we could identify speakers by their voice. But one man's faith in his own ability should not be accepted as proof that it is currently possible". Ainsi, l'analogie entre empreintes digitales et vocales est selon eux relativement fausse (1).

Pour clore ce paragraphe, indiquons qu'il a été demandé l'avis de spécialistes sur la parole des Etats-Unis, et aussi d'autres pays. Parmi les nombreuses réponses, il n'en eut pas une qui eut été en faveur de l'utilisation des spectrogrammes à des fins légales.

Il faut également citer une lettre de BOLT, COOPER, DAVID, DENES, PICKETT & STEVENS (1973), qui expose l'état de l'art en reconnaissance à l'aide de spectrogrammes, et discute la possibilité d'application de ces méthodes. La conclusion réitère une mise en garde déjà formulée : la méthode n'a pas été clairement établie. Une réponse immédiate a été faite par BLACK, LASHBROOK, NASH, OYER, PEDREY, TOSI & TRUBY (1973), qui met en cause la non-expérience personnelle des auteurs précédents, et affirme qu'un personnel entraîné est au contraire qualifié pour mener à bien les enquêtes.

---

(1) Les empreintes digitales sont invariantes et persistantes dans la vie d'un individu, et la probabilité d'avoir deux empreintes parfaitement identiques est évaluée à $1{,}6 \cdot 10^{-11}$. Il n'est pas possible d'estimer de même la probabilité d'égalité de deux voix pour une raison a priori : la voix est toujours sujette à modification.

2.2. Le problème de l'imitation :

La facilité d'imitation est liée à l'opportunité de parodier des caractéristiques articulatoires grossières :

"Les imitateurs professionnels semblent devoir leur succès à ce qu'ils trouvent l'intonation et le rythme des personnages dont ils contrefont la voix, sans pour autant reproduire leurs attitudes articulatoires" (GUIBERT, 1979). Ils agissent sur les caractéristiques prosodiques (intonation et accentuation), ce qui nécessite des séquences parlées suffisamment longues : on peut rendre l'imitation difficile en utilisant des mots isolés.
Pour autant que la perception de la parole soit concernée, il n'est pas facile de déterminer quels aspects du signal sont utilisés par l'individu pour communiquer des émotions spécifiques. Il faudrait étudier les habitudes vocales de l'individu en relation avec ce qu'elles sont supposées signifier. Ceci rend le cas de l'imitation d'une investigation malaisée. Néanmoins, il faudrait être capable de déterminer la résistance d'un système à l'imitation (ce qui suppose connue la probabilité d'imitation pour le locuteur courant). Cette remarque vaut lorsqu'on désire conserver un taux d'erreur stable, alors que l'on a des imposteurs fortuits. Le déguisement de la voix, donc l'imitation, peut causer desfaux rejets à l'audition (cf. CARBONEIL, 1965).

Comme le fait remarquer ROSENBERG (1977), "Les systèmes qui sont plus résistants a priori à l'imitation, sont ceux dont les paramètres sont fortement corrélés avec la physiologie du locuteur, à l'inverse d'une corrélation avec le comportement. On a souvent pensé à s'intéresser aux membres d'une même famille et en particulier aux jumeaux. Une expérience de cet auteur montra que, pour 2 jumeaux, le jumeau-récepteur ne parvient pas à imiter les caractéristiques comportementales de son frère. On sait en outre que l'on peut imiter la voix de quelqu'un tout en gardant des valeurs moyennes du fondamental et des formants, différentes de celles de la personne imitée. LUMMIS et ROSENBERG (1972), conseillent l'utilisation des formants en plus de l'intensité et du fondamental pour le recours à des paramètres résistants. Selon WOLF (1971), le fondamental est peut-être le paramètre le plus facile (et le plus évident) à modifier lors d'un déguisement de la voix.

Au moins quatre études essentielles concernent l'imitation :

1 - LUMMIS et ROSENBERG (Bell Labs, 1972), ont essayé la Vérification avec quatre imitateurs particulièrement entraînés, imitant huit locuteurs. Ce qui les conduisit à chercher à améliorer le système de Vérification.

2 - DODDINGTON (Texas Instruments, 1974), a testé le système de Texas avec deux imitateurs, imitant six locuteurs, et confirmé l'augmentation du taux d'erreur en présence d'imposteurs.

3 - HAIR et REKIETA (1972), utilisent un professionnel imitant six locuteurs. L'emploi de la totalité des paramètres retenus pour la Vérification, a seul permis l'élimination des impostures dues à l'imitation.

4 - LUCK (1961), a utilisé 3 locuteurs non entraînés, imitant un même locuteur : aucune fausse acceptation supplémentaire, mais le plus mauvais des trois est parvenu à améliorer son score d'imitations.

L'imitation et le déguisement de la voix (la différence est claire), ont aussi été étudiées par ENDRES, BAMBACH & FLOSSER (1971),
FANT et ÖHMAN ont réalisé quelques expériences pour connaître quels aspects de la parole peuvent être imités : un résultat essentiel est que la dynamique de la source vocale est d'une imitation aisée. Souvent, les expériences d'imitation sont faites immédiatement après l'écoute de la voix de référence à imiter, afin de chercher à maximiser le succès. Le système de Texas Instrument - où les imitateurs sont des professionnels - montra que, sur une phrase, la probabilité d'acceptation d'un imposteur-imitateur était le double de celle d'un imposteur normal.
Le problème de l'imitation est très mal cerné. Peut-être est-ce un art, ou bien faut-il que les expérimentateurs soient eux-mêmes de bons imitateurs.

2.3. L'émotivité du locuteur

L'état émotionnel d'une personne influence outre la valeur de la fréquence fondamentale, la position des formants (surtout le premier de la première syllabe des mots). Parfois les auteurs (comme KRAUSE, 1976) acceptent la confusion : étant émotionnel et débit moyen en distinguant le parler rapide du parler normal (tranquille). Une corrélation entre les mesures des deux premières qualités devrait pouvoir être mise en évidence. Cette influence émotionnelle semble plus forte chez les sujets jeunes : la variabilité de la voix chez une personne âgée reste plus limitée. En outre, les observateurs notent q'un sujet se trouvant état de forte émotivité recherche dans son langage des mots très courants. En outre, ceux-ci sont d'autant moins soumis à des modifications des habitudes articulatoires qu'ils sont prononcés quotidiennement : noms toponymiques, salutations, mots "sémantiquement neutres" indépendants du contexte ; KRAUSE cite les mots : 'enfin', 'c'est tout'.

II

LES FONDEMENTS DU PROBLEME

I - LA VARIABILITE DE LA VOIX

1 - Quelques rappels sur certains aspects de la production de parole :

Nous renvoyons directement à FANT (1960) et FLANAGAN (1972), pour des descriptions du processus de la production de parole.

Définition 1 :

On appelle fréquence fondamentale courante, la valeur de la fréquence laryngienne à un moment donné, c'est-à-dire le nombre de cycles par seconde dans la vibration des cordes vocales. Lorsque cette vibration existe, les sons sont dits voisés. Dans le cas contraire (et complémentaire), ils sont dits sourds. On se reportera au tableau I pour une répartition des principaux phonèmes suivant le trait (*) de voisement.

Habituellement, la fréquence fondamentale de moyenne (moyenne des valeurs courantes sur un intervalle de temps donné ou sur une phrase donnée), varie entre :

- 90 et 150 Hz pour une femme,
- 170 et 255 Hz pour un homme.

Définition 2 :

Pour une représentation machine convenable du voisement, nous adoptons une nouvelle définition : la fréquence fondamentale à un instant donné est la valeur échantillonnée, correspondant à l'inverse du temps qui sépare deux passages par zéro du signal de parole (voir IV-III-5.).

Note : par la suite, nous appellerons son voisé, toute réalisation de parole produite avec vibration des cordes vocales, et son non voisé, dans le cas contraire. On peut ainsi obtenir le classement traditionnel :

sons voisés	voyelles	constrictives (1)	occlusives (1)	semi-voyelles liquides nasales
sons non voisés		constrictives (2)	occlusives (2)	variante de /R/

voyelles [i, e, ɛ, a, a, ɔ, o, u, y, ø, œ, ə, ɛ̃, ɑ̃, ɔ̃, œ̃]

const. voisés [v, z, ʒ] semi-voyelles [w, ɥ] nasales [m, n, ɲ, ŋ]

non voisés [f, s, ʃ] liquides [l, R] variante non voisée de /R/ : [R_0]

(*) Voir glossaire.

2 - Ebauche d'une idiolectologie (*) :

Nous parlerons dans les quelques lignes qui suivent, en faveur d'une méthodologie de l'individuation (*), car c'est, à notre avis, une façon correcte d'envisager la reconnaissance du locuteur. On a dit que dans sa démarche, la linguistique a jusqu'ici éliminé les diversités inter et intra-individuelles des descriptions phonétiques. Ainsi, la variabilité inter-locuteurs a été, par hypothèse, considérée comme hors-système phonologique (*), et dans le cas où elle serait intégrée au système phonologique étudié, elle le modifierait (par exemple, par un changement des règles phonologiques). Mais parallèlement à cette attitude, de très nombreuses études - d'ailleurs indépendantes de toute recherche phonématique (*) - ont décrit et exploité la variabilité inter-locuteurs.

Cet éclaircissement a été finement dégagé dans ABRY & BOË (1979), qui constatent que cette alternative a été pendant longtemps un obstacle épistémologique de taille au traitement structural de la variation dialectologique (*), "avant que le concept de diasystème (*) ne soit instauré (par WEINRECH, 1954)". Convenant que la variabilité intrasystématique (à l'intérieur d'un même système) n'est pas intégrée par les systèmes dialectaux ou idiolectaux, ABRY et BOË (1979), estiment qu'elle doit être pensée au niveau de la réalisation de chacun des systèmes, dans laquelle doivent être maintenues des constantes relationnelles. Par exemple, "la plupart des processus de normalisation des espaces vocaliques tendent à se débarrasser de la variance inter-locuteurs", par des transformations linéaires, le plus souvent.

La question fondamentale de notre problème se pose alors en ces termes :

> Existe-t-il une variabilité irréductible entre les différentes réalisations par différents locuteurs, d'un même système ?

Les auteurs précédents répondent par l'affirmative, et essaient avec élégance de montrer que ce sont en fait de véritables variantes inter-individuelles intrasystématiques. Proposons-nous de décrire brièvement leur pensée.

(*) Voir ce mot dans le glossaire.

Il existe deux sortes de différences individuelles dans le signal acoustique de la parole :

- celles dues à l'anatomie et à la physiologie ; elles sont relativement mieux connues, et constituent la part de la réalisation physique individuelle.
- celles dues à l'utilisation individuelle de son appareil vocal propre (1), et qui peut amener à des réalisations individuées du système phonologique.

Ainsi est dégagée la réalité d'une variation individuelle intrasystématique. Il reste à proposer (ibid., op.cit.), des stratégies expérimentales mettant en évidence le fonctionnement de cette variation. La plupart de celles-ci semblent avoir couru à la Vérification ou à l'Identification par des procédures totalement aveugles à l'organisation de la communication linguistique individuée. Pourtant, la variabilité de la parole est très liée aux caractéristiques linguistiques "extérieures" de la situation, ainsi qu'à la réaction du locuteur à elles. Ceci fait proposer à FRANCESCATO (in STEVENS, 1971), en remplacement de la distinction :

variations intra et inter-locuteurs,

la distinction entre les deux plans de variation suivants :

- variations inhérentes ou permanentes,
- variations conditionnées par la situation.

Cette dernière distinction est cependant plus orientée vers les études de reconnaissance (Identification) de la situation, que de reconnaissance du locuteur. En outre, elle laisse, semble-t-il, peu de prise à une étude rationnelle : il ne faut pas donner au contexte une importance excessive lors de la détermination de la forme acoustique d'un signal.

---

(1) C'est le correspondant des habitudes posturales et gestuelles (ou hexis corporelle, suivant le terme de BOURDIEU). Selon les auteurs, c'est sans doute le concept de base de l'articulatoire individuelle qui rendrait le mieux compte, au niveau phonétique, "de cet aspect de l'hexis (*) qu'est l'habitus (*) vocal" (ibidem).

3 - Les sources de variabilité intra et inter-locuteurs :

Le signal de parole fait partie des signaux physiologiques qui varient le plus. D'abord, parce que les cavités et autres composantes de l'appareil vocal ont des tailles et des formes différentes, ensuite parce que les mots, en des occasions (contextes) différentes, sont prononcés de façons différentes, avec des énergies à chaque fois différentes.
La gestuelle articulatoire a très peu de probabilité de se réaliser identiquement d'une phrase à une autre.

Comme le fait remarquer STEVENS, la variabilité des sons intervenant dans une phrase est liée au fait que le contexte est suffisamment redondant et peut permettre une personnalisation de la production du son, qui ne nuise pas à sa reconnaissance. Cet auteur indique quelques exemples comme le prévoisement des occlusives initiales (qui est laissé libre en anglais), et des degrés divers de nasalisation pour des voyelles précédant une consonne nasale.

Selon LOWERRE (1977), la variabilité des locuteurs pour la parole continue, revêt trois formes :
- dialectale : changement de prononciation des mots, suivant le locuteur,
- contextuelle : changement de prononciation des mots, suivant le contexte de ces mots,
- acoustique : modification du conduit vocal,

qui peuvent se superposer. La variabilité dialectale peut être codée au niveau du lexique (cf. le système de reconnaissance de la parole HARPY), la variabilité contextuelle peut être prise en compte par des règles de jonction entre mots, mais la variabilité acoustique reste toutefois de nature différente : c'est un phénomène qui dépend du locuteur. Il y a eu plusieurs tentatives de modélisation de ces dernières variations. Citons par exemple : la normalisation des formants vocaliques la détermination des caractéristiques individuelles.
Ces deux moyens relèvent de deux approches relativement duales.

Comme sources de variabilité inter-locuteurs, GARVIN et LADEFOGED (1963), distinguent :
- les différences organiques,
- les différences acquises.

En premier lieu, on relève que les dimensions et propriétés du système respiratoire inférieur, varient suivant les individus : dimension des conduits, nature de leurs

Sur le plan acoustique, les différences précédentes introduisent des variations pour :

a) la fréquence fondamentale, son intensité et la durée des pauses phonatoires (résultant essentiellement de différences sur la mécanique ventilatoire et des dimensions des cordes vocales),

b) la fréquence centrale des formants pour les voyelles (STEVENS & HOUSE, 1961) : à une fréquence basse correspond - toutes choses égales par ailleurs - un conduit vocal long et inversement ; la largeur des formants et la position des zéros.

c) les évolutions des formants (suivant la gestuelle articulatoire en vigueur).

On peut trouver dans WOLF (1969), l'affirmation de cette correspondance. En ce qui concerne la parole continue, on peut s'attendre à une plus forte différentiation inter-locuteurs, par rapport aux mots isolés ou aux logatomes (*), eu égard aux phénomènes suprasegmentaux qui s'ajoutent : intonation, distribution et durée des pauses, accent d'insistance,...

STEVENS (1971), rapporte que la structure qui semble la plus responsable de la variabilité intra et aussi inter-locuteurs, est le larynx et plus particulièrement, les cordes vocales : "la source de l'excitation acoustique du conduit vocal pour les sons voisés, est déterminée directement par l'élasticité, la masse et la forme des cordes vocales". Toute asymétrie ou toute autre irrégularité (telle que le degré d'humidification des parois suivant l'intensité de la salivation) dans les cordes vocales, se traduit par des irrégularités dans le phénomène vibratoire, quant à sa période et sa forme. L'ouverture cyclique des cordes vocales entraîne, toujours selon STEVENS, une modulation de la fréquence et la largeur de bande du 1er formant et peut-être du second. La troisième composante du système de production de parole - la portion du conduit vocal comprise entre la glotte et les lèvres - est également à considérer comme une importante source de variabilité inter-locuteurs, alors que la variabilité pour un même locuteur y est relativement réduite. En effet, le conduit vocal joue le rôle de filtre de la source vocale. Ses dimensions étant différentes suivant les personnes (quant aux longueurs et sections respectives de la portion pharyngée et de la cavité buccale), il est à prévoir, dans chaque cas, une particularisation de la fréquence des formants.

parois, élasticité des poumons [2], différences de capacité vitale, et de capacité résiduelle.

Ces différences organiques atteignent les quatre mécanismes acoustiques fondamentaux qui sont, rappelons-le, les poumons, le larynx, le conduit oral et les cavités nasales, et incluent des éléments comme la langue et les dents.

Par exemple, les cavités nasales, qui jouent un rôle dans la production des consonnes et des voyelles nasales, varient fortement suivant les individus, quant à leur taille et quant à leur configuration. Il faut aussi compter avec l'élasticité des membranes internes, qui est, elle, une source de variabilité pour un même locuteur. STEVENS (1971), indique que toutes ces variations peuvent s'observer directement dans le spectre du murmure nasal, qui a lieu lors de l'émission d'une consonne nasale, particulièrement en ce qui concerne la fréquence et la largeur de bande.

Trois facteurs "externes" conditionnent ces différences organiques. Ce sont : l'hérédité, le sexe et l'âge. La dysphonie, la dysarthrie, certaines neuropathologies (parkinsonisme) et psychopathologies (hystérie), peuvent amplifier les différences organiques. Les différences acquises sont le résultat de différences dans les commandes de coordinations neurales, apprises par chaque individu : contrôle des muscles du larynx, gestuelle articulatoire [1]. Elles expliquent les variations dans la dynamique du conduit vocal (taux de transition formantique et effet de coarticulation).

Comme facteurs "extérieurs" agissant sur ces différences acquises par l'individu, on note les facteurs d'origine géographique, et les facteurs sociaux et culturels. Les différences organiques et acquises, sont évidemment présentes et mélangées dans chaque manifestation phonique.

---

(1) La mécanique ventilatoire a un effet sur la pression subglottique et le taux de reconnaissance, variation de cette pression. On sait que celui-ci est en partie responsable de la variation de la fréquence fondamentale de la voix, pendant la phonation. Ainsi, si un locuteur se trouvant dans un état physiologique et psychologique approprié, voit son rythme d'inspiration-expiration; s'accélérer à la suite d'une cause quelconque, la pression subglottique durant la phonation, s'accro et sa fréquence fondamentale tend à s'accroître également.

(2) Rappelons que l'acquisition et la production de sons d'une langue particulière dépendent vraisemblablement de la formation d'habitudes, et il faut considérer que certains attributs individuels de la parole restent fixes lors de la production des sons.

La variabilité intra-locuteurs peut être augmentée par l'effort, la fatigue, la tension nerveuse, la peur. On observe surtout des modifications dans l'amplitude, la fréquence et la forme des impulsions glottales et aussi des changements dans la gestuelle articulatoire. Il semble par ailleurs qu'il y ait une augmentation du fondamental, et une diminution du débit avec l'âge. On note aussi l'apparition du chevrottement.
Respectivement à un même individu, rappelons que l'on peut définir, d'après leur durée, plusieurs types de variations :

- cycles circadiens : variations ayant lieu dans l'intervalle d'une journée, et sujettes à se répéter chaque jour,

- cycles plus longs, portant sur une période de une ou plusieurs semaines, lesquels ne sont pas clairement identifiés (variations dues à l'état physique : surmenage, fatigue..., et émotionnel du locuteur, aux variations climatiques...),

- enfin, les variations les plus étendues peuvent porter sur plusieurs années, et sont plutôt en rapport avec l'âge.

4 - Mode d'utilisation des variabilités inter et intra-locuteurs :

La variabilité de la voix d'une personne à l'autre, est donc un fait, et on peut abandonner l'idée que les sons de parole puissent être reproduits identiquement à eux-mêmes. On pourrait même affirmer ceci a priori, alors que nos moyens de perception (l'ouïe), nous en donnent une preuve subjective permanente. Outre cela, comment ne pas supposer qu'il y a une correspondance bijective entre les voix (les manifestations) et les caractéristiques individuelles (les causes) ? On aimerait que le schéma suivant soit vérifié :

Etat des caractéristiques individuelles — Etat du signal vocal

$C_1 \longrightarrow V_1$

$C_2 \longrightarrow V_2$

Deux complications nous empêchent donc de le faire :
1 - on ne peut répéter deux fois exactement le même signal de parole,
2 - deux personnes distinctes peuvent produire des signaux relativement voisins (par exemple : la discrimination des spectrogrammes n'est pas toujours possible

Il resterait donc à prouver que les habitudes vocales sont des conséquences de condition, et/ou d'états physiologiques des organes vocaux individuels. Les systèmes de reconnaissance automatique de la parole confirment ces points. Assez tôt, des auteurs (dont HECKER, 1971), ont indiqué que le succès de toute méthode de reconnaissance du locuteur dépend du rapport de la variabilité inter-locuteurs à la variabilité intra-locuteurs. Le problème consiste alors précisément à pouvoir mesurer ces variations.

Il s'agit donc d'évaluer des différences quantitatives contenues dans le signal de la parole, pour un même groupe socio-culturel homogène en âge et sexe. Des études précoces (KAISER, 1938), se sont intéressées aux corrélations :

qualités somatiques / qualités psychiques
d'un individu.

Si nous groupons en des classes abstraites des sons voisins, selon des similarités qui peuvent être de nature phonétique (physiologique, acoustique), allophonique, phonologique, prosodique, psychologique.

ces groupes contiennent probablement des clés pour l'identité du locuteur.

Prenons maintenant l'exemple de la reconnaissance des empreintes digitales (avec semblablement les opérations de Vérification et d'Identification), sans affirmer pour autant qu'il soit un homologue de notre problème, au point de vue de la méthodologie (ce qui est une question difficile et assez stérile). On sait que les deux assertions suivantes sont expérimentalement vraies :

1 - la variation des formes (stries, vortex, ...) est pratiquement nulle au cours du temps pour un même sujet,

2 - cette variation est très grande entre personnes différentes, à tout moment.(1)

Dans le cas de la parole, l'assertion 2 reste vraie, mais on vient d'indiquer que l'assertion 1 est tout à fait fausse.

KRAUSE (1971) affirmait que les résultats en reconnaissance du locuteur étaient d'une précision condisérablement moindre que pour la dactyloscopie (examen des dessins de la surface des doigts).

---

(1) Surtout si l'on ajoute la récente analyse des pores de la peau.

Figure 1 - Les différences inter-individuelles de la voix

Différences anatomiques

Forme, dimension des organes, (larynx, pharynx langue, dents, lèvres, cavités nasales, cavité orale).

↓

Différences physiques

Fréquence fondamentale valeur et largeur de bande des formants

Différences d'apprentissage

Influence de l'environnement (dès la petite enfance) sur la prononciation des sons, syllabes et mots.

↓

Variations dialectales

Différences dans le contrôle nerveux

Commande des articulations

↓

Différences dans la gestuelle articulatoire

Organisation temporelle provoquant des différences segmentales (transition formantique) et segmentales (débit, pause, ...).

II - FORMALISATION DU PROBLEME

1 - Introduction

Les auteurs ont cherché à formaliser les tâches de la Vérification et d'Identification de façon à mieux les distinguer (ROSENBERG, 1976 ; DODDINGTON 1970). PFEIFER (1978) suggère une Identificationpar Vérification : chaque référence est évaluée séparément des autres lors de la comparaison avec l'échantillon témoin. Cette approche a l'avantage de supprimer l'hypothèse non réaliste, mais jusqu'ici nécessaire dans l'Identification, à savoir la considération d'u population close (choix fermé). Mais BRICKER et PRUZANSKY (1971) déplorent le manque de théorie générale pour la reconnaissance du locuteur.

Les formalisations proposées ici (une pour la Discrimination, une pour l'Identification et une pour la Vérification) sont très étroitement liées.

Elles permettent de clarifier les différences entre les trois contextes d'application. En elles-mêmes elles ne nécessitent par d'application, étant de outils théoriques descriptifs.

2 - Première formalisation

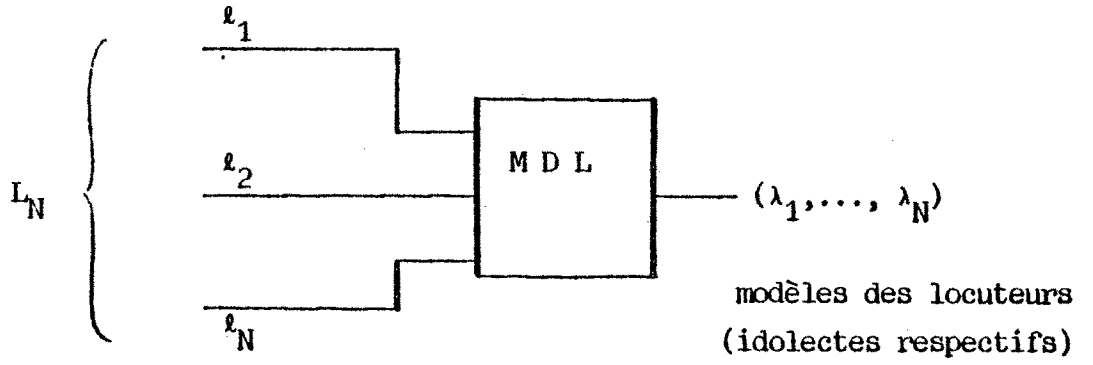
Soit $L_N = \{\ell_i, i = 1,\ldots,N\}$ l'ensemble des locuteurs. ℓ_o désignera un locuteur extérieur à la population et ℓ_x le locuteur à reconnaître.
$\{P_j, j = 1, J\}$ désigne l'ensemble des paramètres effectivement utilisés.

Le but de l'étude est double : étant donnée la population de référence L_I à "traiter", on doit parvenir à :

1 - classer les différents idiolectes en présence,
2 - reconnaître ensuite tout locuteur se présentant (que ce soit pour une vérification ou une identification d'identité).

Cela revient à construire deux machines parmi les trois suivantes, la première étant toujours présente.

2.1. Machine à discriminer les locuteurs

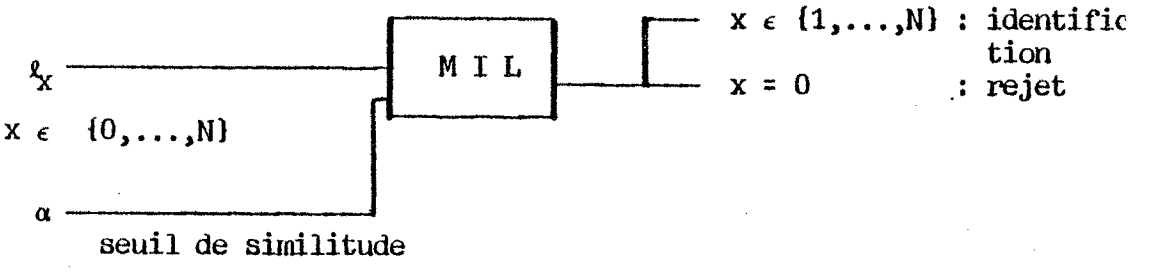


La structure de cette machine dépendra en général de N.

2.1. Machine à identifier un locuteur

C'est la machine à décider suivante :

Locuteur entrant

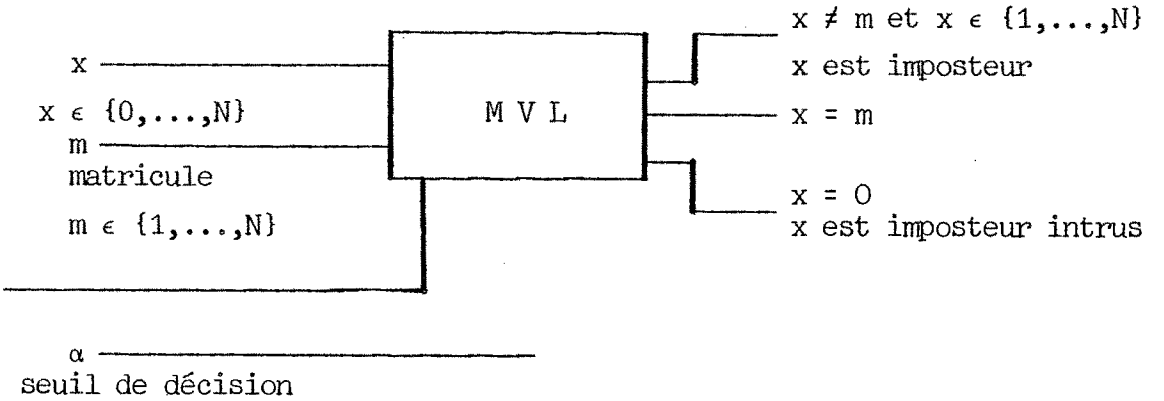


Les réponses possibles sont les suivantes :

Cas possibles	décision et action
Comparaison réussie (avec identification vraie ou fausse) : $\exists\ p \in \{1,\ldots,N\}$ tant que ℓ_x est semblable à ℓ_p d'au moins α	$x = p$ et ℓ_x est identifié à ℓ_p
Comparaison non réussie : $\forall\ p \in \{1,\ldots,N\}$ ℓ_x est semblable à ℓ_p d'au plus α	décider $x = o$ et déclarer qu'il s'agit d'une personne extérieure à la population $(\ell_x = \ell_o)$

2.3. Machine à vérifier un locuteur

On admet que le locuteur à vérifier est quelconque. Cette machine de décision est la suivante :



et caractérisée par le tableau suivant :

Cas possibles	décision et action
ℓ_x est semblable à ℓ_m d'au moins α	ℓ_m et ℓ_x condonfus Vérification acceptée
ℓ_x semblable à ℓ_m d'au plus α	Vérification refusée

2.4. Définition interne du locuteur

Pour un système automatique la relation binaire R entre locuteurs potentiels (ensemble $\mathcal{L}$) :

"est reconnu comme",

est fondée sur la relation de similitude S :

"est semblable, au seuil α, à ",

et est réflexive (cas de la Vérification) et symétrique. Elle doit être aussi antisymétrique car :

si seulement : $\ell_1 \, R \, \ell_2$ alors : $\ell_2 \, R \, \ell_1$ (symétrie)

$\ell_1 = \ell_2$ (identité)

S n'est pas antisymétrique ni transitive car sa définition n'appelle pas à ur décision globale. La transitivité est vraie pour R qui se place à un niveau décisionnel. Pour le système, l'ensemble des locuteurs est l'ensemble des cla ses d'équivalence $\mathcal{L}/R$. Si il coïncide avec L_N, le modèle de la population est

alors parfait. Il ne reste plus qu'à étiquetter convenablement (dans le bon ordre) les éléments pour éviter les erreurs de substitution. En pratique, les classes d'équivalences sont imparfaites car d'une part, les variations de la voix prise en compte ne sont pas exhaustives et d'autre part, on ne possède pas de paramètre idéal disjoignant les locuteurs obtenus lors de l'apprentissage.

3 - Descriptions détaillées des tâches

Nous retenons ici une procédure de type descendant, c'est-à-dire divisive ou opérant par fractionnements successifs. Elle permet avant tout une visualisation des phénomènes.

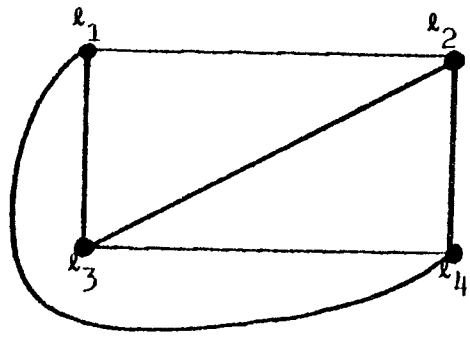
3.1. Description de la machine à discrimer les locuteurs

Soit le graphe $G_o = (L_N, U)$ non orienté dont les sommets sont les locuteurs et dont tout arc $u \in U$ est tel que :

$$u = (\ell_i, \ell_j) \in U \qquad \text{ssi :}$$

la machine confond les locuteurs ℓ_i et ℓ_j, à un moment donné de l'étude. Formellement, on confond au départ tous les locuteurs entre eux :

Exemple : N = 4



$G_o = K_4$, graphe complet d'ordre 4

Pour être efficace, chaque paramètre doit enlever au moins un arc au graphe de départ. Il se peut, évidemment, qu'après utilisation d'un paramètre p_1, le paramètre p_2 n'enlève aucun arc au graphe obtenu après p_1, alors qu'il est quand même efficace sur le graphe de départ. Tout le problème est donc, étant donnée une liste de paramètres, d'organiser les choix successifs des paramètres.

A chaque paramètre P_i on associe les arcs (U_{i_j}) $1 \leq j \leq k_{p_i}$ qu'il enlève au graphe G_o (ceci est relatif à la population L_N) et formant l'ensemble U_{p_i}.

r_{p_i} est l'ordre de P_i. Ainsi, au paramètre P_i, on associe l'application :

$$\Pi_{p_i} = (L_N, U) \rightarrow (L_N, U - U_{p_i}).$$

A la séquence ordonnée de paramètres :

$$P_{i_1,,,i_q} = (P_{i_1},\ldots,P_{i_q}),$$

correspond l'application composée :

$$\pi_{i_1,\ldots,i_q} = \pi_{i_q} \circ \ldots \circ \pi_{i_1}.$$

(la composition des paramètres est une loi interne commutative associative, idempotente).

On cherche une application $\pi_{i_1,\ldots,i_q}$ qui déconnecte le plus de sommets possible.

Si l'on cherche seulement des typologies de locuteurs, il suffira de déconnecter G_o en un certain nombre de composantes connexes.

En d'autres termes, on cherche un q-uplet ordonné (nous l'appellerons base) :

$$(i_1,\ldots,i_q) \quad \in \quad \mathbf{N}_+^q$$

tel que :

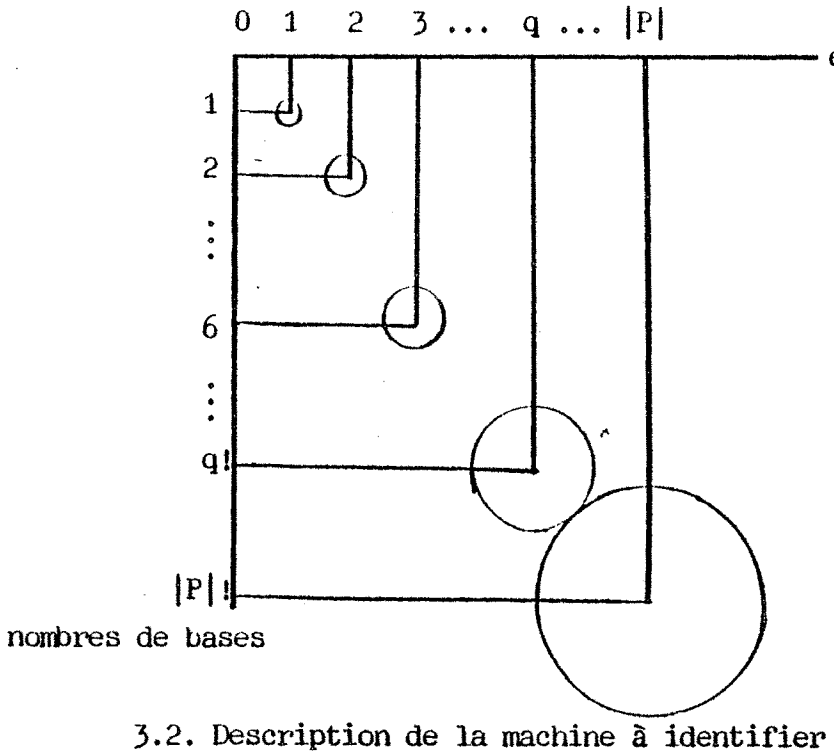
$$q \leq |P|,$$

$$\underset{P \leq |P|}{\text{Min}} \; |U - U_{j_1} - \ldots - U_{j_p}| \; = \; |U - U_{i_1} - \ldots - U_{i_q}|$$

$$(j_1,\ldots,j_p) \in \sigma_p$$

(σ_p est le groupe des permutations de $\{1,\ldots,p\}$).

Il n'y a pas unicité de la solution. On notera la croissance de complexité de l'algorithme théorique lorsque q croît.



3.2. Description de la machine à identifier un locuteur

L'heuristique est la suivante : on cherche d'abord à rejeter l'identification de ℓ_x avec tout élément de la population. En cas d'échec, ℓ_x est identifié à le ou les locuteurs ℓ_i tel que (ℓ_x,ℓ_i) soit un arc restant. Considérons le graphe $G' = (I_N \cup \{\ell_x\}, U\ \Omega(\ell_x))$, où $\Omega(\ell_x)$ est le cocycle du sommet Ω_i dans G' , c'est-à-dire l'ensemble des arcs joignant ℓ_i à un autre sommet :

$$\Omega(\ell_x) = \{(\ell_x,\ell_j) \in U,\ \ell_j \in I_N\}$$

$\Omega(\ell_x)$ est un déconnectant trivial de ℓ_x. Le problème de l'identification est de trouver $(i_1,\ldots,i_q)$ tel que :

$$\bigcup_{k=1}^{q} U_{i_k} \supset \Omega(\ell_x) \quad \text{dans } G'.$$

On ne retiendra parmi les bases possibles que (l'une de) celle qui optimise l'un des critères :

- minimiser q
- minimiser le coût de l'identification : le coût de calcul du paramètre P_i étant C_i, le coût de l'identification sera $\sum_{i=1}^{q} C_i$.

Ce schéma privilégie l'identification parfaite plutôt qu'une identification rapide. C'est du reste en accord avec le type d'application que l'on peut en attendre.

3.3. Description de la machine à vérifier un locuteur

La tâche consiste en une comparaison entre :

$$\ell_x \quad \text{et} \quad \ell_m$$

où m est le code $\in \{1,\ldots,N\}$ annoncé par ℓ_x. Considérons l'arc (ℓ_x,ℓ_m) de G'

L'heuristique choisie obéit à l'un des deux principes suivants :

- pour accepter la vérification de ℓ_x avec ℓ_m, il est nécessaire que tous les paramètres déconnectent ℓ_x et ℓ_m :

$$\bigcap_{i=1}^{|P|} U_i \quad (\ell_x,\ell_m),$$

- pour rejeter la Vérification de ℓ_x d'avec ℓ_m, il est suffisant de trouver un paramètre p_i tel que

$$U_i \supset (\ell_x,\ell_m).$$

En pratique on ne peut être aussi catégorique et il faudra décider d'un seuil en deçà duquel on décide l'acceptation et en delà duquel on décide le rejet. Ce seuil dépend de la pertinence des paramètres.

3 - Conclusion

Nous avons ramené le problème de la Discrimination à un problème de recherche d'ensembles déconnectants minimum, celui de l'Identification à un problème de minimisation, et celui de la Vérification à la recherche d'un ensemble de paramètre déconnectant.

Même si elles sont éloignées des détails théoriques exposés ici, les procédures pratiques proposées par la suite tendent à suivre même imparfaitement et incomplètement, à cause de l'introduction de seuils, les directives données ici.

Nous avons décrit des principes d'identification de haut en bas, ce qui correspond au point de vue suivant : on favorise le rejet des locuteurs comparés au locuteur examiné, c'est-à-dire que l'on préfère dire qu'un locteur n'est pas telle personne plutôt que le contraire.

Les procédures indiquées ci-dessus peuvent donner lieu à l'existence de plusieurs locuteurs répondant à la question : c'est que l'on n'a pas inclu dans le modèle de notion de métrique (de similitude, de proximité) permettant, par exemple, de décider du plus proche.

D'autre part, l'assertion "le paramètre p déconnecte les locuteurs ℓ_1 et ℓ_2" sous-entend, lors de l'application, la définition d'un seuil, en deçà (resp. au-delà) duquel ℓ_1 et ℓ_2 sont donfondus (resp. déclarés distincts) La valeur de cette limite n'a rien d'arbitraire, dépend de nombreux facteurs et nécessite beaucoup d'expérimentations.

III - CONSTITUTION D'UNE BASE DE DONNEES

1 - L'acquisition d'un corpus de taille supérieure :

1.1. Considérations générales :

Les lignes suivantes traitent de considérations générales et particulières en vue de la constitution d'une véritable base de données vocales.

Ces considérations ont été dégagées grâce à l'expérience acquise lors de la première étude décrite dans la partie IV-I.

1.1.1. Considérations générales :

Le fait de conduire une expérimentation sur des données préalablement enregistrées, c'est-à-dire n'évoluant pas, constitue bien l'hypothèse de travail de base. C'est là une hypothèse limitative dans la mesure où l'on fige les voix, c'est-à-dire où l'on réduit le nombre de variables du problème.
Cependant, quels sont les buts d'une base données vocales générale ?

1 - Fournir une grande variété d'évènements parlés, susceptibles de mettre en évidence (par des mesures plus ou moins directes), les caractéristiques de la voix. Celles-ci incluent :

- Une variabilité intra-locuteurs la plus exhaustive possible, sous des conditions normales (le locuteur ne modifie pas sa voix, il jouit d'un bon état physiologique).

- Une variabilité inter-locuteurs s'appuyant sur la considération d'un échantillon assez représentatif de la population à considérer.
  Les enregistrements doivent permettre d'identifier les éléments pertinents du domaine temporel et/ou fréquentiel de la parole, avec un haut niveau de confiance.

- La possibilité d'évaluation des variations qualitatives et quantitatives du signal de parole à l'intérieur d'un même groupe homogène en âge et caractéristiques sociales (voir les travaux précurseurs de Louise KAISER, 1938 ...).

2 - Permettre la simulation d'un système de reconnaissance du locuteur (essentiellement la tâche de Vérification, puisque le nombre de locuteurs y est limité), l'évaluation de méthodes différentes par la comparaison des différentes performances.

3 - Tester la résistance des caractéristiques à la dérive vocale : les enregistrements doivent être distribués dans le temps afin d'autoriser des études sur la variation de la parole dans le temps à locuteur fixé.

4 - Permettre des études stratifiées : par sexe, âge, variantes régionales.

5 - Tester des systèmes de reconnaissance de la parole (qualités d'adaptation au locuteur courant, normalisation du locuteur courant).

D'une façon générale, elle ne comporte pas d'algorithmes de prétraitement [1].

1.1.2. Justifications : pourquoi un fichier de données vocales ?

Un tel fichier de données vocales portant sur différents locuteurs, doit être fondé sur les variations inter-locuteurs [2]. Il constitue un matériau indispensable pour la réalisation d'un prototype expérimental (les chercheurs américains et allemands en ont déjà réalisé plusieurs), et doit aider à la modélisation du locuteur. Il est couramment utilisable pour la phase de test, étant donné que le test in vivo est peu pratique lorsque l'on fait intervenir de nombreux locuteurs.

1.1.3. Limitations de la base de données

Les données recueillies conditionnent la suite des travaux.
La population de référence devant être linguistiquement homogène, les études ne pourront être valables que pour une langue donnée. Y-a-t-il une durée de validité des données d'apprentissage ? D'après les auteurs (GRENIER, 1977), la durée de validité croît fortement avec l'étendue des données d'apprentissage. On peut même avancer le tableau suivant :

---

(1) L'avis d'un auteur ayant réalisé une telle base de données :
"It was not desirable to incorporate elaborate algorithms into the data-base creation programs, as the perfection of a segmentation routine is in itself a major project challenging the current state of the art of speech understanding systems research" (COLLINS, 1977).

(2) Rappelons la phrase bien connue de HECKER (1971) : "Le succès de toute méthode de reconnaissance du locuteur dépend de combien la variabilité inter-locuteurs dépasse la variabilité intra-locuteurs."

NOMBRE DE SESSIONS D'ENREGISTREMENT	PORTEE
1 session	3-4 jours
4 sessions sur 4 jours	1-3 semaines
des sessions réparties sur 5-7 semaines et espacées d'une semaine	plusieurs mois

D'autre part, on ne peut gonfler les données à volonter : elles occupent une place mémoire précieuse, d'où la nécessité de réduire le nombre de phrases enregistrées, ainsi que leur longueur (également celle de supprimer les silences entre les phrases).
Une autre limitation est constituée par l'influence du contexte, due au texte précis utilisé, et qui ne peut être éliminée.

1.1.4. Quelques questions auxquelles il n'est pas clairement répondu.

a) Combien d'échantillons par locuteur retenir ? On répond indirectement par la nécessité d'avoir des enregistrements étendus dans le temps, donc leur nombre sera déterminé par une couverture suffisante des variations temporelles intra-locuteurs.
En fait, si l'on réalise les enregistrements au cours d'une même session, on ne sait pas très bien combien de répétitions considérer.

b) On n'a pas de moyen permettant d'apprécier le degré de coopérativité des locuteurs.

c) Y-a-t-il des enregistrements bons (à conserver) et d'autres insuffisants qui doivent être rejetés ? Comment interpréter : silences, hésitations, toussotements, lapsus, répétitions de mots ? Avec un entraînement du locuteur, ils deviennent rares, de faible importance et ne nécessitent peut-être pas de recommencer.

d) Quelle importance donner d'ailleurs à l'entraînement du locuteur ?
Ceci constitue une forme d'adaptation du locuteur à l'expérience qui peut aller au-delà d'une simple accoutumance et qui peut accroître les résultats de reconnaissance.

1.2. La constitution de la base :

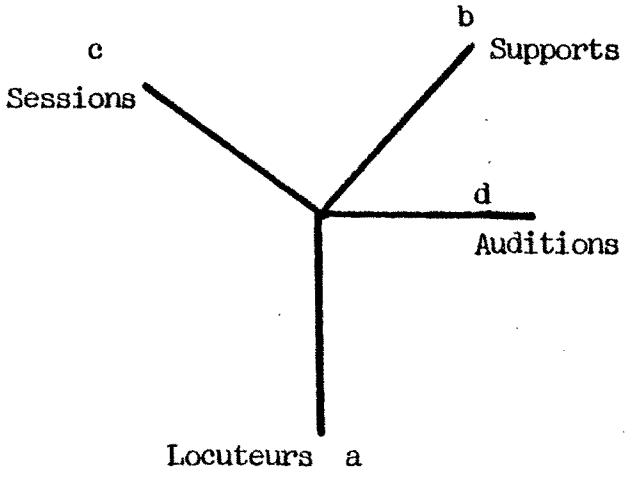
D'une façon générale, les bandes analogiques sont numérisées sur une piste une deuxième piste sevant à repérer - souvent par signal carré- la première. La finesse de la graduation dépend du taux d'échantillonnage utilisé pour la première piste. Ce sont les études fréquentielles qui le déterminent, puisqu'elles requièrent un taux beaucoup plus élevé que pour les études de type purement temporel (par exemple, 12 KHz contre 100 Hz).

Les étiquettes sont codées de façon à en réserver l'usage.

1.2.1. Les éléments principaux constituant la base :

Ils sont des représentations abstraites des quatre facteurs principaux de l'étude, répondant aux questions : qui, quoi, où, comment, à savoir respectivement :

a) les locuteurs,
b) les supports phonétiques (les phrases par exemple),
c) les répétitions (ou sessions),
d) les conditions d'enregistrement.



On peut faire les hypothèses suivantes :

1- Le même nombre de répétitions vaut pour chaque phrase et chaque personne.

2- Il existe un nombre R de répétitions (ou de sessions), qui est nécessaire et suffisant pour étudier le problème.

1 - Les locuteurs :

a) Il faut se fixer des choix sur l'homogénéité ou les variétés linguistiques, dialectables, socio-culturelles, etc...[1], mais il restera à apprécier le degré d'homogénéité. L'âge pourra varier entre 20 et 50 ans environ. Les deux sexes doivent être représentés en nombre égal, à concurrence d'une trentaine de représentants pour chacun. Mais le problème de la taille en locuteurs se heurte bien-sûr à des considérations pratiques[2]. En tout état de cause, les locuteurs pourront être disponibles ultérieurement, une année ou plusieurs années après, pour des études de dérive de la voix, et des expériences de reconnaissance (d'où une estimation de la dégradation du taux de reconnaissance si elle est constatée). De façon évidente, pour toutes ces études de laboratoire, on évitera les affections ORL ainsi que les gros fumeurs. Une question semble se poser : peut-on, ou doit-on utiliser des professionnels de la voix (acteurs, speakers, imitateurs) ?

En phonétique, on élimine systématiquement les professionnels de la voix, qui constituent une catégorie en soi.

b) Mise en oeuvre :

A chaque instant de l'utilisation, les locuteurs seront divisés en deux classes : des locuteurs coopératifs et des imposteurs chargés de déguiser et/ou d'imiter d'autres voix que les leur. Bien-sûr un même locuteur pourra être tour à tour l'un et l'autre pour augmenter la "productivité" de la base.

Il est loisible que chaque locuteur dispose en échange du don de sa voix, d'une fiche individuelle décrivant sommairement les caractéristiques de sa propre voix, surtout de façon comparée aux autres, à la moyenne de la population étudiée. Ceci fait partie de conditions psychologiques entourant la constitution de la base (cf. COLLINS, 1977 a) et b), et § 4).

---

(1) Au sujet de la prise en compte du passé dialectal, voir FANT & GRANDSTROM 1972, p. 10.

(2) Il nous semble que les études à taille restreinte (moins de 15 locuteurs), devraient désormais laisser la place aux études sur larges populations uniquement.

2 - Les supports phonétiques :

Ils doivent d'abord être représentatifs de la langue parlée concernée, être les mêmes pour chaque locuteur. Ils seront organisés en phrases plutôt qu'en mots, logatomes ou sons isolés, compte-tenu du fait que d'assez nombreuses études ont porté sur ces unités minimales et que le taux de reconnaissance se dégrade lorsqu'on plonge les unités phonétiques étudiées dans un contexte tel que celui de la phrase, qui a par ailleurs l'avantage d'être le contexte naturel, donc celui à utiliser lors des applications (1). Il convient toutefois de neutraliser quelque peu a priori le débit, dans la mesure du possible et lorsque ce n'est pas le débit qui est étudié : on préfèrera alors des phrases et des textes neutres (qui n'impliquent pas le lecteur), sans interrogations ni exclamations.

L'ordre des phrases pourra être imposé, non aléatoire. Les phrases seront séparables à l'oreille.

Quant aux sons intentionnellement insérés dans les syllabes, mots et phrases, on se souviendra que :

- le contour des voyelles (très étudié) est caractéristique du locuteur dans le domaine fréquentiel,
- le [ʁ] est très polymorphe (possède une grande latitude de réalisation et ne saurait être investi d'une stabilité phoné ique toujours recherchée)
- le [p] est moins sensible au contexte, [s] et [ʃ] varient beaucoup moins en durée que [p] et [k].

Sur le choix d'un corpus pour la détermination du fondamental, on pourra consulter l'article de HORII (1975).

Il faut aussi tenir compte du problème de la segmentation phoné.
tique (pour autant que cela reste possible), et à un échantillonnage point par point, ou par déplacement d'une fenêtre.

Parmi les modes d'énonciation suivants :
- parole libre (l'effet de répétition la diminue),

---

(1) Le mot isolé est une production particulière, à étudier en tant que telle.

- lecture extemporanée,
- lecture sans préparation,
- déclamation,
- conversation extemporanée,
- conversation libre.

L'intonation sera de type énonciatif. Les intonations de type non énonciatif (interrogatif, exclamatif, ...) sont très difficilement reproductibles, de très nombreuses variantes étant acceptables en situation de communication. On se reportera au tableau 7 pour les caractéristiques propres d'un texte.

Les paramètres qui seront mesurés sont évidemment fonction des supports phonétiques sélectionnés (cf. WOLF, 1972). Par exemple, la mesure de la durée d'occlusion nécessite la position intervocalique ; pour les mesures de durées segmentales, il vaut mieux choisir des syllabes accentuées et en position non finale. Ce conditionnement réciproque :

choix des supports phonétiques $\longleftrightarrow$ choix de paramètres a priori

fait qu'une base de données à vocation universelle reste peu réalisable.

TABLEAU 7 - Les caractéristiques d'un texte destiné à être lu.

Organisation :	- Mots courts, simples. - Phrases courtes, un seul paragraphe, (Un texte éducatif utilisé pour apprendre le français par exemple).
Contenu :	- N'implique pas affectivement le locuteur (on pourra préférer les phrases isolées aux mêmes phrases plongées dans le texte d'origine), - Vocabulaire facile non spécialisé, pas de noms propres, ni d'abréviations.
Présentation :	- Claire et dactylographiée. Emploi des virgules et des tirets ? Lignes pas trop longues. - Typographie très lisible (pour souplesse de lecture et évitement de retours-arrière).

Le texte sera testé avant d'être utilisé. Il sera fait une statistique (nombre de mots, de syllabes, longueur moyenne des mots, des phrases).

3 - Les sessions :

Les diverses répétitions à intervalles réguliers ou non, sont rendues nécessaires pour des raisons d'exhaustivité des variations intra-locuteurs. Elles introduisent, lorsqu'elles se suivent au cours d'une même session, diverses altérations connues sous le nom d'effet de liste (ou de série) : par exemple, baisse incomplète en amplitude et en intensité de la fin de phrases. Il semble y avoir deux solutions pour éviter un effet ou effet de liste abusif : soit faire davantage de répétitions et rejeter les dernières (éventuellement les premières aussi), soit appliquer des corrections(plus ou moins arbitraires).

Dans le cas idéal, les sessions sont parfaitement distribuées dans le temps : il y a plutôt de nombreuses sessions courtes que le contraire. Les sessions seront placées de façon à pénétrer les divers types de variations :

circadiennes et autres. Ainsi, en retenant divers sous-ensembles de ces sessions, on peut directement étudier l'un des types de variation. Il serait utile de définir un facteur de dissémination qui indiquât la pénétration.
La formule est du type :

$$D = \sum_{v=1}^{V} a_v \; S_v \; n_v$$

où V est le nombre de types de variabilités étudiées (circadiennes , hebdomadaires, mensuelles, trimestrielles ou annuelles pour fixer les idées [1]), S_v est le nombre de sessions systématiquement programmées pour chaque type de variation, (par exemple : 3 sessions par jour ou 5 par semaine), n_v est le nombre de cas disjoints où ce type de variation est étudié (exemple : un enregistrement par jour, pendant un mois, donne quatre cas de variations hebdomadaires disjoints). a_v est le vecteur $(0,\ldots,\underset{v}{1},\ldots,0)$ à autant de composantes que de types.
Cas particulier : il arrive que le nombre de sessions lors de chaque cas et pour un même type de variation, ne soit pas le même. Il faut alors remplacer $S_v \; n_v$ par sa généralisation : $\sum_{k=1}^{K} S_v^k$ où K est le nombre total de cas et S_v^k le nombre de sessions ayant lieu dans chaque cas.

4 - Le matériel et les conditions d'enregistrement :

On pourra choisir une chambre sourde, des bandes magnétiques à faible écho, un magnétophone 2 pistes à vitesse élevée (9,5 ou 19 cm/s). Mais l'applicabilité des résultats au monde réel sera d'autant moins envisageable. La question d'utiliser ou non le réseau téléphonique usuel avec emploi d'un microphone courant) est en grande partie conditionnée par l'importance que l'on donne aux applications possibles de la généralisation des techniques de reconnaissance du locuteur.

---

(1) On peut admettre pour simplifier, que l'étude d'une unité supérieure se fait en ignorant les variations pour l'unité inférieure : par exemple, pour étudier les variations hebdomadaires, il suffit d'un enregistrement, par jour, mensuel, un par semaine, etc...

Toutefois, l'enregistrement se fera chaque fois dans les mêmes conditions acoustiques. Le niveau sonore, l'intensité du signal, le niveau de bruit moyen et la distance micro-bouche, seront préalablement observés et réglés. Les conditions psychologiques ne seront pas négligées (confort, anxiété,..., des locuteurs). Certains auteurs (cf. COLLINS, 1977, 1979), réalisent une préparation en commun. Le mode de prononciation est fixé : parole libre, récitation, lecture,.. Les locuteurs doivent appartenir à un type donné : coopératifs, naïfs, exercés. La coopérativité des locuteurs peut s'obtenir en leur demandant d'être aussi naturels que possible. DAS & al. (1972), indiquent de parler "clairement et distinctement" sans donner de consignes plus spécifiques nuisibles. Il est commode d'avoir un étiquettage immédiat : le locuteur annonce le numéro de la phrase, etc. , à chaque fois. Il y a enfin un problème de présélection des enregistrements : faut-il parfois éliminer des phrases à cause d'une mauvaise diction ? Un essai préalable peut être utile.

3 - Présentation de SYSIPHE

Il a été commencé dans le cadre du présent travail la constitution d'une collection de données vocales en vue de la reconnaissance du locuteur. Ses caractéristiques résultent des considérations générales ci-dessus. Nous n'en donnerons qu'une brève description étant donné que son exploitation et sa transformation en base de données n'en sont qu'au début, et sont peu utilisées dans la suite. 38 locuteurs (dont 31 femmes) d'âge adulte ont été enregistrés suivant les modalités résumées à la page suivante : les personnes téléphonent (par le réseau universitaire) au moment de leur choix au cours de la tranche horaire correspondant à leur groupe à un poste donné muni d'un dispositif d'enregistrement automatique des messages. Il est également ajouté quelques locuteurs imposteurs. La fréquence des appels est laissée libre (maximum : 1 par demi-journée).

Si l'on calcule la valeur du facteur de dissémination théorique, en considérant qu'il y ait un appel par jour, sur 25 jours, soit 5 semaines, on trouve :

$D = (D_{circadienne}, D_{hebdomadaire}, D_{mensuelle}) = (1\times25, 5\times5, 1\times5) = (25, 25, 5)$.

GROUPE 3

SEMAINES	MATIN		APRES-MIDI	
				entre 13h et 15h
.du 7 mai au 11 mai	☎	entre 9h15 et 13h	☎	" " "
.du 14 mai au 18 mai			☎	" " "
.du 21 mai au 25 mai			☎	" " "
.du 28 mai au 1er juin			☎	" " "
.du 4 juin au 8 juin			☎	" " "
.du 11 juin au 15 juin			☎	" " "
.du 18 juin au 22 juin			☎	" " "

Groupe	Horaire	
	matin	après-midi
1	9h15 - 11 h	13 h - 15 h
2	11h - 13 h	13 h - 15 h
3	9h15 - 13 h	13 h - 15 h
4	9h15 - 13 h	15 h - 17 h

INSTITUT DE PHONETIQUE

Opération Sisyphe

Petit guide à l'usage du correspondant téléphonique

1 - Se présenter ...	Allo # Sisyphe # (prénom ... nom) # Vous m'entendez ?
2 - Compter à rebours de 10 à 0 en détachant bien chaque nombre ...	10 9 8 7 6 5 4 3 2 1 0
3 - Lire le texte suivant ...	Le café de la rue du Paradis n'a pas changé depuis quinze ans. Depuis quinze ans, les mêmes tables, les mêmes cendrier les mêmes fauteuils qui souffrent quand v vous asseyez, les mêmes clients, pourrait ajouter. Monsieur et Madame viennent ici tous les jours. Tous les jours, à la même heure, ils s'assoient à la même place et demande à la serveuse la même chose. Ils doivent y retrouver toute leur vie.
4 - Lire les deux phrases suivantes en respectant les pauses	"L'avocat # Il est bien connu" "Il le mata # C'est bien connu"
5 - Faire un texte libre comportant dans un ordre qui peut varier d'un appel à l'autre ...	Vos Prénom et Nom, le lieu, le jour, la date, le mois, l'année, l'heure.
6 - Dire ...	"Fin de l'enregistrement"

Pause.

En cas de lapsus, veuillez recommencer le paragraphe concerné.

III

METHODOLOGIE DE LA MISE EN OEUVRE

"Une substance sans forme est un continuum sans contour, inaccessible à la description et à la science".

Bertil MALMBERG

I - Reconnaissance du locuteur par l'audition :

Ce type de reconnaissance possède une limitation particulièrement forte : c'est son entière subjectivité. Quelle que soit la précision et la confiance attribuées aux auditeurs, on sait bien que personne n'est à même de décrire les critères utilisés, et donc ne peut justifier ses décisions (3). Pourtant, il est bien souvent supposé que l'auditeur utilise un nombre de paramètres qui est faible, lui permettant d'ailleurs d'extraire plus d'informations que par des mesures physiques.

Les caractéristiques de ce type de reconnaissance sont les suivantes :

- on enregistre un ensemble de personnes prises parmi une population donnée ; le texte est choisi et fixé à l'avance,
- on fait écouter ces enregistrements à un ensemble d'auditeurs (le jury).

Le but essentiel de ce type est d'évaluer la vraisemblance d'erreur d'un auditeur ou du jury. Mc GEHEE (1937) a étudié la capacité de reconnaissance des voix non familières (1). Le résultat essentiel fut la décroissance rapide de la qualité des réponses au-delà de deux semaines d'intervalle entre l'apprentissage et le test. Par ailleurs, à intervalle de temps égal, la reconnaissance décroît assez rapidement lorsque croît le nombre de locuteurs. Enfin, il a été mis en évidence l'influence de l'imitation et de la modification de la voix (2) sur la baisse du taux de reconnaissance. Ce taux reste d'une façon générale, toujours très loin de 100%.

Mc GEE (1965) a montré que les variations inter-personnelles des prononciation des voyelles isolées sont fondamentalement indépendantes de la voyelle.

Comme on le voit, les recherches de ce type sont les plus anciennes. Selon HECKER (1971), elles ont plusieurs buts :

a) Déterminer les variables affectant la performance de l'auditeur ; développer des tests pour contrôler ces variables.

---

(1) Chaque auditeur a participé à deux sessions (1ère session d'apprentissage), séparées par un jour à 5 mois.

(2) En anglais : vocal disguise.

(3) Des locuteurs entraînés ont pu exposer leur méthode de sélection des voix à variante dialectale, précision articulatoire, résonance nasale,... (CARBONELL& al. 1965).

b) Comprendre les bases perceptives de la reconnaissance du locuteur. (On postule qu'un auditeur n'utilise qu'un petit nombre de paramètres perceptifs pour discriminer la voix).

c) Connaître des correspondances acoustiques de l'identité du locuteur : (en modifiant certains attributs du signal de la parole et en notant l'effet correspondant sur la performance de l'auditeur, on peut décider quelles sont les caractéristiques acoustiques pertinentes du point de vue de l'identité du locuteur). Cet aspect est important d'un point de vue fondamental.

d) Estimer le taux d'erreur de l'auditeur (1).
Il faut pondérer ces études par la notion de faillibilité de l'auditeur : quelle est la limite de résolution de l'oreille ?

e) Parvenir à quelque connaissance de la nature des mécanismes de décision chez l'être humain : il n'est pas certain que deux individus utilisent les mêmes caractéristiques acoustiques ni les mêmes critères, face à l'évaluation d'un même signal de parole. Il faut accepter le fait que deux personnes ne travaillent pas avec le même matériau : par suite de l'entraînement de la mémoire auditive et d'expériences antérieures de voix "semblables". Ceci nous conduit à la notion d'auditeur entraîné à la Vérification et à l'Identification. Du point de vue de l'auditeur, il faut faire la distinction entre voix familières et voix non familières (2).

Pour reconnaître les personnes par l'audition, nous étudions, selon LINDQVIST, l'intonation et la dynamique du spectre glottal. Ceci suffirait pour nous montrer qu'une machine peut difficilement utiliser les mêmes caractéristiques. Les variables de la reconnaissance par l'audition tiennent compte de la taille, du matériau phonétique et de l'homogénéité (3) de la population de locuteurs.

---

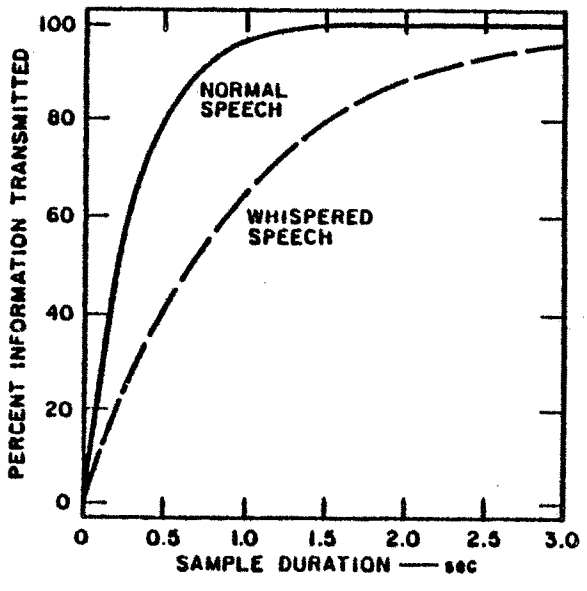
(1) Cette question est d'origine légale. On consultera Mc GEHEE (1937) pour : "The Lindberg Case of 1935".

(2) STEVENS & al. (1968) ont constaté que cette distinction est assez bien faite (90% pour 32 items, 1 échantillon et 8 références librement disponibles).

(3) C'est-à-dire la similitude perceptuelle des voix lors de l'audition. COMPTON (1963) considère le fondamental de la voyelle [i] prononcée isolément, comme une bonne "mesure" de l'homogénéité.

Si des expériences de synthèse (parler à la manière de ...) étaient réalisés, elles pourraient nous renseigner sur de "bons" paramètres perceptifs. STUNTZ(1963) a trouvé que des groupes de locuteurs considérés comme d'autant plus homogènes, donnaient des résultats d'autant plus faibles. En outre, les confusions ne sont pas du tout réparties uniformément.
Le schéma suivant indique outre la contribution majeure de la fonction de voisement à la reconnaissance du locuteur, que lorsque la durée de l'échantillon dépasse 1s, il n'y a pas un gain d'information important,pour la parole normale.

Figure 1- Transmission de l'identité en fonction de la durée de l'échantillon.
Extrait de POLLACK, PICKETT & SUMBY (1954).
(Voix normale et chuchotée).



Selon BRICKER & PRUZANSKY (1966), le nombre de sons contenus dans l'échantillon est plus important pour la performance de l'auditeur que la durée de cet échantillon (cf. les expériences de WILLIAMS (1964) sur les mots monosyllabiques, STEVENS & al. (1968) sur les disyllabiques). CLARKE & al. (1966) confirment ces résultats.

La taille du groupe d'auditeurs est un facteur important et beaucoup d'auteurs ont inclus plus de dix auditeurs, afin de neutraliser la dépendance de leur étude vis-à-vis des auditeurs réellement utilisés.
La façon de présenter le matériau vocal a également son importance ; soit :
- l'ordre des séquences parlées est fixe et déterminé par l'expérimentateur, soit :
- l'ordre des séquences est libre et choisi par l'auditeur : tout le matériau vocal du locuteur à tester est disponible au même moment.

Le matériau vocal fait appel aux mêmes considérations d'ensemble que pour une reconnaissance automatique (cf. section IV-II). On a aussi utilisé les sons sourds : SCHWARTZ (1968) et INGEMANN (1968) ont utilisé avec succès les fricatives [s] et [ʃ] pour faire distinguer le sexe aux auditeurs.

Le temps alloué au test peut être limité ; il n'est pas sûr que les tests en temps non limité donnent de meilleurs résultats.
WILLIAMSON (1961) a étudié les deux modes de présentation : l'ordre imposé donne de meilleurs résultats en temps limité à la question : "les deux échantillons proviennent-ils du même locuteur ?". Lorsque le temps alloué est largement augmenté, la présentation libre offre des résultats meilleurs [1]. Ce formatage des tests d'audition n'a pas donné des résultats tranchés. Remarquons toutefois que, selon le mode de présentation des échantillons, l'expérience des auditeurs peut se trouver renforcée (les tests jouent le rôle de nouvel apprentissage), ou diminuée (nécessité d'un nouvel apprentissage par suite de confusions).

---

(1) Il faut voir que l'on fait appel à deux types de mémoires différentes : mémoire à court terme pour le premier cas, à plus long terme pour le second.

Les tâches proposées aux auditeurs sont variées. Le tableau suivant en détaille quelques-unes :

Description	Nombre d'échantillons en présence
1) Identifier un locuteur connu (mémoire à long terme)	1
2) Identifier un locuteur : quelle référence est la plus proche de l'échantillon (comparaisons d'échantillons)?	1 de test + 1 de comparaison
3) Affecter des attributs à une voix (1) (sur des échantillons assez longs)	1
4) Les deux échantillons proviennent-ils du même locuteur (discrimination)?	2

Tableau 1- Classification de tâches des auditeurs (d'après HECKER, 1971).

Les tâches 2 et 4 sont binaires. La tâche 4 présente des avantages théoriques : rôle de la mémoire à court terme le plus faible possible, une seule stratégie chez l'auditeur. Avantage pratique : c'est la plus facile à mettre en oeuvre. Les résultats sont cependant plus difficiles à interpréter. On utilise souvent les courbes COR de la théorie du signal.

Une question reste à préciser, c'est la stratégie utilisée par l'auditeur. Bien qu'il soit difficile de la fixer, une uniformisation des stratégies permettrait des comparaisons de résultats plus fiables.

---

(1) HOLMGREN (1963) a traité 10 voix par 10 auditeurs avec 12 paires d'attributs. VOIERS (1964) fit une étude générale : 49 paires d'attributs pour 16 voix et 32 auditeurs. Une analyse en composantes principales fit ressortir 4 axes principaux dénommés par l'auteur : clarté, rudesse, grandeur, et vivacité, avec leurs opposés.

Elles sous-entendent l'usage de seuils de décision : on sait que deux auditeurs distincts utilisent des seuils distincts et donnent donc des réponses différentes, et/ou opposées. C'est pourquoi on leur demande d'associer un taux de confiance subjectif à leur décision. Deux réponses distinctes peuvent être parfois ramenées à deux réponses identiques (1). D'autre part, l'auditeur peut changer de stratégie en cours d'analyse, éventualité dont les conséquences n'ont pas été prises en considération.
La taille en locuteurs dépend fortement de la tâche proposée et aussi de la façon dont est présenté le matériau (de 4 locuteurs s'il y a appel à la mémoire à court terme, à beaucoup plus dans les autres cas.

On espère donc de la reconnaissance par l'audition, une meilleure connaissance du sujet, par l'investigation des mécanismes de perception, et surtout, un moyen d'évaluer les performances, donc la fiabilité, des systèmes automatiques. Les bases perceptuelles de la reconnaissance du locuteur, et ses corrélations acoustiques, n'ont pas été vraiment étudiées jusqu'ici. En outre, il se dégagerait la conclusion que les caractéristiques articulatoires seraient plus importantes que les caractéristiques de la source glottale.

Toutes les études sur la reconnaissance du locuteur par ordinateur, devraient donc porter un vif intérêt à cette méthode naturelle constamment utilisé Ces quelques pages ont donc valeur d'introduction au sujet.

Pour une étude détaillée de la reconnaissance du locuteur par l'audition, il faut se reporter à HECKER (1971).

---

(1) Le nombre de réponses possibles, multiplié par le nombre de taux de confiance (de très sûr, à douteux), donne un nombre élevé d'éventualités qui se regroupent en cas moins nombreux.

II - Reconnaissance du locuteur par l'analyse visuelle de spectrogrammes :

La présente méthode est considérée comme davantage objective : le spectrogramme, représentation visuelle d'une analyse du signal sonore fugitif des sons, est une image pouvant se prêter à des analyses objectives. Cependant, son interprétation reste subjective.

La méthode est la suivante : on présente à une personne entraînée spécialement, des spectrogrammes de différentes énonciations d'un même mot ou phrase ; la personne essaie de déterminer si ces prononciations proviennent ou non du même locuteur, ou effectue l'une des diverses tâches exposées au paragraphe précédent.

On sait que la particularité du spectrogramme est de rapporter simultanément les propriétés temporelles et fréquentielles du signal. Cependant, il possède des limitations : seules certaines caractéristiques du signal de parole peuvent apparaître à chaque fois. Seuls les spectrogrammes à larges bandes sont utilisés ici et de nombreuses différences peuvent ne pas apparaître. Ainsi, deux spectrogrammes apparaissant comme identiques, peuvent ne pas provenir de la même personne, ni représenter le même message parlé. Cela est encore plus vrai lorsque les signaux subissent une dégradation ou une distorsion (1) (cf. la mystique du sonagramme, section I-III).

KERSTA (1962), YOUNG & CAMPBELL (1967), STEVENS & al. (1968), ont montré que la difficulté de reconnaissance n'est pas la même suivant les locuteurs. Les femmes sont reconnues plus facilement (KERSTA (1962), ANON, (1965)).

On peut définir, comme lors de l'audition, un concept d'homogénéité, ou similarité perceptuelle. On sait toujours peu de choses sur les corrélations perceptuelles et physiques de cette homogénéité. (Deux voix apparemment proches à l'audition, peuvent produire deux spectrogrammes assez différents).

Les mots utilisés pour les expériences ont été souvent ceux du contexte téléphonique. Les résultats sont souvent nettement moins bons que pour l'audition.

---

(1) Il n'y a pas eu beaucoup d'études sur l'influence de ces distorsions sur la performance. Dans ANON (1965), l'utilisation du téléphone est acceptable. Mais l'observateur ne saura pas distinguer chaque fois ce qui est propre au locuteur et ce qui est dû à la dégradation.

Une expérience de STEVENS & al. (1968) indique les taux d'erreurs comparés suivants :

Longueur	Audition	Spectrogrammes
1 syllabe	12	32
2 syllabes	9	24
phrase	9	18

Ainsi, le taux d'erreur diminue lorsqu'augmente la durée de l'échantillon, ce qui n'est pas le cas de la reconnaissance auditive. Certains observateurs semblent obtenir généralement de meilleurs résultats que d'autres.
Les mots isolés donnent de bien meilleurs résultats[1]. YOUNG & CAMPBELL (1967), invoquent deux raisons : un mot prononcé isolément tend à être plus long et l'environnement d'un mot phonétique, modifie ses propriétés acoustiques (effets de coarticulation). En outre, étant donné que les spectrogrammes de locuteurs différents peuvent se ressembler localement, on tei à y remédier en utilisant un grand nombre de mots pour le même test de reconnaissance.

Le spectrogramme de contour a été utilisé, mais a conduit à une performance légèrement plus faible que le spectrogramme de raies. Dans le premier cas, les règles de décision comparent deux contours à l'aide d'une mesure de similarité et d'un seuil. Cette méthode a l'avantage de limiter les effets des variations du débit du locuteur (cf. SCHROEDER, 1968).

D'une façon générale, l'analyse porte plutôt sur les fréquences basses (inférieures à 3 kHZ), car les formants élevés sont tenus pour être plus stables, d'où l'utilisation fréquente d'une échelle logarithmique. L'observateur doit

---

(1) 78,4% contre 37,3% en moyenne, pour 5 locuteurs et les mots "you" et "it" (expérience de YOUNG & CAMPBELL (1967)).

faire une évaluation de la variabilité intra-locuteur ; pour cela, il dispose souvent simultanément de plusieurs références.

STEVENS & al. (1968) ont comparé l'aptitude à reconnaître les locuteurs familiers et non familiers. Elle est un peu moins bonne que dans le cas de l'audition, surtout pour des locuteurs non familiers, qui sont reconnus comme étant des familiers.

On peut faire une discussion des tâches assez semblablement au cas de l'audition. Notons simplement que les tâches de vérification (comparer deux spectrogrammes), et d'identification, ont été largement utilisées en criminologie. La question de la faillibilité de l'observateur a donné lieu à de nombreuses discussions (cf. la mystique des spectrogrammes, section I-III). Il n'a toujours pas été prouvé que cette méthode puisse atteindre le niveau de confiance caractéristique des empreintes digitales. Il semble même, au vu des remarques précédentes, que l'on doive la considérer comme inférieure à la reconnaissance par l'audition. Souvent, ce sont les mêmes locuteurs qui sont le mieux (resp. le plus mal) reconnus, à la fois dans les deux méthodes [(1)]. Ce qui suggère qu'il n'y a pas de méthode universelle, mais plutôt une pluralité de méthodes de reconnaissance avec toute la généralité que l'on peut donner à cette remarque.

III- Reconnaissance du locuteur par l'analyse automatique de spectrogrammes :

1. Introduction :

Ce type de reconnaissance est à considérer dans la mesure où il fournit une reconnaissance plus ou moins automatisée, et où il est susceptible de dépasser les performances humaines pour la capacité de mémoire et le processus d'analyse (EL CHAFEI, 1978).
Cependant, vues les limitations des deux types précédents, la reconnaissance automatique s'introduit d'elle-même.

---

(1) Mais les meilleurs observateurs ne sont pas les mêmes dans les deux cas. La question de l'entraînement est importante. Celui-ci varie en qualité et surtout en durée, suivant les auteurs.

Il existe deux types d'approches :

- la machine produit et examine des matrices (colonnes : intervalles temporels ; lignes : découpage fréquentiel) qui sont de véritables spectrogrammes discrets.

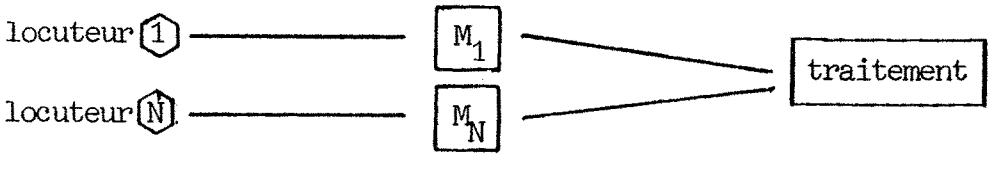
- la machine extrait les paramètres voulus et les soumet à une analyse statistique (reconnaissance du locuteur par ordinateur).

Chacune de ces deux approches a fait l'objet de plusieurs techniques. Il faut pouvoir évaluer la fiabilité de la reconnaissance par une machine. Les paramètres utilisés ici ne sont pas forcément les mêmes que ceux de l'audition. La deuxième approche sera évidemment détaillée dans la section IV et aussi à la fin de la présente.

Dans ce qui suit, le terme "système" désigne l'ensemble du matériel et du logiciel utilisé pour une reconnaissance semi-automatique ou automatique, suivant la définition de EL CHAFEI (1978).

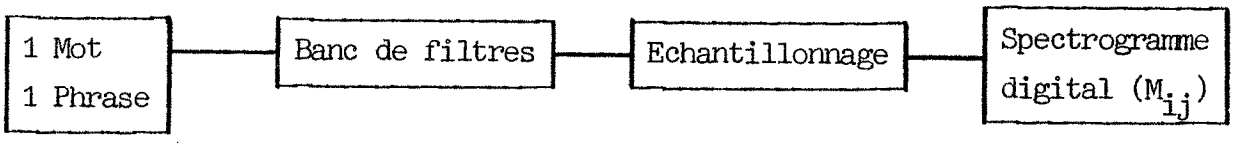
2. Techniques utilisant un spectrogramme discret :

Le schéma du traitement est le suivant :



locuteurs — matrices — comparaison des matrices 2 à 2 et décision.

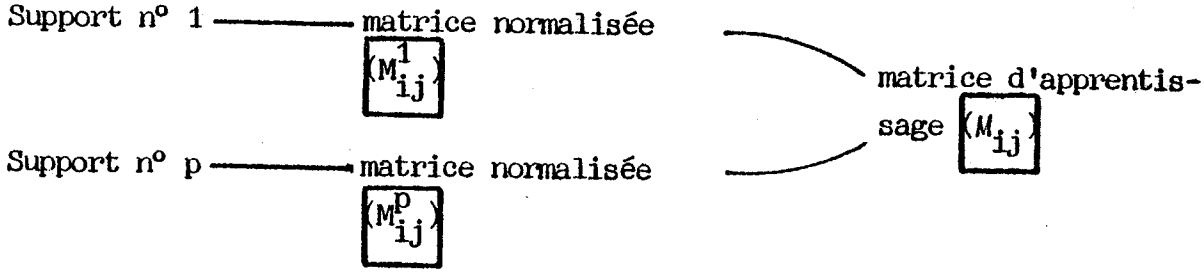
Le matériau linguistique utilisé est composé de mots isolés et de phrases, les mêmes pour l'ensemble des locuteurs. Le traitement du signal de la parole est décrit ci-dessous :



Support : Découpé en intervalles de temps $j = 1, \ldots, m$

Analyseur : Découpé en intervalles de fréquence $i = 1, \ldots, n$

Chaque support (mot, phrase, ...) produit une matrice $M = (M_{ij})$ où M_{ij} est l'amplitude quantifiée de l'énergie dans la bande i pendant l'intervalle de temps j. On dit que la matrice M comporte n x m cellules quantifiées. Avant les comparaisons, il est nécessaire de normaliser M par rapport à l'amplitude maximale. Ceci peut être fait en fixant la somme de ses éléments égale à une constante. (Cette normalisation correspond à la norme de matrice $||A||_1 = \sum_{i,j} |Aij|$). Ainsi, pour un même locuteur, on a autant de matrices que de supports et elles sont de mêmes dimensions; en les mettant bout à bout, suivant l'axe temporel, on obtient la matrice M_{ij} de référence (ou d'apprentissage) du locuteur considéré.



Lorsqu'on veut tester un locuteur, on construit la matrice de test (X_{ij}) correspondante (à partir de l'un des support ou de plusieurs), que l'on compare, en faisant correspondre les colonnes concernées, avec la matrice de référence M du locuteur désigné (cas de la Vérification) ou de chaque locuteur (cas de l'Identification). La comparaison suppose que soit clarifiée la notion de similarité entre telles matrices et la fixation d'un seuil de décision. Le seuil de décision est généralement fixé d'après une marge de tolérance sur les faux rejets.

Quelles règles de décision faut-il appliquer ? Les règles optimales (cas Bayésien), ne peuvent en général pas l'être, car on ne connaît pas les distributions probabilistes relatives à chaque locuteur, pour les supports considérés. On réalise souvent un compromis en utilisant des mesures de similarité qui soient simples, par exemple :

- la distance euclidienne entre X et M : $d_{XM} = (\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - M_{ij})^2)^{1/2}$

- le coefficient de corrélation entre X et M, lequel ne nécessite pas de normalisation préalable des matrices :

$$r_{xM} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - m_X)(M_{ij} - m_M)}{n \cdot m \cdot \sigma_X \cdot \sigma_M}$$

(avec : n . m = nombre de cellules, m_X et m moyenne des valeurs des cellules de X et de M, σ_X et σ_M écart-types des cellules de X et de M).

Les auteurs ont utilisé, outre les configurations temps-fréquence-amplitude les configurations : amplitude-fréquence (obtenues par moyenne sur tous les intervalles temporels). Le tableau 2 décrit les caractéristiques de 6 méthodes mises au point de 1963 à 1968.

PRUZAWSKY & MATHEWS (1964), ont essayé d'utiliser l'information contenue dans une seule cellule de la matrice, en calculant son F-ratio. Le résultat essentiel est que moins de 20% de cellules - celles ayant le F-ratio le plus élevé - suffisent pour autoriser un taux de reconnaissance analogue à celui obtenu par utilisation de la totalité (environ 85%).

Un problème relativement ouvert est celui de la meilleure résolution temporelle.

Tableau 2- Comparaison de techniques de reconnaissance par machine (Extrait de HECKER, 1971).

Study	*Cue Material*	*Matrix Configuration*				*Utterances Incl.*		*Recogn. Task*	*Decision Rules*	*Speakers*	*Perfor %*
		Frequency Bands	*Range kHz*	*Interval msec*	*Amplitude bits*	*Ref. Matrix*	*Test Matrix*				
Pruzansky (1963)	10 Words	17	0.2–7.0	10	10	3	1	Ident.	Cross-Correl.	10	89
Pruzansky and Mathews (1964)	10 Words	17	0.1–10.0	10	10	3	1	Ident.	Σd^2	10	93
Ramishvili (1965)	10 Words	7	0.2–10.0	50	2	10	1	Ident.	Σd^2	20	92
Li et al. (1966)	3 Phrases*	15	0.3–4.0	20	12	10+	1	Discr.	Adaptive	20	90
Glenn and Kleiner (1968)	Conson. [n]	25	1.0–3.5	–	6	10	10	Ident.	Cross-Correl.	30	93
Meeker (1967)	4 Vowels	19	0.2–8.0	40+	†	20	3	Discr.	Σd^2‡	11	95

* Used only first 500 msec of each utterance.
† Used relative frequencies of occurrence of three spectral slopes.
‡ Decision threshold: 1% false rejection.

IV - Les modes opératoires :

1. Tests contemporain et non contemporain :

Dans un système de reconnaissance des formes, où les classes sont définies a priori (on connaît leur nom à l'avance), on distingue très généralement deux phases :

a) Phase d'apprentissage - ou d'élaboration des classes : acquisition et stockage des vecteurs - formes références appartenant chacune à l'une des classes. On construit ici "l'expérience du système".

b) Phase de test (phase opérationnelle) : comparaison et classement de la forme à classer avec les formes mises en mémoire.

Dans notre application, les classes sont les locuteurs, les formes de référence et celles à classer sont respectivement, les échantillons de parole de référence, et du locuteur à tester. Etant donné le manque de commodité intrinsèque de la phase test (faire physiquement entrer sur le système un nouveau locuteur à chaque nouveau test), on prend souvent pour échantillon de test, l'un des échantillons de référence, que l'on ôte de l'ensemble des échantillons de référence précédents.

Deux cas sont alors à considérer :

- la date d'enregistrement de l'échantillon test est antérieure à la date du dernier échantillon de référence ; nous parlerons alors de test de type contemporain ;
- dans le cas contraire, il s'agit pour nous d'un test de type non contemporain.

Evidemment, la pratique du test contemporain convient pour les seules études en laboratoire et possède une limitation intrinsèque : il suppose que les caractéristiques du locuteur n'évoluent pas au cours du temps. C'est une simplification, l'étude doit être complétée par des tests non contemporains, ce qui n'est malheureusement pas toujours fait.

2. Systèmes en ligne :

1) Un système entièrement automatisé peut opérer par l'intermédiaire de connections avec un téléphone. Mais on est alors en présence de nombre de difficultés susceptibles de dégrader la performance du système.

Suivant ROSENBERG (1976), on peut citer :

a) D'abord les conditions acoustiques pendant la session d'enregistrement, à savoir :
   - le bruit de fond, de perturbations diverses dues à la présence de l'équipement téléphonique (manipulations, etc.).
   - également les conditions de transmission : modification du signal sur les lignes branchées. Le téléphone a une bande passante 300-3400 Hz. Il faut aussi compter en général sur des variations et des distorsions d'amplitude et de phase.

b) Ensuite des difficultés "comportementales" : comment prendre en compte, avec des phrases fixes, toutes les variations de la voix ? Une solution pourrait consister en une mise à jour pure et simple des échantillons de référence. On ne peut négliger cette difficulté si l'on considère la performance en pratiquant des enregistrements distribués dans le temps (LI & al., 1968), puisqu'il s'agit d'incorporer dans les références la plus grande partie possible de variations naturelles de la voix d'une même personne.

2) Dans le cas de non-utilisation du téléphone, on a recours à des enregistrements en studio, qui améliorent la qualité de la prise de son, mais ne changent rien à la discussion qui suit .

3) Les avantages de l'utilisation du téléphone concernent de toute évidence :
   - la facilité d'obtention des enregistrement-tests (la personne ne se déplace pas, appelle quand elle veut),
   - la généralisation des récepteurs téléphoniques confère une souplesse d'utilisation inégalable lors de la phase de test (le récepteur en bout de ligne peut commander l'accès à un quelconque système),
   - dans le cas de la vérification, la personne peut donner ses numéros de code par la voix de façons variées : son nom, un numéro de code composé au cadran ou frappe au clavier...

4) Contexte d'utilisation

Un système en ligne convient a priori pour de petites populations. En effet, l'utilisation du téléphone requiert une exécution proche du temps réel, mais le temps de réponse risque d'augmenter fortement avec un grand nombre d'utilisateurs. Si la finalité d'un système de reconnaissance des locuteurs est d'être inséré dans le cadre de travail des personnes, le choix d'un système en ligne est évidemment adapté.

3. Systèmes hors ligne :

1) Ils présentent l'avantage de pouvoir travailler avec des capteurs de qualité supérieure (microphone, magnétophone, bande de type professionnel) et justifient alors l'utilisation d'un studio pour tous les enregistrements.

2) En contrepartie, le système présente moins de souplesse d'utilisation : chaque personne doit se déplacer pour chaque session d'enregistrement, ce qui peut introduire des phénomènes perturbateurs (voir les précautions psychologiques prises par COLLINS (1977).

4. Résumé synoptique : Voir le tableau 3.

V - Reconnaissance du locuteur par ordinateur :

1. La chaîne de traitement :

Comme dans les problèmes classiques de reconnaissance des formes, la reconnaissance du locuteur fait appel à plusieurs phases en lesquelles elle se décompose. Les deux plus importantes sont la mesure des paramètres, et la classification, mais on a besoin des traitements successifs suivants :

a) choisir a priori les paramètres pertinents ou caractéristiques (1),

b) extraire l'information pertinente : obtenir la valeur de ces paramètres sur l'échantillon de population étudié,

c) sélectionner les meilleurs paramètres parmi les précédents, ou extraire de nouveaux paramètres à partir des précédents,

d) classer les vecteurs obtenus en groupes représentant les locuteurs,

e) décider de l'appartenance du vecteur observé dans la phase de test, à un groupe déterminé. Ici intervient la distinction entre Vérification et Identification. Cette dernière phase est absente dans le cas de la Discrimination.

Aucune de ces phases n'est indépendante des autres. Au contraire, les correspondances amont/aval :

échantillons $\longleftrightarrow$ modèle $\longleftrightarrow$ décision

sont en fait une inter-relation de ces parties : la performance du système global dépend de l'adéquation mutuelle de ces phases. Suivant les auteurs et les études, l'une de ces phases est privilégiée. En outre, comme le remarque WOLF (thèse, 1969)

---

(1) Le mot "caractéristique" sera synonyme de paramètre pertinent, descripteur, caractère. Une information pertinente est ici une information qui aide à la distinction des classes (ou discrimination). Nous éviterons le mot "trait", qui reçoit d'autres acceptions en phonétique.

l'effort consacré à la sélection des paramètres ne semble pas être utile à la conception des procédures de classification, ou même à une meilleure compréhension de la production de la parole. Dans un système global idéal, il faut inter-relier ces phases de façon optimale, ce qui, d'un point de vue théorique, n'est pas un problème résolu.

Cependant, remarquons que l'on ne peut compenser un ensemble de paramètres déficients par un classificateur plus sophistiqué. Il semble que la solution théorique soit la suivante :

1) un ensemble de N-paramètres parfaitement pertinents, c'est-à-dire discriminant parfaitement les formes-locuteurs (on peut extraire N-caractéristiques rendant les nuages disjoints dans $\mathbb{R}^N$) ,

2) un classificateur trivial (un ensemble de $\frac{M}{2}(M-1)$ hyperplans séparateurs, séparant les M classes deux à deux).

Pour se rapprocher de cette situation (paramètres adéquats + classification économique), notre effort doit donc porter sur le point 1. Ceci justifie assez l'orientation donnée au début de la section IV, sur le choix de paramètres reliés aux différences inter-individuelles[(1)]. Chaque faiblesse de l'ensemble des paramètres utilisés (qui traduit une faiblesse de nos connaissances), doit être complétée par une complexité du classificateur accrue: laquelle ne peut être augmentée sans qu'elle ne devienne à un certain point inutile ou nuisible ; elle estime les propriétés d'ordre supérieur des distributions sous-jacentes, et requiert alors des données en plus grand nombre, pour qu'elle soit statistiquement consistante (SEBESTYEN, 1962). La complexité de la classification doit être en accord avec le pouvoir discriminant des paramètres, c'est-à-dire ne pas excéder leur pouvoir de résolution de la population étudiée. Observons que la complexité algorithmique de la classification dépend de la complexité des lois de probabilités des données, et du nombre de classes.

---

(1) La possibilité de les extraire d'une façon entièrement automatique, n'a pas été systématiquement étudiée. Chaque mesure est associée à un segment phonétique, qui lui sert de support.

L'exemple simple suivant l'illustre : les schémas de classification optimale linéaire sont mis en défaut dès que les classes ne peuvent être décrites par des régions à la fois disjointes, simplement connexes et convexes.

1) Choix des caractéristiques

Le moyen courant de retenir les informations pertinentes, est d'utiliser des descripteurs des variations inter-individuelles, fondés sur les remarques faites en IV-I , lesquels véhiculent la dite information, simultanément à d'autres types d'informations, qu'il est difficile d'écarter. Le choix des paramètres détermine, ou est déterminé par les appareils de prétraitement à utiliser. Il correspond à des limitations techniques, mais aussi à une limitation du coût global.
Ce choix se fait au vu d'études phonétiques statistiques préalables.

2) Sélection/Extraction des caractéristiques

Le but est de réduire la quantité d'information en diminuant le nombre de descripteurs, sans perdre de vue l'information pertinente. On sera conduit à retenir les meilleures caractéristiques, ou à éliminer les plus mauvaises, au sens d'un critère capable d'ordonner totalement l'ensemble des caractéristiques. On préfère traiter une information redondante en conservant plus d'informations.

3) Classification des données (phase apprentissage)

Il s'agit de mettre en oeuvre des procédures qui permettent, à l'aide des caractéristiques précédentes, de définir les classes de locuteurs, c'est-à-dire de déterminer à quelle classe appartient avec le plus de vraisemblance, telle observation, et de les séparer, en définissant des moyens d'individualiser ces classes.
Si l'on est en présence de classes se séparant bien, il n'y a pas d'assurance qu'elles seraient aussi bien séparées ni placées respectivement de la même façon (en supposant définie une métrique induite), avec un nombre de données plus élevé. Pour obtenir une bonne valeur du pouvoir discriminant des paramètres, il faut en principe s'assurer de suffisamment de données.

Lorsqu'on utilise une distance euclidienne, on peut considérer que chaque classe est représentée par un seul point : son centre. Cette façon d'analyser la situation ne tire pas avantage de toute l'information contenue dans la forme de la classe entière. Si l'on ne fait pas cette hypothèse, il faut recourir à des règles autorisant un découpage de l'espace des observations en classes de locuteurs (cf. WELCH & WIMPRESS, 1961).
Prenons l'exemple de deux classes (locuteurs) à séparer à l'aide de deux caractéristiques. Schématisons la position des deux nuages de points, par la figure 2 suivante :

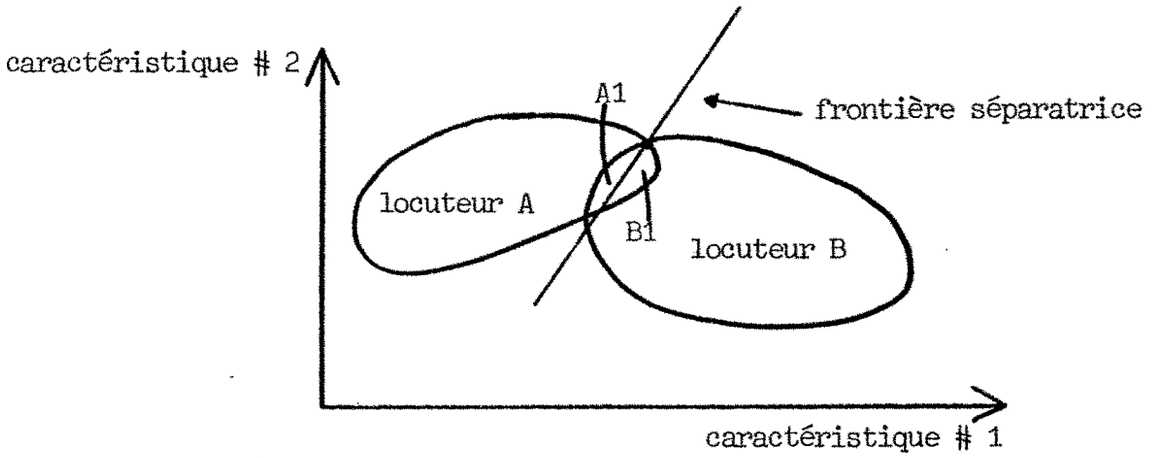


Figure 2. Dissection de l'espace des observations

On cherche habituellement à égaliser les aires A1 et B1, pour tracer la frontière séparatrice. L'aide de A1 (resp. B1), représente les cas d'erreur où B peut être reconnu à tort comme étant A (respectivement A ou B). Cette dissection de l'espace des observations conduit à la recherche d'un meilleur arbre de décision (méthode par dissections successives).

4) Décision (phase de test)

Il s'agit de décider à quelle classe appartient vraisemblablement l'échantillon courant. Ceci ne peut guère se faire par comparaison avec les techniques de décision par l'audition, car celles-ci sont, dans une très large mesure, inconnues. La justesse de la décision repose sur la qualité de toutes les phases précédentes, en particulier : 2 et 3. La décision peut supposer la définition d'une marge de tolérance s'exprimant généralement, comme indiqué plus haut, (I.I.2.6.) en taux de fausses admissions.

Dans le cas où les échantillons de référence se réduisent à un élément par locuteur, on a un problème simple :

- il est facile de mesurer la distance de l'échantillon de test à chacun des autres [1],
- on fait l'hypothèse que cette distance est inversement proportionnelle à la probabilité que les deux échantillons en présence proviennent du même locuteur.

Il est facile de décider en utilisant une règle de plus proche voisin, ou à l'aide d'un seuil. Il y a plusieurs sortes de seuils (rayons d'une boule centrée en le point-test). Les méthodes à seuil variable permettent de construire les courbes COR de la théorie du signal.

Indiquons que la probabilité de faux rejet varie avec le seuil de décision, ce qui offre un moyen de la minimiser respectivement à ce seuil. Ceci peut être fait aux dépens de la probabilité de fausse acceptation. A chaque application, doit pouvoir correspondre un équilibre à définir. Souvent, les auteurs ont traité le cas où ces deux probabilités restent égales.
La décision est souvent fondée sur la densité de probabilité (ddp) de chaque locuteur, qui intègre tous les paramètres à la fois :
on choisit le locuteur dont la ddp est la plus grande en le point d'observation.

---

(1) Dans le cas de l'indépendance de paramètres, et d'une contribution égale à la discrimination (deux hypothèses plus ou moins vérifiées suivant les cas), cette distance est la distance euclidienne entre les deux points. Celle-ci a généralement donné de bons résultats, qui ont été comparés aux résultats de la reconnaissance par l'audition.

En cas de plusieurs ddp maximum et égales, ou très proches, on peut rechercher, si c'est possible, un plus grand nombre d'observations.
A un ensemble de paramètres de même cardinal, et pour des règles de décision d'un type donné, le taux de reconnaissance est influencé par divers facteurs :
- le nombre d'échantillons d'apprentissage, et aussi de test (combien de tests sont nécessaires et suffisants pour obtenir un taux de reconnaissance stable, c'est-à-dire pour atteindre la performance réelle du système ?),
- la taille et l'homogénéité de l'ensemble des locuteurs.

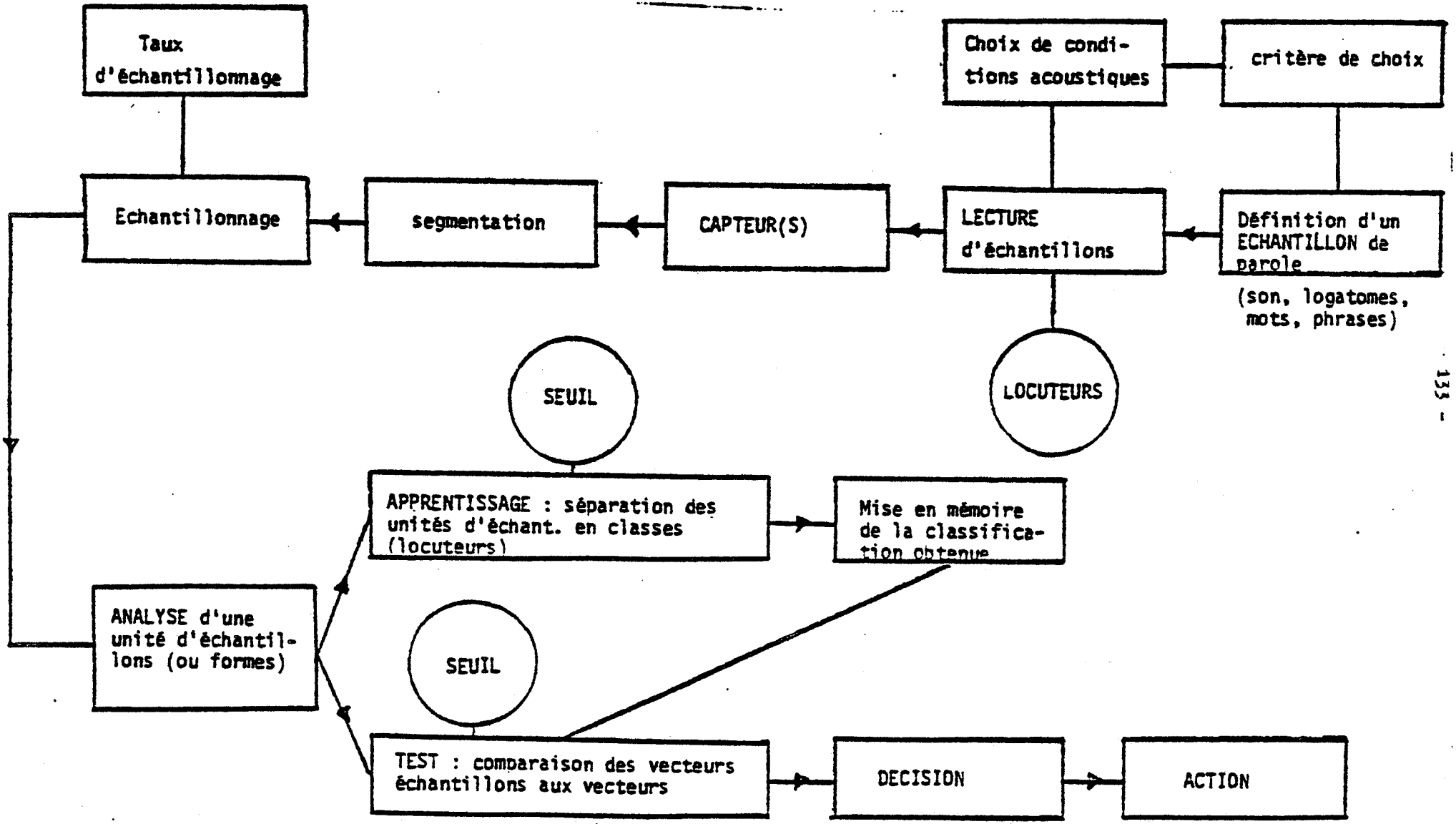
Les règles de décision associées aux paramètres de type spectral, sont généralement plus complexes que celles pour les paramètres de type temporel. Plusieurs auteurs ont obtenu des résultats meilleurs dans ce dernier cas, surtout en cas de dégradation du signal (environnement bruité, limitation de bande : téléphone, ...). C'est parce que les paramètres spectraux sont très sensibles au manque de précision sur les fréquences des formants.

5) Visualisation

C'est un aspect important du traitement et les auteurs se sont attachés à fournir des résultats sous une forme immédiatement perceptible, étant donné que le nombre de paramètres sélectionnés ou extraits dépasse généralement deux.

On trouvera dans le schéma 1 , le détail de l'ensemble des traitements à opérer. Les différentes parties de la chaîne de traitement, et en particulier la sélection des paramètres pertinents, ne seront pas davantage discutées dans cette section.

SCHEMA 1 - LA CHAINE DE TRAITEMENT EN RECONNAISSANCE DU LOCUTEUR PAR ORDINATEUR.



2. Les problèmes fondamentaux :

Ils sont communs à la Vérification, et à l'Identification. Ils justifient donc une étude conjointe préalable de ces deux contextes, pourtant différents par leur cadre d'application.

1) Problèmes théoriques

Comme on l'a vu précédemment, dans un système de reconnaissance de formes, chaque élément de la chaîne n'est pas à traiter séparément, mais de façon compatible avec les autres éléments.

a) Taux d'échantillonnage adapté :

Afin d'avoir une information la plus exhaustive possible (en fréquence et en temps), le taux d'échantillonnage choisi doit être adapté à la caractéristique considérée. Par exemple, l'extraction du troisième formant vocalique requiert un taux d'échantillonnage élevé (supérieur à 7 kHz).

b) Modélisation de locuteur :

"Le problème de la construction des modèles est de premier ordre en reconnaissance du locuteur" (GUBRINOWICZ, 1971).

D'une façon générale, "La reconnaissance est le résultat d'une association de la forme à reconnaître avec un modèle prototypique représentatif d'une classe de formes, sur la base d'un critère de similarité" (DEVIJVER, 1971).

En reconnaissance des formes, le modèle associé à la notion de forme est un point de l'espace vectoriel $\mathbb{R}^n$, où n est le nombre de caractéristiques. Comme dans toutes les applications, les coordonnées de ce vecteur résultent de mesures et sont donc entachées d'erreurs. Ici s'ajoute cependant la variabilité permanente du signal de parole, ce qui empêche de considérer avec certitude les points obtenus. Les hypothèses probabilistes et les méthodes statistiques ne permettent pas de s'affranchir de ce "flou" sur les données. Habituellement, on cherche à compenser ce phénomène, que l'on ne sait pas prendre en compte, en moyennant la valeur sur plusieurs enregistrements disjoints et répartis dans le temps. Il faut voir cependant que cela sous-tend des hypothèses probabilistes réductrices d'information : on accorde plus d'importance à la moyenne qu'aux observations.

c) Mesure de la distance inter-locuteurs :

On a également besoin de porter un jugement quant à la ressemblance de deux formes à l'aide d'un critère qui pourra être simplement une distance sur $\mathbb{R}^n$.

On sait que le choix d'une métrique dépend de l'application considérée. Trois critères guident le choix :

- complexité des calculs,
- forme analytique simple,
- fiabilité dans l'évaluation des caractéristiques.

La distance euclidienne entre deux vecteurs échantillons e_k et e_ℓ appartenant à $\mathbb{R}^n$: $d(e_k,e_\ell) = \left(\sum_{i=1}^{n} |e_k^i - e_\ell^i|^2 \right)^{1/2}$
offre une simplicité d'analyse et de calcul.

d) La séparation des données en données d'apprentissage et données de test :

Nous sommes en présence de trois sortes d'échantillons (répétitions) pour chaque locuteur :

1 - l'échantillon d'apprentissage $\mathcal{A}$: il sert à définir les caractères discriminants,
2 - l'échantillon-test $\mathcal{T}$: il sert à vérifier a posteriori que la discrimination précédente est satisfaite, ce qui conduit à un taux de bonne classification,
3 - l'échantillon-test anonyme $\mathcal{A}$: il concerne les répétitions dont l'auteur est inconnu, et que l'on veut identifier ou vérifier.

Puisque ce dernier échantillon est souvent absent des études en laboratoire, lesquelles portent sur des tests contemporains, nous aurons à partager les données disponibles en les deux échantillons 1 et 2. Le taux de bonne classification a posteriori obtenu, est fonction de ce partage. Nous cherchons donc en quelque sorte un partage "optimal".
Si ROMEDER (1973) conseille un prélèvement de 20 à 30% des données pour les affecter à l'échantillon-test, d'autres auteurs utilisent des méthodes variées. Notant K le nombre total de répétitions disponibles (ensemble $\mathcal{K}$), on distingue :

1 - la méthode R ("R-method") : utiliser les mêmes données pour et (=). Ceci introduit un biais extrêmement favorable à l'obtention d'un taux élevé.

2 - la méthode H ("H-method") : utiliser des informations indépendantes pour et (K est partitionné en a et). Cette méthode a été utilisée par HEIGHLEIRAN, 1962 et KANAL, 1971.

3 - la méthode "d'exclusion d'un élément" ("leave one out method ") : est réduit à un élément k, les autres éléments de forment , et k prend successivement toutes les valeurs 1, ..., K. Cette méthode nécessite donc K sessions (K apprentissages et K tests). Elle utilise toutes les données pour le test et permet en fait le meilleur usage de l'information disponible (utilisée par GRENIER (1977) et FASOLO-MIAN (1978) pour le présent problème, et GLICH (1978)).

Ces méthodes doivent donc être considérées comme des méthodes d'estimation de l'erreur de probabilité.
Nous utiliserons la méthode 2 en variant les ensembles et , ce qui nous rapproche de la méthode 3. Celle-ci n'a pas été mise en oeuvre systématiquement pour des raisons de coût.

e) But recherché :

On peut formuler deux classes de problèmes de vérification. Le choix des algorithmes de décision va être guidé par les contraintes liées à des finalités différentes :

1. La vérification (admission ou rejet) doit se faire très rapidement, si possible en temps réel. On aboutit alors à un seuil de confiance de $\alpha\%$.

2. La vérification doit être réalisée à un niveau de confiance très élevé ($> \alpha\ \%$) et dans ce cas, la rapidité de l'opération n'est pas l'objectif prioritaire. La durée de l'opération n'a pas d'importance. Le système doit, ici, s'adapter au locuteur et organiser un choix de caractéristiques. Une optimisation de ce programme conduit à la minimisation du taux d'erreur.

f) Evaluation des performances de la machine :

Celle-ci peut se faire en fonction du taux d'erreur de classification et de la vitesse de traitement.

g) Détermination du nombre d'échantillons par locuteur :

Il n'existe malheureusement pas de développements théoriques permettant d'évaluer la quantité d'échantillons à considérer par classe, pour un seuil de confiance fixé. Intuitivement, on pourrait penser que les classes sont d'autant mieux spécifiées (au sens statistique du terme), que la taille des échantillons est grande. Effectivement, EL CHAFEI (1978) constate une décroissance asymptotique du taux d'erreur avec le nombre d'échantillons. En fait, on ne peut généraliser un tel résultat.

h) La dépendance du contexte :

Elle est présente dans la mesure où les paramètres sont dépendants du contexte, et où la segmentation et l'isolement des séquences phonétiques ne sont pas rendues automatiques indépendamment du contexte.

2) Problèmes pratiques

a) La qualité des conditions d'enregistrement :

Certains paramètres d'analyse dépendent directement de la qualité de l'enregistrement. Pour fixer les idées (Table Ronde de Padoue, 1978) :

- la qualité "orthophonique": conditions idéales pour la prise de son et l'enregistrement,
- la qualité téléphonique : exemple type de signal de parole dégradé par rapport au précédent.

b) La non-reproductibilité de la parole :

Il s'agit là de la variation temporelle de la voix en fonction de très nombreux facteurs : heure de la journée, fatigue, état émotionnel. Le problème dual de celui-ci est la fiabilité de la voix : quelle confiance peut-on accorder aux références, lors d'un usage ultérieur ? (déjà soulevé ci-dessus).

c) Nécessité d'un alignement temporel, d'une segmentation et d'une normalisation :

Comparer deux échantillons temporels signifie d'abord faire coïncider leurs instants initiaux, ce qui n'est pas toujours évident et peut conduire à des choix arbitraires. Ensuite, vient l'étape de segmentation(1), d'où des séquences retenues. La normalisation est rendue nécessaire pour tenir compte de débits différents. Cette dernière étape ne peut se limiter à une simple homothétie globale, mais doit être opérée segment par segment.

d) La constitution d'une banque de données préliminaires :

Il s'agit d'une opération longue et qui conditionne la suite de l'étude. Elle soulève des problèmes spécifiques : il faut théoriquement se limiter à un échantillon de personnes provenant d'un même milieu linguistique, dialectal et socio-culturel, afin de ne pas introduire de variations inter-personnes exogènes à notre étude. Ces contraintes sont, la plupart du temps, trop fortes pour être appliquées.

3) Problèmes techniques

Il faut tenir compte du gros volume de données à traiter. Ceci est fréquemment le cas en reconnaissance des formes, et pose des problèmes pour aboutir au temps réel.

---

(1) Car la parole n'est pas un phénomène stationnaire.

III - DESCRIPTION SOMMAIRE D'UN SYSTEME DE RECONNAISSANCE DES LOCUTEURS

1) Les entrées du système

Si le traitement se fait en ligne, elles ont lieu à partir d'un poste de saisie, qui peut être un microphone ou un téléphone. Dans le cas d'un traitement hors ligne, on utilise une bande analogique pré-enregistrée.

2) Le prétraitement

La phase préliminaire consistant à digitaliser le signal, est placée en aval de l'acquisition des données précédentes.

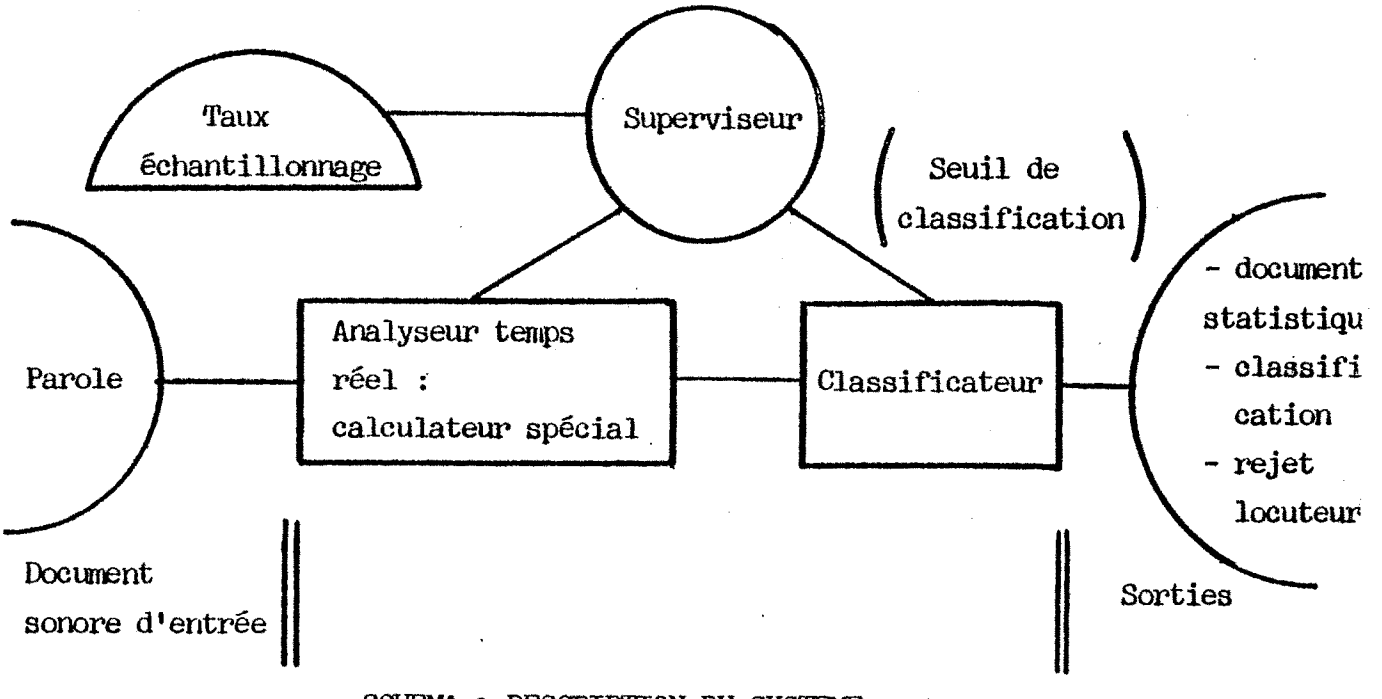
3) Le traitement

On cherche ici à combiner :

- l'analyse en temps réel du signal de parole,
- les techniques de reconnaissance des formes.

BÜNGE (1977), constate qu'il est difficile de trouver un compromis optimal entre un raffinement de l'analyse et la sophistication des procédures de classification et de décision.

Les parties principales du système de traitement sont décrites dans le schéma 2.



SCHEMA 2. DESCRIPTION DU SYSTEME

L'analyseur effectue le calcul des valeurs des paramètres à partir du signal analogique, soit à partir du signal temporel, soit à partir de sa transformée de Fourier : voir BELLISSANT (1978), pour une discussion sur l'utilité relative de ces deux types d'approche.

Le programme superviseur a la charge de fixer les options de travail et aussi d'orienter éventuellement l'analyse et le choix de la procédure de classification en fonction du locuteur, du support linguistique utilisé.

4) Les sorties

Elles consistent en données statistiques décrivant les voix des locuteurs, un taux de reconnaissance global, un taux d'erreur global, un taux de rejet, si l'option éventuelle de "rejet du locuteur" est incluse pour des raisons d'insuffisance de capacité discriminante des paramètres. Une matrice de confusion entre locuteurs, visualise les résultats obtenus par le système pour chacun des locuteurs.

Dans le cas d'un système de vérification en ligne, une réponse est émise en direction de l'utilisateur ; elle fait partie du protocole de dialogue entre l'Unité de Vérification et le locuteur.

IV

CONCEPTION D'UN SYSTEME OPERATIONNEL

"Système : théâtre de la
mythologie informatique :
tous les dieux s'y rencontrent".

E. GIRARD, Bulletin de l'IRIA, 55, 45.

INTRODUCTION : Les facteurs de l'étude et les hypothèses implicites.

Nous avons vu (I) que, placée dans un contexte d'application quelconque, la Vérification du locuteur consiste à décider l'acceptation ou le refus de l'accès d'une personne à un système donné, sur la base de :

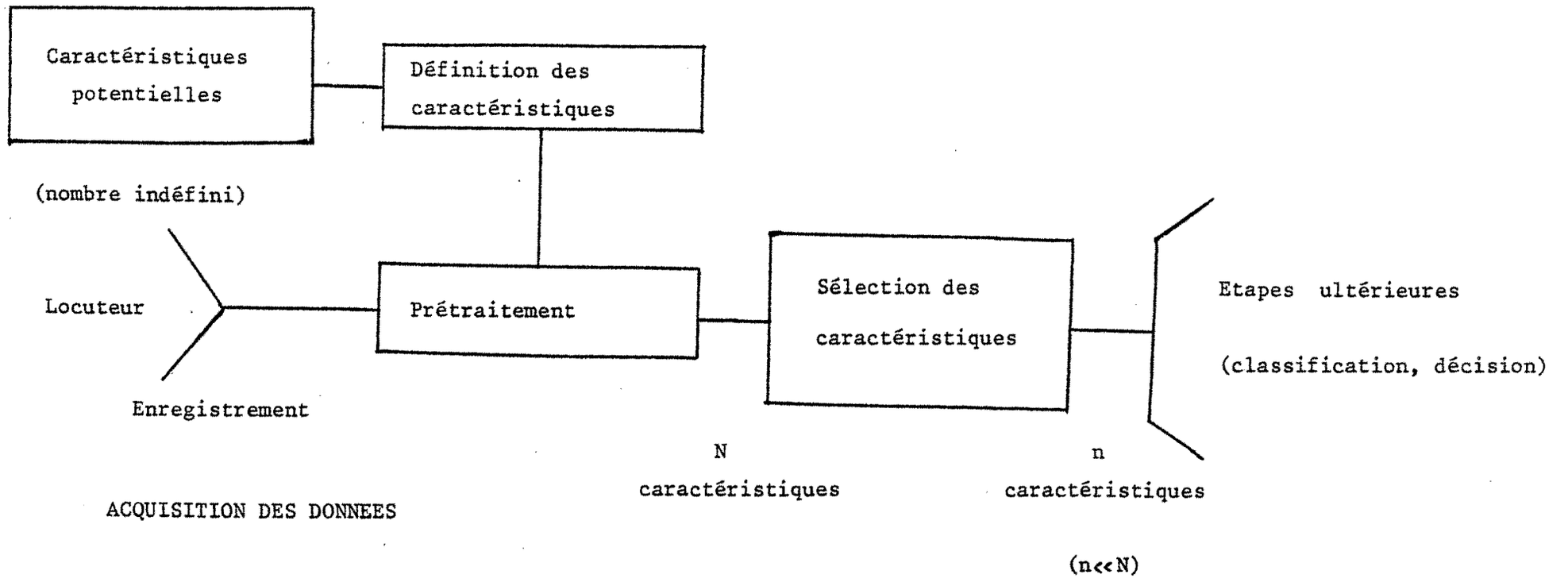
- la déclaration - par elle-même - de son identité (on suppose au départ qu'elle est exacte) ;

- l'analyse de son signal vocal.

Quant à l'Identification, elle relève d'une recherche, parmi une population, d'une référence qui soit suffisamment proche de l'échantillon courant pour pouvoir déclarer l'égalité des deux locuteurs en question.

Dans chacun des cas, la méthode est la suivante : on compare le signal vocal courant au modèle témoin correspondant à la personne dont l'identité est celle proclamée ou recherchée (principe de reconnaissance par comparaison d'une forme à une classe de formes). Ainsi, la décision est fondée sur un apprentissage supervisé du système, préalable à toute opération de reconnaissance d'un certain nombre de formes prototypiques. On s'intéresse d'abord aux moyens de réaliser simplement un tel apprentissage, ce qui conduit à la recherche de bonnes caractéristiques. Il est évident que les diverses parties d'un système global de reconnaissance ne sont pas indépendantes, mais présentent des inter-relations amont/aval. Par exemple, on ne peut pas toujours compenser une déficience des caractéristiques vis-à-vis du problème étudié, par une complexité accrue du classificateur.

FIGURE 1 - LA CHAINE DE RECONNAISSANCE



Le modèle de reconnaissance fait appel à trois étapes essentielles :

- la définition des caractéristiques et l'acquisition des données (prétraitement),
- la sélection des caractéristiques,
- la classification supervisée pour l'apprentissage.

PREMIERE PARTIE : ETUDES PREPARATOIRES

I - METHODE DE CHOIX A PRIORI DES PARAMETRES

1 - Définition d'un paramètre :

Nous nommerons paramètre tout trait idiolectal[1] repérable dans le signal de parole, c'est-à-dire tout descripteur des différences individuelles entre patrimoines vocaux des locuteurs. Classiquement, il y a les paramètres qui portent:
- sur une phrase entière : paramètres supra-segmentaux (intonations, accent),
- sur des sons (réalisation de phonèmes) paramètres segmentaux (formants...),
- sur des éléments infra-segmentaux (tenue des sons,...).

2 - Les qualités d'un paramètre :

WOLF (1972), a mis en relief une liste de critères idéaux a priori, que devraient posséder les paramètres. Nous offrons ici un point de vue plus unifié de ces critères.

2.1. La pertinence (ou efficacité) sera définie par l'aptitude à retenir l'information relative au locuteur lui-même. Etant donnée la présence de variabilités inter-locuteurs et intra-locuteurs, les paramètres utilisés devraient s'appuyer sur les sources de variabilité inter-locuteurs. Plus précisément, elles doivent recéler une variabilité intra-locuteur faible par rapport à une variabilité inter-locuteur forte (les mesures de chacune des deux variabilités n'ayant de sens que l'une par rapport à l'autre, ou bien la même variabilité d'un locuteur par rapport à l'autre [2]). La variabilité intra-locuteurs peut être facilement réduite globalement en demandant la coopérativité des locuteurs (voir I - I), ce qui oblige à se placer dans le cadre de la Vérification. Un autre moyen (couramment utilisé en Vérification), est de fixer une fois pour toutes le texte du message prononcé. Comme l'indique ATAL (1976), il est toujours intéressant de savoir si une telle réduction de variabilité est absolument nécessaire pour l'obtention de bons résultats : pour des raisons pratiques évidentes, on aimerait

---

(1) Idiolecte : ensemble des réalisations propres à un locuteur.

(2) Les différences de variabilité d'un locuteur à l'autre, n'ont pas été étudiées, et sont un aspect de la modélisation des locuteurs. Nous les utiliserons sous une forme descriptive au § III.2.

pouvoir autoriser le changement du message à prononcer (l'auditeur qui écoute les locuteurs n'a pas besoin d'un même support linguistique pour les distinguer). Les auteurs ont en fait décrit jusqu'ici des méthodes adaptées aux deux cas : dépendance et indépendance du contexte, par le choix de paramètres et de méthodes appropriées (par exemple : FURUI, 1972 ; ATAL, 1976 ; LI & HUGUES, 1974, ...).

On ne peut guère compenser un manque de pertinence d'un paramètre par un classificateur plus sophistiqué. En outre, seuls des paramètres pertinents peuvent permettre la mise au point d'un bon modèle du locuteur.
La mesure de cette efficacité nécessite évidemment la définition d'un critère. La mesure idéale est la probabilité d'erreur de reconnaissance d'un locuteur (ou taux d'erreur). Cette mesure est en fait ambiguë, car deux règles de décision différentes donnent deux taux différents. Ainsi, et on voit combien sont interreliées les parties de la chaîne de traitement, l'efficacité d'un paramètre ne peut être déterminée indépendamment des règles de classification (ce qui rend délicate la tâche de comparaison des méthodes fondées sur des paramètres distincts Pour résoudre ce point (1), on peut s'en tenir à des critères évaluant un simple pouvoir discriminateur entre deux locuteurs arbitraires. Alors la nature des règles de décision ne joue pas. Nous avons vu en I-I-2, que la Discrimination était une tâche de base en reconnaissance du locuteur, et qu'elle se diversifiait en Vérification et Identification. On retiendra que la probabilité d'erreur est une mesure idéale de pertinence, à règle de décision fixée.

Pour terminer, disons que les avantages de paramètres efficaces sont évidents : volume de données traitées plus faible, procédures de classification plus rapides et moins complexes, obtention de taux d'erreur plus faible.

---

(1) Le recours à une règle de décision optimale n'est pas réalisable : il faudrait ranger toutes les règles de décision et tous les paramètres possibles (commentaire de ATAL, 1976).

2.2. La disponibilité :

On nomme ainsi l'apparition fréquente et sans condition préalable, d'un paramètre dans le signal de parole analogique (éventuellement le signal transformé, pour d'autres paramètres). Le choix de la phrase support peut donc être à considérer. Ceci est également lié à une mesure facile du paramètre (adaptation au téléphone, microphone, ...). La disponibilité suppose donc une adéquation :

phrase support - paramètre - matériel,

sensée améliorer la performance du système de reconnaissance global, notamment par des simplifications dans la conception du classificateur.

2.3. La résistance :

Elle exprime une non-sensibilité à une modification de l'environnement physique ou émotionnel : modification volontaire de la voix, humeur, fatigue... et avant tout, une faible dégradation dans le temps (stabilité temporelle pour une personne adulte).

On peut évidemment considérer des estimateurs de tendance, de dispersion de ce type de variation de la voix, mais souvent, on ne saura pas établir la carte de ces variations.

2.4. L'interprétabilité :

Etablissement des correspondances physiologiques. Remarquons que les mesures physiologiques ou anthropométriques des parties de l'appareil phonatoire conduiraient à une meilleure compréhension de la variabilité de la voix en général, mais elles ne sont pas disponibles. Force est de recourir à une analyse directe du signal vocal. On ne s'intéressera pas ici aux sources de variabilité intra- et inter-locuteurs (cf. HECKER(1971) et STEVENS(1971). Contrairement à ce qui se fait assez souvent, nous donnerons une grande importance à l'interprétabilité. On peut considérer comme inutiles les caractéristiques qui n'ont pas de signification intrinsèque, indépendamment des critères mathématiques sur lesquels elles se fonder

2.5. La polyvalence :

Aptitude des paramètres à être traités par des techniques diverses. Nous verrons au III, un exemple de paramètre non polyvalent. Mais la méthode de traitement devra être adaptée au paramètre, par exemple, à sa dispersion. On peut également ajouter l'indépendance d'un paramètre vis-à-vis des langues (contexte linguistique), et l'indépendance vis-à-vis du contexte phonétique.

Dans beaucoup d'applications, cependant on pourra relaxer une ou plusieurs de ces contraintes.

3 - Choix de paramètres temporels et acquisition des données :

Tout d'abord, nous justifierons le choix de la catégorie de paramètres utilisés : ils sont en effet en général de type temporel. Les paramètres temporels ont été :

- peu étudiés,
- sont faciles à extraire,
- peuvent être extraits en temps réel.

Ces points les opposent aux paramètres fréquentiels. En outre, pour le fraçais, ils ont été systématiquement étudiés à l'Institut de Phonétique de Grenoble : ces études préliminaires ont montré que la fréquence fondamentale moyenne était relativement stable dans les mêmes conditions d'une année sur l'autre, que l'écart-type sur cette fréquence pouvait aider à caractériser le locuteur, que la durée des pauses séparait bien les sexes (le débit peut être corrélé au nombre de pauses, à leurs durées et à la vitesse d'élocution des groupes de phonation(1); les pauses dépendent en partie de la respiration phonatoire, outre leur action sur l'intelligibilité du message parlé(2)). En outre, un appareillage spécialisé y a déjà été développé.

---

(1) cf. BOË & al. (1975)

(2) cf. SAINT-BONNET & BOË (1977). Pour l'essentiel des recherches dans ce domaine des faits prosodiques, on se reportera à l'ouvrage "Recherches sur la Prosodie du Français, Institut de Phonétique de Grenoble, Publication de l'Université des Langues et Lettres de Grenoble (1979).

3.2. Liste des paramètres retenus a priori :

Les variables fondamentales sont constituées par les catégories supra-segmentales et segmentales suivantes :

- fréquence fondamentale moyenne (identificateur FM), taux de voisement (TV) et débit (DB) sur le texte, et chacune des trois premières phrases du texte. (On détecte le voisement par le nombre de passages par zéro du signal, dans une bande de fréquence : voir ABRY & BOË (1975)).

On sait que la réalisation des occlusions non voisées se décompose en plusieurs phases :

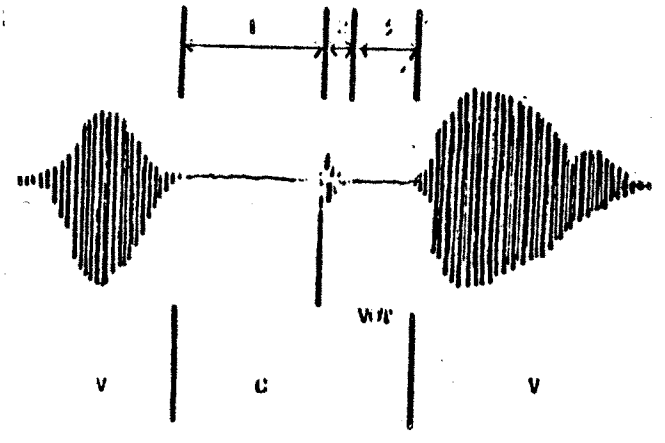


Figure 2

Réalisation d'une occlusive non voisée
entre deux voyelles V

1 : occlusion

2 : explosion

3 : phase intermédiaire entre l'explosion et le début de la voyelle, on désigne par VOT (voice on set time) ou DEV (durée d'établissement du voisement), le temps mesuré entre le début de l'explosion et le début de la voyelle (2+3).

On caractérise les occlusives sourdes par leur durée à l'intérieur d'une phrase de structure particulière et en position inter-vocalique V_1 C V_2 accentuée où V_2 n'est jamais élidée ni diminuée ([a] convient), et qui tend à diminuer l'influence du contexte : la coarticulation entre V_1 et C entre C et V_2, doit être réduite au minimum (transitions brèves) ; l'accentuation est obtenue par des phrases clivées.

- durée d'occlusion "stricte", occlusion seule (sans l'explosion), (OS), voir figure,

- durée d'occlusion "large" (occlusion + durée d'établissement du voisement), (voir figure), (OL),(1) (le délai d'établissement du voisement étant de 10 à 30 ms).
- durée de sons (PH) dans certaines des six mêmes phrases.

On considère également des :
- durées de phonation (sur une phrase) et de segments phonétiques,
- durées globales : durées d'énonciation de phrases de texte,
- durées de pauses : pauses intermédiaires entre deux groupes de phonation.

Les pauses ont des origines variées. Elles contribuent à rendre plus compréhensible la phrase à l'auditeur (niveau sémantique), c'est-à-dire introduisent un lien :

temps d'assimilation - intelligibilité,

SAINT-BONNET & BOË (1977), observent la relation :

nombre de pauses - variation du débit - durée moyenne des groupes de phonation

et les différences dans le nombre et la durée des pauses entre hommes et femmes (plus courtes chez celles-ci). Physiologiquement, les pauses sont corrélées avec la respiration phonatoire. On obtient :
- pour les pauses : la durée, le nombre,
- pour l'ensemble (pauses ≠ phonation) : durée totale d'énonciation (les pauses considérées ont une durée supérieure à 200 ms).

---

(1)IBBA & al. (1979), viennent de reprendre une étude des occlusives dans le même contexte.

durée de chaque groupe de phonation dans la troisième phrase du texte (G1 à G4),

- durée de la pause intermédiaire (PI), durée totale de phonation (PN), dans les six phrases isolées, clivées,

Notant par Q l'ensemble des supports phonétiques (phrases, texte, ...), prononcés par les locuteurs, on définira un paramètre d'après la combinaison - lorsqu'elle est retenue possible -

$$(p, q)$$

d'une variable $p \in P$ et d'un support $q \in Q$. Ceci pour exprimer la dépendance par rapport au contexte. Un paramètre est donc une application :

$$\pi : P \times Q \rightarrow R+$$

La structure des identificateurs de paramètres est :

< identificateur de support > < identificateur de variable >

Par exemple :

P5PN : durée d'énonciation de la 5ème phrase,

T3G1 : durée du 1er groupe de phonation dans la 3ème phrase du texte TX.

On se reportera au tableau 1 pour la construction et l'identification des paramètres.

On note que :

1) La durée voisée divisée par la durée d'énonciation, donne une mesure du taux de voisement, sur une phrase.

2) Une mesure du débit est obtenue par division du nombre de mots (ou mieux de syllabes) prononcés par la durée d'énonciation, avec quelquefois - si nécessaire - une correction suivant l'origine géographique du locuteur (unités : syll./mn). La durée d'énonciation doit être voisine d'une minute, au minimum.

3) On obtient une mesure du fondamental moyen, sur une phrase ou un texte, en divisant le nombre des impulsions glottiques par la durée de voisement.

Nous avons ainsi un ensemble de paramètres probablement fortement redondant et nous recherchons un bon sous-sensemble de paramètres indépendants.

La segmentation est toujours un problème difficile en reconnaissance automatique de la parole, mais nous ne sommes pas concernés dans la mesure où l'on peut faire appel à un même contexte linguistique lors des expériences.

3.3. Appareillage utilisé et conditions d'enregistrement :

Le matériel disponible [1] comporte : un studio d'enregistrement (chambre sourde), et magnétophone de haute qualité, un analyseur de la fréquence fondamentale permettant la localisation des impulsions glottiques et du temps de voisement en ms, connecté à un détecteur de sonorité [2] (entrées analogiques et fonctionnement temps réel), et un oscillographe à jet d'encre, réalisant l'inscription du signal de parole, de la fonction de voisement et des variations de la fréquence fondamentale. Celui-ci permet de faire des mesures à la main d'une précision suffisante : $\pm$ 5 ms.

En outre, compte-tenu des erreurs de mesure, toutes les valeurs sont considérées comme ayant une décimale exacte. Notons que la plupart des variables utilisées dépendent évidemment du contexte phonétique.

3.4. Acquisition des données :

L'étude porte sur douze hommes adultes qui lisent d'une façon naturelle 10 fois au cours d'une même session, un texte (identificateur TX) d'une durée d'une minute environ, et 6 phrases courtes isolées, clivées (P1 à P6). On écarte l'étude des mots isolés car l'absence d'une phrase porteuse rend leur diction a priori trop variable.

---

(1) Institut de Phonétique de Grenoble.

(2) Selon l'appareillage de ABRY, BOË & ZURCHER (1974), BOË & al. (1975)

On se reportera aux tableaux suivants pour les caractéristiques du corpus de données.

Chaque mesure effectuée n'est pas à retenir par sa valeur absolue, mais essentiellement relativement aux autres (on utilise des données centrées réduites). En ce qui concerne la précision des mesures faites à la main, on peut affirmer qu'elles possèdent toutes le maximum de précision, compte-tenu de la difficulté intrinsèque de la segmentation des sons. On observe pour quelques locuteurs, des particularités : "mauvaise réalisation" d'occlusives, bruits de friction durant l'occlusion, occlusive sourde réalisée avec voisement. Elles nécessitent parfois des règles arbitraires de décision des instants de début et de fin des sons. Elles resteront toutefois les mêmes pour l'ensemble des locuteurs.
Le texte et les phrases sont lus plutôt que récités. Toutes les répétitions sont conservées, étant donné que les locuteurs sont invités à prendre connaissance du texte, et à faire un essai préalable d'enregistrement d'une dizaine de secondes (réglage de l'intensité).

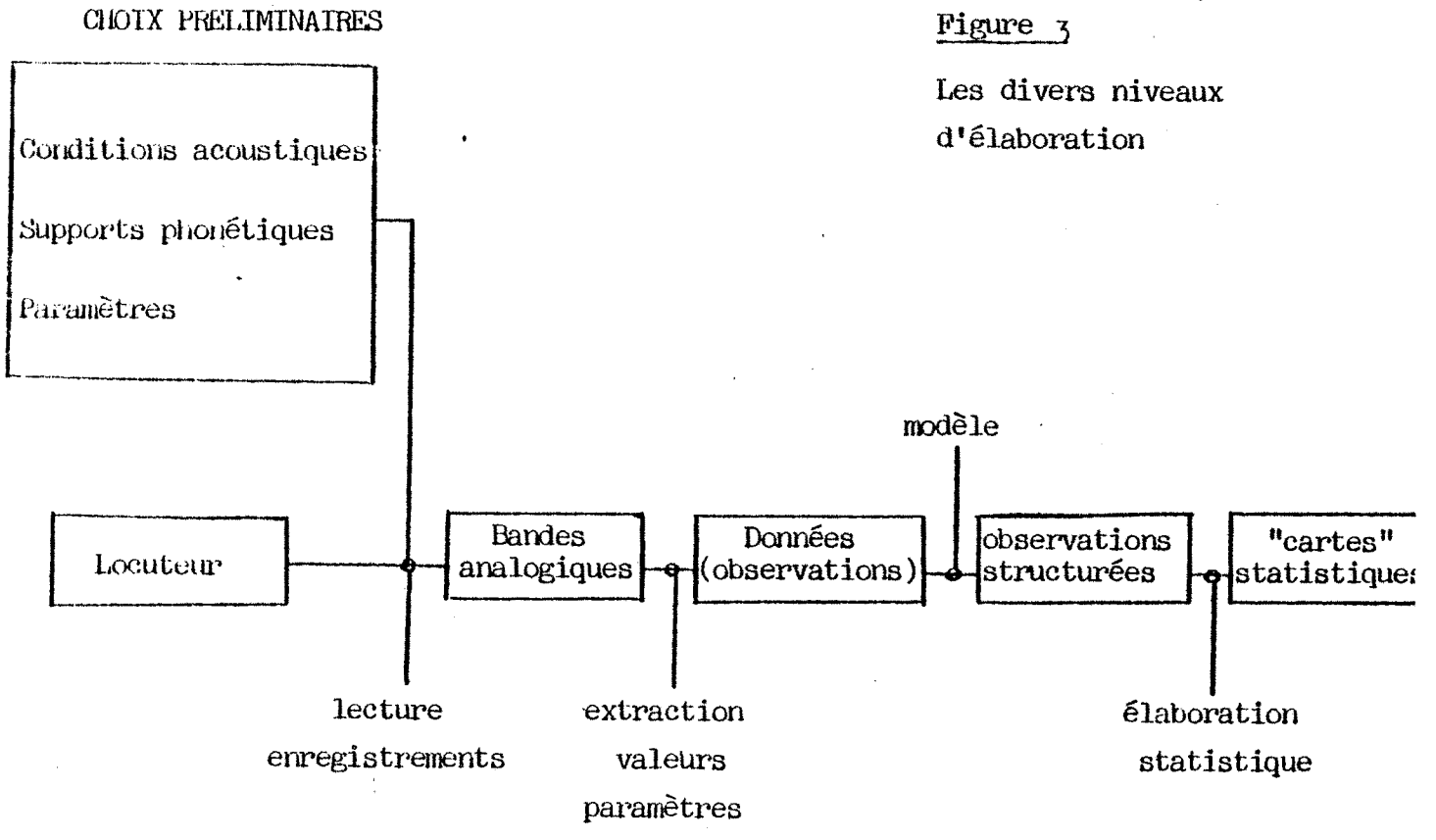


Figure 3

Les divers niveaux d'élaboration

TEXTE RETENU :

Identificateurs : TX : texte complet
T1 : 1ère phrase
T2 : 2ème phrase
T3 : 3ème phrase.

La troisième phrase est subdivisée en :

1er groupe de phonation : G1
1ère pause : P1

2ème groupe de phonation : G2
2ème pause : P2

3ème groupe de phonation : G3
3ème pause : P3

4ème groupe de phonation : G4
4ème pause : P4

A cet effet, l'Etat a cru trouver une solution fiscale au problème en s'orientant vers la récupération des plus-values foncières. Une loi du 3 juillet 61 avait, à cet effet, institué une redevance d'équipement payable par tous les propriétaires dont les terrains se trouveraient valorisés par la construction d'équipements publics. Pratiquement, cette loi sur la redevance d'équipement, laissée à la discrétion des communes, n'a jamais été appliquée. Elle mettait en effet les collectivités locales dans une position assez délicate vis-à-vis de leurs administrés et, en outre, son assiette était fort difficile à déterminer. C'est pourquoi elle a été remplacée par une taxe de régularisation des valeurs foncières instituée par l'article 8 de la Loi de Finances . Cette taxe de régularisation frappe les terrains nus ou bâtis, et situés dans les zones en voie d'urbanisation ou de rénovation, dans lesquelles sont réalisés des équipements publics d'infrastructure. Elle est fonction des besoins et de la capacité des bâtiments existants ou à réaliser sur la parcelle.

Caractéristiques du texte :
Le vocabulaire est peu spécialisé et le contenu sémantique relativement neutre. Les 168 mots, de deux syllabes en moyenne, forment quelques phrases longues (jusqu'à 29 mots), demandant un certain contrôle de la respiration. La durée totale est d'environ une minute.

PHRASES :

Identificateur	Texte	Réalisation phonème	Occurrence
P1	"Il rapa, c'est bien connu"	[/p/]	dans : rapa
P2	"Il le mata, c'est bien connu"	[/t/]	mata
P3	"L'avocat, il est bien connu"	[/k/]	avocat
P4	"Il l'agaça, c'est bien connu"	[/s/]	agaça
P5	"C'est un chat, c'est bien connu"	[/ʃ/]	chat
P6	"Quelle cacophonie"	[/k/]	/cacophonie/ ↑

LOCUTEURS :

numéro	1	2	3	4	5	6	7	8	9	10	11	12
identificateur	AB	BC	BE	BO	BU	CN	CO	HA	LC	LU	MA	WI

Moyenne d'âge : 35 ans (6 enseignants + quelques étudiants + autres)
Origine géographique variée. Pas de symptômes ORL apparents.

GROUPES DE VARIABLES :

Identificateur	FM	TV	DB	GI	PI	OS	OL	PH	PN
Definition	Fondamental moyen	Taux de voisement	Débit	Durée du groupe de phonation numéro I	Durée de la pause intermédiaire	Durée d'occlusive stricte	Durée d'occlusive large	Durée son (durée phonème)	Durée de phonation

TABLEAU 1 - CONSTRUCTION ET IDENTIFICATION DES PARAMETRES

IDENT. de con-texte	contenu linguis-tique	son évent. recher-ché	PARAMETRE	Variable	IDENT. de varia.	IDENT. de param.	nombre ordinal para-mètres	nombre param. par catég.	form FORTR
PHRASES ISOLEES				durée occlusion stricte	OS	P10S	17		
						P10L	18		
P1		[p]				P20S	19		
						P20L	20		
P2	tableaux	[t]		durée oc-clusion large	OL	P30S	21	8	F5.1
						P30L	22		
P3	annexes	[k]				P60S	23		
P4		[s]		durée to-tale son	PH	P60L	24		
						P4PH	25	2	F5.1
P5		[ʃ]				P5PH	26		
				durée pause intermé-diaire	PI	P1PN	27		
P6		[k]				P1PI	28		
						P2PN	29		
						P2PI	30	11	F6.1
				durée totale phrase	PN	P3PN	31		
						P3PI	32		
						P4PN	33		
						P4PI	34		
						P5PN	35		
						P5PI	36		
						P6PN	37		
PHRASES DU TEXTE									
TX				fondam. moyen		T1FM	04		
						T1TV	05		
T1	1ère phase					T2FM	06	6	F5.1
					FM	T2TV	07		
						T3FM	08		
T2	2ème phase			taux voisement	TV	T3TV	09		
T3	3ème phase			1 durée	G1	T3G1	10		F4.1
				2 groupe	G2	T3G2	11		
				3 phonat.	G3	T3G3	12		
				4	G4	T3G4	13	7	
				1 durée	P1	T3P1	14		F5.1
				2 pause	P2	T3P2	15		
				3 interméd.	P3	T3P3	16		
TEXTE LONG			variables intermé-diaires NIG	fondamental moyen	FM	TXFM	01		F5.1
			DV						
TX	Annexe		DT	taux voisem.	TV	TXTV	02	3	F5.1
				débit	DB	TXDB	03		F4.1

II - MODELISATION DES LOCUTEURS

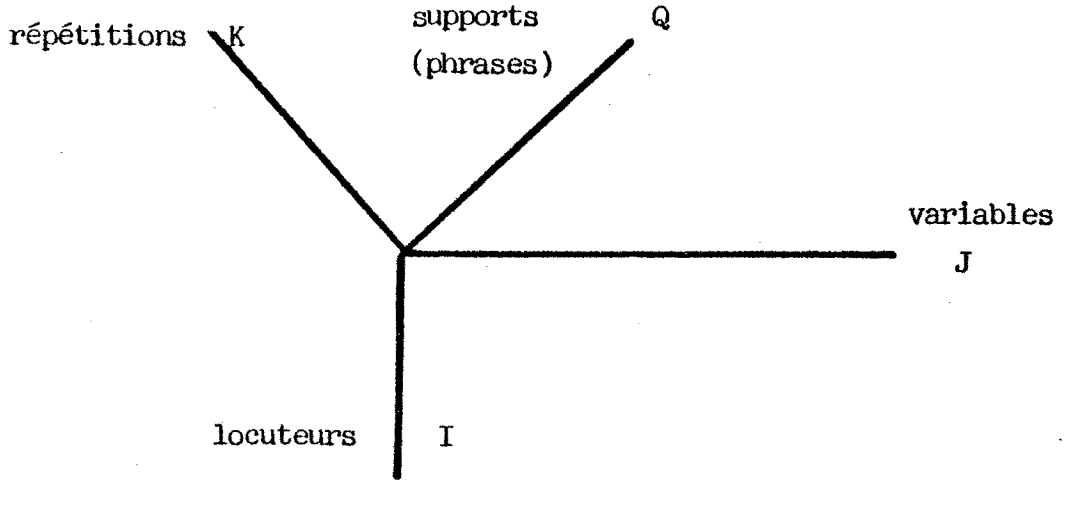
- Les facteurs en présence :

Un locuteur $i \in I$ est modélisé par un ensemble de paramètres J prenant diverses valeurs réelles positives pour K répétitions du support phonétique imposé. ${}^{k}x_{i}^{j} \in R_{+}$ désigne la valeur du $j^{ème}$ paramètre pour le locuteur i lors de la $k^{ème}$ répétition.

Le nombre de locuteurs $|I|$ est donné à l'avance (spécialement dans le contexte de la Vérification). On suppose choisies les $|Q|$ phrases, ainsi que les $|J|$ paramètres a priori (variables). On remarquera que le rapport $|J|/|I|$ doit être suffisamment grand.

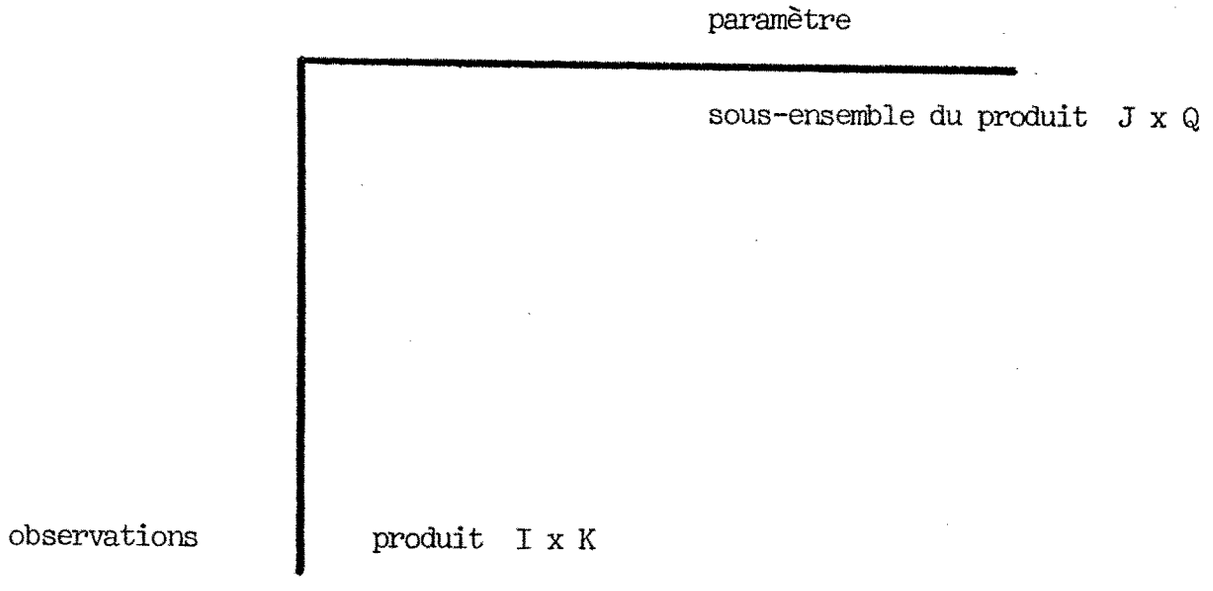
Les diverses répétitions servent - théoriquement - à épuiser les variations temporelles de la voix d'un même locuteur. On simplifie en faisant répéter un même nombre de fois chacun. Il est connu que, lorsqu'on ne dispose que d'un nombre fini d'échantillons pour l'apprentissage, la performance du classificateur croît vers un maximum, puis décroît quant on augmente le nombre d'échantillons ; le phénomène a une action d'autant plus négative que les échantillons sont plus dépendants (JAIN & DUBES, 1978). Ainsi, la classification sera d'autant meilleure que l'on aura utilisé des caractéristiques non corrélées. Ces mêmes auteurs donnent une règle empirique : le rapport du nombre d'échantillons à celui des caractéristiques, doit être supérieur à 5 pour que le problème soit bien posé. Augmenter le nombre de caractéristiques oblige à faire passer la difficulté essentiellement au niveau du classificateur (mais trop peu de caractéristiques sont peut-être insuffisantes pour discriminer les locuteurs). Il y a là une correspondance dont l'optimisation n'est pas résolue.

On obtient donc le tableau à 4 dimensions suivant :



Pour analyser plus facilement l'ensemble des données, on le réduit au tableau bidimensionnel obtenu en groupant :

- en colonnes : les variables et les supports (axe des paramètres),
- en lignes : les locuteurs et leurs répétitions (axe des observations) :



Le signal de parole fournit - plus ou moins facilement - de très nombreuses caractéristiques, dont beaucoup pourront être fortement corrélées. Ni le théorème d'échantillonnage, ni les considérations de coût d'obtention (temps et facilité), ne constituent des règles heuristiques pour limiter convenablement le nombre de caractéristiques. Inversement, si on ne dispose que d'un nombre très limité d'échantillons, enregistrer de plus en plus de locuteurs et chacun au cours de plus en plus de sessions, devient rapidement difficile. C'est là une situation assez contradictoire et peu étudiée en reconnaissance des formes.

DEUXIEME PARTIE :
ETUDES EXPLORATOIRES ET SIMULATION D'EXPERIENCES

III - MISE EN OEUVRE DE PARAMETRES PERTINENTS

1 - Sélection de caractéristiques :

Appelons caractéristique (ou primitive) un paramètre pertinent. Nous allons effectuer une sélection de caractéristiques à partir des paramètres de base, en deux temps, par la :
- création de sous-groupes décorrélés,
- sélection des éléments les plus représentatifs de chaque groupe au sens d'un critère. Ceci est justifié dans la mesure où il n'existe pas de méthode directe pour l'estimation des dépendances entre paramètres.

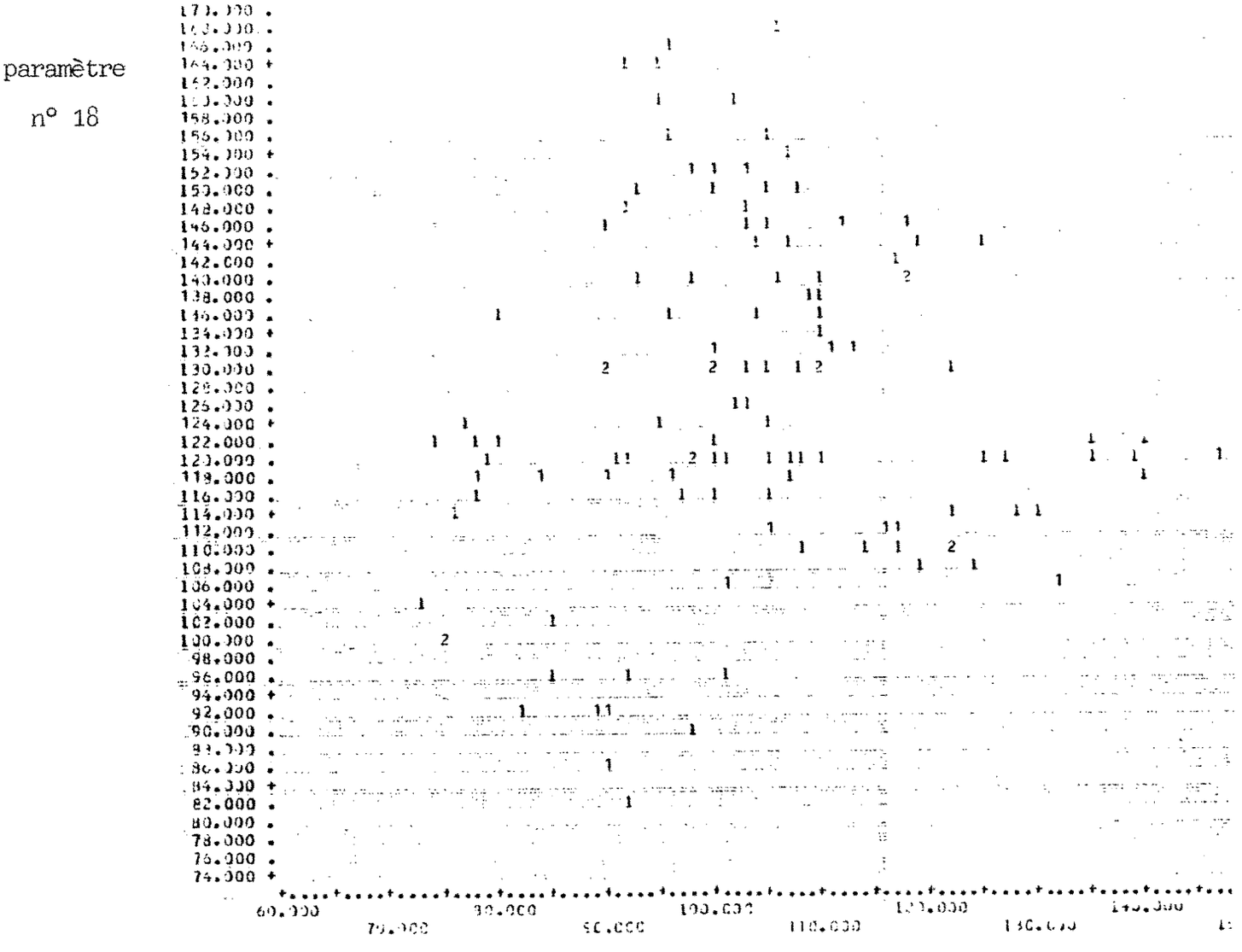
1) Niveau 1 : étude des corrélations entre les paramètres

Notre but est de résumer les 37 caractéristiques par leur matrice de corrélation C pour ne pas avoir à postuler de lois statistiques. Leur sélection est ainsi fondée sur la décorrélation linéaire deux à deux sur un certain sous-ensemble (tous locuteurs confondus). Ainsi, C est considéré comme une matrice de proximité fondée sur la mesure de la valeur absolue du coefficient de corrélation. Les mots corrélation et variance sont de simples quantités, non pas des paramètres de lois statistiques. On construit le graphe symétrique G = (J,U) sur l'ensemble des locuteurs, où J est l'ensemble des caractéristiques et u = (i,j) ϵ U, (i et J ϵ J), si : la corrélation $r(i,j) \geq 0{,}50$. On observe en effet dans C, d'une part des valeurs élevées (supérieures à 0,70), de très nombreuses valeurs faibles (inférieures à 0,30 [1]) et des valeurs intermédiaires relativement rares. Il y a : 12 locuteurs x 10 répétitions = 120 échantillons ; on sait que l'erreur sur la valeur du coefficient de corrélation décroît lorsque croît le nombre d'échantillons. La procédure classique serait la mise en oeuvre d'un algorithme ascendant de classification hiérarchique : on aurait à décider du seuil s choisi pour la classification ($s \in [0,30 ; 0,70]$). En fait, on observe que les composantes simplement connexes de G sont à peu près les groupes des variables définis en II/. Quelques corrélations sont inattendues et semblent fortuites (voir Fig. 4). Ainsi, on peut

---

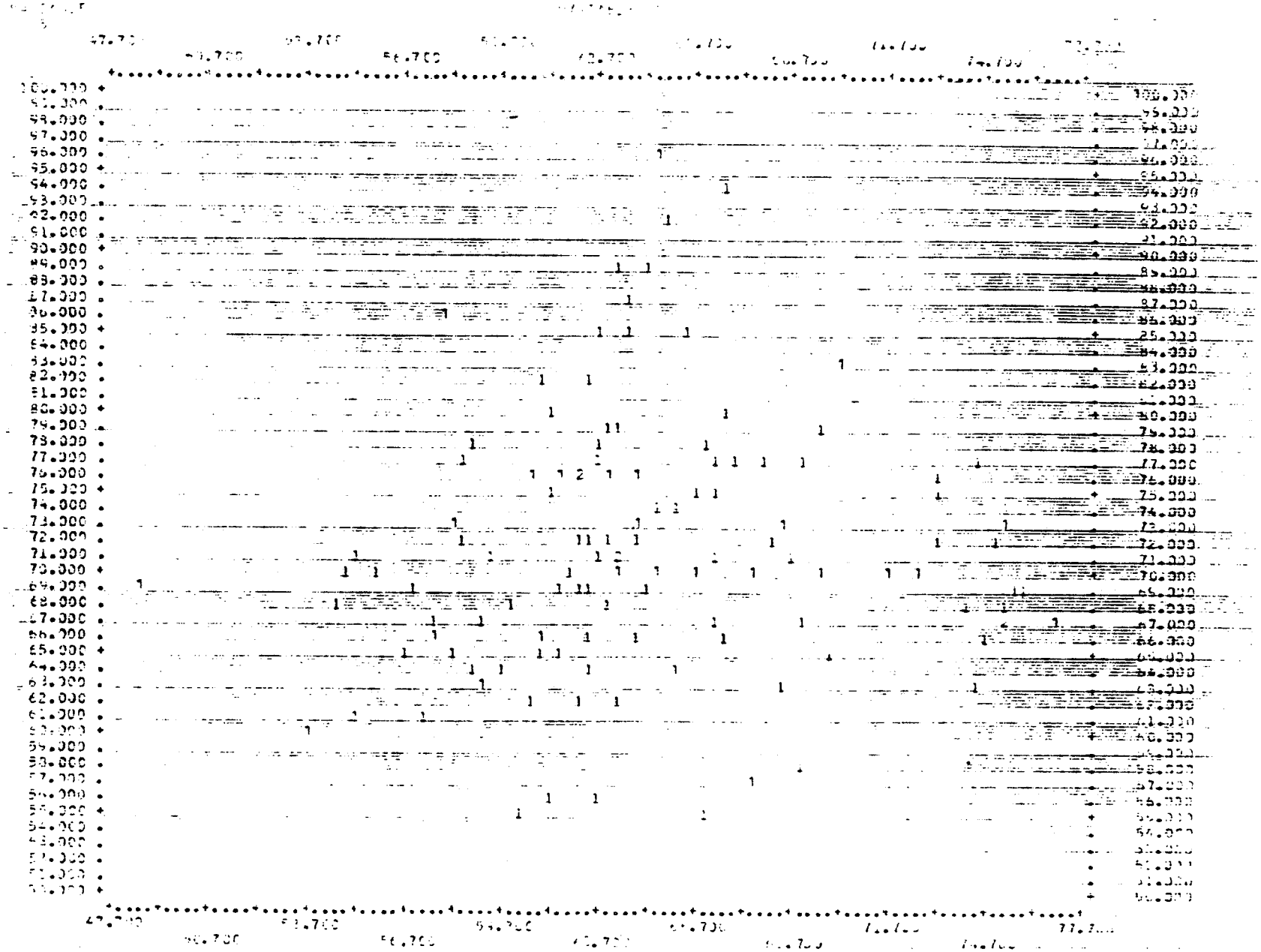
(1) En supposant que les distributions des variables aléatoires discrètes $({}^{k}x_{i}^{j})$ $1 \leq k \leq 10$, sont normales pour tous les paramètres $j \in \{1,\ldots,37\}$, les corrélations observées sont significatives pour le test de Student au seuil 1%, à partir de 0,23. Nous ne retenons que les corrélations fortement significatives à ce seuil.

Figure 4 - Examen de corrélations "douteuses" entre éléments de chaque paire de variables. En toute rigueur, si $r \neq \pm 1$, il faut observer directement le nuage de points, ce qui fait près de $C_{37}^2 = 666$ graphiques. Ci-dessous, sont montrés les nuages pour les couples (3TXDB, 5T1TV) et (18P10L, 24P40L) : on n'observe pas de corrélation entre le débit sur le texte et le taux de voisement sur la 1ère phase du texte, ni entre les durées "larges" de deux occlusives dans les phrases 1 et 6.



paramètre n° 5

TXDA



paramètre n° 3

TXDB

prévoir qu'un algorithme de classification hiérarchique fondé sur la mesure de similarité (i,j), construit d'abord ces groupes.
Dans la suite de l'étude, étant donné cette configuration, nous conserverons ces groupes et chercherons à leur ôter des paramètres afin de les décorréler complètement, plutôt que d'établir de nouveaux groupes. Cette attitude exprime la contrainte du matériel qui est spécifique à chaque groupe.

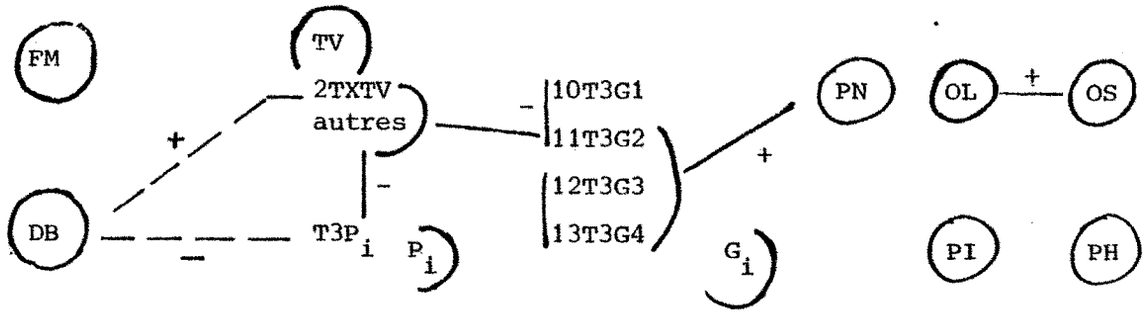
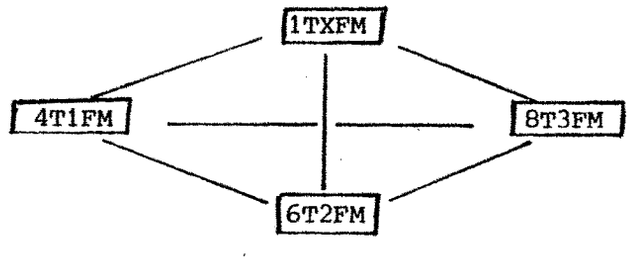
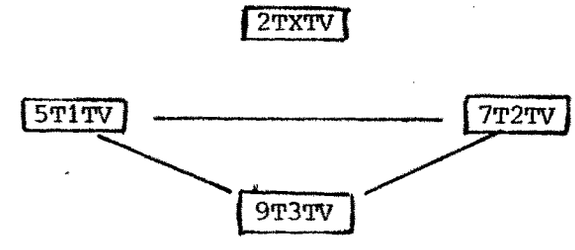


Figure 5 - Examen des corrélations entre groupes de variables. Les pointillés expriment des corrélations faiblement significatives.
D'autre part, étant donné la non-connexité forte à l'intérieur de chaque groupe de variables (excepté pour le groupe FM), (voir fig. 5), nous sommes amenés à vérifier une éventuelle régression quadratique ou autre entre deux caractéristiques faiblement corrélées. Ceci n'est pas fait systématiquement mais pour des cas douteux (figures 4 et 5).
Enfin, on remarque une relative décorrélation DB - TV (débit - taux de voisement

Figure 6 - Examen des corrélations à l'intérieur de chaque groupe. On observe une transitivité seulement en cas de fortes corrélations comme c'est le cas pour le sous-graphe complet construit sur le groupe FM.

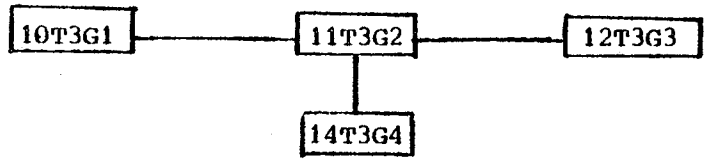


a) Très fortes corrélations à l'intérieur du groupe FM.

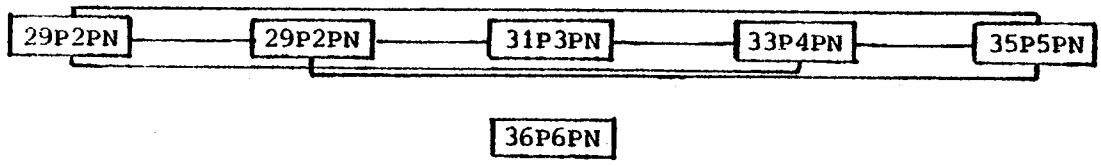


b) Connexité entre les trois phrases seulement.

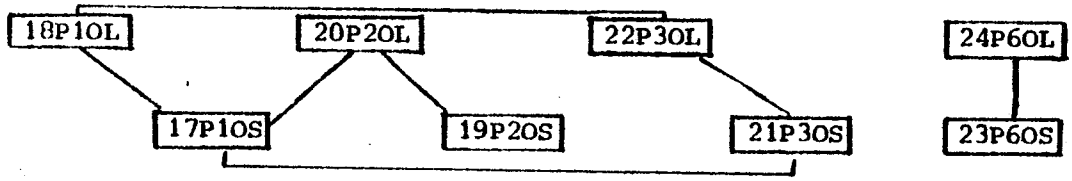
c) Connexité du groupe GI



d) Nombreuses corrélations à l'intérieur du groupe PN

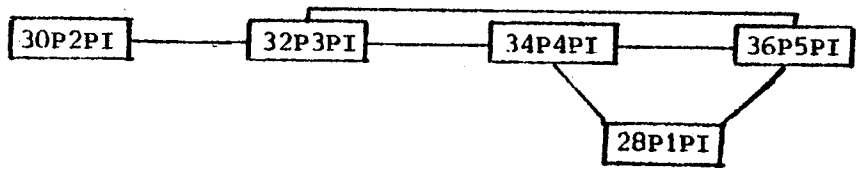


e) Groupes OS et OL : fortes corrélations (verticales) entre caractéristiques correspondantes ;

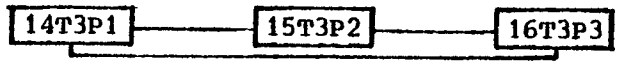


(la phase P6 a une structure différente des 5 autres, ce qui peut expliquer les non-corrélations horizontales avec 23 et 24).

f) Groupe PI



g) "Groupe" $T3P_i$



Toutefois, nous garderons bien à l'esprit que la décorrélation ne signifie pas que les variables apportent des informations indépendantes. Mais dès que l'on fera des hypothèses de normalités, cela pourra être supposé.
L'indépendance des paramètres signifie que l'on peut les utiliser séparément plutôt que conjointement lors de la classification. L'information globale est alors la somme des informations obtenues à l'aide de chacun des paramètres.

Figure 7 - Corrélations entre paramètres, observées pour le locuteur W_i

a) Construction

Supposant que, pour tout $i \in I$, $X_i = (X_i^1, \ldots, X_i^{37})$ suive une loi normale à 37 dimensions, chaque couple (X_i^{j1}, X_i^{j2}) pour $j_1, j_2 \in \{1,\ldots,37\}$, suit une loi normale bidimensionnelle. On peut appliquer le test de Student : observant les 10 couples $({}^kX_i^{j1}, {}^kX_i^{j2})$, $(k = 1,\ldots,10)$ et choisissant le seuil 1%, les corrélations r_i observées sont significatives dès que

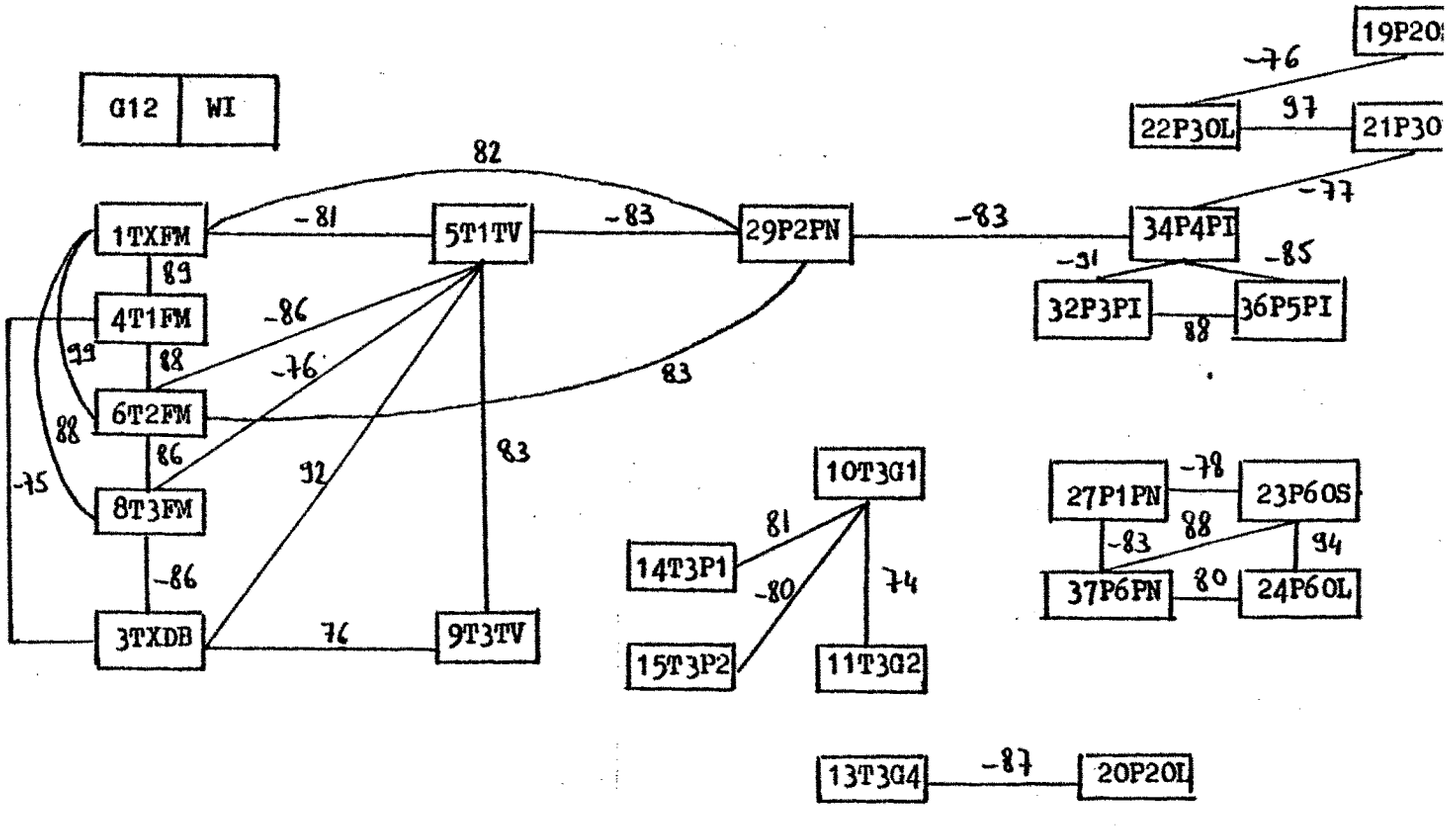
$$r_i \geq 0{,}76.$$

On construit les graphes $G_i = (J, U_i)$ relatifs aux locuteurs, avec :

J : ensemble des sommets-paramètres π_j,

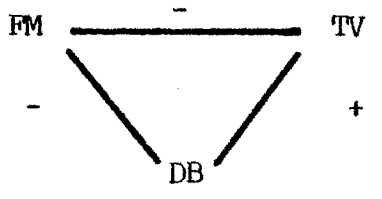
U_i : ensemble des arcs (π_{j_1}, π_{j_2}) construits lorsque $r_i(\pi_{j_1}, \pi_{j_2}) \geq 0.76$.

Ci-dessous, est montré le graphe G_{12} relatif au locuteur 12 WI. Les sommets isolés ne sont pas indiqués, les valeurs des arcs sont $100.r_i$.



b) Observations

On remarque, et cela reste vrai chez tous les locuteurs, des corrélations positives très fortes dans le groupe FM (sous-graphe complet d'ordre 4). Plus généralement, des corrélations toujours positives à l'intérieur de chaque groupe (ce qui établit une sorte de cohérence de la définition de ceux-ci), sauf pour le groupe PI (pauses intermédiaires dans les phrases). Ce qui pourrait s'expliquer par un phénomène de compensation interne dans les durées des pauses et des groupes de phonation d'une même phrase.
Un point important est la présence de corrélations largement significatives entre les groupes : FM, TV et DB, selon le schéma :



Ce dernier résultat n'est pas vrai dans le graphe G construit sur l'ensemble de tous les locuteurs, et ne se retrouve d'ailleurs pas chez la plupart des autres locuteurs. Ceci pourrait servir à mettre en valeur la nécessité d'un grand nombre de locuteurs pour la démonstration d'une corrélation TV - DB, laquelle a été tentée par certains auteurs.

Rappel du tableau 1 - Construction et identification des caractéristiques

SUPPORT PHONETIQUE ET IDENTIFICATEUR	GROUPE DE VARIABLES D'APPARTENANCE	IDENTIFICATEUR DES CARACTERISTIQUES
Texte long TX	- fondamental moyen FM - taux de voisement TV - débit DB	01TXFM 02TXTV 03TXDB
Phrases extraites du texte 1ère phrase T1 2ème phrase T2 3ème phrase T3	- fondamental moyen FM - taux de voisement TV - débit DB - durée de groupes de phonation G1 à G4 - durée de la pause intermédiaire P1 à P3	04T1FM 05T1TV 06T2FM 07T2TV 08T3FM 09T3TV 10T3G1 11T3G2 12T3G3 13T3G4 14T3P1 15T3P2 16T3P3
Phrases isolées /p/ "Il rapa, c'est bien connu." P1 /t/ "Il le mata, c'est bien connu." P2 /k/ "L'avocat, il est bien connu." P3 /s/ "Il l'agaça, c'est bien connu." P4 /ʃ/ "C'est un chat, c'est bien connu." P5 /k/ "Quelle cacophonie." P6	- durée occlusion stricte OS - durée occlusion large OL - durée son PH - durée pause intermédiaire PI - durée totale phrase PN	17P1OS 18P1OL 19P2OS 20P2OL 21P3OS 22P3OL 23P6OS 24P6OL 25P4PH 26P5PH 27P1PN 28P1PI 29P2PN 30P2PI 31P3PN 32P3PI 33P4PN 34P4PI 35P5PN 36P5PI 37P6PN

Ce tableau concerne l'ensemble des locuteurs.

TABLEAU 2 - Indication des valeurs du rapport F pour chaque paramètre, selon leur groupe d'appartenance. Par exemple, la variable 3 du groupe DB a un F-ratio de 6 environ. Se reporter au tableau 1 pour la définition des paramètres indiqués ici par leur numéro d'ordre.

F	FM	TV	DB	GI	PI	OS	OL	PH	PN
16- 15- 14- 13- 12- 11- 10-	1						22		31 37
9- 8- 7- 6- 5-	8 4	2	3			17 21	18 20 24		27 29
4- 3- 2- 1- 0-	6	9 5 7		10 11 12 13	32 28 34 14 36 15 16 30			26 25	35 33

2) Niveau 2 : compression

Afin de choisir un sous-ensemble petit des 37 caractéristiques qui ne comporte aucune corrélation, on sélectionne l'une des caractéristiques de chaque composante simplement connexe. Le critère de sélection peut être (se référant toujours aux qualités idéales) :

- le coût et le temps d'obtention suivant la disponibilité d'appareils délivrant les mesures en temps réel avec un coût pratiquement nul,

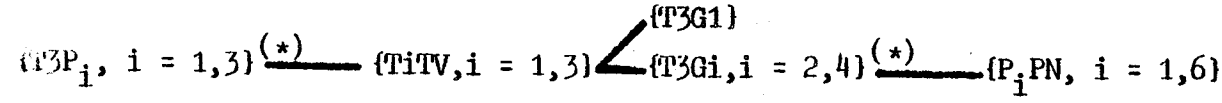
- la discrimination opérée sur la population des locuteurs (critère obligatoire). On sait que le meilleur n-uplet de caractéristiques n'est pas en général obtenu en choisissant les n meilleures, même si elles sont indépendantes (COVER, 1974). Cependant, on se limitera au rangement des caractéristiques à l'intérieur des groupes, suivant le critère :

$$F = \frac{\text{variance des moyennes des locuteurs}}{\text{moyenne des variances des locuteurs}},$$

valable dans le cas où le nombre de répétitions le même pour chacun et qui donne une estimation globale de l'efficacité. Il a été beaucoup utilisé dans ce contexte (ROSENBERG, 1976), mais reste relativement insuffisant vu son insensibilité à la dispersion des diverses variances.
Le F-ratio est défini à un facteur multiplicatif près ; il possède la propriété d'invariance par homothétie et aussi par translation.
F ne possède pas de propriété d'optimalité puisqu'il ne minimise pas d'erreur de propbabilité. Simplement, plus les distributions entre deux locuteurs sont éloignées et/ou plus les distributions entre répétitions d'un même locuteur sont proches, plus le rapport est grand.

D'autres critères sont possibles : variance maximum (déterminer la part de variance expliquée par chaque caractéristique et prendre celle correspondant à la plus forte part), taux de mauvaise classification sur l'ensemble d'apprentissage, analyse discriminante pour déterminer une combinaison des caractéristiques.
D'autres statistique, qui nécessitent la donnaissance des lois de probabilités sous-jacentes, ont été étudiées : la divergence (MARILL & GREEN, 1963) et l'information mutuelle (LEWIS, 1962 ; KAMENTSKY & LIU, 1963).

L'examen de la figure 6 montre qu'il suffit de déconnecter la chaîne :



durées pauses intermédiaires 3ème phrase — taux de voisement

Paramètres :	(14,15,16)	(5,7,9)	(11,12,13)	(27,29,31,33,35,37)
Nombre :	3	3	3	6
Somme F-ratio :	4,7	4,8	5,4	38,5
Moyenne F-ratio :	1,6	1,6	1,8	6

Pour les raisons suivantes :
- calcul du taux de voisement sur des phrases plutôt que sur le texte,
- proximité des F-ratio dans les trois ensembles de gauche ci-dessus,

on peut proposer d'ôter les paramètres marqués (*). Il reste alors le 9-uplet :

FM, TV, DB, Gi, PI, OS, OL, PH, PN)
(1, 2, 3, 10, 32, 17, 22, 26, 31)

qui, ordonné par F-ratio décroissants, donne :

(FM, OL, PN, OS, DB, TV, Gi, PH, PI
(1 , 22, 31, 17, 3 , 2 , 10, 26, 32).

Si l'on remplace les paramètres 1 et 2 par 8 et 9 de la même catégorie, et si l'on ôte le débit 3, cela permet de se limiter à des phrases sans recours au texte. Mais l'on mesure la perte de pertinence subie dans le fondamental et le taux de voisement, qui ont un relatif avantage à être mesurés sur un texte long.
Enfin, le critère empirique de JAIN & DUBES se trouve largement vérifié (120/9 > 13, alors que précédemment : 120/37 ~ 3,2 < 5).

2 - Description statistique de la population :

2.1. Suivant un point de vue quelque peu complémentaire du § 1, nous essayons ici de dresser des cartes des locuteurs sur l'ensemble des paramètres choisis. Chaque observation élémentaire est un nombre réel positif, image de la variable aléatoire :

$$ {}^{k}X_i^j : [a,b] \to \mathbb{R}_+, $$

correspondant au paramètre j, lors de la $k^{ième}$ répétition du locuteur i (qui a lieu durant [a,b]). Les ensembles de variation des indices sont :

locuteurs : $i \in I = \{1,\ldots,12\}$,
paramètres : $j \in J = \{1,\ldots,37\}$,
répétitions : $k \in K = \{1,\ldots,10\}$.

La matrice X décrit l'ensemble I x K des observations (120 lignes), par les paramètres de J. Pour chaque locuteur i, on observe le vecteur :

$$ X_i = (X_i^1,\ldots, X_i^{37}) = \sum_{j=1}^{37} X_i^j \; e_j \qquad (1) $$

où X_i^j est la notation pour $({}^{1}X_i^j, \ldots, {}^{10}X_i^j)$ et prend ses valeurs dans $\mathbb{R}_+^{10}$.

La structure statistique qui correspond à l'observation des paramètres pour un locuteur donné i est la structure d'échantillon :

$$ (\mathbb{R}_+^{37}, B(\mathbb{R}_+^{37}), P_i)^{10}; $$

où $P_i = (P_i^1, \ldots, P_i^{37})$ est la loi conjointe des lois de probabilités relatives aux divers paramètres, et qui dépendent - a priori - du locuteur.

---

(1) X_i est considérée à la fois comme une application linéaire et une matrice 10 x 37, et $e_j = (0,\ldots,1,\ldots,0) \in \mathbb{R}^{37}$ est le $j^{ème}$ élément de la base canonique de l'espace vectoriel construit sur les paramètres.

Globalement, à l'observation des valeurs de l'application linéaire $X = \begin{pmatrix} X_1 \\ \vdots \\ X_{12} \end{pmatrix}$, et en supposant que les composantes X_i sont indépendantes

(ce qui semble naturel), correspond le produit de structures du type précédent :

$$\bigotimes_{i=1}^{12} (\mathbb{R}_+^{37}, B(\mathbb{R}_+^{37}), P_i)^{10},$$

qui regroupe les 37.10.12 = 4440 observations de X. A chaque répétition, locuteur et paramètre, est affecté un même poids (resp. $\frac{1}{10}$, $\frac{1}{12}$, $\frac{1}{37}$). De même, on associe à chaque locuteur une même probabilité de demande de vérification.

2.2. Calcul des indices de dispersion :

Pour chaque i et j, on calcule :

$$\cdot X_i^j = \frac{1}{10} \sum_{k=1}^{10} {}^kX_i^j \quad \text{(moyenne empirique des répétitions)}$$

$$S^{2j}_i = \frac{1}{9} \sum_{k=1}^{10} (({}^kX_i^j)^2 - 10\,(\cdot X_i^j)^2) \quad \text{(estimateur de la variance)},$$

$$S_i^j = (S^{2j}_i)^{1/2} \quad \text{(écart-type)},$$

$$r_i^j = \bar{X}_i^j - \underline{X}_i^j \quad \text{(étendue empirique, avec } \bar{X}_i^j = \operatorname{Max}_k {}^kX_i^j \text{ et } \underline{X}_i^j = \operatorname{Min}_k {}^kX_i^j\text{)}.$$

r_i^j présente l'inconvénient de dépendre uniquement des valeurs extrêmes : on a une confiance médiocre en celles-ci, à cause de l'effet de série dans les répétitions successives ; mais celui-ci peut être réduit par un entraînement de la voix. On calcule aussi :

$$V_i^j = \frac{S_i^j}{X_i^j} \quad \text{(taux de dispersion relative empirique intra-locuteurs)}.$$

Pour chaque j, on calcule :

$$F^j = \frac{\sigma^j}{S^{2j}_{\cdot}} \quad \text{où} \quad \sigma^j = \frac{1}{11}\left(\sum_{j=1}^{12} ({}^{\cdot}X_i^j)^2 - 12\,({}^{\cdot}X_{\cdot}^j)^2\right),$$

$$\text{avec} \quad {}^{\cdot}X_{\cdot}^j = \frac{1}{12}\sum_{i=1}^{12} {}^{\cdot}X_i^j \quad \text{et} \quad S^{2j}_{\cdot} = \frac{1}{12}\sum_{i=1}^{12} S^{2j}_i$$

Ce rapport de la variante inter-locuteurs à la variance intra-locuteurs (variance des moyennes des locuteurs rapportée à la moyenne des variances des locuteurs), donne, on l'a vu, une estimation globale de l'efficacité. Mais il ne tient pas compte des différences entre les variances intra-locuteurs. Afin de définir un indice fondé sur le locuteur "le plus défectueux", on utilise la valeur :

$$E^j = \frac{\sigma^j}{S^{2j}_i} \quad (1)$$

(variance inter-locuteurs sur maximum des variances des locuteurs).

2.3. Tableau des dispersions relatives intra-locuteurs :

On fait ici une analyse globale et locale de la population qui repose sur l'idée suivante : une condition nécessaire pour l'utilisation d'un paramètre est que sa dispersion intra-locuteurs ne soit pas trop grande pour la plupart des locuteurs. Cette approche affaiblie de la pertinence des paramètres sera complétée par l'examen des variabilités inter-locuteurs et permet de pallier aux défauts du F-ratio, vus ci-dessus.

Notant $\overline{V_i^j} = \underset{i}{\text{Max}}\ V_i^j$, on attribue au paramètre j un indicateur variant de 0 à 9 et d'autant meilleur que la dispersion maximum $\overline{V_i^j}$ est faible, par :

$K^j = \text{Max}\,(0,\ 9 - d(\overline{V_i^j}))$ où d est le chiffre des dizaines du pourcentage entre parenthèses.

---

(1) Cet indice n'est pas utilisé dans les tests puisqu'il est défini hors du contexte linéaire.

On se reportera au tableau 4 et à sa légende pour les commentaires résultant de son observation.

Résultats de l'expérience :

La "force" de chaque paramètre varie suivant le locuteur. Ceci est localement mis en valeur par les indicateurs K^j et G^j (voir explications).

De l'examen comparatif des résultats obtenus, nous dégageons les résultats suivant

- parmi les paramètres choisis, de nombreux peuvent être reconnus comme pertinents et nous avons indiqué un exemple de sélection.
- l'utilisation de plusieurs paramètres appartenant à un même groupe de définition est superflue dans la mesure où ils sont corrélés (et significativement sous des hypothèses de normalité).

D'une façon générale, les paramètres construits sur une moyenne conduisent à des résultats plus fiables sur une phrase que sur des segments plus courts.

Les différences entre les débits suivant les locuteurs n'ont pas selon HECKER (1971) une origine organique mais sont dues au milieu socio-culturel de l'individu.

CLASSIFI. SELON E^j	$E^j = \frac{j}{S^{2j}}$	CLASSIFI. SELON F^j	$F^j = \frac{j}{S^{2j}}$	PARAMETRE
2	5.3	I	I6.4	I
I7	I.0	I4	5.4	2
9	2.0	I2	5.6	3
I4	I.5	I3	5.5	4
	0.35	33	I.2	5
	0.I9	24	2.2	6
	0.65	30	I.6	7
I3	I.55	5	7.8	8
I	7.I	29	I.7	9
II	I.6	I6	4.5	I0
	0.6	22	3.I	II
	0.3	3I	I.5	I2
	0.34	32	I.5	I3
	0.6	25	2.I	I4
	0.23	28	I.7	I5
	0.21	35	0.9	I6
5	3.I	7	7.5	I7
3	3.8	6	7.5	I8
	0.02	36	0.3	I9
II	I.6	II	5.9	20
I0	I.9	8	6.7	21
3	3.8	2	II.I	22
	0.2	34	I.0	23
I5	I.3	I5	5.I	24
	0.7	I9	3.4	25
I6	I.2	I7	3.5	26
8	2.6	9	6.5	27
	0.6	20	3.3	28
	I.05	I0	6.5	29
	0.0I	37	0.I	30
7	2.8	3	II.I	31
	0.8	I8	3.4	32
	0.I5	27	I.7	33
	0.5	2I	3.2	34
	0.2	23	2.4	35
	0.27	26	I.9	36
6	2.9	4	I0.3	37

TABLEAU 3 -

Classement des paramètres d'après les critères du F-ratio et du critère E^j.

TABLEAU 4 - Dispersions relatives et interprétations

Légende : les locuteurs sont désignés par leur numéro.
+ signifie : tous les locuteurs, C. signifie : complement ensembliste pris dans les lignes 6, 7 et 8 réunies, - signifie : aucun locuteur. (2) en ligne 8 signifie que le locuteur 2 est proche de la frontière de doute. Les cases barrées indiquent que le cas ne se présente pas.

ENSEMBLES	-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
DISP. MOY.	V_i^j %	2.2	3.8	3.5	2.9	7.4	3.3	3.0	2.6	5.8	6.6	6.2	9.0	9.3	34.5	52.9	91.5	6.3	5.4	18.9
DISP. MAX.	$\overline{V_i^j}$ %	3.8	10.0	7.0	4.7	13.0	15.9	7.4	4.1	9.2	10.5	12.7	19.7	20.4	210.8	154.0	153.0	9.8	8.8	131.5
INDIC. GLOBAL	k^j	9	8	9	9	8	8	9	9	9	8	8	8	7	0	0	0	9	9	0
INDIC. LOCAL étendue :	G^j																			8/11
LOCUTEURS TYPE R						1,5,6	2				5	4,11	1,5,6	7	+	C.	+			2,8
LOCUTEURS TYPE A		+	C.	+	+	C.	C.	+	+	+	C.	C.	C.	C.	-	4	-	C.	+	4,6 9,10 11,12
LOCUTEURS TYPE D		-	2	-	-	-	-	(2)	-	(1)	7,12	-	8	5,8 9,12	-	7,10	-	1,10)	10	3,5 7,12

PARAMETRES	j	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
DISP. MOY.	V^j %	6.7	7.1	12.1	9.1	6.3	8.2	5.9	4.2	59.2	3.3	1034	3.3	35.8	6.4	73.0	6.1	87.8	3.8
DISP. MAX.	$\overline{V_i^j}$ %	11.2	11.3	8.7	21.3	13.4	19.8	10.5	6.5	36.2	8.1	316.2	6.1	107.3	32.7	77.4	27.0	146.3	5.7
INDIC. GLOBAL	k^j	8	8	9	7	8	8	8	9	0	9	0	9	0	6	0	7	0	9
INDIC. LOCAL étendue :	G^j														9/11			9/11	
LOCUTEURS TYPE R					12					C.		C.		C.	11	C.	5	C.	
LOCUTEURS TYPE A		C.	C.	+	C.	C.	C.	C.	+	-	C.	3	+	-	C.	4	C.	-	+
LOCUTEURS TYPE D		5,7 8	7,8 12	(7)	3,4	12	6	6	-	3,4	(8)	4,10	-	4	-	-	-	4	-
IBC			-	-	-	-	-	-	-	1,2 5,7	-	2,7	-	2,5 7,9	-	2,7	-	1,2 5,6	-

Explications sur le tableau 4 -

1) K^j est l'indicateur défini à partir de $\overline{V_i^j}$: $K^j = 9$ indique les meilleurs paramètres au sens d'une dispersion intra-locuteur faible (inférieure à 10%) ; lorsque $K^j = 0$, il existe au moins un locuteur i pour lequel la dispersion V_i^j est très forte (supérieure à 90%). Par exemple, $K^j = 8$ signifie que $10\% \leq V_i^j < 20\%$.

2) G^j est comme K^j, mais ne porte que sur un sous-ensemble $I' \subsetneq I$ de locuteurs (on élimine quelques locuteurs pour lesquels le paramètre semble - pour des raisons pouvant être très diverses - inadapté. Ainsi, on peut obtenir un score très supérieur à K^j et qui garde tout son sens s'il porte sur presque tous les locuteurs (11 sur 12 le plus souvent). On dira alors que l'étendue de G^j est $|I'|$.

3) Pour donner une idée de la pertinence des paramètres face à chaque locuteur d'après le critère V_i^j, on range pour chaque j les locuteurs en trois groupes, suivant l'application :

$$J = \{1,\ldots,37\} \longrightarrow (\mathcal{P}(\{1,\ldots,12\}))^3$$

$$j \overset{f}{\longmapsto} (R(j), A(j), D(j))$$

satisfaisant à :

$$|R(j)| + |A(j)| + |D(j)| = 12 \quad \text{(les ensembles R, A et D partitionnent P(I))}$$

et définie par :

$R(j) = \{i \in I,\ V_i^j \geq 20\}$: locuteur de type R : ne pas utiliser le paramètre j pour le locuteur i.

$A(j) = \{i \in I,\ V_i^j < 10\}$: locuteur de type A : admettre l'utilisation du paramètre j pour i.

$D(j) = \{i \in I,\ 10\% \leq V_i^j < 20\%\}$: locuteur de type D : le paramètre j est douteux pour i.

Les seuils de 10 et 20% sont ici relativement arbitraires et peuvent être modifiés la méthode restant la même.

4) Quelques locuteurs ont présenté la propriété (binaire) de ne jamais réaliser les pauses aux endroits considérés (i.e. les paramètres 28, 30, 32, 34 ou 36 sont nuls). Cette information binaire complémentaire est relevée (IBC).

TABLEAU 5 : localisation du maximum des dispersions relatives.

Pour chaque paramètre classé d'après son groupe d'appartenance, on indique le numéro du locuteur correspondant à la plus grande dispersion : $S^{2\bar{J}}_{i} = \text{Max}\ S^{2j}_{i}$ $i=1,\ldots,12$

		1	2	3	4	5	6	7	8	9	10	11	12
FM	1		2										
	4						6						
	6		2										
	8					5							
TV	2		2										
	5						6						
	7		2										
	9	1											
DB	3		2										
GI	10					5							
	..												
	11				4								
	..												
	12					5							
	13							7					
PI	14	1											
	15	1											
	16	1											
	28									9			
	30					5							
	32						6						
	34								8				
	36									9			
OS	17	1											
	19		2										
	21								8				
	23												12
OZ	18	1											
	20								8				
	22							7					
	24												12
PH	25						6						
	26						6						
PR	27									9			
	29								8				
	31								8				
	33										10		
	35					5							
	37								8				
TOTAL		6	6	0	1	5	5	2	6	3	1	0	2

TABLEAU 6 : Répartition des paramètres en classes (rejet, admission, doute) pour chaque locuteur

Pour chaque locuteur, on indique le nombre de paramètres de chaque type (tableau du haut), et la même valeur en pourcentage sur les 37 paramètres (tableau du bas). Les résultats indiquent que, en utilisant la totalité des 37 paramètres, il existe en moyenne 26 paramètres (37 x 0.70) de classe A pour chaque locuteur, le minimum étant 23, le maximum 29, et 8 paramètres à rejeter (classe R) (max. 11, min. 3).

valeur somme

R	10	10	7	3	11	10	8	9	8	6	10	9	Total 101
A	27	26	27	29	23	25	24	24	28	29	27	24	313
D	0	1	3	5	3	2	5	4	1	2	0	4	30
Tot.	37	37	37	37	37	37	37	37	37	37	37	37	444
LOCUTEURS	1 AB	2 BC	3 BE	4 BO	5 BU	6 CN	7 CO	8 HA	9 LC	10 LU	11 MA	12 WI	

pourcentages

R%	27	27	19	8.1	29.7	27	21.6	24.3	21.6	16.2	27	24.3	Total 22.75
A%	73	70.3	73	78.4	62.2	67.6	64.9	64.9	75.7	78.4	73	64.9	70.5
D%	0	2.7	8	13.5	8.1	5.4	13.5	10.8	2.7	5.4	0	10.8	6.75
Tot.	100	100	100	100	100	100	100	100	100	100	100	100	100.

3 - Expérience de discrimination de l'Identification

Nous rappelons que nous disposons des données suivantes : 12 locuteurs hommes, 37 paramètres et 10 répétitions pour chaque locuteur et paramètre. Nous avons, au paragraphe précédent, précisé les liaisons entre les paramètres à l'aide de la population de référence.

Notre but est maintenant de travailler (discriminer) sur cette population grâce aux paramètres dont nous disposons, sans sélection de certains d'entre eux a priori. Seule une synthèse de ces deux approches permettra de tirer des conclusions non hâtives.

Précisons que les classes (locuteurs) ont une probabilité a priori égale (1/12). Les locuteurs sont étiquettés sur trois caractères de la façon suivante :

<caractère ordinal> <identificateur sur 2 caractères>.

0 est mis pour 10, $ pour 11 et # pour 12.

Le programme utilisé est celui d'analyse discriminante pas à pas du Biomedical Computer Program.

Rappel sur BMD07M

Le programme effectue une analyse discriminante sur plusieurs groupes. On choisit ici comme critère d'entrée d'un paramètre dans la base son F-ratio. Les possibilités a priori des groupes sont utilisées ainsi qu'un ensemble de classification linéaires, pour calculer les probabilités a posteriori.

Au début, les paramètres sont rangés par F-ratio décroissants ; à chaque étape, on prend en compte celui dont le F-ratio est le plus élevé parmi ceux restants. A chaque étape les F-ratios sont recalculés, et nous appellerons base l'ensemble des paramètres pris en compte. Un paramètre peut à chaque étape, être exclu de la base si son F-ratio devient trop petit.

Pour la mise en oeuvre, il a été utilisé un IBM 360-67 sous OS-MVT. Il ne sera pas donné à chaque fois l'estimation de temps de calcul étant donné que les spécifications sont valables sur IBM 7094 et que les résultats présentés ici ne concernent qu'une partie des calculs effectués. Cependant le tableau suivant peut donner des indications au lecteur. On rappelle que l'on traite 37 paramètres et que l'on va extraire les 10 meilleurs.

Expérience	Temps de calcul
	secondes
Discrimination	Apprentissage 2.03
A	Apprentissage et test 1.27
B	" 2.27
C	" 1.37
D	" 2.41

3.1. Expérience de Discrimination

La collection des données forme un unique ensemble d'apprentissage. Nous étudions la finesse de la discrimination obtenue avec 10 "meilleurs" paramètres parmi les 37 : le critère de sélection des paramètres est le F-ratio, précédemment décrit et utilisé.

La séquence ordonnée des 10 paramètres obtenus est la suivante :

1,22,37,31,17,3,10,25,20,29

(à comparer avec le 9-uplet choisi en 2 : voir le 4).

Détail de l'expérience

a) Tableau des F-ratio avant traitement :

1	16.4409	7	1.6295	13	1.4683	19	0.2598	25	3.4106	31	11.0539	37	10.2709
2	5.3625	8	7.7760	14	2.0736	20	5.9131	26	3.5041	32	3.4246		
3	5.5586	9	1.6567	15	1.6773	21	6.6971	27	6.5071	33	1.7211		
4	5.5165	10	4.4823	16	0.9051	22	11.1455	28	3.2919	34	3.1803		
5	1.2487	11	3.0667	17	7.5206	23	1.0276	29	6.4727	35	2.3799		
6	2.2205	12	1.4740	18	7.5260	24	5.1067	30	0.1065	36	1.8907		

b) Valeurs des seuils :

Seuil inférieur d'acception dans la base d'un paramètre $F \leq 0.01$

Seuil supérieur de rejet de la base d'un paramètre $F \geq 0.005$

c) Résultats.

Tableau des carrés des distances de MAHALANOBIS de chaque répétition locuteur, et probabilité a posteriori des locuteurs.

Exemple : La 2ème répétition du locuteur 1AB est à une distance élevée au carré de 0,636 du centre de sa classe et à 28.562 unités du locuteur 9LC. Les densités de probabilités $P_i(X)$ respectives sont : 0,995 et 0,000. Ainsi, avec une règle de décision a posteriori de Bayes, on décide d'affecter la répétition 2 de AB au locuteur AB, c'est-à-dire d'écrire $\delta(X) = 1$, puisque : $\frac{1}{12} > \rho_1(X)$ $\frac{1}{12}$ $\rho_i(X)$, $i = 2,\ldots,12$

EXP. DISC.

GROUP WITH LARGEST PROB.		SQUARE OF DISTANCE FROM AND POSTERIOR PROBABILITY FOR GROUP -											
GROUP		1AB 8HA		2BC 9LC		3BE OLU		4BD $MA		5BU #WI		6CN	7CD
CASE	J(X)												
1	1AB	1.338	0.984,	15.389	0.001,	22.392	0.000,	28.019	0.000,	11.101	0.007,	18.127 0.000,	11.109 0.007,
		45.839	0.000,	24.916	0.000,	26.535	0.000,	26.170	0.000,	24.845	0.000,		
2	1AB	0.636	0.995,	18.772	0.000,	25.320	0.000,	29.911	0.000,	12.074	0.003,	19.570 0.000,	13.054 0.002,
		52.797	0.000,	28.562	0.000,	30.883	0.000,	29.171	0.000,	28.724	0.000,		
3	1AB	0.292	0.996,	17.546	0.000,	24.376	0.000,	27.469	0.000,	12.905	0.002,	19.030 0.000,	13.107 0.002,
		49.830	0.000,	27.522	0.000,	29.865	0.000,	27.124	0.000,	28.149	0.000,		
4	1AB	0.332	0.994,	16.235	0.000,	21.112	0.000,	23.226	0.000,	12.270	0.003,	16.858 0.000,	12.102 0.003,
		46.665	0.000,	26.850	0.000,	28.200	0.000,	24.027	0.000,	25.553	0.000,		
5	1AB	0.940	0.990,	17.460	0.000,	18.921	0.000,	22.363	0.000,	13.130	0.002,	16.826 0.000,	10.986 0.007,
		45.910	0.000,	28.374	0.000,	29.258	0.000,	22.325	0.000,	25.599	0.000,		
6	1AB	1.165	0.959,	19.837	0.000,	28.956	0.000,	26.441	0.000,	20.359	0.000,	22.631 0.000,	16.997 0.000
		50.801	0.000,	29.715	0.000,	30.737	0.000,	24.753	0.000,	31.017	0.000,		
7	1AB	0.806	0.999,	21.277	0.000,	24.063	0.000,	25.927	0.000,	15.925	0.001,	21.077 0.000,	16.096 0.000
		54.849	0.000,	34.861	0.000,	35.293	0.000,	27.845	0.000,	29.895	0.000,		
8	1AB	0.648	0.999,	17.029	0.000,	25.832	0.000,	25.609	0.000,	16.784	0.000,	21.071 0.000,	15.938 0.000
		48.154	0.000,	31.360	0.000,	32.059	0.000,	24.146	0.000,	28.650	0.000,		
9	1AB	0.610	0.999,	19.895	0.000,	25.535	0.000,	25.831	0.000,	15.562	0.001,	21.575 0.000,	14.864 0.001
		53.072	0.000,	32.105	0.000,	32.263	0.000,	26.067	0.000,	29.044	0.000,		
10	1AB	0.237	0.998,	19.656	0.000,	25.495	0.000,	27.185	0.000,	14.987	0.001,	19.209 0.000,	14.506 0.001
		52.873	0.000,	31.457	0.000,	33.187	0.000,	27.222	0.000,	29.688	0.000,		

On observe dans tous les cas que les répétitions sont a posteriori affectées, au sens de cette règle, au locuteur qui l'a prononcée. Ce résultat est valable pour tous les locuteurs.
Le tableau récapitulatif de confusion ci-dessous indique que, par exemple, le locuteur 1 a été classé 10 fois (il y a 10 répétitions) comme étant le locuteur 1.

EXP. DISC.	NUMBER OF CASES CLASSIFIED INTO GROUP -											
	1AB	2BC	3BE	4BD	5BU	6CN	7CD	8HA	9LC	OLU	$MA	#WI
1AB	10	0	0	0	0	0	0	0	0	0	0	0
2BC	0	10	0	0	0	0	0	0	0	0	0	0
3BE	0	0	10	0	0	0	0	0	0	0	0	0
4BD	0	0	0	10	0	0	0	0	0	0	0	0
5BU	0	0	0	0	10	0	0	0	0	0	0	0
6CN	0	0	0	0	0	10	0	0	0	0	0	0
7CD	0	0	0	0	0	0	10	0	0	0	0	0
8HA	0	0	0	0	0	0	0	10	0	0	0	0
9LC	0	0	0	0	0	0	0	0	10	0	0	0
OLU	0	0	0	0	0	0	0	0	0	10	0	0
$MA	0	0	0	0	0	0	0	0	0	0	10	0
#WI	0	0	0	0	0	0	0	0	0	0	0	10

Le tableau suivant dresse la liste des paramètres adoptés ainsi que la variance expliquée cumulativement. Ainsi, les quatre premiers :

1,22,37,31

réunissent 88% de la dispertion totale pour tous les locuteurs.

EXP. DISC.

STEP NUMBER	VARIABLE ENTERED REMOVED	F VALUE TO ENTER OR REMOVE	NUMBER OF VARIABLES INCLUDED
1	1	16.4419	1
2	22	10.5645	2
3	37	5.9043	3
4	31	7.0636	4
5	17	5.4457	5
6	3	4.5651	6
7	10	3.0878	7
8	25	2.8715	8
9	29	2.5090	9
10	24	3.2702	10

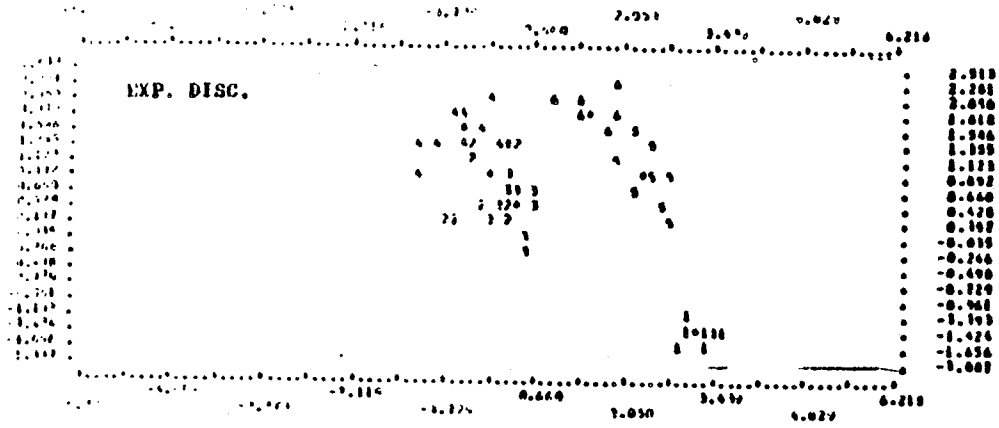
CUMULATIVE PROPORTION OF TOTAL DISPERSION

0.39381	0.62331	0.76453	0.87540	0.92753
0.96459	0.98536	0.99580	0.99913	1.00000

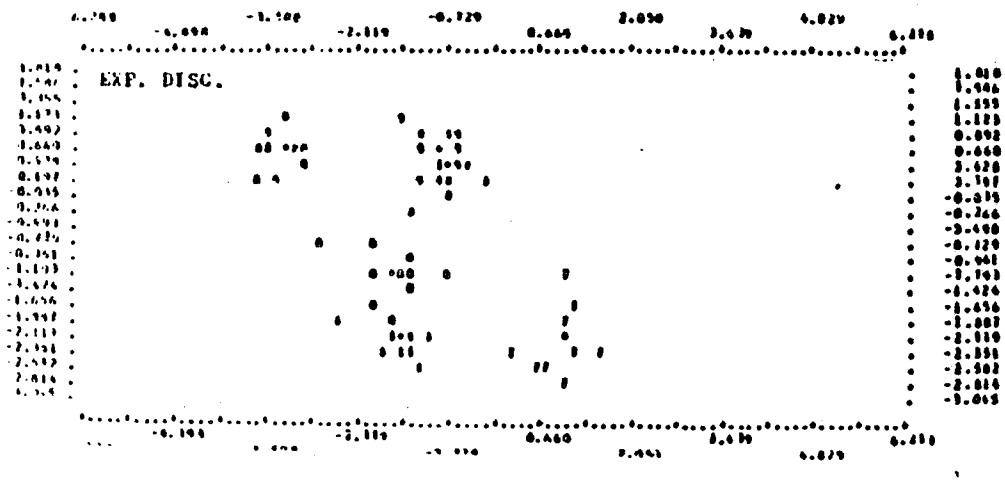
Les deux figures suivantes, qui se superposent, offrent une représentation visuelle des locuteurs (chaque point est une répétition, et "*" indique le centre) dans le plan d'inertie. Ces schémas permettent d'apprécier la variabilité intra-locuteur d'un locuteur à l'autre : faible pour les locuteurs 1, 8,9,11 et 12) elle est beaucoup plus importante pour les locuteurs 5 et 7 (1 les tableaux de l'étude du paragraphe précédent le laissaïent prévoir. On s'aperçoit que seules se recouvrent vraiment, les répétitions des locuteurs 2 et 3, 2 et 4, 4 et 9, 9 et 12. On peut considérer que la discrimination est très bonne avec ces dix paramètres sans toutefois chercher ici à la quantifier.

---

(1) Nous noterons dans les expériences A à D de grandes difficultés pour identifier le locuteur 5.



Représentation bidimensionnelle optimale de la séparation des groupes. Les deux graphiques doivent être superposés.



3.2. Expérience d'Identification

3.2.1. Présentation

Nous décrivons quatre expériences, différant par :
- le nombre de répétitions d'appprentissage (mais en conséquence directe le nombre de répétitions de test) et/ou
- le numéro de ces répétitions,

suivant le tableau récapitulatif ci-dessous :

	Expérience	A	B	C	D
Répétitions	Nombre de répétitions de test	2	2	2	4
	Numéros des répétitions de test	9,10	5,10	4,8	réparties uniformément

Comme au paragraphe 3.1. les paramètres entrent dans la base suivant le critère du F-ratio.

3.2.2. Contrôle des erreurs et évaluation de performance

Définissons, étant donnée une base de paramètres, la matrice de confusion d'apprentissage.

$E_a : J \times J \rightarrow \mathbf{N}$,

par :

$E_a(j_1,j_2)$ = nombre de fois où le locuteur j_1 a été classé a posterior comme étant le locuteur j_2 (fausse identification).

La matrice de confusion de test se définit comme (K est l'ensemble de 10 répétitions)

$E_t : \mathcal{P}(K) \times J \rightarrow \mathbf{N}$,

$E_t(\{k_1,\ldots,k_p\},j)$ = nombre de fois ou les répétitions de test $k_1,\ldots,k_n$ sont classées comme provenant du locuteur j.

Pour les deux raisons suivantes :

- faible nombre des répétitions (10) qui oblige à prendre peu de répétitions pour le test,
- lisibilité des matrices, et étude de l'effet de série dans les répétitions

nous prendrons spécialement le cas où p = 1 ; ainsi :

$$E_t(k,j) = \begin{cases} 1 \text{ si la répétition } k \text{ est classée comme provenant du locuteur } j, \\ 0 \text{ sinon.} \end{cases}$$

Etant donné que nous sommes dans un contexte de simulation et que nous connaissons le bon classement par avance, nous pouvons définir les indices de (bonne) reconnaissance suivants, en ordonnant également pour des clartés de lecture de la même façon les lignes et les colonnes des matrices. Ainsi, les matrices idéales sont diagonales.

Nous distinguons :

1 - Pour l'apprentissage les :

a) taux de reconnaissance a posteriori (TRAP) global :

$$t_g = \frac{\sum_{j=1}^{12} E_t(j,j)}{\sum_{j_1,j_2=1}^{12} E_t(j_1,j_2)} = \frac{\sum_{j=1}^{12} E_t(j,j)}{96},$$

b) taux de reconnaissance a posteriori (TRAP) local (pour le locuteur j_1) :

$$t_e(j_1) = \frac{E_t(j_1,j_1)}{\sum_{j_2=1}^{12} E_t(j_1,j_2)} = \frac{E_t(j_1,j_1)}{8}$$

qui sont les pourcentages pouvant prendre les valeurs $(n.\frac{100}{m})n = 0,m$ où m est le nombre de répétitions d'apprentissage.

Evidemment : $\frac{1}{12} \sum_{j_1=1}^{12} t_e(j_1) = t_g$.

2 - Pour le test, la définition des deux indices est la même moyennant la prise en compte de parties $\{k_1,\ldots,k_p\}$ homogènes c'est-à-dire provenant du même locuteur pour le taux local.

Nous les appellerons les taux d'identification (TI) global [illegible]

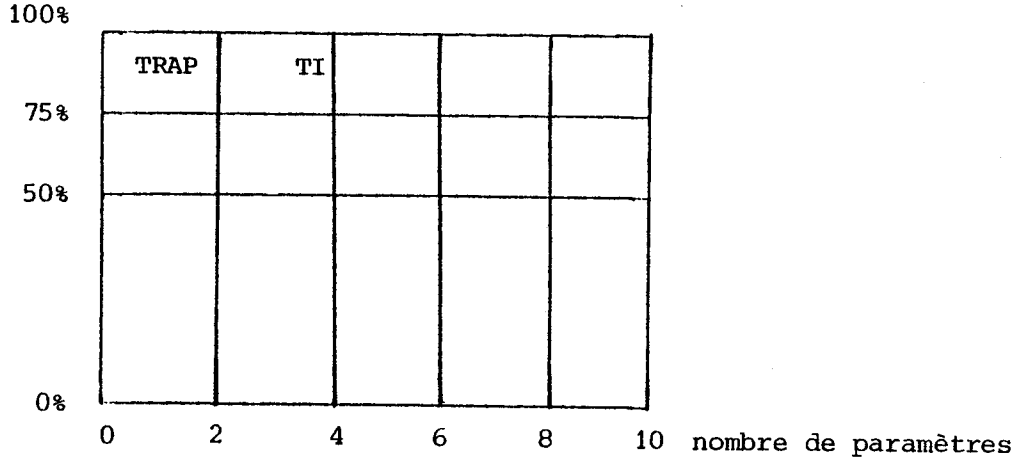
3.2.2. Expérience A

Données de test : 9ème et 10ème répétitions.

Données d'apprentissage : les 8 premières répétitions.

Un taux de reconnaissance de 100% est atteint dès le 8ème paramètre ; les deux tableaux suivants résument les valeurs du taux toutes les deux étapes.

Etape	Eléments de la base	Définitions paramètres	TRAP global %	TI global %
2	1,31	TXFM,P3PN	85	79
4	37,22	P6PN,P3OL	99	92
6	17,3	P10S,TXDB	99	100
8	4,10	T1FM,T3G1	100	100



Taux de reconnaissance a posteriori et d'identification cumulés selon le cardinal de la base courante.

On trouvera ci-dessus les matrices de confusion d'apprentissage et de test pour les mêmes étapes. Chaque ligne correspond à une seule répétition, ce qui permet d'étudier les différences entre deux répétitions (les mêmes numéros de répétions interviennent tout au long de l'expérience à la fois pour le test et l'apprentissage).

(Enfin, on présente le meilleur plan de discrimination avec la position des deux échantillons de référence par locuteur).

Tableau des F-ratio après l'expérience.

STEP NUMBER	VARIABLE ENTERED	REMOVED	VALUE TO ENTER OR REMOVE	NUMBER CP VARIABLES INCLUDED	U STATISTI
1	1		13.2800	1	0.3651
2	31		0.1006	2	0.1646
3	37		6.7520	3	
4	22		5.6667	4	0.0488
5	17		3.7423	5	0.0322
6	3			6	0.0217
7	4		2.5044	7	0.0160
8	10		2.6270	8	0.0117
9	25			9	0.0088
10	29		2.1508	10	0.0067

Etape 2 A1 et A2 sont les tests BC ou locuteur BC
B1 et B2 du locuteur B , etc.

A2

GROUP	1AB	2BC	3BE	4BC	5BU	6CN	7CC	8HA	9LC	0LU	$MA	#WI
1AB	8	0	0	0	0	0	0	0	0	0	0	0
2BC	0	8	1	0	0	0	0	0	0	0	0	1
3BE	0	1	4	0	0	0	0	0	3	0	0	0
4BC	0	0	0	8	0	0	0	0	0	0	0	0
5BU	2	0	0	0	6	0	0	0	0	0	0	0
6CN	0	0	0	0	0	8	0	0	0	0	0	0
7CC	0	0	0	0	0	0	8	0	0	0	0	0
8HA	0	1	0	0	0	0	0	7	0	0	0	0
9LC	0	0	2	0	1	0	0	0	5	0	0	0
0LU	0	0	0	0	0	0	1	0	0	7	0	0
$MA	0	0	0	0	0	0	0	0	0	0	8	0
#WI	0	0	0	0	0	0	0	1	0	0	0	7
A1	1	0	0	0	0	0	0	0	0	0	0	0
A2	1	0	0	0	0	0	0	0	0	0	0	0
B1	0	1	0	0	0	0	0	0	0	0	0	0
B2	0	1	0	0	0	0	0	0	0	0	0	0
C1	0	0	1	0	0	0	0	0	0	0	0	0
C2	0	0	0	0	1	0	0	0	0	0	0	0
D1	0	0	0	1	0	0	0	0	0	0	0	0
D2	0	0	0	1	0	0	0	0	0	0	0	0
E1	1	0	0	0	0	0	0	0	0	0	0	0
E2	1	0	0	0	0	0	0	0	0	0	0	0
F1	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	0	0	0	0	1	0	0	0	0	0	0
G1	0	0	0	0	0	0	1	0	0	0	0	0
G2	0	0	0	0	0	0	1	0	0	0	0	0
H1	0	0	0	0	0	0	0	1	0	0	0	0
H2	0	0	0	0	0	0	0	1	0	0	0	0
I1	0	0	0	0	0	0	0	0	1	0	0	0
I2	0	0	0	0	0	0	0	0	1	0	0	0
J1	0	0	0	0	0	0	0	0	0	1	0	0
J2	0	0	0	0	0	0	0	0	0	1	0	0
K1	0	0	0	0	0	0	0	0	0	0	1	0
K2	0	0	0	0	0	0	0	0	0	0	1	0
L1	0	0	0	0	0	0	0	0	0	0	0	1
L2	0	0	0	0	0	0	0	0	0	0	0	1

Etape 4

A4

GROUP	1AB	2BC	3BE	4BC	5BU	6CN	7CC	8HA	9LC	0LU	$MA	#WI
1AB	8	0	0	0	0	0	0	0	0	0	0	0
2BC	0	8	0	0	0	0	0	0	0	0	0	0
3BE	0	0	8	0	0	0	0	0	0	0	0	0
4BC	0	0	0	8	0	0	0	0	0	0	0	0
5BU	0	0	0	0	7	1	0	0	0	0	0	0
6CN	0	0	0	0	0	8	0	0	0	0	0	0
7CC	0	0	0	0	0	0	8	0	0	0	0	0
8HA	0	0	0	0	0	0	0	8	0	0	0	0
9LC	0	0	0	0	0	0	0	0	8	0	0	0
0LU	0	0	0	0	0	0	0	0	0	8	0	0
$MA	0	0	0	0	0	0	0	0	0	0	8	0
#WI	0	0	0	0	0	0	0	0	0	0	0	8
A1	1	0	0	0	0	0	0	0	0	0	0	0
A2	1	0	0	0	0	0	0	0	0	0	0	0
B1	0	1	0	0	0	0	0	0	0	0	0	0
B2	0	1	0	0	0	0	0	0	0	0	0	0
C1	0	0	1	0	0	0	0	0	0	0	0	0
C2	0	0	0	0	1	0	0	0	0	0	0	0
D1	0	0	0	1	0	0	0	0	0	0	0	0
D2	0	0	0	1	0	0	0	0	0	0	0	0
E1	0	0	0	0	1	0	0	0	0	0	0	0
E2	1	0	0	0	0	0	0	0	0	0	0	0
F1	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	0	0	0	0	1	0	0	0	0	0	0
G1	0	0	0	0	0	0	1	0	0	0	0	0
G2	0	0	0	0	0	0	1	0	0	0	0	0
H1	0	0	0	0	0	0	0	1	0	0	0	0
H2	0	0	0	0	0	0	0	1	0	0	0	0
I1	0	0	0	0	0	0	0	0	1	0	0	0
I2	0	0	0	0	0	0	0	0	1	0	0	0
J1	0	0	0	0	0	0	0	0	0	1	0	0
J2	0	0	0	0	0	0	0	0	0	1	0	0
K1	0	0	0	0	0	0	0	0	0	0	1	0
K2	0	0	0	0	0	0	0	0	0	0	1	0
L1	0	0	0	0	0	0	0	0	0	0	0	1
L2	0	0	0	0	0	0	0	0	0	0	0	1

Etape 6

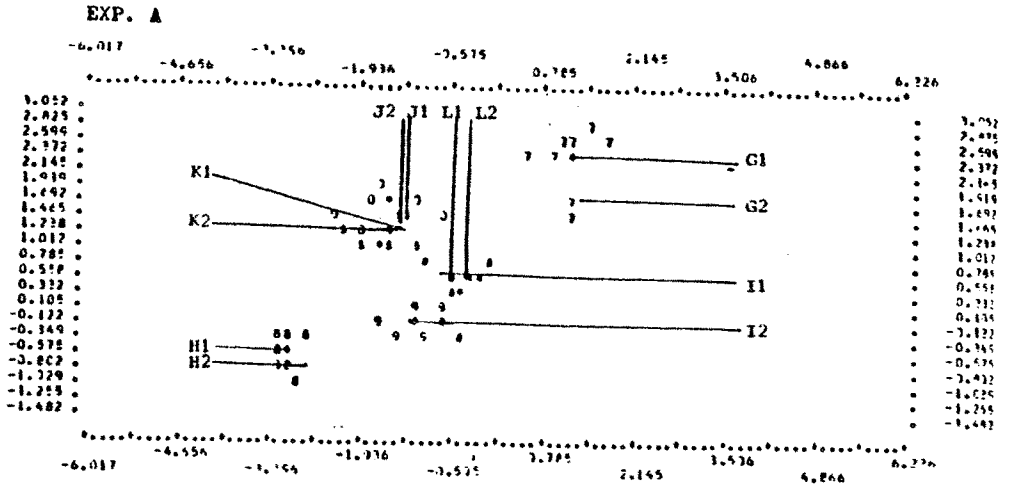
A6

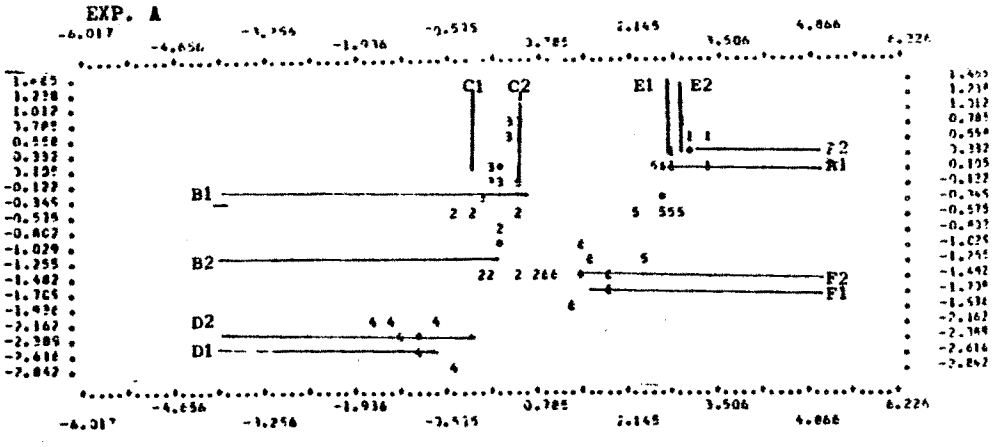
GROUP	NUMBER OF CASES CLASSIFIED INTO GROUP -											
	1AB	2BC	3BE	4BD	5BU	6CN	7CO	8HA	9LC	0LU	$MA	#WI
1AB	8	0	0	0	0	0	0	0	0	0	0	0
2BC	0	8	0	0	0	0	0	0	0	0	0	0
3BE	0	0	8	0	0	0	0	0	0	0	0	0
4BD	0	0	0	8	0	0	0	0	0	0	0	0
5BU	0	0	0	0	7	1	0	0	0	0	0	0
6CN	0	0	0	0	0	8	0	0	0	0	0	0
7CO	0	0	0	0	0	0	8	0	0	0	0	0
8HA	0	0	0	0	0	0	0	8	0	0	0	0
9LC	0	0	0	0	0	0	0	0	8	0	0	0
0LU	0	0	0	0	0	0	0	0	0	8	0	0
$MA	0	0	0	0	0	0	0	0	0	0	8	0
#WI	0	0	0	0	0	0	0	0	0	0	0	8
A1	1	0	0	0	0	0	0	0	0	0	0	0
A2	1	0	0	0	0	0	0	0	0	0	0	0
B1	0	1	0	0	0	0	0	0	0	0	0	0
B2	0	1	0	0	0	0	0	0	0	0	0	0
C1	0	0	1	0	0	0	0	0	0	0	0	0
C2	0	0	1	0	0	0	0	0	0	0	0	0
D1	0	0	0	1	0	0	0	0	0	0	0	1
D2	0	0	0	1	0	0	0	0	0	0	0	0
E1	0	0	0	0	1	0	0	0	0	0	0	0
E2	0	0	0	0	1	0	0	0	0	0	0	0
F1	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	0	0	0	0	1	0	0	0	0	0	0
G1	0	0	0	0	0	0	1	0	0	0	0	0
G2	0	0	0	0	0	0	1	0	0	0	0	0
H1	0	0	0	0	0	0	0	1	0	0	0	0
H2	0	0	0	0	0	0	0	1	0	0	0	0
I1	0	0	0	0	0	0	0	0	1	0	0	0
I2	0	0	0	0	0	0	0	0	1	0	0	0
J1	0	0	0	0	0	0	0	0	0	1	0	0
J2	0	0	0	0	0	0	0	0	0	1	0	0
K1	0	0	0	0	0	0	0	0	0	0	1	0
K2	0	0	0	0	0	0	0	0	0	0	1	0
L1	0	0	0	0	0	0	0	0	0	0	0	1
L2	0	0	0	0	0	0	0	0	0	0	0	1

Etape 8

A8

GROUP	NUMBER OF CASES CLASSIFIED INTO GROUP -											
	1AB	2BC	3BE	4BD	5BU	6CN	7CO	8HA	9LC	0LU	$MA	#WI
1AB	8	0	0	0	0	0	0	0	0	0	0	0
2BC	0	8	0	0	0	0	0	0	0	0	0	0
3BE	0	0	8	0	0	0	0	0	0	0	0	0
4BD	0	0	0	8	0	0	0	0	0	0	0	0
5BU	0	0	0	0	8	0	0	0	0	0	0	0
6CN	0	0	0	0	0	8	0	0	0	0	0	0
7CO	0	0	0	0	0	0	8	0	0	0	0	0
8HA	0	0	0	0	0	0	0	8	0	0	0	0
9LC	0	0	0	0	0	0	0	0	8	0	0	0
0LU	0	0	0	0	0	0	0	0	0	8	0	0
$MA	0	0	0	0	0	0	0	0	0	0	8	0
#WI	0	0	0	0	0	0	0	0	0	0	0	8
A1	1	0	0	0	0	0	0	0	0	0	0	0
A2	1	0	0	0	0	0	0	0	0	0	0	0
B1	0	1	0	0	0	0	0	0	0	0	0	0
B2	0	1	0	0	0	0	0	0	0	0	0	0
C1	0	0	1	0	0	0	0	0	0	0	0	0
C2	0	0	1	0	0	0	0	0	0	0	0	0
D1	0	0	0	1	0	0	0	0	0	0	0	0
D2	0	0	0	1	0	0	0	0	0	0	0	0
E1	0	0	0	0	1	0	0	0	0	0	0	0
E2	0	0	0	0	1	0	0	0	0	0	0	0
F1	0	0	0	0	0	1	0	0	0	0	0	0
F2	0	0	0	0	0	1	0	0	0	0	0	0
G1	0	0	0	0	0	0	1	0	0	0	0	0
G2	0	0	0	0	0	0	1	0	0	0	0	0
H1	0	0	0	0	0	0	0	1	0	0	0	0
H2	0	0	0	0	0	0	0	1	0	0	0	0
I1	0	0	0	0	0	0	0	0	1	0	0	0
I2	0	0	0	0	0	0	0	0	1	0	0	0
J1	0	0	0	0	0	0	0	0	0	1	0	0
J2	0	0	0	0	0	0	0	0	0	1	0	0
K1	0	0	0	0	0	0	0	0	0	0	1	0
K2	0	0	0	0	0	0	0	0	0	0	1	0
L1	0	0	0	0	0	0	0	0	0	0	0	1
L2	0	0	0	0	0	0	0	0	0	0	0	1


EXP. A
J2
J1
L1
L2
K1
K2
G1
G2
I1
I2
H1
H2


EXP. A
C1
C2
E1
E2
F2
F1
B1
B2
D2
D1

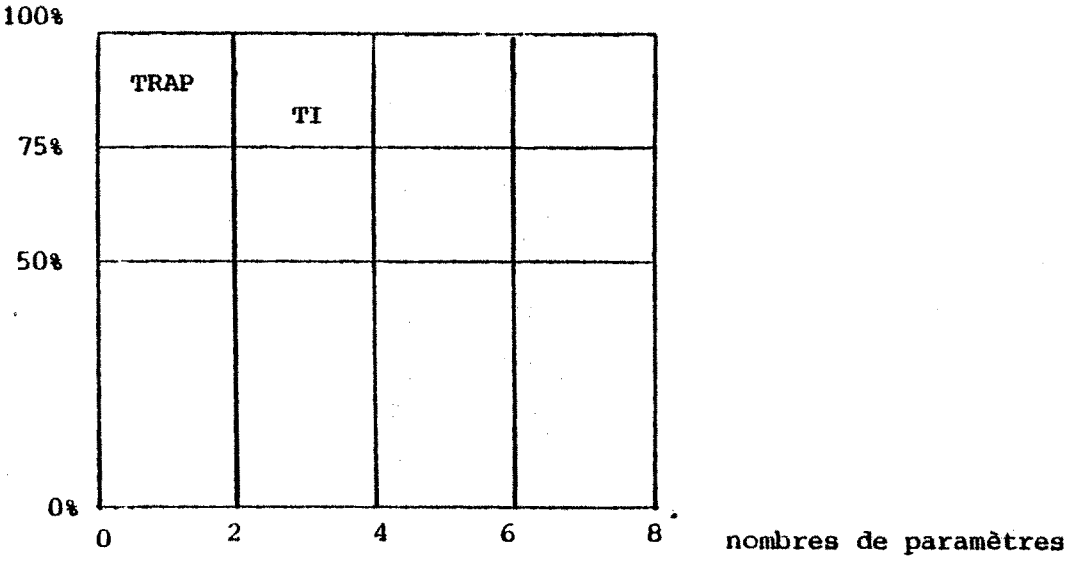
3.2.3. Expérience B

Données de test : les 5ème et 10ème répétitions

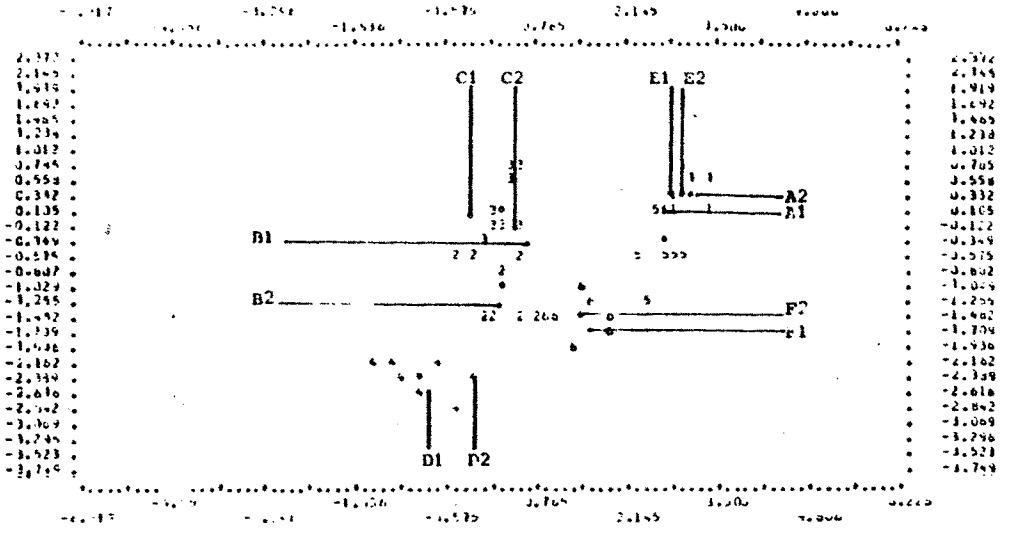
Données d'apprentissage : les 8 autres répétitiont

Le taux de reconnaissance de 100% est atteint lorsque la base a 6 éléments :

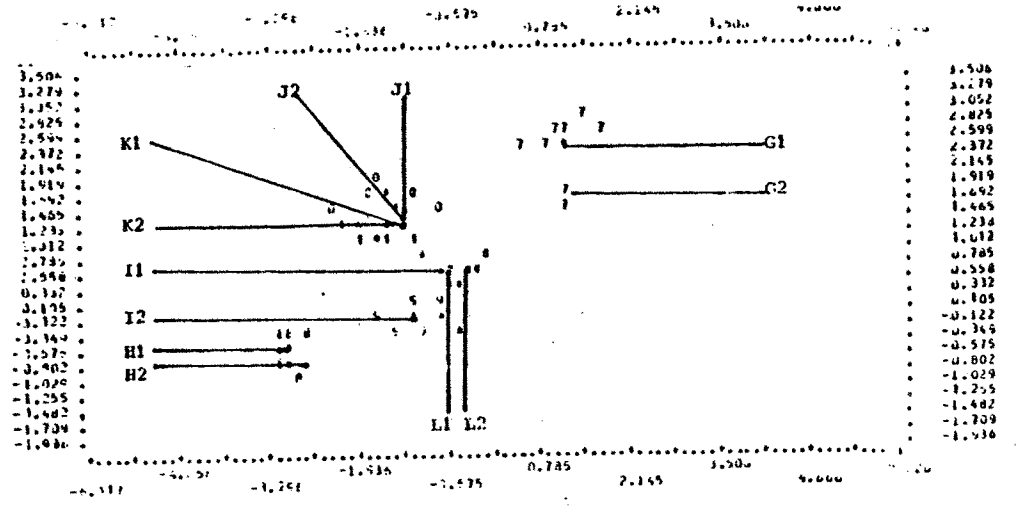
Etape	Eléments de la base	Définition paramètres	TRAP global %	TI global %
2	1,31	TXFM,P3PN	85	79
4	37,22	P6PN,P3OL	99	92
6	17,3	P10S,TXDB	99	100
8	4,10	T1FM,T30L	100	100



EXP. B



EXP. B

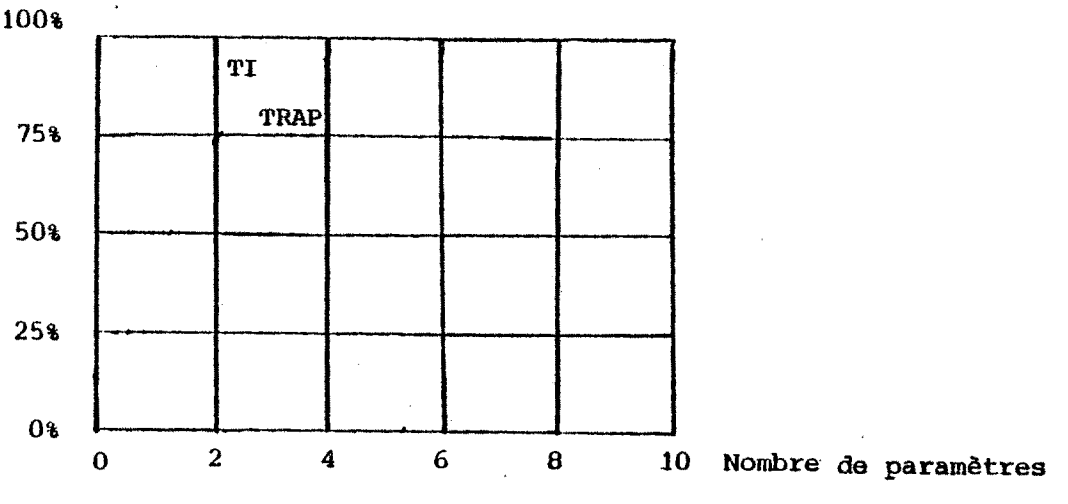


3.2.4. Expérience C

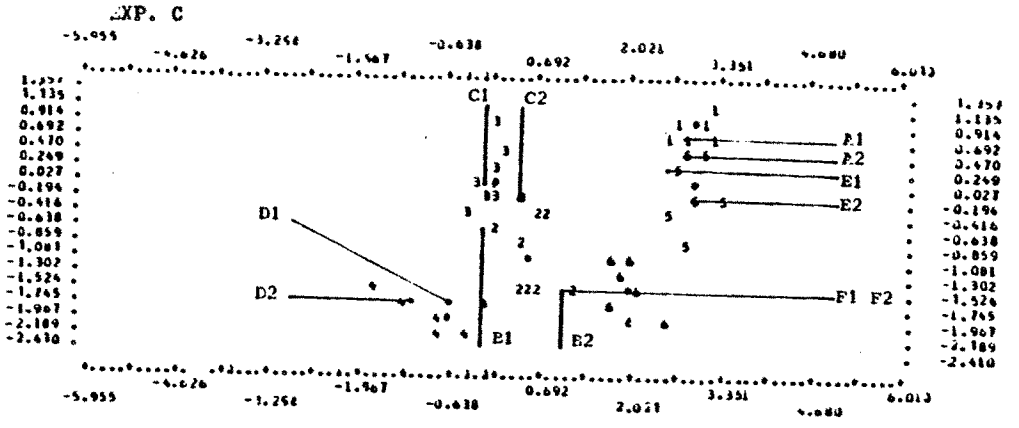
Données de test : 4ème et 8ème répétitions

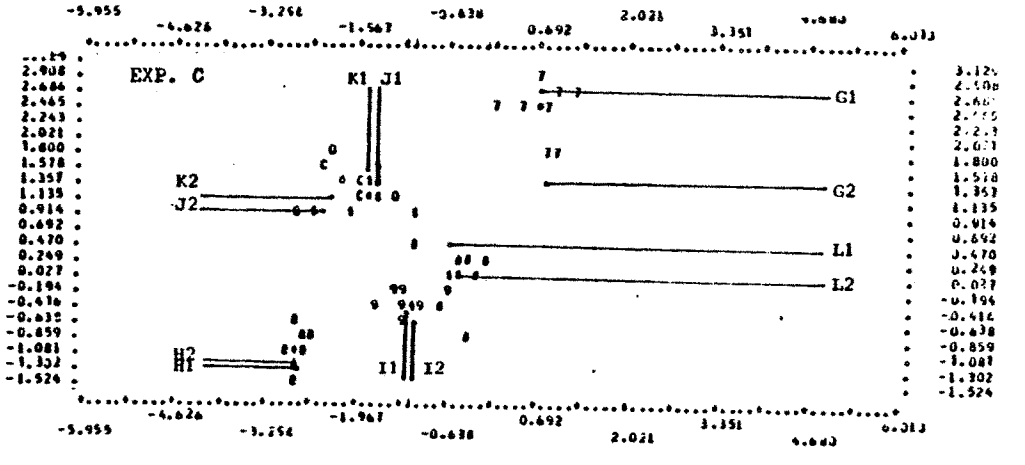
Données d'apprentissage : les 8 autres répétitions

Etape	Eléments de la base	Définition paramètres	TRAP %	TI %
2	1,22	TXFM,P3OL	75	71
4	37,31	P6PN,P3PN	98	100
6	17,3	P10S,TXDB	99	100
8	25,28	P4PH,P1PI	99	100
10	10,32	T3G1,P3PI	100	100



La classification parfaite a posteriori est atteinte au bout de 10 étapes alors qu'il en avait fallu seulement 6 dans l'expérience C. On voit ici l'influence du mode de la séparation des données.


EXP. C
C1
C2
D1
D2
E1
E2
F1
F2
B1
B2


EXP. C
K1
J1
K2
J2
G1
G2
L1
L2
H2
H1
I1
I2

3.2.5. Expérience D

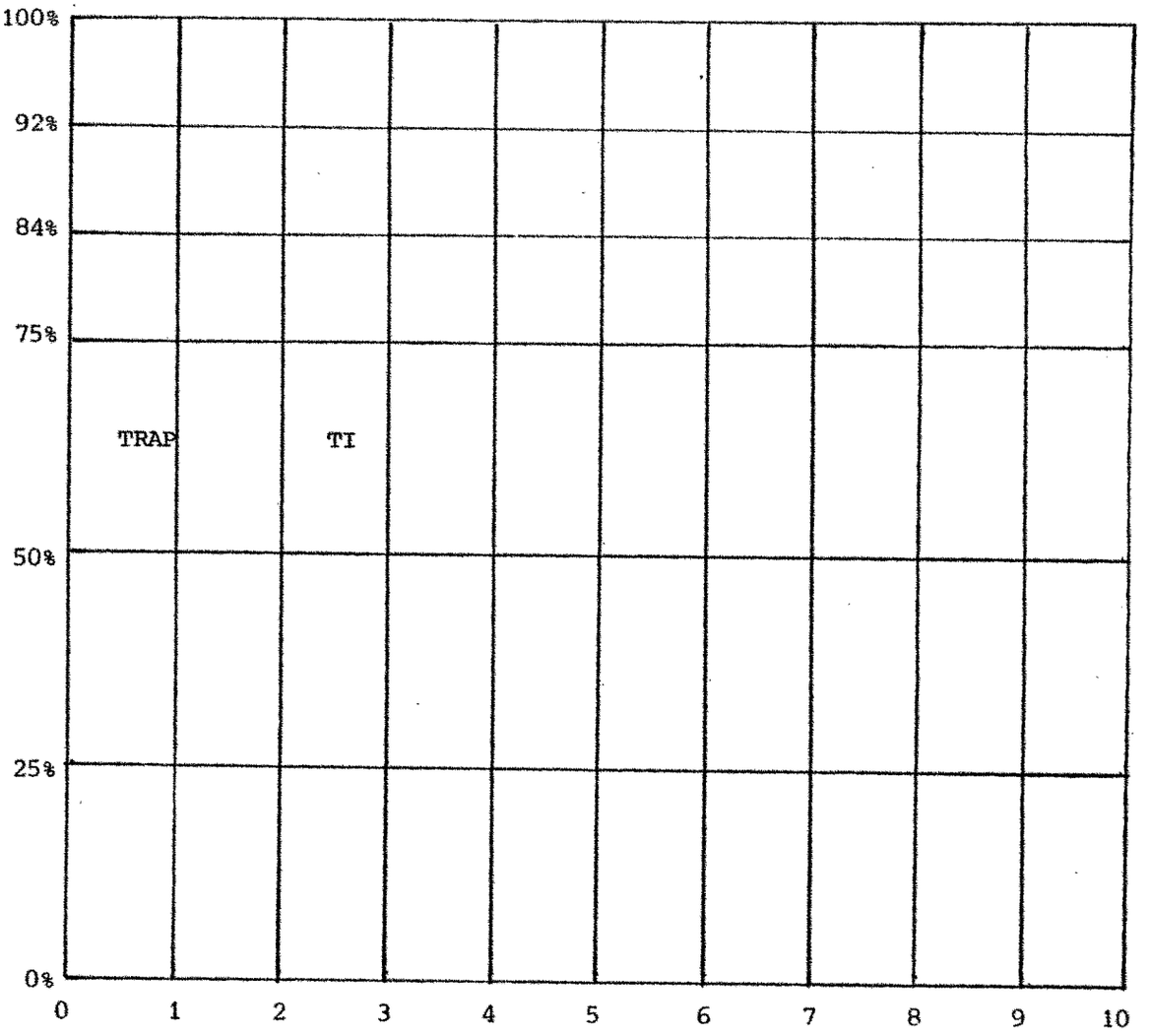
Ici les données d'apprentissage sont 6 répétitions et les données de test les 4 autres, réparties uniformément.

A l'apprentissage, les paramètres sont choisis dans cet ordre :

1,22 37 31,2,18,24,32,29,30

Etape	Elément de la base	Définition paramètre	TRAP %	TI %
1	1	TXFM	50	48
2	22	P3OL	72	62
3	37	P6PN	89	92
4	31	P3PN	99	98
5	3	TXDB	97	94
6	18	P1OL	97	98
7	24	P6OL	100	98
8	32	P3PI	100	96
9	29	P2PN	100	98
10	30	P2PI	100	94

On observe dans cette expérience un phénomène remarquable. C'est la non croissance de TI en fonction du nombre de paramètres utilisé. Nous en donnons l'explication suivante : la complexité du classificateur croît au-fur et à-mesure que la base a plus d'éléments et cette complexité nuit à la performance globale. Il y aurait donc un nombre optimal de paramètres par exemple les 4 ou 7 premiers.


100%
92%
84%
75%
50%
25%
0%
TRAP
TI
0
1
2
3
4
5
6
7
8
9
10
Nombre de paramètres

F. Ratios

1	11.0803	7	1.0357	13	1.1148	19	2.7240
2	3.8517	8	4.7404	14	1.0529	20	4.1359
3	4.2794	9	1.2120	15	1.0110	21	4.9533
4	2.5802	10	3.3669	16	0.7869	22	5.9860
5	0.5902	11	1.4834	17	4.0838	23	0.9392
6	0.8790	12	0.7226	18	1.7615	24	3.8622

25	1.5728	31	6.3437	37	5.6455
26	2.8970	32	2.2838		
27	4.0502	33	0.8093		
28	1.0549	34	1.4835		
29	4.0038	35	1.0234		
30	1.3348	36	2.5008		

Etape I (fondamental Seul)

DI

NUMBER OF CASES CLASSIFIED INTO GROUP -

GROUP	1AB	2BC	3BE	4BD	5EL	6CN	7CD	8HA	9LC	0LD	1HA	2HI
1AB	2	0	0	0	0	0	0	0	1	0	2	1
2BC	0	3	0	0	2	1	0	0	0	0	0	0
3BE	2	1	0	0	2	0	0	0	1	0	0	0
4BD	0	0	0	5	0	1	0	0	0	0	0	0
5EL	0	2	0	0	6	0	0	1	1	0	0	0
6CN	0	0	0	2	0	4	0	0	0	0	0	0
7CD	0	0	0	0	0	0	6	0	0	0	0	0
8HA	0	2	0	0	0	0	0	4	0	0	0	0
9LC	1	0	1	0	0	0	0	0	2	0	0	0
0LD	0	0	0	0	0	0	0	0	0	0	0	0
1HA	1	0	0	0	0	0	0	0	0	0	3	0
2HI	1	1	0	0	0	0	0	0	1	0	0	1
AB1	0	0	0	0	0	0	0	0	0	0	0	0
AB2	0	0	0	0	0	0	0	0	0	0	0	0
AB3	1	0	0	0	0	0	0	0	0	0	0	0
AB4	0	0	0	0	0	0	0	0	0	0	0	0
BC1	0	0	0	0	0	0	0	1	0	0	0	0
BC2	0	0	0	0	0	1	0	0	0	0	0	0
BC3	0	0	0	0	0	1	0	0	0	0	0	0
BC4	0	0	0	0	0	1	0	0	0	0	0	0
BE1	0	1	0	0	0	0	0	0	0	0	0	0
BE2	0	1	0	0	0	0	0	0	0	0	0	0
BE3	0	1	0	0	0	0	0	0	0	0	0	0
BE4	0	0	0	0	0	0	0	1	0	0	0	0
BD1	0	0	0	1	0	0	0	0	0	0	0	0
BD2	0	0	0	1	0	0	0	0	0	0	0	0
BD3	0	0	0	1	0	0	0	0	0	0	0	0
BD4	0	0	0	1	0	0	0	0	0	0	0	0
BE1	0	1	0	0	0	0	0	0	0	0	0	0
BE2	0	1	0	0	0	0	0	0	0	0	0	0
BE3	0	1	0	0	0	0	0	0	0	0	0	0
BE4	0	0	0	0	0	0	0	0	0	0	0	0
CN1	0	0	0	0	0	1	0	0	0	0	0	0
CN2	0	0	0	0	0	1	0	0	0	0	0	0
CN3	0	0	0	0	0	1	0	0	0	0	0	0
CN4	0	0	0	0	0	1	0	0	0	0	0	0
CD1	0	0	0	0	0	0	1	0	0	0	0	0
CD2	0	0	0	0	0	0	0	0	0	1	0	0
CD3	0	0	0	0	0	0	0	0	0	1	0	0
CD4	0	0	0	0	0	0	0	0	0	1	0	0
HA1	0	0	0	0	0	0	0	1	0	0	0	0
HA2	0	0	0	0	0	0	0	1	0	0	0	0
HA3	0	0	0	0	0	0	0	1	0	0	0	0
HA4	0	0	0	0	0	0	0	1	0	0	0	0
LE1	0	0	0	0	1	0	0	0	0	0	0	0
LE2	0	0	0	0	1	0	0	0	0	0	0	0
LE3	0	0	0	0	1	0	0	0	0	0	0	0
LE4	0	0	0	0	1	0	0	0	0	0	0	0
LD1	0	0	0	0	0	0	0	0	0	1	0	0
LD2	0	0	0	0	0	0	0	0	0	1	0	0
LD3	0	0	0	0	0	0	0	0	0	1	0	0
LD4	0	0	0	0	0	0	0	0	0	1	0	0
MA1	0	0	0	0	0	0	0	0	0	0	1	0
MA2	0	0	0	0	0	0	0	0	0	0	1	0
MA3	0	0	0	0	0	0	0	0	0	0	1	0
MA4	0	0	0	0	0	0	0	0	0	0	1	0
HI1	0	0	0	0	1	0	0	0	0	0	0	0
HI2	0	0	1	0	0	0	0	0	0	0	0	0
HI3	0	0	0	0	0	0	0	0	1	0	0	0
HI4	0	0	0	0	0	0	0	0	0	0	0	1

Etape 4

D4

GROUP	NUMBER OF CASES CLASSIFIED INTO GROUP - 1AB	2BC	3BE	4BO	5BL	6CN	7CC	8HA	9LC	OLU	$MA	#WI
1AB	6	0	0	0	0	0	0	0	0	0	0	0
2BC	0	6	0	0	0	0	0	0	0	0	0	0
3BE	0	0	6	0	0	0	0	0	0	0	0	0
4BO	0	0	0	6	0	0	0	0	0	0	0	0
5BU	0	0	0	0	5	1	0	0	0	0	0	0
6CN	0	0	0	0	0	6	0	0	0	0	0	0
7CO	0	0	0	0	0	0	6	0	0	0	0	0
8HA	0	0	0	0	0	0	0	6	0	0	0	0
9LC	0	0	0	0	0	0	0	0	6	0	0	0
OLU	0	0	0	0	0	0	0	0	0	6	0	0
$MA	0	0	0	0	0	0	0	0	0	0	6	0
#WI	0	0	0	0	0	0	0	0	0	0	0	6
AB1	1	0	0	0	0	0	0	0	0	0	0	0
AB2	1	0	0	0	0	0	0	0	0	0	0	0
AB3	1	0	0	0	0	0	0	0	0	0	0	0
AB4	1	0	0	0	0	0	0	0	0	0	0	0
BC1	0	1	0	0	0	0	0	0	0	0	0	0
BC2	0	1	0	0	0	0	0	0	0	0	0	0
BC3	0	1	0	0	0	0	0	0	0	0	0	0
BC4	0	1	0	0	0	0	0	0	0	0	0	0
BE1	0	0	1	0	0	0	0	0	0	0	0	0
BE2	0	0	1	0	0	0	0	0	0	0	0	0
BE3	0	0	1	0	0	0	0	0	0	0	0	0
BE4	0	0	1	0	0	0	0	0	0	0	0	0
BO1	0	0	0	1	0	0	0	0	0	0	0	0
BO2	0	0	0	1	0	0	0	0	0	0	0	0
BO3	0	0	0	1	0	0	0	0	0	0	0	0
BO4	0	0	0	1	0	0	0	0	0	0	0	0
BU1	0	0	0	0	1	0	0	0	0	0	0	0
BU2	0	0	0	0	1	0	0	0	0	0	0	0
BU3	0	0	0	0	1	0	0	0	0	0	0	0
BU4	1	0	0	0	0	0	0	0	0	0	0	0
CN1	0	0	0	0	0	1	0	0	0	0	0	0
CN2	0	0	0	0	0	1	0	0	0	0	0	0
CN3	0	0	0	0	0	1	0	0	0	0	0	0
CN4	0	0	0	0	0	1	0	0	0	0	0	0
CO1	0	0	0	0	0	0	1	0	0	0	0	0
CO2	0	0	1	0	0	0	0	0	0	0	0	0
CO3	0	0	0	0	0	0	1	0	0	0	0	0
CO4	0	0	0	0	0	0	1	0	0	0	0	0
HA1	0	0	0	0	0	0	0	1	0	0	0	0
HA2	0	0	0	0	0	0	0	1	0	0	0	0
HA3	0	0	0	0	0	0	0	1	0	0	0	0
HA4	0	0	0	0	0	0	0	1	0	0	0	0
LC1	0	0	0	0	0	0	0	0	1	0	0	0
LC2	0	0	0	0	0	0	0	0	1	0	0	0
LC3	0	0	0	0	0	0	0	0	1	0	0	0
LC4	0	0	0	0	0	0	0	0	1	0	0	0
LU1	0	0	0	0	0	0	0	0	0	1	0	0
LU2	0	0	0	0	0	0	0	0	0	1	0	0
LU3	0	0	0	0	0	0	0	0	0	1	0	0
LU4	0	0	0	0	0	0	0	0	0	1	0	0
MA1	0	0	0	0	0	0	0	0	0	0	1	0
MA2	0	0	0	0	0	0	0	0	0	0	1	0
MA3	0	0	0	0	0	0	0	0	0	0	1	0
MA4	0	0	0	0	0	0	0	0	0	0	1	0
WI1	0	0	0	0	0	0	0	0	0	0	0	1
WI2	0	0	0	0	0	0	0	0	0	0	0	1
WI3	0	0	0	0	0	0	0	0	0	0	0	1
WI4	0	0	0	0	0	0	0	0	0	0	0	1

Etape 7

D7

GROUP	NUMBER OF CASES CLASSIFIED INTO GROUP - 1AB	2BC	3BE	4BD	5EC	6CN	7CO	8HA	9LC	0LO	$MA	#N1
1AB	6	0	0	0	0	0	0	0	0	0	0	0
2BC	0	6	0	0	0	0	0	0	0	0	0	0
3BE	0	0	6	0	0	0	0	0	0	0	0	0
4BD	0	0	0	6	0	0	0	0	0	0	0	0
5EC	0	0	0	0	6	0	0	0	0	0	0	0
6CN	0	0	0	0	0	6	0	0	0	0	0	0
7CO	0	0	0	0	0	0	6	0	0	0	0	0
8HA	0	0	0	0	0	0	0	6	0	0	0	0
9LC	0	0	0	0	0	0	0	0	6	0	0	0
0LO	0	0	0	0	0	0	0	0	0	6	0	0
$MA	0	0	0	0	0	0	0	0	0	0	6	0
#N1	0	0	0	0	0	0	0	0	0	0	0	6
AB1	1	0	0	0	0	0	0	0	0	0	0	0
AB2	1	0	0	0	0	0	0	0	0	0	0	0
AB3	1	0	0	0	0	0	0	0	0	0	0	0
AB4	1	0	0	0	0	0	0	0	0	0	0	0
BC1	0	1	0	0	0	0	0	0	0	0	0	0
BC2	0	1	0	0	0	0	0	0	0	0	0	0
BC3	0	1	0	0	0	0	0	0	0	0	0	0
BC4	0	1	0	0	0	0	0	0	0	0	0	0
BE1	0	0	1	0	0	0	0	0	0	0	0	0
BE2	0	0	1	0	0	0	0	0	0	0	0	0
BE3	0	0	1	0	0	0	0	0	0	0	0	0
BE4	0	0	1	0	0	0	0	0	0	0	0	0
BD1	0	0	0	1	0	0	0	0	0	0	0	0
BD2	0	0	0	1	0	0	0	0	0	0	0	0
BD3	0	0	0	1	0	0	0	0	0	0	0	0
BD4	0	0	0	1	0	0	0	0	0	0	0	0
BE1	0	0	0	0	1	0	0	0	0	0	0	0
BE2	0	0	0	0	1	0	0	0	0	0	0	0
BE3	0	0	0	0	1	0	0	0	0	0	0	0
BE4	0	0	0	0	1	0	0	0	0	0	0	0
CN1	0	0	0	0	0	1	0	0	0	0	0	0
CN2	0	0	0	0	0	1	0	0	0	0	0	0
CN3	0	0	0	0	0	1	0	0	0	0	0	0
CN4	0	0	0	0	0	1	0	0	0	0	0	0
CO1	0	0	0	0	0	0	1	0	0	0	0	0
CO2	0	0	0	0	0	0	1	0	0	0	0	0
CO3	0	0	0	0	0	0	1	0	0	0	0	0
CO4	0	0	0	0	0	0	1	0	0	0	0	0
HA1	0	0	0	0	0	0	0	1	0	0	0	0
HA2	0	0	0	0	0	0	0	1	0	0	0	0
HA3	0	0	0	0	0	0	0	1	0	0	0	0
HA4	0	0	0	0	0	0	0	1	0	0	0	0
LC1	0	0	0	0	0	0	0	0	1	0	0	0
LC2	0	0	0	0	0	0	0	0	1	0	0	0
LC3	0	0	0	0	0	0	0	0	1	0	0	0
LC4	0	0	0	0	0	0	0	0	1	0	0	0
LO1	0	0	0	0	0	0	0	0	0	1	0	0
LO2	0	0	0	0	0	0	0	0	0	1	0	0
LO3	0	0	0	0	0	0	0	0	0	1	0	0
LO4	0	0	0	0	0	0	0	0	0	1	0	0
MA1	0	0	0	0	0	0	0	0	0	0	1	0
MA2	0	0	0	0	0	0	0	0	0	0	1	0
MA3	0	0	0	0	0	0	0	0	0	0	1	0
MA4	0	0	0	0	0	0	0	0	0	0	1	0
#E1	0	0	0	0	0	0	0	0	0	0	0	1
#E2	0	0	0	0	0	0	0	0	0	0	0	1
#E3	0	0	0	0	0	0	0	0	0	0	0	1
#E4	0	0	0	0	0	0	0	0	0	0	0	1

Etape IO

GROUP \ DIO	NUMBER OF CASES CLASSIFIED INTO GROUP - 1AB	2BC	3BE	4BO	5BL	6CN	7CE	8HA	9LC	OLU	$MA	#WI
1AB	6	0	0	0	0	0	0	0	0	0	0	0
2BC	0	6	0	0	0	0	0	0	0	0	0	0
3BE	0	0	6	0	0	0	0	0	0	0	0	0
4BO	0	0	0	6	0	0	0	0	0	0	0	0
5BU	0	0	0	0	6	0	0	0	0	0	0	0
6CN	0	0	0	0	0	6	0	0	0	0	0	0
7CO	0	0	0	0	0	0	6	0	0	0	0	0
8HA	0	0	0	0	0	0	0	6	0	0	0	0
9LC	0	0	0	0	0	0	0	0	6	0	0	0
OLU	0	0	0	0	0	0	0	0	0	6	0	0
$MA	0	0	0	0	0	0	0	0	0	0	6	0
#WI	0	0	0	0	0	0	0	0	0	0	0	6
AB1	1	0	0	0	0	0	0	0	0	0	0	0
AB2	1	0	0	0	0	0	0	0	0	0	0	0
AB3	1	0	0	0	0	0	0	0	0	0	0	0
AB4	1	0	0	0	0	0	0	0	0	0	0	0
BC1	0	1	0	0	0	0	0	0	0	0	0	0
BC2	0	1	0	0	0	0	0	0	0	0	0	0
BC3	0	1	0	0	0	0	0	0	0	0	0	0
BC4	0	1	0	0	0	0	0	0	0	0	0	0
BE1	0	0	1	0	0	0	0	0	0	0	0	0
BE2	0	0	1	0	0	0	0	0	0	0	0	0
BE3	0	0	1	0	0	0	0	0	0	0	0	0
BE4	0	0	1	0	0	0	0	0	0	0	0	0
BO1	0	0	0	1	0	0	0	0	0	0	0	0
BO2	0	0	0	1	0	0	0	0	0	0	0	0
BO3	0	0	0	1	0	0	0	0	0	0	0	0
BO4	0	0	0	1	0	0	0	0	0	0	0	0
BU1	0	0	0	0	1	0	0	0	0	0	0	0
BU2	0	0	0	0	1	0	0	0	0	0	0	0
BU3	0	0	0	0	0	0	0	0	0	1	0	0
BU4	0	0	0	0	1	0	0	0	0	0	0	0
CN1	0	0	0	0	0	1	0	0	0	0	0	0
CN2	0	0	0	0	0	1	0	0	0	0	0	0
CN3	0	0	0	0	0	1	0	0	0	0	0	0
CN4	0	0	0	0	0	1	0	0	0	0	0	0
CO1	0	0	0	0	0	0	1	0	0	0	0	0
CO2	0	0	0	0	0	0	1	0	0	0	0	0
CO3	0	0	0	0	0	0	1	0	0	0	0	0
CO4	0	0	0	0	0	0	1	0	0	0	0	0
HA1	0	0	0	0	0	0	0	1	0	0	0	0
HA2	0	0	0	0	0	0	0	1	0	0	0	0
HA3	0	0	0	0	0	0	0	1	0	0	0	0
HA4	0	0	0	0	0	0	0	1	0	0	0	0
LC1	0	0	0	0	0	0	0	0	1	0	0	0
LC2	0	0	0	0	0	0	0	0	1	0	0	0
LC3	0	0	0	0	0	0	0	0	1	0	0	0
LC4	0	0	0	0	0	0	0	0	1	0	0	0
LU1	0	0	0	0	0	0	0	0	0	1	0	0
LU2	0	0	0	0	0	0	0	0	0	1	0	0
LU3	0	0	0	0	0	0	0	0	0	1	0	0
LU4	0	0	0	0	0	0	0	0	0	1	0	0
MA1	0	0	0	0	0	0	0	0	0	0	1	0
MA2	0	0	0	0	0	0	0	0	0	0	1	0
MA3	0	0	0	0	0	0	0	0	0	0	1	0
MA4	0	0	0	0	0	0	0	0	0	0	1	0
WI1	0	0	0	0	0	0	0	0	1	0	0	0
WI2	0	0	0	0	0	0	0	0	0	0	0	1
WI3	0	0	0	0	0	0	0	0	0	1	0	0
WI4	0	0	0	0	0	0	0	0	0	0	0	1

4 - Synthèse des résultats

L'ensemble des paramètres a d'abord été résumé par l'examen des liens entre eux : on obtient des groupes non corrélés très voisins des groupes de définition. Les durées d'occlusion large et stricte sont toujours fortement corrélées entre elles ainsi que très souvent les paramètres à l'intérieur d'un même groupe de définition. Ce qui montre a posteriori la validité de la définition de ces groupes.

Du point de vue du critère du F-ratio, la fréquence fondamentale moyenne donne les meilleurs résultats et a intérêt à être obtenue sur un texte assez long. La remarque vaut pour le taux de voisement. Les durées d'occlusives et la durée totale de la 3ème et de la 6ème phrase obtiennent ensuite les meilleures valeurs. Le débit et le taux de voisement semblent enfin être des paramètres meilleurs que les durées des groupes de phonation, les durées de pauses intermédiaires et les durées de sons. Il a enfin été suggéré diverses façons de sélectionner un "bon" n-uplet de paramètres au vu de ces observations.

Les expériences d'identification pas à pas ont mis en valeur la pertinence de la fréquence fondamentale, toujours choisie en premier. Le tableau ci-joint résume la situation. On doit noter le taux d'identification élevé (voisin de 100 %) obtenu avec environ 6 paramètres en moyenne. Ce taux est parfois obtenu sans que la classification a posteriori soit parfaite (expériences A,B et D).

SIMILITUDE DES RESULTATS DES EXPERIENCES

Expérience		Effectif de la base donnant un taux de 100 %			Expérience				
		TRAP	TI	Liste ordonnée paramètre	A	B	C	D	
	A	8	6	1,31,37,22,17,3,4,10,25,28	1				
	B	8	6	1,31,37,22,17,3,4,10,25,28	1	1			
	C	10	4	1,22,37,31,17,3,25,28,10,32	0,1 0,7	0,1 0,7	1		
	D	7	9	1,22,37,31,3,18,24,32,29,30	0,1 0,5	0,1 0,5	0,4 0,5	1	

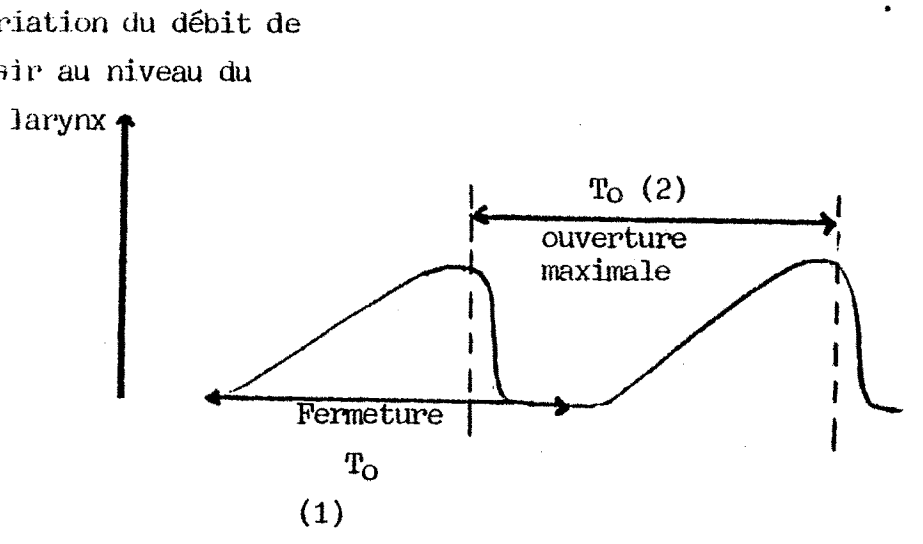
coefficient de similitude entre les 10-uplets de chaque expérience

Le 1er coefficient de similitude est par définition la proportion d'éléments égaux jusqu'au premier non égal.
Il indique clairement que la fréquence fondamentale doit être considérée à part, puis en même temps, les durées de phase (31,37) et les durées d'occlusion (22).
Le 2ème coefficient est le cardinal de l'intersection divisé par 10.
Il est toujours supérieur à 50 % ce qui indique une certaine stabilité des paramètres relativement au découpage des données en apprentissage et test.

5 - La fréquence fondamentale moyenne : modélisation, traitements, discriminance et stabilité.

5.1. Les modélisations de la fréquence laryngienne :

La fréquence fondamentale doit être en principe repérée directement à la source de vibration laryngienne : les cordes vocales.
On suppose que, par une méthode donnée, on retrouve dans le signal de parole, des instants correspondant à une position déterminée des cordes vocales (ouverture maximale et fermeture surtout).
Fo devient calculable (i.e. repérable dans le signal de parole), en ces instants privilégiés.
Le schéma suivant indique ces deux possibilités de mesure :



(1) : à T_o correspond la mesure entre deux passages successifs par zéro t_i et t_{i+1} du signal de parole filtré.

(2) : à T_o correspond la mesure entre deux valeurs crête du signal de parole filtré.

Notons $(t_i)_{1\leq i\leq N}$ la suite strictement ordonnée des temps de passage par zéro du signal sur un intervalle [a,b] donné, (en faisant abstraction ici des problèmes de mesure).

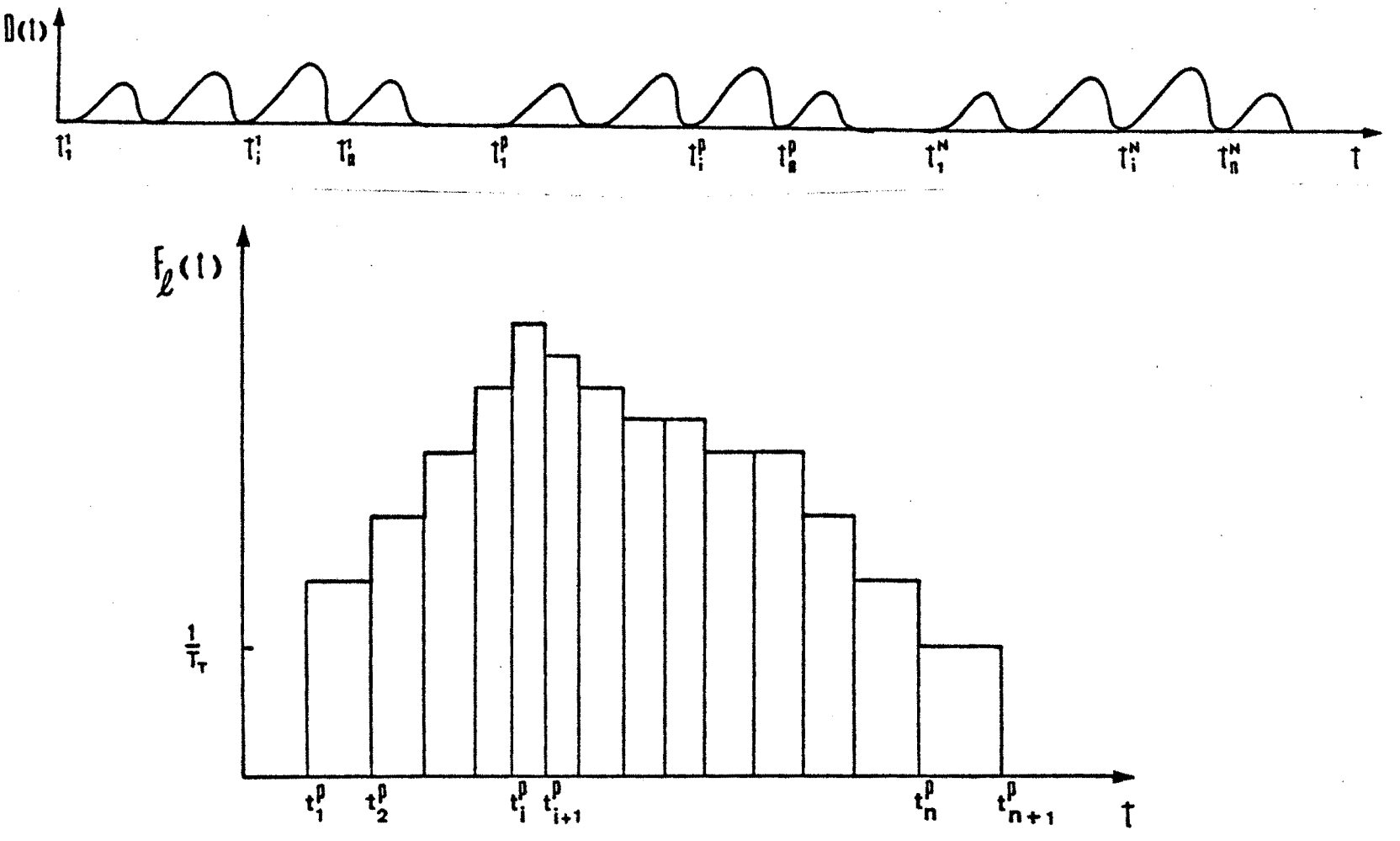


Fig 4 - Variations du débit de l'air au niveau du larynx et variations de $F_1(t)$ pour les valeurs de la séquence p (extrait de BOË(1975)).

La fréquence fondamentale est la fonction :

$$F_0 : [a,b]^2 \ni (t_i,t_{i+1}) \longmapsto \frac{1}{t_{i+1}-t_i} \geq \frac{1}{\bar{T}} \geq \frac{1}{\bar{\bar{T}}}$$

On peut prendre : 1/T = 50 Hz soit T = 20ms, pour l'acte parlé.

La fonction de voisement détecte la présence de cycles :

Si : $\frac{1}{t_{i+1}-t_i}$ < 50 Hz, par exemple, on dira qu'il n'y a pas de voisement.

C'est donc la fonction :

$$V : [a,b]^2 \ni (t_i,t_{i+1}) \longmapsto \begin{cases} 1 & \text{si } \frac{1}{t_{i+1}-t_i} \geq \frac{T}{\Delta} \\ 0 & \text{sinon} \end{cases}$$

On doit donc en pratique s'assurer de deux points :

- la présence d'un cycle
- la durée de ce cycle.

On doit considérer les périodes $t_{i+1} - t_i$ comme indépendantes du système utilisé (des comparaisons ont été conduites : il apparaît une variation de l'ordre de 1% sur les données, ce qui, en analogique, est pratiquement la meilleure précision possible. C'est que le bruit accompagnant le fondamental est très faible).

Le seuil constant T permet de délimiter les différentes séquences et de décider la présence ou l'absence de voisement. Sa valeur est celle de la plus basse fréquence provoquant une impression de voisement. (Objectivement, ce sera la durée de la plus grande période observable dans la mesure de la fréquence fondamentale). La tolérance sur la valeur de T est grande car les valeurs observées sont pratiquement supérieure à 80 Hz.

Soit maintenant une suite de séquences voisées $(v^p)_{1\leq p\leq V}$.

Les abscisses t_i correspondantes sont alors notées :

$$(t_i^p)_{1\leq i\leq N^p}$$

pour chaque séquence p. Au cours de la production de parole, l'évolution de la

fréquence fondamentale est le plus souvent lente : les intervalles de temps $t^p_{i+1} - t^p_i$ successifs sont voisins (LIEBERMAN (1963)), a montré qu'il y a peu de corrélation entre les valeurs $t^p_i - t^p_{i-1}$ et $t^p_{i+1} - t^p_i$, et qu'il y a par contre une certaine corrélation entre les valeurs $t^p_i - t^p_{i-1}$ et $t^p_{i+2} - t^p_{i+1}$.

Il serait intéressant de modéliser le processus non stationnaire :

$$(t^p_i)_{1\leq i\leq N^p} \ ,$$

par une chaîne de Markov.

On appelle donc fréquence laryngienne instantanée de l'impulsion i de la séquence la valeur :

$$F^p_{oi} = \frac{1}{t^p_{i+1} - t^p_i} = \frac{1}{T^p_i}$$

Lorsqu'il y a par contre absence de vibration laryngienne, on dit que l'on a une réalisation sourde ou une pause. La distinction entre ces deux éventualités se faisant à l'aide d'un second seuil constant de durée U, fixé à 200ms (il y a pause si l'on ne détecte pas de signal pendant une telle durée minimale : le cas complémentaire inclue entre autres, la réalisation d'occlusives).

Soit la séquence voisée p. La fréquence laryngienne moyenne sur cette séquence est la moyenne de F_{oi}^{p}, c'est-à-dire aussi :

$$\overline{F_o^p} = \frac{1}{t_{N_p+1}^p - t_1^p} \sum_{i=1}^{N_p} (t_{i+1}^p - t_i^p) = \frac{N_p}{t_{N_p+1}^p - t_1^p}$$

La moyenne sur l'ensemble des séquences d'une phrase est donc :

$$\overline{F_o} = \frac{\sum_{p=1}^{V} N_p}{\sum_{p=1}^{V} t_{N_p+1}^p - t_1^p} = \left(\sum_{p=1}^{V} N_p\right) \frac{1}{\sum_{p=1}^{V} \frac{N_p}{\overline{F_o}^p}} = \frac{\sum_{p=1}^{V} N_p}{DTV},$$

où DTV est la durée totale de voisement de la phrase.
Les procédures de mesure sont décrites dans BOË & RAKOTOFIRINGA (1972, 1975).

La fréquence fondamentale est un paramètre de base caractérisant le sexe. (Figure 2). Elle peut aussi être utilisée pour distinguer les enfants. Elle détermine la hauteur de la voix et en quantifie le spectre en association avec les harmoniques. Sa variation dans le temps détermine l'intonation particulière de la phrase, paramètre d'un certain intérêt pour la reconnaissance du locuteur.

D'après BOË (1975), l'intervalle de confiance au seuil 4% est pour les locuteurs français [79 Hz, 157 Hz], c'est-à-dire :

$$\text{Prob } (F_o < 79 \text{ ou } F_o > 157) \leq 0{,}04.$$

5.2. Un programme de traitement statistique de la fréquence fondamentale.

1) Implémentation :

On considère maintenant que l'on dispose d'un appareillage A délivrant une valeur tous les 1/40èmes de secondes :

- a) Si le signal est sourd (pas de voisement), cette valeur est 0.
- b) Si le signal est voisé, cette valeur est celle de la fréquence fondamentale courante.

- c) En cas d'absence de parole, la valeur délivrée est 0.

On inclue les dispositions et conventions suivantes :
- le locuteur prononce une phrase dont il débute la diction lorsqu'il le veut,
- il peut interrompre la réception de son message vocal à la sortie de A, à tout moment par une commande d'interruption prioritaire,
- il peut effectuer les pauses qu'il juge utiles, à l'intérieur de la phrase,
- une absence de parole de durée suffisante est interprétée comme une fin de phrase.

Notons que certaines valeurs nulles peuvent correspondre à la courte durée d'occlusion du conduit vocal, qui se produit pour les consonnes occlusives :

[b], [d], [g], [p], [t], [k].

Le schéma de décision de la catégorie courante en cas d'absence de parole, est le suivant :

durées :	0	U	W
décision :	sourde	pause	fin de phrase

Différents calculs statistiques sont alors entrepris : nombre d'échantillons voisé non voisés, fondamental, moyen, écart- type et quadratique, etc...

5.2. Un programme de traitement statistique :

1 - Implémentation

Les données, actions élémentaires permises et règles de terminaison figurent dans le tableau suivant

donné	tableau F d'entiers dernière valeur fictive non nulle appelée sentinelle (1)
manipulation du tableau F	
entier	indice I
démarrer	I ← 1
avancer indice	I ← I+1
élément courant	F(i)
Terminaison :	
	1/ occurrence d'une suite de 0 de longueur > W 2/ occurrence de la sentinelle
Règles de terminaison	
	Elles sont un choix par rapport au problème de la récupération des erreurs. On choisit la règle qui est la plus proche du contexte d'application de l'algorithme à savoir : "Lorsque la sentinelle apparaît, après une suite de 0 de taille inférieure à W, cette suite de 0 est considérée comme la pause finale". En d'autres termes, la sentinelle équivaut à W+1 zéros.

(1) L'implémentation d'une sentinelle pour l'interruption du programme n'est pas obligatoire du point de vue intelligence artificielle.

La traduction des définitions utilisées est la suivante :

Segment voisé	Suite d'entiers non nuls, non vide
Segment non voisé	Suite de k zéros, $k \in [1, V]$
pause	suite de k zéros, $k \in]U, V]$
pause finale	suite de k zéros, $k > W$
pause initiale	suite des premiers zéros consécutifs, si le 1er élément des données est nul

Le programme qui suit est écrit - après simplifications - d'après les règles de la programmation structurée (1). On ne détaille pas les traitements des séquences eux-mêmes mais simplement la "gestion" de la séquence de valeurs issues du vocodeur.

(1) Cette collaboration avec P.C. SCHOLL de l'I.M.A.G. a permis d'en faire un exemple pédagogique tiré donc d'une situation très réelle. Le programme lui-même est implémenté en FORTRAN sur LSI-11.

2 - Algorithme :

```
Initialiser-traitement
{amorce : lecture de la 1ère suite de 0 si elle existe}
acquérir données
longueur-suite ← 0
tantque él. courant = 0 et longueur-suite < W faire
        longueur-suite ← longueur-suite+1
        avancer él. courant
ftantque
        {él. courant ≠ 0 et longueur-suite = W}
si él. courant = (0 ou sentinelle)
alors écrire (fichier vide) ; STOP
sinon
{él. courant = 1er non nul du fichier}
itérer
{él. courant = 1er non nul du prochain segment voisé (lequel existe)}
        acquérir-segment-voisé
        calculer-statistiques-segment-voisé
        {él. courant = (0 ou sentinelle)}
        co : acquérir segment non voisé
        si él. courant = sentinelle
        alors segment final ← vrai
        sinon {él. courant = le 1er zéro}
              ℓ ← 1          {ℓ = nb de 0 déjà rencontrés}
              itérer avancer indice
              arrêt él. courant ≠ 0 ou ℓ ≥ W
                    ℓ ← ℓ+1
              fitérer
              si él. courant = (0 ou sentinelle)
              alors segment final ← vrai
              sinon  segment final ← faux
        fsi
        {segment final ou exclusif courant 1 ≠ (0 et sentinelle)}
arrêt : segment final
        calculer-statistiques-segment-non-voisé.

Imprimer-résultats
```

L'algorithme est réalisé en vue de l'adjonction d'un coupleur vocodeur-mini-ordinateur. Nous présentons ici quelques exemples où les données ont été entrées au clavier. TE est le taux d'échantillonnage, MAX1 = U, MAX2 = W.

```
ECHANTILLONS
 0 FICHIER VIDE
```

```
MAX1 =              5
MAX2 =              10
TE =                2 500000000E-02
NB ECH=             69
ECHANTILLONS
 0   0   0   0   0   0   0   0   0   100   102   10   101   104   0   0   0   0
 0   0   100   102   101   105   104   106   108   109   0   0   0   0   0
 100   102   0   100   0   0   101   0   0   0   100   0   0   0   0   100   0
 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1000   1000   0   0
DUREE PAROLE ·    2 500000000E-02
DUREE VOISEE ·    2 500000000E-02
NOMBRE ECHANTILLONS VOISES·       2
FONDAMENTAL MOYEN ·               101
ECART TYPE ·      1
ECART QUADRATIQUE MOYEN ·         4
DUREE SOURDE ·    0
NOMBRE DE SOURDS ·                0
NOMBRE DE PAUSES ·                0
DUREE TOTALE DES PAUSES ·         0
```

```
DUREE MOYENNE DES PAUSES ·        MAX1 =              5
MAX2 =              10
TE =                2 500000000E-02
NB ECH=             69
ECHANTILLONS
 0   0   0   0   0   0   0   0   0   100   102   10   101   104   0   0   0   0
 0   0   100   102   101   105   104   106   108   109   0   0   0   0   0
 100   102   0   100   0   0   101   0   0   0   100   0   0   0   0   100   0
 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1000   1000   0   0
DUREE PAROLE ·    2 500000000E-02
DUREE VOISEE ·    2 500000000E-02
NOMBRE ECHANTILLONS VOISES·       2
FONDAMENTAL MOYEN ·               101
ECART TYPE ·      1
ECART QUADRATIQUE MOYEN ·         4
DUREE SOURDE ·    0
NOMBRE DE SOURDS ·                0
NOMBRE DE PAUSES ·                0
DUREE TOTALE DES PAUSES ·         0
```

```
DUREE PAROLE :      35
DUREE VOISEE :    7 500000000E-02
NOMBRE ECHANTILLONS VOISES :        10
FONDAMENTAL MOYEN :                 104 5
ECART TYPE :        8 25
ECART QUADRATIQUE MOYEN :            428571428571 4
DUREE SOURDE :      1
NOMBRE DE SOURDS :                  4
NOMBRE DE PAUSES :                  2
DUREE TOTALE DES PAUSES :            175
DUREE MOYENNE DES PAUSES :          8 750000000E-02
```

```
MAX1 =              5
MAX2 =              10
TE =                2 500000000E-02
NB ECH=             69
ECHANTILLONS
  0   0   0   0   0   0   0   0   0  100  102  103  101  104   0   0   0   0
  0   0  100  102  101  105  104  106  108  109   0   0   0   0   0
  100  102   0  100   0   0  101   0   0   0  100   0   0   0   0  100   0
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  1000  1000   0   0
DUREE PAROLE :      675
DUREE VOISEE :      3
NOMBRE ECHANTILLONS VOISES :        19
FONDAMENTAL MOYEN :                 102 5263157895
ECART TYPE :        7 40720158
ECART QUADRATIQUE MOYEN :           7 571428571429
DUREE SOURDE :      25
NOMBRE DE SOURDS :                  5
NOMBRE DE PAUSES :                  1
DUREE TOTALE DES PAUSES :            125
DUREE MOYENNE DES PAUSES :           125
```

Même test avec présence de sentinelle :

```
MAX1 =              5
MAX2 =              10
Te =                2 500000000E-02
NB ECH=             69
ECHANTILLONS
  0   0   0   0   0   0   0   0   0  100  102  103  101  104   0   0   0   0
  0   0  100  102  101  105  104  106  108  109   0   0   0   0   0
  100  102 [10] 100   0   0  101   0   0   0  100   0   0   0   0  100   0
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  1000  1000   0   0
DUREE PAROLE :      525
DUREE VOISEE :      3
NOMBRE ECHANTILLONS VOISES :        15
FONDAMENTAL MOYEN :                 103 1333333333
ECART TYPE :        7 58222267
ECART QUADRATIQUE MOYEN :           17 66666666667
DUREE SOURDE :      1
NOMBRE DE SOURDS :                  1
NOMBRE DE PAUSES :                  1
DUREE TOTALE DES PAUSES :            125
DUREE MOYENNE DES PAUSES :           125
```

5.3. Quelques résultats

1. Discriminance

Le tableau suivant résume les taux de bonne discriminance a posteriori des 12 locuteurs, obtenus par la seule fréquence fondamentale au cours des expériences précédentes :

Expériences	TRAP
A	50
B	50
C	50
D	50
Moyenne	50

Sachant que l'on est en présence d'un même sexe, on examine le taux de bonne discrimination séparément sur les hommes et sur les femmes (schéma suivant (1)).

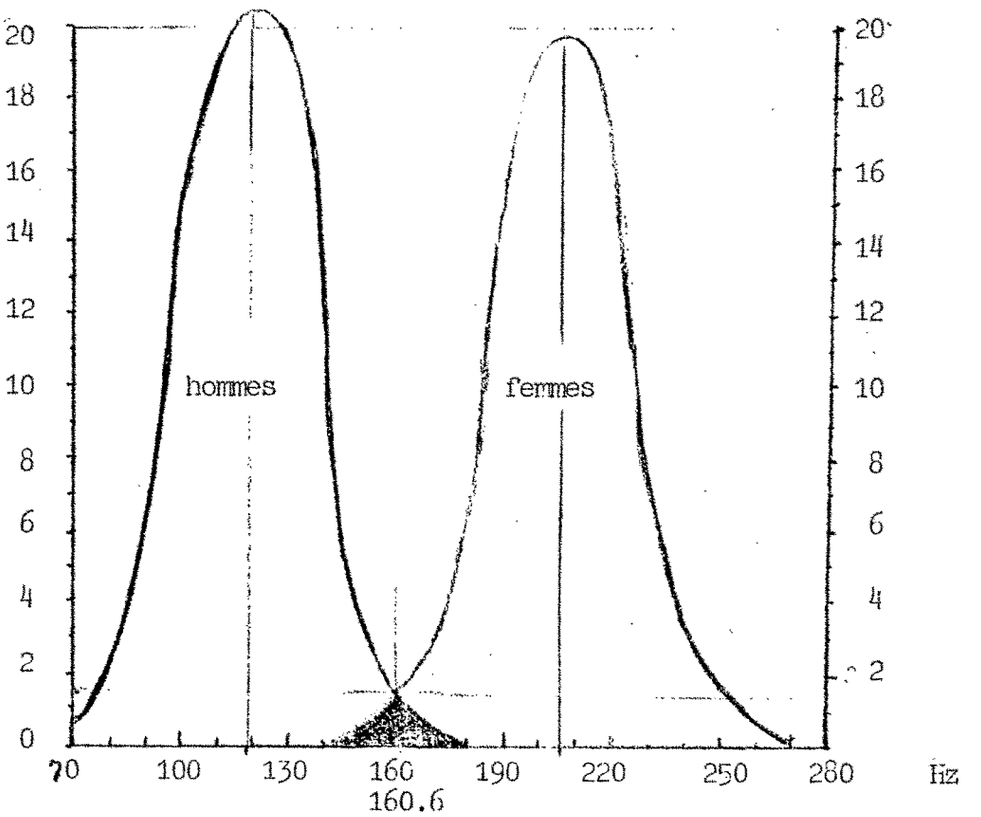


Fig.2

(1) extrait de BOE (1975)

Selon BOË & RAKOTOFIRINGA (1972) JASSEM & al. (1973), etc. , la fréquence laryngienne peut être considérée comme une variable aléatoire gaussienne.

Si l'on extrapole les résultats de cette expérience à tout sous-ensemble aléatoire de locuteurs de langue française, le classificateur :

$$\delta_{F_0} : \ell_i \longmapsto \begin{cases} \mathcal{L} \longrightarrow \{H,F\} \\ H \text{ si } \overline{F_0}(\ell_i) < 160.6 \\ F \text{ sinon} \end{cases}$$

minimise le taux de mauvaise classification (aire hachurée rapportée à l'aire totale). En effet, la fréquence d'occurrence des hommes est alors approximativement la même que celle des femmes (1/2), et les deux distributions correspondantes peuvent être tenues pour égales. On applique alors la règle de décision de BAYES de taux d'erreur minimum.

Combinant les deux étapes, on a le modèle suivant :

	population aléatoire (sexes représentés de façon homogène).		discrimination des sexes
a)	$\mathcal{L} \xrightarrow{\delta_{F_0}}$	$(\mathcal{L}_H, \mathcal{L}_F)$	Taux d'erreur expérimental $= t_{\delta_{F_0}}$
b)	$\mathcal{L}_H \xrightarrow{\mathcal{L}_H}$	$\{\ell_1, \ldots, \ell_N\}$	Taux d'erreur expérimental $= t_{\delta_H}$
	$\mathcal{L}_F \longrightarrow$	$\{\ell_1, \ldots, \ell_N\}$	Taux d'erreur $= t_{\delta_F}$
c)	$\mathcal{L} \xrightarrow{\delta_H \circ \delta_{F_0}}$	$\{\ell_1, \ldots, \ell_N\}$	

Au total, puisque (notations abrégées) :

$p = P(\{\ell_1, \ldots, \ell_N\} \text{ correct}) =$

$P(\{\ell_1, \ldots, \ell_N\} \text{ correct} / \mathcal{L}_H \text{ correct}) \times P(\mathcal{L}_H \text{ correct}) +$

$P(\{\ell_1, \ldots, \ell_N\}) \text{ correct} / \mathcal{L}_H \text{ erroné}) \times P(\mathcal{L}_H \text{ erroné})$

Il vient

$$P = \alpha \times 0.99 + \beta \times 0.01$$

β peut être considéré comme très faible, et α peut être assimilé à la valeur moyenne obtenue au début. Ainsi :

$P = \alpha \times 0.99$ avec α voisin de 95 - 100 %

qui représente la probabilité de bonne discrimination de l'ensemble des locuteur pour un effectif de 2 x 12 = 24. Ces dernières valeurs n'ont évidemment pas de rigueur, mais servent seulement à guider la pensée.

2. Stabilité

Nous n'avons pas réalisé d'étude de stabilité, mais il convient d'inclure ici les remarques importantes suivantes. D'abord, la variabilité intra-locuteur de la fréquence fondamentale instantanée, est loin d'être négligeable et approche même, selon ATKINSON (1976), la variabilité inter-locuteur. En d'autres termes, une normalisation de ce paramètre par l'emploi d'une moyenne, efface les caractéristiques prosodiques .

Si la fréquence fondamentale varie avec l'âge (à la vieillesse correspond une diminution du registre vocal et une valeur maximale plus faible sur les voyelles (PTACEK & al. 1966)), il existe des variations sur une même année pouvant rendre ce paramètre non fiable.

Cependant, sous l'hypothèse - assez largement vérifiée - de normalité, l'écart-type prend une signfication particulière.

6 - Exploitation de l'intensité du Signal

1 - Description de l'expérience

Parmi les paramètres temporels du signal de parole - par opposition aux paramètres spectraux - l'évolution du niveau d'intensité n'a fait l'objet que de très peu d'études par comparaison à celles sur le fondamental (ROSENBERG, 1976). Cependant, l'examen du signal de parole dans le domaine temporel a déjà révélé de nombreuses informations utiles à la reconnaissance (BAKER, 1975).

On dispose ici d'une suite de valeurs correspondant à l'échantillonnage du niveau d'intensité du signal de parole pour une phrase de 8 syllabes (durée : 600 à 850 ms) prononcée d'une façon naturelle. La mesure du niveau d'intensité est définie de la façon suivante :

Si S(t), pour $t \in [0,T]$, est le signal continu de parole correspondant à l'énonciation de la phrase, le niveau d'intensité à l'instant t est :

$$NI(t,\tau) = 10 \log\left[\frac{1}{\tau}\int_{t-\tau/2}^{t+\tau/2} S^2(\theta)\, d\theta\right]$$

La constante de temps d'intégration τ est fixée à 20 ms. L'évolution temporelle X(t) du niveau d'intensité est obtenue analogiquement et échantillonnée toutes les 10 ms. L'utilisation du logarithme permet de s'affranchir de l'intensité avec laquelle parle le locuteur, laquelle dépend de la distance du microphone aux lèvres. Cette distance est gardée relativement constante. Le "contour" X(t) du niveau d'intensité représente et contient l'information sur laquelle nous fonderons la discrimination et la vérification du locuteur. La durée d'élocution T est utilisée implicitement en ce sens que, lors de la comparaison de deux des courbes précédentes, pour deux répétitions ou deux locuteurs différents, il n'est pas fait de normalisation en durée préalable. Ainsi, il est fait un usage combiné des deux paramètres : niveau d'intensité et durée. (Il suffit d'un échantillonnage faible du signal, puisque ces deux quantités décrivent la macro-évolution du signal). Tout ceci conduit à la modélisation qui suit.

L'étude concerne une population de 10 hommes adultes, homogène sur le plan linguistique et d'âge moyen 35 ans. Chaque locuteur répète 5 fois la même phrase au cours de la même session. Les indices i et j désignant respectivement le numéro de la répétition et celui du locuteur, on obtient la famille des courbes d'intensité continues :

$$(X_i^j(t))_{\substack{1\le i\le I = 5 \\ 1\le j\le J = 10}}$$

où pour X_i^j, t varie de 0 à T_i^j. Le locuteur j est donc modélisé par les 5 courbes $(X_i^j(t))$ et les 5 valeurs T_i^j, i = 1,... I = 5.

Il n'est fait qu'un usage implicite des T_i^j. En effet, pour permettre la comparaison des images des fonctions X_i^j, et la définition d'une norme commune, on prolonge leur domaine de définition à l'intervalle $[0,T_{max}]$ défini comme :

$$T_{max} = \underset{i}{\text{Max}}\ \underset{i}{\text{Max}}\ T_i^j.$$

en posant X(t) = 0 pour les nouvelles valeurs (prolongement par continuité). La séquence de traitement lors de l'apprentissage, est rappelée à la figure 1.

L'analyse d'une courbe donnée ne présente pas, de notre point de vue, d'intérêt en soi. C'est l'écart au sens de proximité intuitive, qui est le fondement des techniques de discrimination ci-dessous. En fait, on prend pour mesure approchée de l'écart entre deux courbes, une mesure de l'écart entre les deux suites[1] de points résultant de l'échantillonnage (on notera S cet espace vectoriel muni d'une métrique).

On utilisera, dans notre étude, la norme suivante :

Si ${}^{k}X_i^j$ est la notation abrégée pour $X_i^j(k) = X_i^j(t=k)$,

- $||X_i^j||_1 = \sum_{k=1}^{T_{max}} P_k \, |{}^{k}X_i^j|$, avec $P_i \geq 0$, P_i entier, $\forall$ i.

Le vecteur poids (P_k) est une pondération des différentes contributions à la norme de la suite. La distance induite entre deux éléments de S est alors :

$$d_1(X_i^j, X_{i'}^{j'}) = \sum_{k=1}^{T_{max}} P_i \, |{}^{k}X_i^j - {}^{k}X_{i'}^{j'}|$$

Afin de calculer une estimation de la distance entre deux locuteurs, on calcule les distances moyennes entre répétitions de ces deux locuteurs :

a) pour un même locuteur (dispersion moyenne interclasse) :

$$dm(j) = \frac{2}{I(I-1)} \sum_{\substack{i,i'=1 \\ i<i'}}^{I} d(X_i^j, X_{i'}^j)$$

$$= \frac{1}{I(I-1)} \sum_{i=1}^{I} \sum_{i'=1}^{I} d(X_i^j, X_{i'}^j)$$

b) entre deux locuteurs distincts (distance moyenne interclasse) :

$$dm(j,j') = \frac{1}{I^2} \sum_{i,i'=1}^{I} d(X_i^j, X_{i'}^{j'}) = d\left(\frac{1}{I} \sum_{i=1}^{I} X_i^j \, , \frac{1}{I} \sum_{i'=1}^{I} X_{i'}^{j'}\right)$$

$$= d(X_\cdot^j, X_\cdot^{j'})$$

---

(1) On préfère le mot suite à la locution série temporelle, car la dynamique l'élocution n'est pas conservée.

On établit d'autre part un histogramme des distances entre deux locuteurs dont le but est d'évaluer la confiance que l'on peut avoir dans la mesure de la distance moyenne et aussi d'évaluer leur éloignement (figure 2). On considère aussi les distances maxima et minima suivant les répétitions (les 2 locuteurs sont proches d'au moins leur distance maxima, loin d'au moins leur distance minima). Le problème est de mesurer les deux types fondamentaux de dispersion : intra-locuteur et interlocuteurs, et de définir alors une méthode mettant en évidence leurs relatives valeurs. Intuitivement, plus les groupes formés des répétitions d'un même locuteur sont distants les uns des autres, relativement à leurs distances à l'intérieur d'un même groupe, plus il est vraisemblable qu'on puisse reconnaître les locuteurs lors de la phase de test.
Cette technique de traitement, qui ne fait intervenir que les paires d'éléments, n'est pas adaptée à une visualisation du recouvrement de deux classes, car on ne peut représenter dans le plan un ensemble de I points, dès que $I \geq 4$ avec leurs éloignements respectifs, et d'autre part, selon les conventions utilisées pour représenter les positions relatives des répétitions, deux locuteurs peuvent apparaître plus ou moins recouverts. On utilise au contraire des mesures de la vraisemblance de non-recouvrement. On dira d'abord qu'il y a vraisemblance de la classe j (c'est-à-dire le locuteur j est "relativement séparé des autres") si la moyenne des distances entre les points de la classe j et les points appartenant à une classe $j' \neq j$, ceci pour tout $j' \neq j$. C'est donc dire que la matrice symétrique :

D MOY(j,j') = moyenne des distances du locuteur j au locuteur j'

est à diagonale dominée en ligne (ou colonne).
L'évaluation de l'écart entre deux locuteurs j et j' conduit à leur catégorisation à l'aide de la règle suivante :

étant donné deux seuils α et $\beta(\alpha < \beta)$,
si dissimilarité (locuteur j, locuteur j') > β alors j et j' disjoints,
sinon si dissimilarité (locuteur j, locuteur j') >> α alors j et j proch
sinon j et j' confondus.

L'intervalle $[\alpha\ \beta]$ apparaît donc comme un intervalle de rejet correspondant à l'incertitude de la situation.

On observe que chaque locuteur est vraisemblable (figure 3), ce qui indique que le mode d'utilisation du paramètre intensité proposé est susceptible de développements.

Tous les histogrammes de la distance entre deux locuteurs sont unimodal excepté pour les couples de locuteurs :

(2,2,) , (9,6) , (9,9)

où la distribution est bimodale, et les valeurs des deux modes, différentes (6,7 % des cas). Dans les cas suivants :

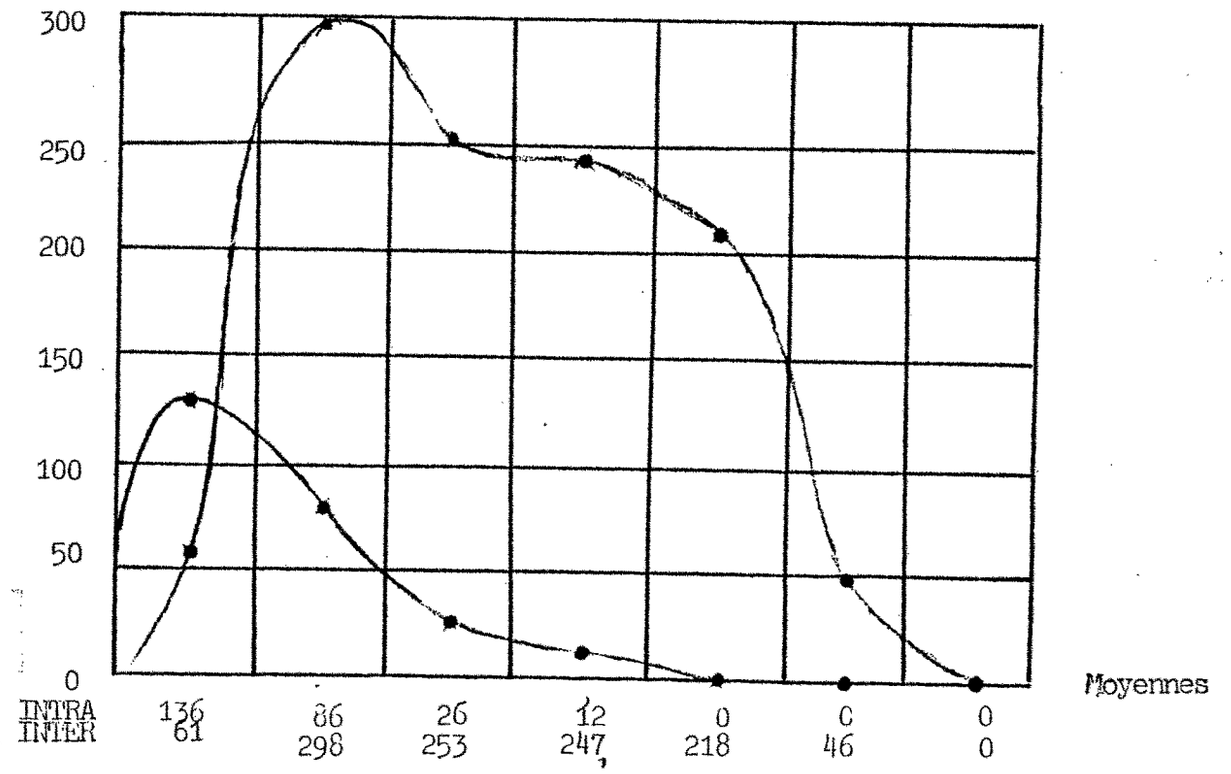
(5,3), (5,4), (7,1), (7,4), (9,7), (10,9)

soit dans 13,3 % des cas, le maximum est atteint sur deux cases successives d l'histogramme. Pour ces dernières distributions bimodales, le mode sera considéré comme étant la plus faible valeur des deux. D'une façon générale, l'analyse visuelle des histogrammes illustre bien les deux points fondamentaux :

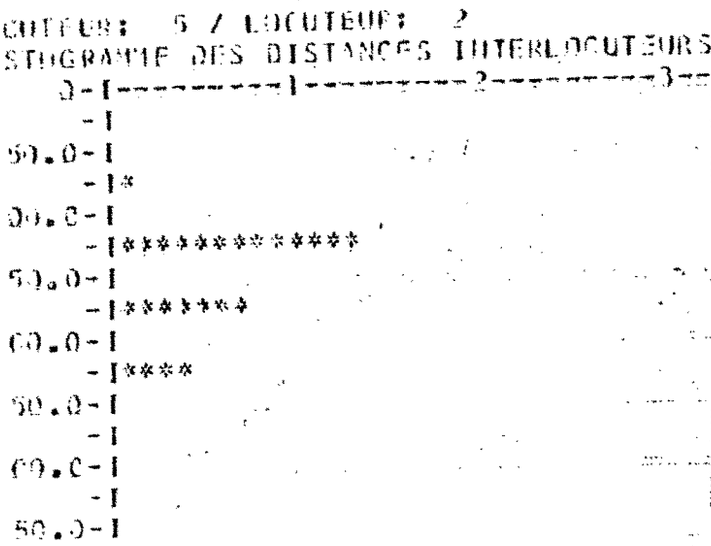
- étant donné un locuteur,
  1. Les valeurs se rangent vers les faibles valeurs de l'histogramme pou les distances entre les répétitions de ce locuteur,
  2. Les valeurs sont assez nettement décalées vers la droite dès que les mesures sont hétérogènes (comparaison avec les répétitions d'un autr locuteur),

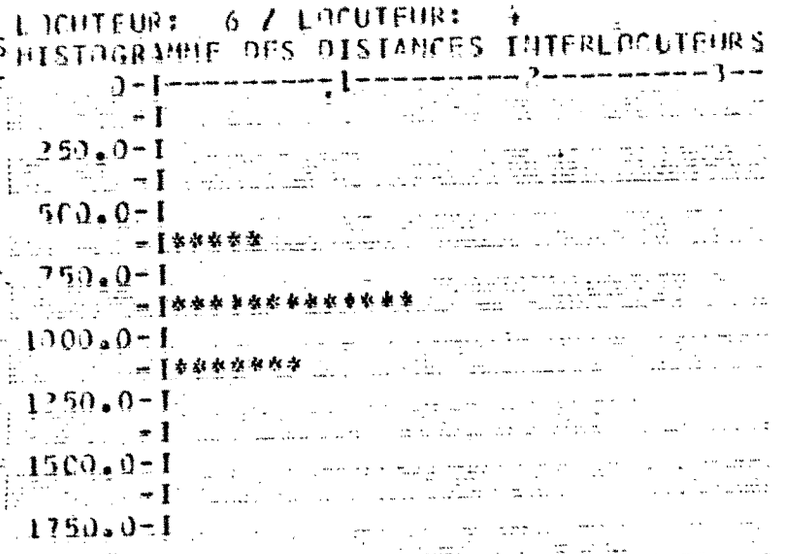
mais montre que l'on ne saurait prendre pour critère la valeur absolue de ces dispersions : elles n'ont de sens que par rapport à un locuteur donné, c'est-dire à sa dispersion intra-locuteur.

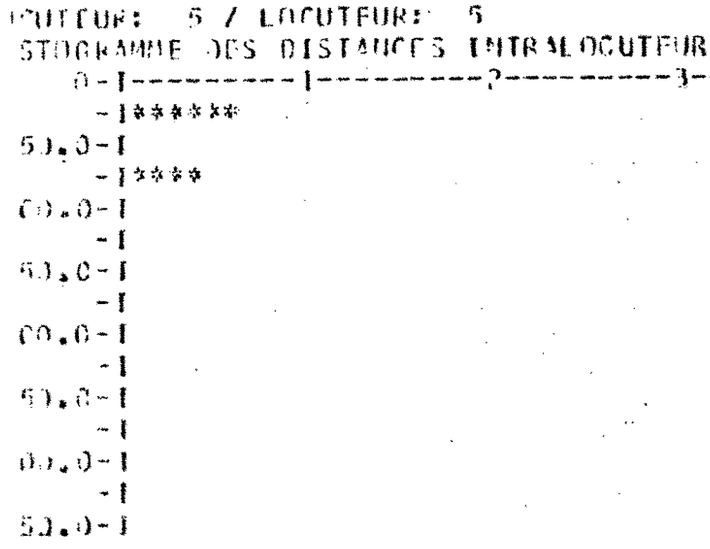
Histogramme des inter et intra-distances entre répétitions

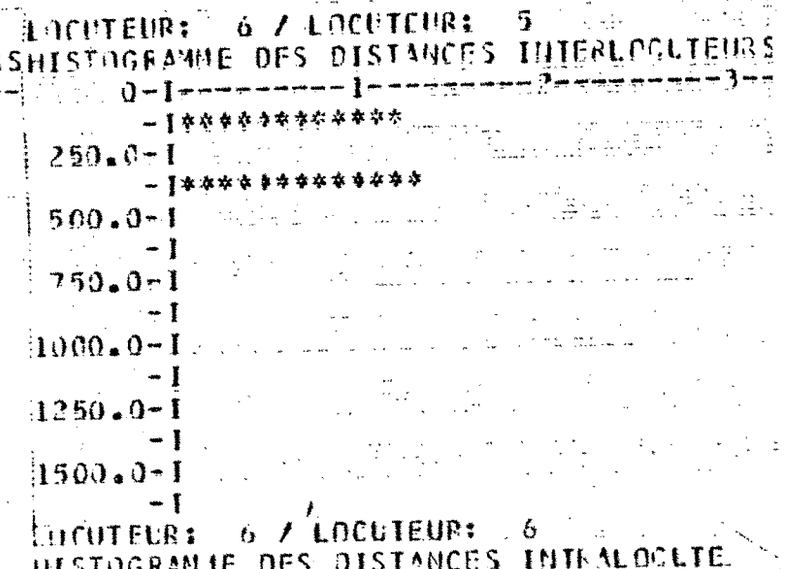


Utilisation des histogrammes des distances inter-répétitions. Les distances entre deux répétitions sont distribuées selon un axe divisé en sept cases. Pour deux locuteurs différents, il y a I x I = 25 valeurs, tandis que pour un même locuteur, il y a $\frac{1}{2}$ I(I-1) = 10 valeurs. La plus petite (resp. grande) valeur n'est autre que la distance minimum (resp. maximum) entre les deux répétitions concernées.









Vraisemblance des locuteurs. Les lignes (colonnes) de la matrice D MOY(j,j') = moyenne des distances du locuteur j au locuteur j' atteignent leur minimum sur la diagonale.

MATRICE DES PROXIMITES DES LOCUTEURS

	LOC1	LOC2	LOC3	LOC4	LOC5	LOC6	LOC7	LOC8	LOC9	LOC10
LOC 1:	372.7									
LOC 2:	537.3	279.3								
LOC 3:	423.1	667.7	206.4							
LOC 4:	716.7	412.7	846.0	256.2						
LOC 5:	438.4	754.5	233.2	931.2	173.3					
LOC 6:	435.4	673.1	255.6	885.3	269.2	118.7				
LOC 7:	1198.4	1000.7	1184.7	882.0	1250.3	1249.4	466.9			
LOC 8:	741.5	503.5	848.6	388.8	918.0	856.6	920.4	224.5		
LOC 9:	640.7	481.2	746.8	428.6	801.9	790.9	1020.1	449.3	378.2	
LOC10:	818.5	481.5	952.4	345.2	1057.8	976.1	848.6	460.2	531.1	236.2

Figure - Visualisation de l'écart entre distance minimum, moyenne et maximu entre deux locuteurs.

```
          LOCUTEUR:  5 -LOCUTEUR:  1
  155.5*I+++++++
  438.4*I++++++++++++++++++++++
  879.5*I++++++++++++++++++++++++++++++++++++++++++++++++
          LOCUTEUR:  5 -LOCUTEUR:  2
  489.5*I++++++++++++++++++++++++
  754.5*I++++++++++++++++++++++++++++++++++++
 1122.0*I++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
          LOCUTEUR:  5 -LOCUTEUR:  3
  169.5*I++++++++
  283.2*I++++++++++++++
  507.0*I++++++++++++++++++++++++
          LOCUTEUR:  5 -LOCUTEUR:  4
  543.5*I+++++++++++++++++++++++++++

  931.2*I++++++++++++++++++++++++++++++++++++++++++++++++++
 1161.5*I++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
          LOCUTEUR:  5 -LOCUTEUR:  5
    0.0*I+
  173.3*I++++++++
  310.5*I++++++++++++++++
```

IV - CONCLUSION D'ENSEMBLE

Comme l'indique la succession des chapitres, nous avons opté pour une approche pluridisciplinaire- à notre avis nécessaire - et préféré asseoir les bases de la reconnaissance du locuteur et envisagé le sujet d'une façon systématique, face à la diversité et la disparité des études antérieures.

Bien qu'une population de 12 locuteurs ne soit pas une taille triviale car ils sont de même sexe, il apparaît que, pour des études ultérieures, une taille largement supérieure est nécessaire. C'est une des raisons pour lesquelles le projet SYSIPHE a été commencé. Cependant, pour des applications très précises, un tel effectif peut suffire. Nous pensons alors avoir démontré la faisabilité d'un système opérationnel sur la base des paramètres choisis, lesquels sont simples, accessibles et disponibles.

Les résultats obtenus - dans des conditions idéales d'enregistrement, il est vrai - sont de bonne qualité. Les paramètres choisis peuvent être utilisés en reconnaissance de la parole. Cependant, l'ensemble des résultats doit être a priori remis en cause dans les deux contextes plus larges suivants :

- étude d'une population de grande taille (une centaine de locuteurs),
- étude de la stabilité des mêmes paramètres au cours du temps et étude de leur résistance à l'imitation.

Il faut en outre tenir compte qu'un ensemble de K caractéristiques pertinentes prises individuellement ne forme pas le meilleur ensemble de taille K.

Nous pensons que le sujet requiert toujours des recherches fondamentales sur :

- la variabilité de la voix : causes,conditions et limites,
- le choix de méthodes simples de classification et de décision, et la détermination de la dimensionalité intrinsèque du problème (combien de caractéristiques ?)
- l'applicabilité au téléphone et plus généralement à un signal de parole,
- l'évaluation de la performance d'un système, autrement que par l'indication d'un simple taux de reconnaissance. Il faut évaluer aussi la résistance aux imposteurs et à la déformation de la voix.

En conclusion, nous dirons qu'il est temps que soient menées des études à plus grande échelle en langue française avec toute l'infrastructure nécessaire (base de donnée adaptée, appareillages,...)

La durée allouée à la présente thèse (17 mois effectifs parmi 24) ne pouvait permettre leur aboutissement, mais seulement leur préparation.

Nous espérons que ce travail puisse être de quelque utilité.

V

LES PERSPECTIVES D'APPLICATION

"Je ne me défie pas de la machine
que je regarde avec curiosité
sur son socle ou sous la verrière
Je me défie de la machine qui
est en moi, je me défie de ma
façon d'utiliser la machine".

Georges DUHAMEL

I - LES TENDANCES SOCIO-ECONOMIQUES ET LES BESOINS ACTUELS

1 - Le contexte de la Communication homme-machine (bis) :

Les moyens modernes de communication donnent une importance accrue au signal sonore : téléphone, radio, magnétophone. Le seul indice identificateur reste alors la voix. Traditionnellement, les méthodes utilisées pour reconnaître des individus sont fondées sur des caractéristiques externes : empreintes digitales, photographies et diverses techniques anthropométriques. Le contour dentaire et les cavités nasales ont été quelque peu utilisées par certaines compagnies d'assurance (BOLT & al., 1970).

Or, selon plusieurs auteurs , l'informatique n'a pas atteint un degré de développement tel qu'elle soit accessible à tous et d'une façon simple : la réalisation de terminaux banalisés, semble impérative. La parole semble être un vecteur particulièrement bien adapté à ce but. Les caractéristiques fondamentales sont, du point de vue de l'utilisateur :

- de permettre à des entités distantes d'agir les unes sur les autres,
- d'offrir aux utilisateurs locaux l'accès au service réseau.

Et beaucoup d'auteurs de prédire le développement accru des possibilités de dialogue entre l'homme et la machine, par l'intermédiaire de réseaux d'ordinateurs. La communication orale avec l'ordinateur est alors une des clés de cette évolution.
D'ailleurs, le dialogue avec une machine en vue d'une tâche précise, a lieu dans l'un des trois cas suivants : demande de renseignements, entrée d'informations, commande.
Ces tâches s'inscrivent dans le contexte général de la Communication homme-machine, et tendent à donner à la machine, à la fois une utilisation et des moyens de perception proches de ceux de l'homme.

On notera que la technologie de la vérification, bénéficie de plusieurs avantages spécifiques remarquables (BEEK, NEUBERG, HODGE, 1977) le locuteur est coopératif : il offre le meilleur comportement vocal possible, les mots et phrases sont choisis de façon à permettre une discrimination élevée; le locuteur est connu du système et l'on peut contrôler l'environnement acoustique. Plus généralement, l'usage de la parole impose un fonctionnement aussi proche que possible du temps réel. Ceci incite au recours aux structures de machines parallèles tout autant qu'aux algorithmes à la fois simples et performants.

Par ailleurs, la parole peut être facilement transmise par téléphone, radio ou satellite, et le développement des réseaux d'ordinateurs est susceptible d'offrir l'extension la plus large de l'usage de la parole (se reporter à GARDAN (1978)pour une extension aux téléconférences).

Ainsi, l'usage de la parole est par exemple susceptible d'amener à une conception très différente du bureau, qui va dans le sens d'un supplément de confort dans le poste de travail. D'une façon générale, un système de reconnaissance de la parole ou du locuteur, doit être accessible à des non-informaticiens. Ce n'est pas un outil pour initiés : une de ses qualités sera d'être transparent pour l'utilisateur.

Un point extrêmement important est le suivant : la reconnaissance du locuteur ne se fait en général pas indépendamment du contexte. Ainsi, la langue parlée, et donc l'adaptation du traitement à celle-ci, prend toute son importance GUIBERT (1979) prédit : "les ordinateurs, dans quelques années, ne comprendront que l'anglais, si aucun effort, même modeste, n'est entrepris et maintenu dans ce domaine, de ce côté-ci de l'Atlantique...".

2 - Une analyse du marché potentiel :

2.1. Quels avantages attendre d'une communication orale avec une machine ?

D'abord, une certaine facilité ou aisance que permet la parole pour communiquer (simplicité). Ensuite, le débit de communication est notablement plus rapide que lors de la frappe au clavier (vitesse). On estime que le rapport

pourrait atteindre la valeur 10. Une certaine fiabilité due au contrôle auditif parallèle, et la simultanéité possible avec d'autres tâches : en effet, la vue reste disponible, ainsi que les mains. Il s'agit donc d'améliorations en productivité et en mobilité de l'utilisateur.

Notons la disparition des coûts dûs à la frappe, la perforation et les corrections, et évidemment, une réduction du volume de documents écrits. Ce dernier point provoque de fortes motivations à l'heure actuelle. Il sous-entend que l'accès à l'information doit se faire, en partie, d'une façon autre que par des textes imprimés. Ce qui est réalisé ici par l'utilisation de messages sonores.

2.2. Que peut permettre un système de vérification d'identité ?

Par son automatisation même, un tel système ne pose plus le problème de personnel entraîné, spécialisé (guichets, vérification des signatures). Dans le cas de l'accès à distance par liaison téléphonique, radio, un système de vérification du locuteur peut être d'un secours apprécié : actuellement, vérifier une personne, c'est avoir la preuve visuelle de son identité (pièces d'identité, signature manuscrite, ...) ou l'utilisation d'une carte magnétique. Ces systèmes classiques sont mis en défaut par contrefaçon, vol. D'autre part, un système de vérification du locuteur peut contribuer à la généralisation de l'accès à distance à des systèmes d'usage courant.

D'autre part, la sécurité des données en général, qu'il s'agisse de protection de données ou d'accès réservé, n'est pas due aux ordinateurs, et n'est pas non plus une question nouvelle."Il est nécessaire de trouver des méthodes de protection de l'information enregistrée[..]L'une consiste à vérifier l'identité de ceux qui utilisent le système"(MAISON ROUGE, 1973).(1) Cette affirmation d'ordre très général, trouve son application par la répartition des rôles, suivante (même auteur) :

"[..]La responsabilité du législateur est de décider qui a accès à quoi, celle du constructeur de mettre au point les techniques de préservation de la [illegible] d'informations, afin de pouvoir réserver l'accès à ceux-là seuls qui y [illegible] droit : la protection de la vie privée est la responsabilité de tous".

(1) "Sur la Sécurité en Informatique", Triangle, 3, 1973.

Il y a deux problèmes principaux à résoudre :

1) le contrôle des données mises en fichier,
2) le contrôle de l'accès au fichier et son corollaire : la dissémination de l'information.

L'un des moyens de protéger des données, est de définir une hiérarchie d'accès aux informations, suivant le niveau d'autorisation de la personne. On peut aussi autoriser/interdire l'entrée de nouvelles données, la modification des données existantes, etc.

2.3. Quelles performances doit-on exiger en reconnaissance du locuteur ?

Sans répondre à cette question de façon générale, indiquons qu'il existe des situations de Vérification qui ne requièrent pas un taux de performance parfait. Soit, en effet, une situation de communication avec locuteur coopératif, dans laquelle :

a) plusieurs énonciations du code locuteur sont disponibles (celui-ci n'en est pas rebuté),

b) les autres locuteurs ont peu de chances d'utiliser le même code,

c) le locuteur n'est pas rebuté par de nouveaux essais de Vérification s'il vient à être rejeté.

Alors le taux de faux rejet pourra être élevé.

2.4. Quelles applications envisager en reconnaissance du locuteur ?

Globalement, on peut distinguer les deux types d'applications suivantes :

a) la personne recherchée est identifiée par sa voix ;
b) la personne qui se présente subit une infirmation/confirmation (vérification) d'identité.

La planche I détaille par secteurs ces types et fournit quelques remarques spécifiques. Elle porte sur les deux contextes suivants (HODGE et al. 1978) :

ELEMENT I - Classification par secteurs des applications potentielles de la reconnaissance du locuteur.

(PU) Public

(PR) Privé

(DM) Domestique

(LC) Local

Secteurs/Remarques	Exemples types d'application	Type	Personnel intermédiaire Degré d'automaticité requise	Avantages spécifiques
I. Contrôle d'accès à des processus physiques PB/PR	- toutes zones contrôlées : bases militaires coffres-forts dépts. industriels - surveillance canaux communication	Vérif.		- sécurité militaire - protection industrielle
II. Accès privilégié à des informations (réservées) PB/PR (validation de l'accès par niveau d'autorisation)	- centres de renseignements documentation, archives, banques de données médicales, bancaires, renseignements généraux, fichiers scientifiques.	Vérif. et Ident.		- une solution aux problèmes de la sécurité des banques de données, - possible par téléphone - sont couplées les 3 tâches : - vérif/ident. - reconnaissance des mots - réponse vocale (év.)
III. Validation des transactions et opérations bancaires PR	- guichets externes et internes	Vérif.	- un opérateur intermédiaire ou non suivant automatique ou semi-automatique	- peuvent s'effectuer par téléphone (pas de déplacement ni de courrier).

(PU) Public

(PR) Privé

(DM) Domestique

(LC) Local

IV. Affectation de tâches nominatives et confidentielles (utilisation d'appareils ne devant servir qu'à une seule personne)	- manutentions spéciales (objets manufacturés, tri postal)	Vérif.	automatique	
V. Applications domestiques DM (fait partie de I)	- protection de domiciles, garages, voitures, par verrou électronique	Vérif. seul.		- une solution à la protection des biens importants
VI. Reconnaissance de la parole PB (aspect complémentaire des recherches sur la reconnaissance de la parole)	- adaptation d'un système de reconnaissance de la parole à plusieurs locuteurs	Ident.	automatique	
VII. Téléconférence PR	- accès réservé aux membres de la réunion (que les messages soient ou non signés) - informer les correspondants de l'auteur du message - surveillance des canaux de communication	Ident.	automatique	- évite intrusion sur le réseau - téléphone recommandé - n'évite pas les écoutes
VIII. Criminologie PR	- recherche de suspects	Ident.		
IX. Applications connexes	- détermination de l'état émotionnel du locuteur			

a) tâches de sécurité

- Identification du locuteur parmi une population donnée, souvent de grande taille .
- Accès vocal à un système (informatique, bureautique), qui peut être combiné aux techniques courantes de contrôle d'identité .
- Surveillance des canaux de communication. La reconnaissance du locuteur est une méthode douce d'espionnage : elle permet de livrer des informations sans préparation spéciale de la personne et à son insu.

b) tâches de communication et de contrôle de la communication

- Contrôle de systèmes (machines outils, ...), tâches de manutention, dans la mesure où une seule personne est autorisée à commander la machine (le contrôle se fait par exemple à chaque début de session). Ceci inclut certaines opérations de levage ou de manutention dangereuses ou délicates, ou hors de portée de l'opérateur, dès que le contact avec le pupitre de commande devient mal commode ou dangereux. Un des intérêts majeurs est que l'opérateur garde les mains libres.
- Réception de messages à partir d'un émetteur éloigné (on désire une exclusivité de la personne réceptrice). Liaison téléphonique ou radio.
- Transmissions de données dans le contexte bancaire et financier : opérations bancaires à distance, recherche de numéro de compte, vérification de cartes de crédit, paiement de factures par téléphone. Il y a un cas où les opérations bancaires (par exemple), à distance, sont irremplaçables : celui des personnes isolées (loin du domicile et/ou de tout centre bancaire), et des handicapés (1). Il leur suffit de disposer d'un terminal et d'une ligne téléphonique. On ne laissera pas non plus de côté le contexte de la plongée sous-marine à grande profondeur et de

---

(1) Le nombre des handicapés est en France très supérieur à un million, dont 40% d'handicapés sensoriels et moteurs (aveugles et sourds totaux ou partiels essentiellement). Il faut y inclure les personnes privées de l'usage des mains, qui ne peuvent pas signer, et ajouter les illétrés, qui peuvent eux aussi prononcer leur nom, mais ne pas signer explicitement.

longue durée, et des missions spatiales... Toutes les techniques développées pour l'analyse, la synthèse de la parole et la reconnaissance du locuteur, sont ici mises en valeur.

On peut aussi considérer une classification des intérêts par "niveau de confidentialité" :

a) Cas de la vérification du locuteur pour l'accès à des informations réservées (par exemple : accès payant à des banques de données).

Intérêt pour tout centre de stockage de l'information : centres de documentation, archives, banques de données, centres de renseignements divers... (automatisés ou non), mais nécessitant encore la présence d'un opérateur intermédiaire. On peut remarquer que de tels centres ont souvent déjà des systèmes d'accès par "niveau d'autorisation".

b) Cas de la vérification du locuteur pour l'accès à des informations confidentielles (par exemple, de type bancaire ou médical). Dans ce cas, la vérification de l'identité permettra la divulgation des renseignements par téléphone et non plus par courrier, si l'intéressé ne veut pas se déplacer. Ceci va dans le sens du "courrier de l'avenir" (développement du télex, des téléconférences, ...) télématique.

Lorsque l'utilisateur se branche sur le système, il peut être soumis à un contrôle vocal automatique d'identité (vérification du locuteur), qui peut être combiné à d'autres formes plus classiques de contrôle d'identité (frappe du numéro de code, ...).

Décrivons maintenant la possibilité de deux autres applications :

- La détermination de l'état émotionnel. Elle a été étudiée par l'extraction de la fréquence fondamentale essentiellement, et de son évolution. Outre les applications criminologiques, GUIBERT (1979) indique qu'en enseignement assisté par ordinateur, il y a là le moyen de juger de l'efficacité d'une méthode pédagogique (ennui, tristesse, refus ou réceptivité, intérêt, plaisir, joie...) et donc de l'adapter au sujet.

Plus généralement, il est possible d'extraire du signal vocal des paramètres détectant des maladies des organes de production ou des défauts dans la production de parole. Au sens le plus large, nous pouvons considérer les applications en reconnaissance du locuteur comme un catalyseur des recherches sur la parole.

- Les organismes de vente par correspondance avec commandes par téléphone, ceux qui livrent à domicile, les agences de réservation de places, s'accommoderaient facilement de systèmes de communication vocale, la "signature vocale" étant pour le vendeur la garantie de sérieux dans la commande. Les clients constituent un fichier qui se met à jour avec chaque nouveau client.

2.5. Etude économique du terminal vocal et du terminal financier :

Nous allons détailler deux appareils spécialisés, car ils se retrouvent dans plusieurs contextes d'application, et peuvent être combinés avec des opérations de reconnaissance du locuteur : le terminal vocal d'acquisition de données, et le terminal vocal financier.

1) Le terminal vocal :

On sait qu'il y a trois avantages majeurs à l'entrée de données sous forme vocale :

a) Elle augmente la productivité de l'opérateur, et le coût de l'équipement est devenu inférieur au coût d'un opérateur humain, ce qui a élargi ses classes d'application [1]. L'apprentissage est réduit, surtout si le codage est très simplifié.

b) Il y a une interaction avec l'opérateur.

c) L'opérateur est entièrement mobile dès qu'il utilise un porte-microphone léger, et a les mains et les yeux libres pour des tâches complémentaires et simultanées.

Le terminal vocal ne nécessite aucun outil extérieur au corps humain, sauf un microphone, qui peut être utilisé à tout moment de la journée et de la nuit, en tout lieu (la voix contourne les obstacles...). S'il résiste bien aux poussières,

---

(1) On ne peut plus faire la liste des systèmes actuellement utilisés aux Etats-Unis, tant ils sont nombreux (reconnaissance par mots isolés sans vérification du locuteur). Il y a de bonnes chances pour que le terminal vocal soit bon marché, surtout s'il est combiné avec un récepteur téléphonique existant (facteur de généralisation et de banalisation de l'accès vocal aux ordinateurs).

fumées, conditions atmosphériques particulières (variation de température, ...), on ne peut toutefois pas en dire autant de la voix.

L'action rendue possible par l'analyse de la parole peut être des plus simples (appuyer sur un bouton), ou des plus complexes (exécuter un sous-programm Voici un exemple de succession des phases opératoires dans le cas d'un système de type interactif :

1) le locuteur énonce les données ou la commande (le message)

2) il vérifie visuellement à l'écran le résultat de l'analyse vocale,

3) il confirme vocalement (oui/non) l'envoi du message pour son exécution.

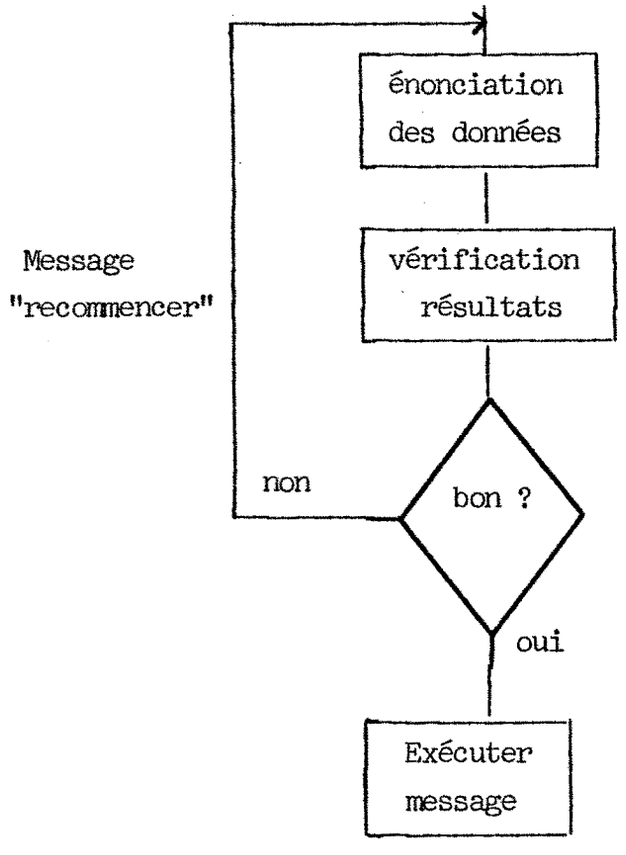


Figure 1- Exemple de système : dialogue d'accès à la commande d'une machine.

Le terminal vocal est un très fort facteur de décentralisation lorsqu'il est rendu intelligent.
La demande de confirmation de la voix (répétition) lors d'une opération de Vérification, est superflue, sauf en cas de :
- score faible et locuteur non satisfait (*)
- ambiguïté sur la décision à prendre (*),
- demande expresse de la part du locuteur (annulation de l'opération).

Dans les cas (*), la ré-initialisation est demandée par la machine.

On doit remarquer que les systèmes opérationnels actuels n'admettent généralement qu'un seul locuteur (et celui qui teste le système doit être celui qui a fait son apprentissage), ou peu de locuteurs ; ce qui poserait pour l'instant dans une moindre mesure, la nécessité d'une vérification d'identité qui soit automatique.

2) Le terminal financier :

Ici, le but est de s'assurer de l'identité du client tout en effectuant une transaction bancaire. Celle-ci se simplifie dans le sens où elle s'accompagnera d'une réduction des échanges de papier.

On est en présence des deux tendances suivantes :

Si le paiement en liquide conserve la primauté (tendance de continuité),

- Les cartes de crédit sont d'un usage en augmentation (première tendance d'évolution). L'une des raisons semble être le besoin à moyen terme de réduire les échanges de papier : il y a à cela des motifs économiques et administratifs.
- Nette orientation vers le guichet automatique, interne ou externe à la banque, capable d'effectuer l'ensemble des transactions financières courantes, sans intervention d'un guichetier.

Des banques ont expérimenté des systèmes à clavier multi-fréquences, avec les codes d'accès successifs suivants : code de liaison avec l'ordinateur, code d'agence, numéro compte client, code interrogation (les opérations possibles). Les avantages observés ont été le faible coût des transactions et leur raccordement au réseau public, les inconvénients, la lenteur du codage et sa complexité.

Regardons de plus près un terminal financier spécialisé. Tout d'abord, s'il est extérieur, il possède des dispositifs de sécurité beaucoup plus élaborés que dans le cas où il est intérieur à la banque. Ensuite, il nécessite la définition et l'évaluation du degré d'exposition aux manipulations malhonnêtes.

Classiquement, il se compose de :

a) Extérieurement, (côté client) :

- un écran-type, qui indique au client les opérations successives à effectuer ;
- un clavier de touches sélection. Les libellés sont variables suivant les types de terminaux ;
- un clavier numérique qui permet d'introduire :
  . le numéro d'identification personnel, le montant du dépôt/retrait en unités préétablies, une demande de chéquier, un distributeur de billets.

b) Intérieurement (côté banque) :

- une imprimante qui liste les transactions effectuées avec les indications habituelles ;
- un moniteur de contrôle à distance, qui, en cas de non-fonctionnement, de manque de papier, interdit les transactions et alerte la banque.

Suivant le contexte d'utilisation, on distingue :

1. Le terminal autonome :

Il y a alors une liaison à un contrôleur qui n'agit sur le terminal qu'en cas de mauvais fonctionnement.

2. Le terminal connecté à un ordinateur :

Lequel met à jour les comptes clients en temps réel.

En ce qui concerne l'utilisation de réseaux d'ordinateurs, les progrès de ceux-ci ne permettent pas encore une diffusion à grande échelle de tels équipements interconnectés. D'une façon générale, les clients appartiennent à un même groupe linguistique (*). La figure 2 montre le schéma de principe d'un terminal vocal.

(*) Bien qu'il existe des guichets automatiques fonctionnant pour 4 langues ("bankomats").

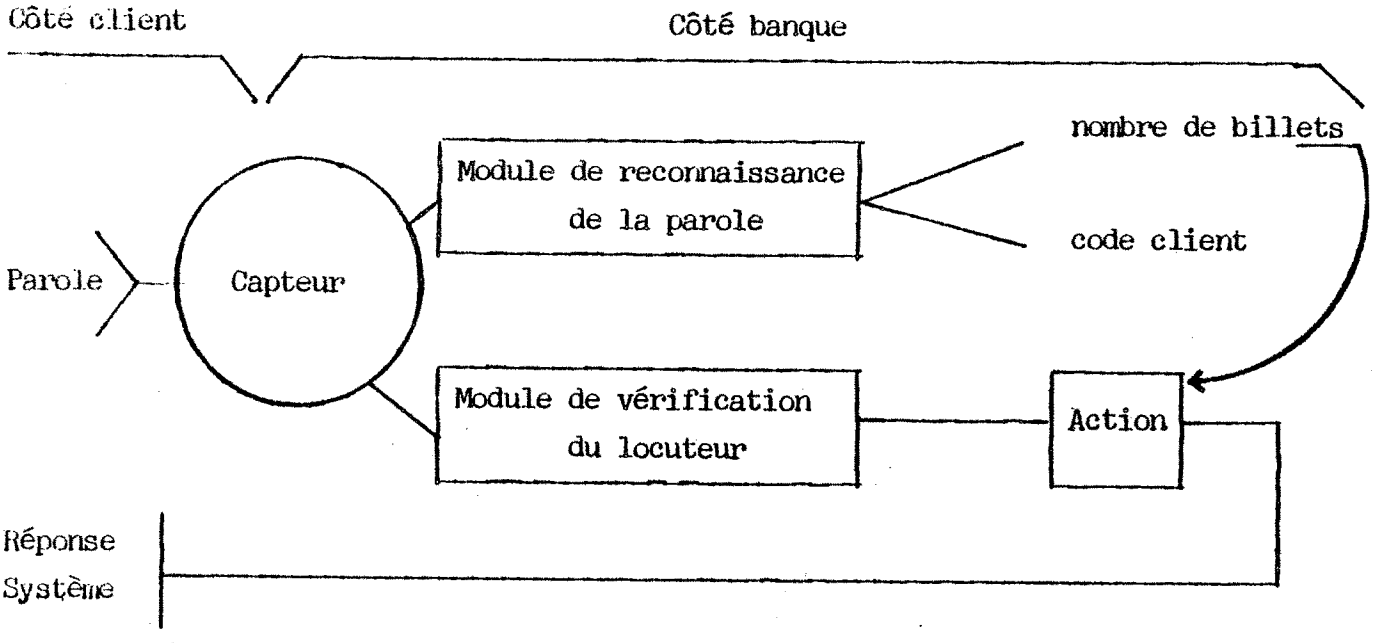


Figure 2 - Schéma de principe d'un terminal financier fonctionnant à partir de la voix.

Une contrainte essentielle est que les systèmes doivent fonctionner en temps réel.
Il faut évoquer le problème de la file d'attente et du temps de réponse du système. Le temps de réponse est perçu par les répercussions qu'il peut avoir sur le client. Par exemple, il n'y a aucune différence entre 0,1 et 0,05 seconde de temps de réponse. Donc, lorsque le client est seul concerné (i.e. lorsque la sollicitation du système est faible ou bien dans le cas d'un guichet automatique), il est suffisant que ce temps soit du même ordre de grandeur que les manipulations manuelles du client.

Le problème de file d'attente se pose lorsque la centralisation de la gestion impose de traiter simultanément un grand nombre de clients situés chacun à un site distinct (cas de forte sollicitation du système). La solution informatique, permettant de gérer ces demandes presque simultanées, doit prendre en compte la façon dont est perçue chaque seconde supplémentaire d'attente par le client (satisfaction marginale d'attente : c'est une valeur négative).

En conclusion, on observera que ces perspectives vont dans le sens de la bureautique, lorsqu'on considère la diminution des échanges de papier et dans le sens des réductions de transport et des déplacements. Ces raisons pourraient être suffisantes pour commander une étude plus officielle - sur l'impact de ces nouvelles techniques sur la société ...

II - PREMIERE EVALUATION DES COUTS D'EXPLOITATION

1 - Evaluation du coût d'exploitation :

Ce coût se décompose en les coûts élémentaires suivants :

- coût du processeur spécialisé (verrou, terminal financier, ...) ;
- coût de connexion à l'ordinateur ;
- coût de préparation de l'application : apprentissage du système ;
- coûts administratifs ;
- coûts d'utilisation, de maintenance, d'amortissement, et de frais de consommables (bandes, disquettes, ...).
- coût de stockage des références des locuteurs (par exemple, tème de Texas Instrument requiert 1200 octets par locuteur.

2 - Evaluation du coût d'utilisation :

- fixation du tarif d'utilisation. Deux tarifs à considérer :
  - . liaison directe ;
  - . liaison à distance (par réseau téléphonique, par réseau ordinateur, par radio).

La possibilité de multiplexage doit être envisagée.

- un mode de facturation : la facturation est principalement fonction du temps d'analyse du signal de parole, du taux de confiance demandé. Elle dépend moins de la distance locuteur-ordinateur.

Il faut en tout cas pouvoir fixer :

- . un coût moyen par opération de vérification,
- . un coût moyen par heure/client.

III - ESTIMATION DE L'IMPACT SUR L'INDIVIDU ET LA SOCIETE

Proposons, pour terminer, quelques conséquences pouvant résulter des applications décrites précédemment. Il y a plusieurs sortes de problèmes :

comportementaux :(personnels et psychologiques;:l'homme face à la machine ,

politiques :(conséquences des "techniques de masse"} et enfin, économiques et sociaux :(changements dans le travail).

Tout d'abord, les applications de Vérification demanderont aux locuteurs de contrôler leur élocution [1], et bien davantage que pour les applications de reconnaissance de la parole. Mais plus généralement, la situation de communication d'homme à machine crée des contraintes quant au comportement de l'individu : sa réaction peut poser des problèmes [2, 3]. Sur les aspects politiques, la construction de systèmes de reconnaissance du locuteur ne peut manquer de poser les problèmes de monopole et de concurrence (choix des paramètres pour chaque système, étendue de l'usage qui peut en être fait).
Evidemment, la question des libertés individuelles fait naître de sérieuses réserves quant à leur diffusion. Mais, "la nécessité d'une moralisation de toutes les activités de surveillance et de mise en fichier, n'est pas nouvelle" (GUIBERT, 1979). Cependant, la tendance peut être à une sectorisation plus forte (hiérarchisation ?) de la société : tel individu se voit autoriser tel niveau d'accès à des informations privilégiées ou à tout autre bien.

En outre, l'Identification sous sa forme la plus générale, pose un problème théorique de taille : étant donné qu'on ne peut que limiter la population d'étude, on ne saura jamais s'il existe un autre individu plus proche de l'échantillon témoin que la meilleure référence obtenue dans la population d'étude ; on ne sait

---

(1) Adoption éventuelle de plusieurs voix, comme on peut adopter plusieurs signatures, suivant le contexte, désir de jouer un rôle d'imposteur (cf. I-I-2.).

(2) La NASA a ouvert des recherches sur de telles réactions.

(3) Le sentiment de souffrir d'une affection ORL ou l'emprise d'une émotion, peut inquiéter le locuteur et le faire échouer dans ses opérations. Le recours aux moyens classiques doit alors pouvoir être autorisé.

constater une égalité de deux échantillons de parole, mais seulement une probabilité d'égalité.

Au point de vue économique et social, le développement des systèmes de communication vocale comporte des risques réels de réduction des effectifs (dans notre cas : suppression de guichets, ou d'opérateurs, réduction de transports, ...). Si la question du reclassement du personnel concerné est faite à l'avance, le secteur tertiaire (et notamment la distribution, les services : transports, télécommunications, banques, assurances, hôtellerie ; administration : gestion du personnel, gestion des stocks), peut entraîner une efficacité accrue, par une réduction de formules administratives en papier, et une productivité plus forte. En supposant qu'elle soit socialement acceptée (mais on peut avoir de petits systèmes), il reste encore à résoudre la question fondamentale : l'informatisation de la tâche de repérage de l'identité et de sa vérification peut-elle contribuer à une garantie de sécurité ?

TABLEAU I- LES DIVERSES CATEGORIES DE PROBLEMES NES DES APPLICATIONS

(adapté de GUIBERT, 1979).

ASPECTS	QUESTIONS SOUS-JACENTES
psychologiques	Rapports de l'homme avec la machine comparés aux rapports inter-humains (plus courants que les précédents, dans la société actuelle).
politiques	- Développement orienté des moyens de masse de reconnaissance du locuteur, ou développement excessif. - Liberté de l'individu face aux moyens de reconnaissance de son identité à son insu.
socio-économiques	- Suppression de documents matériels (clés, cartes magnétiques, preuves d'identité), - Suppressions d'emplois (guichets de banques, surveillance d'accès, ...) et disparition des tâches correspondantes - Facilité d'accès et permanence d'accès à des biens (compte bancaire, ...).

- VI -

1 - MODE DE CLASSEMENT DES REFERENCES

Il se fait par ordre lexicographique, défini sur le quadruplet :

(AT, AN, CO, TT),

où : AT = premier auteur,
AN = année de publication,
CO = liste vide ou non des co-auteurs,
TT = titre de la publication en langue originale.

Ainsi, les publications d'un même auteur sont classées par année, et à année égale par co-auteurs s'il y a lieu, puis par titres. Si ces quatre champs sont les mêmes, un indice (a,b,...) les distingue.

On notera que "Mac GEE" est classé à "M".
La première page regroupe des références sans auteur.

II - GLOSSAIRE DES TERMES PHONETIQUES

Certaines des définitions sont inspirées de G. MOUNIN : Dictionnaire de la linguistique, PUF, 1974.

Accent Accentuation	Phénomène prosodique affectant la syllabe et occasionnant un contraste entre cette dernière et les autres au sein de la même unité accentuelle.
Articulation	Activité musculaire entrant dans la réalisation de la parole.
Coarticulation	Dynamique de liaison entre phonèmes (inter-pollution de phonèmes), par exemple d'une consonne encadrée entre deux voyelles. La coarticulation est liée à des variations situationnelles, dont l'état émotif du locuteur. Elle est explicable par des règles mécaniques.
Dialectologie	Etude de la variation linguistique spatiale et sociale.
Diasystème	Système linguistique permettant à deux variétés dialectales possèdant chacune leur propre système de communiquer.
Formant	Fréquence de résonnance du conduit vocal. Correspond à une zone d'énergie dans le spectre. Un formant est défini par une fréquence centrale, une amplitude et une largeur de bande. On attribue généralement aux formants supérieurs des caractères distinctifs des locuteurs.
Hexis Habitus vocal	(du grec : se tenir) Ensemble des habitudes posturales et gestuelles (l'hexis corporelle de BOURDIEU).

Idiolecte Idiolectal	Ensemble des caractéristiques propres à un locuteur
Individuation	Réalisation proprement individuelle d'un Système
Logatome	Mot sans signification
"Orthophonique"	Mis pour un signal de qualité idéale, par opposition au signal téléphonique, qui est l'exemple même du signal dégradé.
Phonématique	Partie de la phonologie qui étudie les phonèmes
Phonème	Entité linguistique abstraite correspondant aux éléments sonores du langage articulé.
Phonologie	Etude du signifiant phonique.
Polymorphisme	Grande latitude de réalisation pour un phonème
Son	Réalisation de phonèmes.
Spectrogramme	Représentation plane en trois dimensions, du signal de parole : temps, fréquence et amplitude.

REFERENCES EN RECONNAISSANCE DE LA PAROLE

ET DIVERS

BAKER Janet M. (1975)
A New Time Domain Analysis of Human Speech and other Complex Waveform
Ph.D. Dissertation, Mellon Institute of Science, Carnegie-Mellon University, Pittsburg.

BELLISSANT (C.), 1978, *Contribution à l'analyse et la reconnaissance automatique de la parole.* - Thèse d'Etat, IMAG, Grenoble.

BOË (L.J.) & RAKOTOFIRINGA (H.), 1972a , *Une méthodologie systématique de la mesure de la fréquence laryngienne, de l'intensité et de la durée de la parole.* - Bulletin de l'Institut de Phonétique de Grenoble, 1, pp. 1-9.

BOE L.J. (1975)
Quelques Remarques et Précisions Concernant l'Etude Statistique de la Fréquence Laryngienne
Bulletin de l'Institut de Phonétique de Grenoble, IV, 67-84.

CAELEN J. (1979)
Un Modèle d'oreille; Analyse de la Parole Continue, Reconnaissance Phonémique
Thèse d'Etat d'Informatique, Univ. Paul Sabatier, Toulouse.

CHAFCOULOFF M. (1976)
Vingt-cinq Années de Recherches en Synthèse de la Parole
Ed. du CNRS, Paris.

CHOMSKY N. (1976)
La Langage et la Pensée

DAS S.K. & STANAT D.F. (19)
Segmentation of Utterances of a known Phrase Using Linear Threshold Techniques
IEEE Trans. on Audio and Electroacoust.

DAVID E.E. & DEWES P.B. (1972)
Human Communication : A Unified View
Mc Graw-Hill, New York

DE COTRET (1978)
Le modèle linéaire de production de parole utilisé pour la reconnaissance automatique de voyelles non nasalisées de la langue française
9èmes J.E.P., Lannion.

DELATTRE P. (1951)
The Physiological Interpretation of Sound Spectrograms
Proc. Mod. Lang. Assoc., 66, 864-875.

DIXON N.R. & MARTIN T.B. (1979) :
Automatic Speech and Speaker Recognition
I.E.E.E. Press.

FAIRBANKS G. (1960)
Voice and Articulation Handbook
New-York, Harper.

FALL SIDE F. (1976)
Speaker Identification by Multiple Linear Predictive Analysis
I.E.E.E. A.S.S.P, 739.

FANI G. (1960)
Acoustic Theory of Speech Production
Royal Institute of Technology, Stockholm.

FLANAGAN J.L. (1972)
Speech Analysis, Synthesis and Perception
Springer Verlag, Berlin.

FUJISAKI H. & OMURA T. (1971)
Characteristics of Durations of pauser and Speech Segments in Connected Speech
Ann. Rep. of the Eng. Res. Insty Fac. of Eng., University of Tokyo

GRENIER Y. & MAURIN J.C. (1979)
Adaptation au locuteur par analyse canonique des corrélations
Congrès AFCET-IRIA, Reconnaissance des formes et Intelligence Artificielle, Toulouse, 12-14 Sept.

GUIBERT J. (1979)
La Parole ; Compréhension et Synthèse par les Ordinateurs
PUF, Coll. Le Physicien,

HATON J.P. & PERENNOU G. (1978)
La reconnaissance automatique de la parole
VIII ème Ecole d'été de l'AFCET, Namur, Juillet.

JELINEK F. (1976)
Continous Speech Recognition by Statistical Methods
Proc. IEEE, 64, 4, 532-555.

KASHYAP R.L. & MITTAL M.C. (1978)
Recognition of Spoken Words and Phrases in Multitalker Environment Using Syntactic Methods
IEEE Trans. Comp., 27, 5, 442-452.

LIENARD J.S. (1977)
Les processus de la communication parlée
Masson.

LOWERE B.T. (1977)
Dynamic Speaker Adaptation in the Harpy Speech Recognition System
Cours IRIA-SEFI, Reconnaissance et compréhension du dialogue écrit et parlé, 21-25 Nov.

MALMBERG B. (1968)
La Phonétique
P.U.F., 637.

MARKEL J.D. & GRAY A.H. (1976)
Linear Prediction of Speech
Springer Verlag, Berlin.

MOSHIER S.L. (1970)
Some Algebraic Properties of Speech Signals
IEEE, ICASSP, 264-5 (L).

OPPENHEIM A. & SCHAFER R. (19)
Digital Signal Processing
Prentice Hall, Inc., Englewood Cliffs, New Jersey.

PFEIFFER L.L. (1971)
Contributions of Vowel and Nasal Sounds to Speaker Identification
JASA, JJ, 2, 462 (A)

POTTER R. & al. (1966)
Visible Speech
Dover, New-York.

RAMSAY R.W. (1968)
Speech Patterns and Personality
Language and Speech, 11, 54-63

SCHAFER R.W. & RABINER L.R. (1975)
Digital representations of speech signals
Proc. IEEE, 63, 662-677

WELCH P.D. & WIMPRESS R.S. (1961)
Two Multivariate Statistical Computer Programs and Their Application to the Vowel Recognition Problem
J.A.S.A., 33, 426-434.

WILLIAMS (C.E.) & STEVENS (K.N.), HECKER (M.H.), 1970, *Acoustical Manifestations of Emotional Speech*. - Journal of the Acoustical Society of America 47, 66 (A).

REFERENCES EN RECONNAISSANCE DES FORMES

AGRAWALA A.K. Ed. (1977)
Machine Recognition of Patterns
IEEE Press. New-York.

ANDERSON T.W. (1958)
An Introduction to Multivariate Statistical Analysis
J. Wiley. New-York.

ANDREWS . (1972)
Introduction to Mathematical Techniques in Pattern Recognition
Wiley-Interscience, New-York.

BATCHELOR, B.G. (1974)
Practical approach to Pattern Classification,
Plenum Press-London & New-York.

BENZECRI J.C.
Leçons sur la reconnaissance des formes
Cours de l'ISUP, Paris VI.

CAILLEZ F. & PAGES J.P. (1976)
Introduction à l'Analyse des Données
SMASH, Paris.

CHEN C.H. (1978)
A review of Statistical Pattern Recognition
NATO, ASI, PARIS.

CHEN C.H. (1973)
Statistical Pattern Recognition
Rochelle Park, Sparten.

CHEN C.H. (1978)
Finite Sample Considerations in Statistical Pattern Recognition
Proc. of Pattern Recognition and Image Processing.

C.K. CHOW
An optimum recognition error and reject tradeoff, IEEE Trans. Inform. Theory, Vol. IT-16, pp. 41-46, Jan. 1970.

CORAY, G. & STAMON G.
Eléments de Reconnaissance des Formes
Ecole Internationale d'Eté d'Informatique A.F.C.E.T., Tarbes, Juillet 1974.

CORSI P. (1979)
Qu'est-ce que la Reconnaissance des Formes ?
Informatique et Gestion,
(à paraître)

CORSI P. (1979)
Introduction à la Reconnaissance des Formes Statistique
Cours de 3ème Année ENSIMAG, DEA Informatique USMG.

T.M. COVER and P.E. HART

Nearest neighbor pattern classification, IEEE Trans. Inform. Theory,

COVER T.M. (1968)

Rates of convergence of nearest neighbor decision procedures

Hawaï Int. Conf. Systems Theory 413-415.

COVER T.M. (1969)

Learning in Pattern Recognition

Methodologies of Pattern Recognition, S. WATANABE Ed., Academic Press, New-York, 111-132.

COVER T.M. (1972)

A Hierarchy of Probability Density Function Estimates

in Frontiers of Pattern Recognition , S. WATANABE Ed., Academic Press, New-York, 83-98.

COVER T.M. (1973)

Recent Books on Pattern Recognition

IEEE Trans. Inform. Theory, IT-19, 827-833.

COVER T.M. (1974)

The best two measurements are not the two best

IEEE Trans. Syst. Man Cybern. (Corresp.), SMC-4, 116-117.

COVER T.M. & WAGNER T.J. (1976)

Topics in Statistical Pattern Recognition

in Digital Pattern Recognition, K.S. FU Ed., Springer-Verlag, Berlin, 15-46.

DAS S.K. & MOHN W.S. (1969)

Pattern Recognition in Speaker Verification

Fall Joint Computer Conf.

DAS S. K. (1971)

Feature Selection With Linear Dependence Measure

IEEE Trans. on Comp. (corresp.), 20, 9, 1106, 1109.

P.A. DEVIJVER

Relationships between statistical risks and the least-mean-square-error design criterion in pattern recognition, Proc. 1 st. Int. Joint Conf. Pattern Recognition, IEEE Special Publication CHO 821-9c, 1973, pp. 139-148.

DEVIJVER P. (1977)

Reconnaissance des formes par la méthode des plus proches voisins

Thèse Doct.Ing., Univ. Paris VI.

DIDAY E. (1972)
Optimisation en Classification Automatique et Reconnaissance des Formes
IRIA, Note Scientif. 6.

DIDAY E. (1974)
Recent Progress in Distance and Similarity Measures in Pattern Recognition
Proc. 2nd Int. Joint Conf. on Pattern Recognition, Aug. 13-15.

DUBES R., JAIN A.K. (1976)
Clustering Techniques
Pattern Recognition, 8, 4.

DUDA R.O. & HART P.E. (1973)
Pattern Classification and scene analysis.
Wiley, N.Y.

S.C. FRALICK and R.W. SCOTT, Nonparametric Bayes-risk estimation,
IEEE Trans. Inform. Theory, Vol. IT-17, pp. 440-444, 1971.

FRANCES R. (1975)
La Perception des Formes et des Objets
in Traité de Psychologie Expérimentale, P. FRAISSE & J. PIAGET Ed., VI, 199-254, PUF, Paris.

FU K.S. (1968)
Sequential Methods in Pattern Recognition and Machine Learning
Academic Press., New-York.

FU K.S. (1974)
Syntactic Methods in Pattern Recognition.

K. FUKUNAGA and D.L. KESSEL
Application of optimum error-reject functions, IEEE Trans. Inform. Theory, Vol. IT-18, pp. 814-817, Nov. 1972.

FUKUNAGA K. (1973)
Introduction to Statistical Pattern Recognition
Academic Press, New-York.

K. FUKUNAGA and D.L. KESSEL
Nonparametric Bayes error estimation using unclassified samples,
IEEE Trans. Inform. Theory, Vol. IT-19, pp. 434-440, July 1973.

GUILLAUME P. (1937)
La Psychologie de la Forme
Flammarion, Paris.

HELLMAN M.E. (1970)
The nearest-Neighbor classification rule with a reject option
IEEE Trans. Syst. Sci., Cybern., 6, 179-185.

HENRICHON E.G. & FU K.S. (1968)
On mode estimation in pattern recognition
Proc. 7th Symp. Adaptive Processes, UCLA, 3-a-1.

HENRICHON E.G. & FU K.S. (1969)
A non parametric partitionning procedure for pattern classification
IEEE Trans. Comp., 18, 614-624.

HORTON I.F., RUSSEL J.S. & MOORE A.W. (1968)
Multivariate Covariance and Canonical analysis : a method for selecting the most effective discriminators in a multivariate situation.
Biometrics 845-858.

JAIN A.K. & DUBES R. (1978)
Feature Definition in Pattern Recognition with Small Sample Size.
Pattern Recognition, 10, 85-97.

KANAL L. (1968)
Pattern Recognition
Washington, DC : Thompson.

KANAL L. & CHANDRASEKARIAN B. (1971)
On Dimensionality and Sample Size in Statistical Pattern Recognition
Pattern Recognition 3, 225-234.

L.N. KANAL.
Patterns in Pattern Recognition : 1968-1974, IEEE Trans. Inform. Theory
Vol. IT-20, pp. 697-722, Nov. 1974.

KASHYAP R.L. (1977)
Optimal Feature Selection and Decision Rule in Classification Problems With Time Series
Proc. Conf. Inform. Sciences ans Systems, John Hopkins University, Baltimore, MD.

KITTLER J. (1975)
Mathematical Methods of Feature Selection in Pattern Recognition
Int. Jnl. Man-Machine Studies, 7, 603-637

KITTLER J. (1978)
Une généralisation de quelques algorithmes sous-optimaux de recherche d'ensembles d'attributs.
Congrès AFCET-IRIA, Reconnaissance des Formes et Traitement des Images, Chatenay-Malabry, 678-686.

KOVALEVSKY V.A. (1978)
Recent Advances in Statistical Pattern Recognition 4th IJCPR, Kyoto, Nov. 1978.

LEBARD, MORINEAU & TABARD (19)
Techniques de la description statistique
DUNOD

LERMAN I.C. (1971)
Cours de Classification Automatique
Université de Paris VI.

LERMAN I.C. (1971)
Les bases de la classification automatique.
Gauthier-Villars, Paris.

LERMAN I.C. (1976)
Formal Analysis of a General notion of proximity between variables
Proc. Europ. Congress of Statisticians.

LORD & NOVI
Theory of Mental Tests

MEISEL W. (1972)
Computer-Oriented Approaches to Pattern Recognition
Academic Press, New-York.

MENDEL J.M. & FU K.S. Ed. (1970)
Adaptive, Learning and Pattern Recognition Systems : Theory and Applications.

MURPHY R.B. (1948)
Nonparametric Tolerance Limits
Ann. Math. Stat., 17, 377-408.

NAGY G. (1968)
State of the Art in Pattern Recognition
Proc. IEEE, 56, 836-862.

N.J. NILSSON.
Learning Machines : Foundations of Trainable Pattern Classifying Systems,
New-York : Mc Graw-Hill, 1965.

NILLSON (1972)
Introduction to Artificial Intelligence
Academic Press.

PATRICK E.A. (1972)
Fundamentals of Pattern Recognition.
Prentice Hall, N.J.

PAU L.F. (1978)
On finite Learning Sample Size Problems in Pattern Recognition
NATO, ASI, PARIS.

PYKETT C.E. (1977)
Intuition and Empiricism in Pattern Recognition
Seminar on Pattern Recognition, Nov. 1977, Liège.

RAUDYS S. & PIKELIS V. (1978)
On Dimensionality, Sample Size and Classification error in Discriminant Analysis.
Submitted.

ROCHE C. (1972)
Information Utile en Reconnaissance des Formes et en Compression des Données: Application à la Génération Automatique de Systèmes de Reconnaissance Optique et Acoustique
Thèse d'Etat, Univ. Paris VI.

ROMEDER J.M. (1973)
Méthodes et programmes d'analyse discriminante
DUNOD

ROSENBLATT F. (1962)
Principles of Neurodynamics
Spartan Books.

RUSPINI E.H. (1969)
A new Approach to Clustering
Information and Control 15, 22-32.

SEBESTYEN G.S. (1962)
Decision-Making Processes in Pattern Recognition
Macmillan, New-York.

SKYVINGTON W. (1976)
Machina Sapiens - Essai sur l'intelligence artificielle.
Seuil.

SOKAL R.R. & P.H.A. SNEATH (1963)
Principles of Numerical Taxonomy
W.H. Freeman, San Francisco, Calif.

STEARNS S.D. (1976)
On selecting features for Pattern Classifiers
Proc. 3rd Int. Joint Conf. Pattern Recognition, Coronado, 71-75.

THOM R. (1974).
Modèles mathématiques de la morphogénèse.
Collection 10/18, 887.
Union Générale d'Editions, Paris.

TOUSSAINT G.T. (1974)
Bibliography on Estimation of classifications
IEEE Trans. I.T., 20, 472-479.

TOUSSAINT G.T. (1974)
Recent Progress in Statistical Methods applied to Pattern Recognition
Proc. 2nd Int. Joint Conf. Pattern Recognition, Aug. 13-15.

TOUSSAINT G.T. (1978)
The use of context in Pattern Recognition
Pattern Recognition Journal, 10, 3, 189-204.

TULMAN H.G. (1967)
Automatishe Identifikation Von Sprechern
NTZ, Helf 12.

ULLMAN J.R. (1973)
Pattern Recognition Techniques
Crane, Russak & Co.
London, Butterworth.

UHR L. (1973)
Pattern Recognition, Learning & Thought
Prentice Hall, Inc., Englewood Cliffs, New-Jersey.

VERHAGEN C.J.D.M. (1975)
Some general remarks about Pattern Recognition ; Its definition ; its relations with other disciplines ; A litterature survey.
Pattern Recognition, 7, Sept.

WALD A. (1952)
Sequential Analysis
John Wiley & Sons, Inc., New York.

WATANABE S. (1969)
Knowing and guessing
John Wiley.

WATANABE S. Ed. (1969)
Methodologies of Pattern Recognition
Academic Press, New-York.

WATANABE S. (1972)
Frontiers of Pattern Recognition
Academic Press, New-York.

YOUNG I.T. (1974)
The prediction of performance in multi-class pattern classification
Proc. 2nd Int. Joint Conf. Pattern Recognition, Aug. 13-15

YOUNG T.Y. & CALVERT J.W. (1974)
Classification, Estimation and Pattern Recognition.
Elsevier. New-York.

Computer Oriented Learning Processes (1974)
Ed. J.C. SIMON
Nato Advanced Study Institute on Computer Oriented Learning Processes
Bonas, Aug. 26-Sept. 5.

REFERENCES EN RECONNAISSANCE DU LOCUTEUR (*)

(1938-1979)

(*) Documentation réunie grâce à l'appui
de l'Institut de Phonétique de Grenoble.

) - REFERENCES GENERALES (SANS AUTEURS)

AEROSPACE CORP. (1977)
Speaker Identification
Final Program Report, Aerospace Report No. ATR-77(7617-07)- 1.

ANON (1965) :

Bibliography : Speech Identification by Eye, Ear and Machine.
Information Center for Hearing, Speech, and Disorders of Human Communication.
Johns Hopkins Medical Inst. Baltimore, MD.

ANON (1965) :

Speaker Authentification Techniques.
Final Rep. Contrat DA-28-043-AMC-00116(A) U.S. Army Electron. Labs. For Monmouth. N.J. .

Anonyme (1965)
Voice Print Identification
Criminalistics, W.W. Turner Ed., Aqueduct Books, Rochester.

L'IDENTIFICAZIONE DELLA PERSONA PER MEZZO DELLA VOCE (1978)
Atti della Tavola Rotonda, Padova, 14-15 Sett.
A cura di F. FERRERO, Collana degli Atti della Rivista Italiana di Acustica, Vol 1, ESA-Edizioni Scientifiche Associate, Roma,1979.

MICHIGAN STATE POLICE (1972)
Voice Identifivation Research
(Submitted by), PR 72-1, Feb., East Lansing, Michigan
US Department of Justice.

On the theory and Practice of Voice Identification.Office of Publications. National Academy of Sciences, 2101 Constitution Avenue, NW, Washington, D.C. 20418.

2 - REFERENCES PAR AUTEURS

ABRY C. & BOE L.J. (1979)
Pour une idiolectologie: Aspects phonétiques de l'Identité
Colloque Internati onal Production et Affirmation de l'Identité, Toulouse, Sept.

ALPERT M., KURTZBERG R.L., PILOT M. & FRIEDHOFF A.J. (1963)
Comparison of the Spectra of the Voices of Twins.
J.A.S.A. 35, 1877 (A)

ALLPORT G.W. & CANTRIL H. (1934) :
Judging Personality from Voice.
J. Soc. Psychol. 5, 37-55.
Journal of Social Psychology

ATAL B.S. (1968)
Automatic Speaker Recognition based on Speech Contours
Ph.D. Diss., Dept. of Elec. Eng., Polytechnic Inst. of Brooklyn.

ATAL B.S. (1972) a.
Text-Independant Speaker Recognition.
J.A.S.A. 52, 181 (A) . 83th ASA Asil New-York.

ATAL B.S. (1972) b :
Automatic Speaker Recognition based on Pitch Contours.
J.A.S.A. 52, 1687-1697.

ATAL B.S. (1974) :
Effectiveness of Linear Prediction Charateristics of the Speech Wave for Automatic Speaker Identification and Verification.
J.A.S.A., 55, 6, 1304-1312

ATAL B.S. (1976) :

Automatic Recognition of Speakers from their Voices.

Proc. IEEE 64 , 460-475.

ATKINSON J.E. (1976)

Inter and Intra-Speaker Variability in Fundamental Voice Frequency
J.A.S.A., 60, 2, 440-455.

BEAKLEY G.W. & TUTEUR F.B. (1972)

Distribution-free pattern Verification using Statistically equivalent blocs

IEEE Trans. Comput., C-21, 1337-1347.

BECKER M.H. , GNANADESIKIAN R. , MATHEWS M.V. , PINKAM R.S. , PRUZANSKY S. & WILK M.B. (1964) :

Comparisons of Some Statistical Distance Mesures for Talker Identification

J.A.S.A. 36 , 1988(A).

BEEK B. & al. (1971) :

Automatic Speaker Recognition System.

Agard conf. on Artificial Intellignece - 94.
Harford House, London.

BEEK B. , GRECH J. & MEEKER F. (1971) :

Automatic Speaker Recognition Systems.

Rome Air Development Center.Report.

BEEK B. , NEUBERG L.P. , HODGE D.C. (1977) :

An Assessment of the Technology of Automatic Speech Recognition for Military Applications.

IEEE Trans. Acoust. Speech & Signal Process.
ASSP-25 , 310-322.

BLACK J.W. , LASHBROOK W. , NASH E. , OYER H.J. , PEDREY C. , TOSI O.I. & TRUBY H. (1973) :

Reply to "Speaker Identification by Speech Spectrograms : some further Observations".

J.A.S.A. 54 , 535-537.

BOGNER R.E (197)
On Talker Verification Via Orthogonal Parameters
Bell Laboratories Report.

BOLT R.H. , COOPER F.S. , DAVID E.E.J. , DENES P.B. , PICKETT J.M. , STEVENS K.N.,(1
Identification of a Speaker by Speech Spectrograms. How do Scientists View its Reliability for Use as Legal Evidence ?
Science 166, 338-343.

BOLT R.H. , COOPER F.S. , DAVID E.E.Jr , DENES P.B. , PICKETT J.M. , STEVENS K.N.,(19
Speaker Identification by Speech Spectrograms : A Scientistist's View its Reliability for Legal Purposes.
JASA 47 , 597-612.
in Human Communication : A Unified view (1972).
ed. by DAVID E.E.Jr. & DENES P.B.
Chap. 10 Mc. Craw Hill Book Co.

BOLT R.H. , COOPER F.S. , DAVID E.E. , DENES P.B. , PICKETT J.M. & STEVENS K.N.,(1973)
Speaker Identification by Speech Spectrograms : some further Observatic
J.A.S.A. 54 , 531-534.

BONAVENTURA M. (1935) :
Ausdruck der Persönlichkeit in der Sprechstimme und im Photogramm.
Arch. Geo. Psychol. 94 , 501-570.

BORDONE - SACERDOTE C. & SACERDOTE G.G. (1969) :
Some Spectral Properties fo Individual Voices.
Acustica 21 , 199-210.

BORDONE C. , SACERDOTE G.G. (1977) :
Some Aspects of Londness in Speaker Identification.
9 Inter. Congr. Acoustics. I28.

BORDONE - SACERDOTE C. & SACERDOTE G. (1978)
Statistical properties of individual voices
6th International Congress on Acoustics, B-2-10,
Tokyo, 21-28 Aug. 1968.

BOULOGNE M. , CARRE R. , CHARRAS J.P. (1971) :
La fréquence fondamentale, les formants, éléments d'identification du locuteur.
Rapport ENSERG - Grenoble.

[illegible] , CARRE R. , CHARRAS J.P. (1973) :

La fréquence fondamentale, les formants, éléments d'identification des locuteurs.

Revue d'Acoustique 23 , 343-350.

[illegible] P.D. & PRUZANSKY S. (1966) :

Effects of Stimulus Content and Duration on Talker Identification

J.A.S.A. 40 , 1441-1449.

BRICKER P.D. , GNANADESIKAN R. , MATHEWS M.V. , PRUZANSKY S. , TUKEY P.A. , WACHTER K.W. & WARNER J.L. (1971) :

Statistical Techniques for Talker Identification.

B.S.T.J. 50 , 1427-1454.

BRICKER P.D. , PRUZANSKY S. (1976) :

Speaker Recognition.
in Contemporary Issues in Experimental Phonetics.

Ed. by N.J. LASS. Chapter 9 , 295-326.
Academic Press. New-York , San Francisco , London.

BROWN B.L. (1969) :

A Social Psychology of Variations in French Canadian Speech Styles

Unpublished Doct. Dissert. Mc Gill Univ.

BROWN B.L. , STRONG W.J. , BENCHER A.C. (1973) :

Perceptions of Personality from Speech : Effects of Manipulations of Acoustical Parameters.

J.A.S.A. 54, 29-35.

BROWNER P. , STENZEL E. (1976) :

Spracherkennung mit Adaptation an Sprechereigen Schaften (Reconnaissance de la parole avec adaptation au locuteur).

Fortsch . Akust. Tag. Daga 76 , 661-664.

BUNGE E. (1976)

Automatishe Sprechererkennung mit Computern (Identification Automatique par la voix à l'aide de l'ordinateur)
5ème Colloque sur l'Identification d'échantillons Acoustiques, Inst. Trait. Données Tech. et Biol.. Karlsruhe. 27 Fév.

[illegible] E. (1976) :

Statistical Techniques for Speaker Recognition.

5th Meeting of the Deutsche Arbeitsgemeinschaft für Akustik.
DAGA 76 , Heideberg.

BUNGE E. (1976) :

Statistische Verfarren zur Automatischen Sprecherkennuns.
(Procédés statistiques de reconnaissance automatique du locuteur).

BUNGE E. (1976)
System zur Automatishen Sprechererkennung
Proc. 5. IITB Kolloquium Mustererkennung.

BUNGE E. (1977) :

AUROS - Automatic Recognition of Speakers by Computers Principles of the Speaker Recognition System.
9th Int. Congr. Acoust. I 103.

BUNGE E. (1977)
Automatic Speaker Recognition by Computers
Proc. of the Seminar on Pattern Recognition, Liege University, Sart-Tilman, Belgique, 19-20 Nov.

BUNGE E. (1977)

Comparative Investigations on Automatic Identification and Verification of Cooperative Speakers.
Dissert. Technical Univ. Darmstdat.

BUNGE E. (1977)

Speaker Recognition by Computer
Philips Technical Review, 37, 207-219, 1977, n° 8.

BUNGE E. (1977)
Vergleichende Systematishe Untersuchungen Zur Automatishen Identifikation Und Verifikation Kooperativer Sprecher
These Docteur-ingénieur, Dem Farbereich Nachrichtentechnik der Technischen Hochshule, Darmstadt.

BUNGE E. , HÖFKER V. , HÖHNE H.D. , JESORSKY P. , KRIENER B. , WESSELING D. , (1977)
Report about Speaker Recognition Investigation with the AUROS System.
FREQUENZ, Dec.

BUNGE E. , HÖFKER U. , JESORSKY P. , KRIENER B. , WESSELING D. (1977) :
Statistical Techniques for Automatic Speaker Recognition.
IEEE Congr. ASSP, 772-775.

BUNGE E. , HÖFKER U. , HÖHNE H.D. , JESORSKY P. , KRIENER B. , WESSELING D. (1977) :
The AUROS System - Automatic Recogn. of Speakers by Computers.
FREQUENZ Dec. .

BUNGE E. & JACKSON J.S. (1977) :

Automatic Speaker Recognition System AUROS for Security Systems and Forensic Voice Identification.

Inter. Conf. on Crime Countermeasures - Science and Engineering 8,1-7.

CARBONELL F.R. , GRIGNETTI M.C. , STEVENS K.N. , WILLIAMS C.E. & WOODS B. (1965) :

Speaker Authentification Techniques. Rpt. 1296. Contract.

DA-28-043-AMC-00116(E) by BOLT, BERANEK & NEWMAN, Inc. Cambridge, Mass.

CARRE R. , LANCIA R. , WAJSKOF W. (1968) :

Sur la production des voyelles par les locuteurs hommes et femmes.

6th Int. Congr. Acout. B-31.

CARRE R. (1971) :

Contribution aux études sur l'analyse et la synthèse de la parole. Rôle et importance des formants.

Thèse Etat Sci. Univ. I. Grenoble.

CARRE R. (1971) :

Identification des locuteurs : exploitation des données relqtives aux fréquences des formants.

7th Int. Congr. Acoustics. 29-32

CHAPMAN W.D. & LI K.P. (1966) :

Speaker Verification.

J.A.S.A. 40, 1282 (A).

CLARKE F.R. , BECKER R.W. & NIXON J.C. (1966) :

Characteristics that Determine Speaker Recognition.

Rept. ESD-TR-66-636. Decision Sciences Laboratory, Hanscom Field, Bedford , Mass.

CLARKE F.R. & BECKER R.W. (1969) :

Comparison of Techniques for Discriminating among Talkers.

J.S.H.R. 12 , 747-761.

COLEMAN R.O. (1971) :

Male and Female Voice Quality and its Relationship to Vowel Formant Frequencies.

J.S.H.R. 14 , 565-577.

COLLINS A.M. (1977) : a/

Computer Speech Processing for Speaker identification.

Publication EP-RR37
Department of Engineering Physics - Research Scool of Physical Sciences, The Australian National University

COLLINS A.M. (1977) : b/

The Acquisition of an Australian Speech Data Base.

To be published in : Proceedings of the Digital Equipment Computer Users Society, Vol. 3, n° 5, Australia.

COLLINS A.M. (1979)

A data base for Digital Speech Processing Research

Institute of Radio and Electronics Engineers of Australia Conference. Aug. 1979.

COLLINS A.M. (1979)

Digital Signal Processing

Conference in Adelaïde, Mai 1979.

COMPTON A.J. (1963) :

Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally.

J.A.S.A. 35 , 1748-1752.

CORSI P. (1978)

Introduction à la Reconnaissance du Locuteur: contexte de l'étude et problèmes fondamentaux
Bulletin de l'Institut de Phonétique de Grenoble, 7.

CORSI P. (1978)

Studi statistici di parametri temporali per la discriminatione di parlatori.

Proc. Tavola Rotonda Sull'Identificazione della Persona per mezzo della sua voce, Padova, 14-15 Sept. 1978.

CORSI P. (1979)

Compréhension et Synthèse Automatiques de la Parole, Reconnaissance Automatique de Locuteurs: Panorama général et Perspectives d'Application
2ème Congrès AFCET-IMAG de Bureautique, Grenoble, 30 Mai-1er Juin.

CORSI P. (1979)

Techniche Descrittive per la Discriminazione del Parlatore su dei Parametri Temporali
Rivista Italiana di Acustica, Nov.

[illegible] & DUBÉ L.J. (1979)
Définition et Sélection de Caractéristiques Temporelles en vue de la Vérification Automatique du Locuteur
Congrès AFCET-IRIA Reconnaissance des Formes et Intelligence Artificiell
Toulouse, 12-14 Sept.

CURRY E.T. : (1939)
An objective study of the pitch characteristics of the adolescent male Voice.
Ph. D. Dissert. State University of Iowa.

CURRY E.T. : (1940)
An objective study of the pitch characteristics of the adolescent male voice.
Speech Monographs 7 : 48-62.

CURRY E.T. (1946) :
Voice Change in Male Adolescents.
Laryngoscope 56, 795-805.

CURTIS T.H. (1974)
Interspeaker V.S. Intra-Speaker Variability of Glottal Pulse Shapes
JASA, 55, 2, 462 (A)

DAS S.K. (1969) :
A method of Decision in Pattern Recognition (With an Application to the Practical Problem of the Speaker Verification).
IEEE Trans. on Computers C-18, 329-333.

DAS S.K. & MOHN W.S. (1969)
Pattern Recognition in Speaker Verification.
Fall Joint Computer Conf. Las Vegas. 35 , 721-732.

DAS S.K. , MOHN W.S. & SALEEBY S.L. (1970) :
Speaker Verifications Experiments.
J.A.S.A. 49 , 138(A).

DAS K.S. & MOHN W.S. (1971) :
A Scheme for Speech Processing in Automatic Speaker Verification.
IEEE Trans. Audio Electr. AU-19, 32-43.

DAS S.K. , MOHN W.S. , WILLETT S.S. & CHAPMAN W.D. (1972) :

Two Speaker Verification Experiments.

IEEE AFCRL
Conf. Speech. Comm. & Process 275-278

DAS S.K., MOHN W.S., WILLET S.S. & CHAPMAN W.D. (1972)

Two Speaker Verification Experiments

Proc. IEEE Conference on Speech Comm. and Processing, Air Force Cambridge Research Lab., Bedford, MA, 275

DODDINGTON G.R. (1970) :

A Method of Speaker Verification

PR.D. Dissert. Univ. Wisconsin.
Dept. EE.

DODDINGTON G.R. (1971) :

A Method of Speaker Verification

J.A.S.A. 49 , 139(A).

DODDINGTON G. , FLANAGAN J.L. & LUMMIS R.C. (1972) :

Automatic Speaker Verification by Non-Linear Time Alignment of Acoustic Parameters.

U.S. Patent 3 700 815.

DODDINGTON G. , HYDRICK B. & BEEK B. (1973) :

Some Results on Speaker Verification using Amplitude Spectra.

86th ASA oct. Los Angeles

DODDINGTON G. (1974) a :

Speaker Verification.

Tech. Rep. RADC-TR-U1-963 700 F.
Rome Air Development Center, Griffis AFB , N.Y.

DODDINGTON G.R. (1974) b.

Speaker Verification - Final Report. Tech. Rept. RADC 74-179

Rome Air Development Center, Griffis AFB , N.Y.

DODDINGTON G.R. (1975) :

Speaker Verification

Tech. Rept. TR-75-274
Rome Air Development Center, Griffis AFB , N.Y.

DODDINGTON G.R. (1976) :

Speaker Verification

Tech. Rept. TR-76-262.
Rome Air Development Center. Griffis AFB N Y

EDIE J. & SEBESTYEN G.S. (1962) :

Voice Identification General Criteria.

Rept. RADC-TDR-62-278. Litton Systems Inc. Rome Air Development Center, Air Force Systems Command, Griffiss AFB.

EL CHAFEI Chérif (1978)

Un système de reconnaissance automatique de locuteurs sur mini-ordinateur.

Thèse Doct-Ingénieur, Paris-Sud, ORSAY.

EL CHAFEI C. (1979)

Un systeme de reconnaissance automatique de locuteurs sur mini-ordinateur
Congrès AFCET-IRIA Reconnaissance des Formes et Intelligence Artificielle, Toulouse, 12-14 Sept.

ENDRES W. (1970) :

Changes of Human Voice Caused by Age, Disguise, and Simulation.

J.A.S.A. 49 , 138(A).

ENDRESS W., BAMBACH W., FLOSSER G. (1971)
Voice Spectrograms as a Function of Age, Voice Disguise and Voice Imitation
J.A.S.A., 49, 6(2), 1842-1848.

FAIRBANKS G. & PRONOVOST W. (1939) :

An Experimental Study of the Pitch Characteristics of the Voice During The Expression of Emotion.

Speech Monographs 6, 87-104.

FAIRBANKS G. (1940) :

Recent Experimental Investigations of Vocal Pitch in Speech.

J.A.S.A. 11 , 457-466.

FAIRBANKS G. & LE MAR HOAGLIN W. (1941) :

An Experimental Study of the Durational Characteristics of the Voicing during the Expression of Emotion.

Speech Monographs 8, 85-90.

FALLSIDE F. (1976)

Speaker Identification by Multivariable Linear Prediction Analysis.

IEEE Congr. ASSP, 739.

FASOLO L. & MIAN G.A. (1978) :

A Comparison between two Approaches to Automatic Speaker Recognition.

IEEE ICASSP 275-277.

FLANAGAN J.L. (1971)

Research on Speaker Verification

National Academy of Sciences, March.

FLANAGAN J.L. , CLARKE F.R. , COOPER F.S. , HOGAN D.L. , HOUSE A.S. , PARRACK H.O. , POLLACK I. , STEVENS K.N. (1971) :

Research on Speaker Verification.

NAS-NRC Comm. on Hry B.acoust. Biomed. Washington D.C.

FLANAGAN J.L. (1973) :

Digital Techniques in Communications Acoustics.

IEEE NEREM Rec. 15, pt 2 , Signal Processing
IEEE Cat.No 73 CHO 840-9 , NEREM , Boston.

FLOYD W. (1964) :

Voice Identification Techniques.

Rept. RADC-TDR-64-312. Rome Air Development Center, Research and Technology Division, Air Force Systems Command. Griffiss AFB

FURUI S. , ITAKURA F. & SAITO S. (1971) :

Talker Recognition Considering the Variation of Long-Time Average Spectrum.

Conv. Rec. Acoust. Soc. of Japan. 3-1-18.

FURUI S. , ITAKURA F. & SAITO S. (1972) :

Talker Recognition by long time Averaged Speech Spectrum.

Electronics and Communications in Japan 55-A , 54-61.

FURUI S. & ITAKURA F. (1973) :

Talker Recognition by Statistical Features of Speech Sounds.

Electronics and Communications in Japan 56 A , 62-71.

FURUI S. (1974) :

An Analysis of Long-Term Variation of Feature Parameters of Speech and its Application to Talker Recognition.

Electronics and Communication in Japan 57A, 34-42.

FURUI S., ITAKURA F. & SAITO S. (1975)
Personal Information in the Long-time Averaged Speech Spectrum
Review Elec. Commun. Lab., 23, 1133-1141.

FURUI S. (1978)

Research on Individual Information in Speech Wave.
Ph. D., Tokyo University, Faculty of Engineering, Math. Eng. and Instrum. Physics Dept.

FURUI S. (1979)

Effects of long-term Spectral Variability on Speaker Recognition
ASA and Acoust. Soc. of Japan Joint Meeting, NNN28.

GARVIN P. & LADEFOGED P. (1963) :

Speaker Identification and Message Identification in Speech Recognition.

Phonetica 9 , 193-199.

GLENN J.W. & KLEINER N. (1968) :

Speaker Identification based on Nasal Phonation.

J.A.S.A. 43 , 368-372.

GNANADESIKAN R. & WILK M.B. (1966) :

Statistical Techniques for Effective Condensation of Talker Identification Data.

Ann. Math. Stat. 37,1415(A).

GOLDSTEIN U.G. (1975)
Some Speaker-Identifying Features based on Formant Tracks
U.S. Navy Office of Naval Research, Contract N 00014-67-A-0204-0069.

GRAY C. & KOFF G. (1944) :

Voiceprint Identification

Rep. Bell Telephone Lab. Inc. 1-14 (unpublished).

GRENIER Y. (1977) :

Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique.

Thèse Docteur Ingénieur ENST. Paris.

GUBRYNOWICZ R. (1969) :

Zastosowanie pasmowej analizy widmowej sygnaly mowy do okseslania cech osobniczych
(Application d'une analyse passe-bande du signal de la parole à la description des caractéristiques individuelles).

Archivum Akustyki 2,4,222-235.

GUBRYNOWICZ R. (1971) :

Méthode d'analyse statistique du spectre de la parole ; application à la reconnaissance automatique du locuteur.

7th Int. Congr. Acoustics 3, 25-28.

GUBRINOWICZ R. (1973) :

Application of a Statistical Spectrum Analysis to Automatic Voice Identification.

Speech Analysis & Synthesis 3, 171-180.

HAIR G.D. & REKIETA T.W. (1972) :

Automatic Speaker Verification using Phoneme Spectra.

J.A.S.A. 51 , 131(A).

HALL M.E. (1975) :

Spectrographic Analysis of Interspeaker and Intraspeaker Variabilities of Professional Mimicry.

Master's Th. Midingan State Univ.

HARGREAVES W.A. & STARKWHEATER J.A. (1963) :

Recognition of Speaker Identity.

Lang. & Speech. 6 , 63-67.

HARTMAN D.E. & DANHAUER J.R. (1976) :

Perceptual Features of Speech for Males in four Perceived Age Decades.

J.A.S.A. 59 . 712-715.

HAYRE H.S. (1976)

Speech - A possible Indicator of Physical Stress.

IEEE Congr. ASSP, 740.

HAZEN B. (1973) :

Effects of Differing Phonetic Contexts on Spectrographic Speaker Identification.

J.A.S.A. 54 , 650 - 660

HECKER M.H.L. (1969) :

Methods for Measuring Speaker Recognition.

Stanford Research Institute, Menlo Park, Calif.

HECKER M.H.L. (1970) :

Speaker Recognition : Basic Considerations and Methodology.

J.A.S.A. 49 . 138(A).

HECKER (M.), 1971, *Speaker Recognition, an Interpretive Survey of the Literature.* - American Speech & Hearing Association, Monographs 16.

HECKER S. & ITAKURA F. (1972)

Talker Recognition Using Dynamic characteristics of the Speech Spectral Pattern

Conv. Record of Acoust. Soc. of Japan, 2, 2, 20.

HERZOG H. (1933) :

Stimme und Persönlichkeit.

2. Psychol. 130 , 300-379.

HOLLIEN H. (1974) :

The Peculiar Case of "Voiceprints".

J.A.S.A. 56 , 1 , 210-213.

HOLLIEN H., MAJEWSKI W., & HOLLIEN P. (1974)

Speaker Identification by Long-Term a Spectra Under Normal, Stress, and Disguise Conditions.

J.A.S.A. 55, S20 (A)

HOLLIEN H. , CHILDERS D.G. & DOHERTY E.T. (1977) :

Semi-Automatic System for Speaker Identification (SAUSI),

IEEE Congr. ASSP , 768-771.

HOLLIEN H. & MAJEWSKI W. (1977) :

Speaker Identification by Long-Term Spectra under Normal and Distorded Speech Conditions.

J.A.S.A. 62 , 975(A).

HOLMGREN G.L. (1963)

Speaker Recognition in Speech Communication Systems.

Rept. 13-73801-13

Rpt. AFCRL-63-119. Air Force Cambridge Research Laboratories

Office of Aerospace Research, Bedford Mass.

HOLMGREN G.L. (1967)

Physical and Psychological Correlates of Speaker Recognition

J.S.H.L. 10, 57-66.

HORII Y. (1975)

Some Statistical Characteristics of Voice Fundamental Frequency

J.S.H.R. 18, 192-201.

HORII Y. & RYAN W.J. (1975)

Fundamental Frequency Characteristics and Perceived Age of Adult Mall Speakers.

J.A.S.A. 57, S 69 (A)

HUNT M.J., YATES J.W. & BRIDLE J.S. (1977)

Automatic Speaker Recognition for Use over Communication Channels.

IEEE Congress ASSP, 764-767.

IBBA G., PAOLONI A. (1977)

Les consonnes occlusives, éléments d'identification des locuteurs.

9th Int. Congr. Acoustics I.30.

IBBA G., PAOLONI A. & SAVERIONE B. (1979)

Significativita' della durata delle consonanti occlusive ai fini del riconoscimento del parlatore

Rivista Italiana di Acustica, III, 1, 23-39.

INGEMAN F. (1968)

Identification of the Speaker's Sex from Voiceless Fricatives

J.A.S.A. 44, 1142 - 1144.

ITAKURA F. (1975)

Minimum Prediction Residual Principle Applied to Speech Recognition

IEEE Trans. ASSP, 23, Feb., 67-72.

JASSEM W., STEFFEN-BATOG M., CZAJKA S. (1973)

Statistical Characteristics of Short - Term Average F_o Distributions as Personnal Voice Features.

JESORSKY P. (1977)

AUROS - Automatic Recognition of Speakers by Computers Feature Extraction and Reduction fort Text-Dependent Speaker Recognition.

9th Int. Congr. Acoustics. I 102.

JESORSKY P. (1976)

Merkmelsgewinnung als Teilaufgabe der Automatischen Sprecherkennung

NTZ, Dec.

JESORSKY P. (1976)

Speaker Recognition

New Communications with Computers, Mc Millan, L. BOLT Ed.

JESORSKY, P. (1978) :

Principles of Automatic Speaker Recognition.

in "Speech Communication with Computer"

Ed. L. BOLC, Car Hanser, Verlag, München.

KAISER, L. (1939) :

Biological and Statistical Research Concerning the Speech of 216 Dutch Students. I.

Archives Néerlandaises de Phonétique Expérimentale 15, 1-76.

KAISER, L. (1940) :

Biological and Statistical Research Concerning the Speech of 216 Dutch Students. II.

16, 77-136.

KAISER, L. (1941) :

Biological and Statistical Research Concerning the Speech of 216 Dutch Students. III.

17, 143-211.

KAISER, L. (1942) :

Biological and Statistical Research Concerning the Speech of 216 Dutch Students. IV.

18, 1-58.

KAISER, L. (1944) :

Biological and Statistical Research Concerning the Speech of 216 Dutch Students. V.

19. 37-78

KASHYAP R.L. (1976) :

Speaker Recognition from an Unknown Utterance and Speaker-Speech Interaction.

IEEE Trans. Acoust. Speech and Signal Process. ASSP-24. 481-488.

KASHYAP R.L. (1976)

The separation of phonemic and speaker components of speech.

School of Electrical Engineering, Purdue University

Report

KERSTA, L.G. (1948) :

Amplitude Cross-Section Representation with the Sound spectrograph.

J.A.S.A. 20, 796-801.

KERSTA, L.G. (1962) a :

Voiceprint Identification.

Nature 196, 4861, 1253-1257.

KERSTA L.G. (1962b)

Voiceprint Identification

J.A.S.A., 34, 725(A).

KERSTA L.G. (1962c)

Voiceprint Identification Infallibility

J.A.S.A., 34, 1078(A).

KERSTA L.G. (1962) d

Voice Spectrograms for Unique Personal Identifications.

Bell Labs Record 40 , 214-215.

KERSTA L.G. (1965) a

Voiceprint Classification.

J.A.S.A. 37. 1217 (A)

KERSTA L.G. (1965) b

Environmental Influence on the Speech of Family Members Shown by Spectrographie Speech Matching.

J.A.S.A. 38, 935 (A)

KERSTA L.G. (1966)

Voiceprint Classification for an Extended Population.

J.A.S.A. 39, 1239-1240 (A)

KERSTA L.G. (1968)

Speaker Identification by Spectrographie Voiceprints.

Of Voice Print Laboratories.

6th Int. Congr. Acoustics. B-4-10.

KERSTA L.G. & COLANGELO J.A. (1969)

The Spectrographie Speech Patterns of Identical Twins.

Voiceprint Laboratories (New Jersey).

KERSTA L.G. (1969)

Automated Acoustic-Signature Verification System.

Voiceprint Laboratories (New-Jersey)

KERSTA L. & COLANGELO J. (1970)

The Spectrographie Speech Patterns of Identical Twins.

J.A.S.A. 47, 58-59 (A)

KERSTA L.G. (1971)

Progress Report on Automated Speaker-Recognition Systems.

J.A.S.A. 49, 139 (A)

KRAUSE H.J.

Possibilités d'identification par la voix et ses limites

Archiv für Kriminologie. 157. 5 & 6. 1976. Pages 154-164.

KRAUSE H.J. & ENDRES W. (1976)

Möglichkeiten Und Grenzen der Specheridentifizierung Durch Anwendung Spektrographisher Verfahren (Possibilitée de l'Identification par la Voixet leurs Limites au moyen de Procédés Spectrographiques)

5ème Coll. sur l'Identification d'Echantillons Acoustiques, 27 Fév., Inst. de Trait. des Données Tech. et Rfnl., Karlsruhe.

LADEFOGED P. & VANDERSLICE R. (1967)

The Voiceprint Mystique.

W.P.P. 7. 126-142.

LASS N.J., ALMERINO C.A., JORDAN L.F. & WALSH J.M. (1979)

The Effect of Filtered speech on Speaker Race and Sex Identification

Journal of Phonetics. (in press)

LEVIN H. & LORD W. (1975)

Speech Pitch Frequency as Emotional State Indicator

I.E.E.E. Trans. Systems, Man and Cyb., 5, 2, 259

LI K.P., DAMMANN J.E. & CHAPMAN W.D. (1966)

Experimental Studies in Speaker Verification using an Adaptative System.

J.A.S.A. 40, 966-978.

LI K.P., HUGUES G.W. & HOUSE A.S. (1970)

Approaches to the Characterization of Talker Differences by Statistical Operation on Speech Spectra.

J.A.S.A. 47, 66 (A)

LI K.P. & HUGUES G.W. (1974)

Talker Differences as they Appear in Correlation Matrices of Continuous Speech Spectra.

J.A.S.A. 55, 833-837.

LIBERMAN A.M., COOPER F.S., SHANKWEILER D.P. & STUDDERT-KENNEDY M. (1968)

Why are Speech Spectrograms Hard to Read ?

Am. Ann. Deaf 113, 127.

LIN W.C. & PILLAY S.K. (1976)

Feature Evaluation and Selection for an On-Line Adaptative Speaker Verification System

IEEE Congr. ASSP, 734-737.

LUCK J.E. (1969)

Automatic Speaker Verification using Cepstral Measurements.

J.A.S.A. 46, 1026-1032.

LUMMIS R.C. (1971)

Real-Time Technique for Speaker Verification by Computer.

J.A.S.A. 50, 106 (A)

LUMMIS R.C. (1972)

Implementation of On-Line Speaker Verification Schema

J.A.S.A. 52, 181 (A)

LUMMIS R.C. & ROSENBERG A.E. (1972)

Test on Automatic Speaker Verification Method with Intensively Trained Professional Mimics.

J.A.S.A. 51, 131-132 (A)

LUMMIS R.C. (1973)

Speaker Verification by Computer using Speech Intensity for Temporal Registration.

IEEE Trans. Audio Electr. AU-21, 80-89.

Mac DADE T. (1968)

The Voiceprint

The criminologist 7, 52-60.

Mac GEE V.E. (1965)

Invariance of Personal Characteristics of Voice over two Vowel Sounds.

Percept. Mot. Skills, 21, 519-529.

Mac GEHEE F. (1937)

The Reliability of the Identification of the Human Voice

J. Gen. Psychol. 17. 249-271.

Mac GEHEE F. (1944)

An Experimental Study in Voice Recognition.

J. Gen. Psychol. 31. 53-65.

Mac GLONE R.E., HOLLIEN P., HOLLIEN H., JACKSON J.S. (1977)

Acoustic Analysis of Voice Disguise related to Voice Identification.

Int. Conf. on Crime Countermeasures- Science and Engineering 31-35.

MAJEWSKI W., HOLLIEN H. (1974)

Euclidean Distance between Long-Term Speech Spectra as a Criterion for Speech Identification.

S.C.S., 3, 303-310.

MAJEWSKI W., HOLLIEN H. (1974)

Speaker Identification by Means of Long-Term Speech Spectra.

8th Int. Congr. Acoust.

MARKEL J.D., OSHIKA B.T., GRAY A.H. JR (1977)

Long-Term Feature Averaging for Speaker Recognition

IEEE Trans. Acoust. Speech & Signal Process.

ASSP-25, 330-337.

MARKEL J.D. & DAVIS S.B. (1979)

Tex-Independent Speaker Recognition From a Large Linguistically Unconstrained Time-Spaced Data Base

IEEE Trans. ASSP, 27, 1, 74-82.

MATSUMOTO H., HIKI S., SONE T. & NIMURA T. (1973)
Multidimensional Representation of Personal Quality and its Acoustical Correlates
IEEE Trans. Audio Electr. AU-21, 428-436.

MIKHEEV Y.V. (1971)
Statistical Distribution of the Periods of the Fundamental Tone In Russian Speech.
Soviet Physics Acoustics 16, 474-477.
Transl. from Akusticheskii Zhurnal 1970, 16, 558-562.

MOHN W.S. (1969)
Statistical Feature Evaluation in Speaker Identification
Ph.D. Dissert. North Carolina State Univ.
Raleigh, NC

MOHN W.S. (1971)
Two statistical Feature Evaluation Techniques Applied to Speaker Identification.
IEEE Trans. on Computers C-20, 979-987.

MYSAK E.D. (1959)
Pitch and Duration Characteristics of Older Males
J.S.H.R. 2, 46-54.

PAUL J.E., RABINOWITZ A.S., RIGANATI J.P. & RICHARDSON J.M. (1974)
Semi-automatic Speaker Identification System(SASIS)-Analytical Studies Final Report, Rockwell International Report No. C74-11841501, Prepared for the Aerospace Corporation.

PETERS R.W. (1954)
Studies in Extra Messages : Listener Identification of Speakers Voices under Conditions of certain Restrictions Imposed upon the Voice Signal.
U.S. Naval School of Aviation Medecine, Joint Project

PETERS R.W. (1956)
Studies in Extra Messages : the Effect of Various Modifications of the Voice Signal upon the Ability of Listeners to Identify Speakers' Voices.
U.S. Naval School of Aviation Medecine, Joint Project
NM 001-104-500. Rpt. 61, Pensacola Fla.

PFEIFER L.L. (1975)

Speaker Identification from a Minimal Set of Training Data
J.A.S.A. 57, S1, S35

PFEIFER L.L. (1977)

Final report on Feature Analysis for Speaker Identification
Prepared for Rome Air Development Center, Griffis Air Base, New York. Contrat n°. F30602-76-C-0157.

PFEIFER L.L. (1978)

New Techniques for Text-Independent Speaker Identification
IEEEICASSP. 283-286.

POLLACK I., PICKETT J.M. & SUMBY W.H. (1954)

On the Identification of Speakers by Voice
J.A.S.A. 26, 403-406.

PRUZANSKY S. (1963)

Pattern-Matching Procedure for Automatic Talker Recognition.
J.A.S.A. 35, 354-358.

PRUZANSKY S. & MATHEWS M.V. (1963)

Talker Recognition Procedure Based on Analysis of Variance
J.A.S.A. 35, 1877 (A)

PRUZANSKY S. & MATHEWS M.V. (1964)

Talker-recognition Procedure Based on Analysis of Variance
J.A.S.A. 36, 2041-2047.

PTACEK P.H. & SANDER E.K. (1966)

Age Recognition from Voice.
J.S.H.R. 9, 273-277

PTACEK P.H., SANDER E.K., MALONEY W.H., & JACKSON C.C.R. (1966)

Phonatory and Related Changes with Advanced Ages.
J.S.H.R. 9, 353-360

RAMISHVILI G.S. (1966)

Automatic Voice Recognition
Engng. Cybernetics 5, 84-90.

RAMISHVILI G.S. (1974)

Experiments on Automatic Verification of Speakers.
2nd Int. Joint Conference on Pattern Recognition.
Copenhague, 389-393.

RAMISHVILI G.S. & TUSHIVILI M.A. (1976)

On the Connection of some Time Characteristics of Speech Signal with the Individuality of Voice.
IEEE Congr. ASSP. 730-733.

REDDY V.C.V.P. (1975)

A Subjective Method of Speaker Identification.
2. Elektr. Inform. U. Energietech. 5, 321-325.

REKIETA T.W. & HAIR G.D. (1972)
Mimic Resistance of Speaker Verication Using Phonemes Spectra
J.A.S.A., S1, 131(A).

ROSENBERG A.E. (1972)

Listener Performances in Speaker Verification Tasks
Proc. IEEE Conf. on Speech Comm. and Processing, Air Force Cambrid
Res. Lab., Bedgord, MA, 283

ROSENBERG A.E. (1973)

Listener Performance in Speaker Verification Tasks.
IEEE Trans. Audio Elect. AU-21, 221-225.

ROSENBERG A.E. & SAMBUR M.R. (1975)
New Techniques for Automatic Speaker Verification
IEEE Trans. ASSP-23,2,169-176.

ROSENBERG A.E. (1976)

Automatic Speaker Verification : A Review.
Proc. IEEE 64, 475-487.

ROSENBERG A.E. (1976)
Evaluation of an Automatic Speaker Verification System Over Telephone Lines
B.S.T.J., 55, 6, 723-743.

RYAN W.J. & BURK K.W. (1972)

Predictors of Age in the Male Voice
J.A.S.A. 53, 345 (A)

SAINT-BONNET M. & BOË L.J. (1977)

Les pauses et les groupes rythmiques : leur durée et distribution en fonction de la vitesse d'élocution.

7° J.E. sur la Parole, Groupe Communication Parlée du G.A.L.F. 337-343.

SAITO S. (1973)

Talker Recognition

Mathematical Sciences, 116, 49.

SAITO S. & FURUI S. (1978)

Personal Information in Dynamic Characteristics of Speech Spectra

4th. IJCPR, Kyoto, 7-11 Nov.

SAMBUR M.R. (1972)

Speaker Recognition and Verification using Linear Prediction Analysis.

Ph.D. Dissert. M.I.T. Cambridge.

SAMBUR M.R. (1973)

Speaker Recognition and Verification using Linear Prediction Analysis.

Q.P.R. M.I.T. 108, 261-268.

SAMBUR M.R. (1975)

Selection of Acoustic Features for Speaker Identification.

IEEE Trans. on Acoust. Speech and Signal Processing, ASSP-23, 176-182.

SAMBUR M.R. (1976)

Speaker Recognition using Orthogonal Linear Prediction.

IEEE Trans. Acoustic Speech & Signal Process. ASSP-24, pp. 283-289, 1632-1645.

SAMBUR M.R. (1976)

Text Independant Speaker Recognition using Orthogonal Linear Prediction.

IEEE Conf. on Acoustics, Speech & Signal Process. 727-729.

SAPIR E. (1927)

Speech as a Personality Trait.

Amer. J. Social 32, 892-905.

SARMA V.V.S., YEGNANARAYANA B., (1975)
A Critical Survey of Automatic Speaker Recognition Systems.
J. Comput. Soc. India 6, 1, 9-19

SARMA V.V.S. & YEGNANARAYANA B. (1976)
Cascade Realization of Digital Inverse Filter for Extracting Speaker Dependent Features.
IEEE Int. Conf. Acoustics, Speech & Signal Process.723-726.

SHEARME J.N. & HOLMES J.N. (1959)
An Experiment Concerning the Recognition of Voices
Language and Speech. 2, 123-131.

SCHWARTZ M.F. (1968)
Identification of Speaker Sex from Isolated Voiceless Fricatives.
J.A.S.A. 43, 1178-1179.

SCHWARTZ M.F. & RINE H.E. (1968)
Identification of Speaker Sex from Isolated, Whispered Vowels.
J.A.S.A. 44, 1736-1737.

SHIPP T. & HOLLIEN H. (1969)
Perception of the Aging Male Voice.
J.S.H.R. 12. 703-710.

STARKWEATHER J.A. (1956)
Content-free Speech as a Source of Information about the Speaker.
J. Abnorm. Soc. Psychol. 52, 394-402.

STEFFEN-BATÓG M., JASSEM W., GRUSZKA-KOŚCIELAK H. (1970)
Statistical Distribution of Short-Term F_o Values as a Personal Voice Characteristic.
Speech Analysis & Synthesis 2, 195-206.

STEVENS (K.N.) & HOUSE (A.S.), 1961, *An Acoustical Theory of Vowel Production and some of its Implications.* - J.S.H.R. 4, pp. 303-320.

STEVENS K.N. & A.L. HOUSE (1963)
Pertubation of Vowel Articulations by Consonantal Context : an Acoustical Study
Journal of Speech and Hearing Research, 6, III-128.

STEVENS K.N., WILLIAMS C.E., CARBONELL J.R. & WOODS B. (1968)
Speaker Authentification and Identification : A comparison of Spectrographic and Auditory Presentations of Speech Materials.
J.A.S.A. 44, 1596-1607.

STEVENS K.N. (1971)

Sources of Inter- and Intra- Speaker Variability in Acoustic Properties of Speech Sounds.

7th Int. Congr. Phonetic Sciences. 206-227.

STUNTZ S.E. (1963)

Speech Intelligibility and Talker Recognition Tests of Air Force Communication Systems.

Rpt. ESD-TDR-63-224 - Electronics Systems Division, Air Force Systems Command, Hanscom Field.

SU L.S., LI K.P., & FU K.S. (1974)

Identification of Speakers by Use of Nasal Coarticulation.

J.A.S.A. 56, 1876-1882.

TAYLOR H.C. (1938)

Social Agreements on Personality Traits as Judged from Speech

J. Soc. Psychol. 5, 244-248.

TILLMAN H.G. (1967)

Automatische Identification von Sprechern.

NTZ 12, 706-713, Moscou

TILLAN H.G., MENNON H.M., UNGEHEUER G. (1968)

Kontursonogramme und "voiceprints"

Forschungsbericht 68-6, Inst. Phonetik U. Kommunicationforschung, Bonn.

TOSI O. (1967)

Evaluation of the Voiceprint Method

Report to the Michigan Depart. of the State Police.

TOSI O., OYER H., PEDREY C., LASHBROOK B., & NICOL J. (1970)

An Experiment on Voice Identification by Visual Inspection of Spectrograms.

J.A.S.A. 49, 138 (A)

TOSI O., OYER H.J., LASHBROOK W.B., PEDREY C. & NICOL J. (1972)

Voice Identification through Acoustic Spectrography.

Department of Audiology and Speech & Hearing Sci. Lab. Michigan State Univ. (East Lansing Michigan). Rept. n° 171.

TOSI O., OYER H. & NASH E. (1972)
Latest Developments in Voice Identification.
J.A.S.A. 51, 132 (A).

VENUGOPAL D. & SARMA V.V.S. (1977)
Performance Evaluation of Automatic Speaker Recognition Schema.
IEEE Conf. ASSP, 780-783.

VIDALON M., SHRIDHAR M., CAÑAS M. (1977)
Speaker Verification using Composite Reference.
IEEE Congr. ASSP, 758-760.

VOIERS W.D. (1964)
Perceptual Bases of Speaker Identity.
J.A.S.A. 36, 1065-1075.

VOIERS W.D. (1965)
Performance Evaluation of Speech Processing Devices II. The Role of Individual Differences.
Report AFCRL-66-24, Air Force Cambridge Research Laboratories, Office of Acrospace Research, Bedford, Mass.

VOIERS W.D., COHEN M.F. & MICKUNAS J. (1965)
Evaluation of Speech Processing Devices I. Intelligibility, Quality, Speaker Recognizability.
Report AFCRL-65-826. Air Force Cambridge Research Laboratories, Office of Acrospace Research, Bedford, Mass.

WACHTER K.W. (1970)
Talker Recognition on Large Populations.
J.A.S.A. 47, 1. 66 (A)

WAKITA H. (1975)
On the use of Linear Prediction Error Energy for Speech and Speaker Recognition.
J.A.S.A. 57, S1,

WAKITA H. (1976)
Residual Energy of Linear Prediction Applied to Vowel and Speaker Recognition.
IEEE Conf. on Acoustics Signal & Speech Process. 274-275.

WEINBERG B. & BENETT S. (1971)
A Study of Talker Sex Recognition of Esophageal Voices.
J.S.H.R. 14. 391-395.

WILLIAMS C.E. (1964)
The Effects of Selected Factors on the Aural Identification of Speakers.
sec. III in "Methods for Psychoacoustic Evaluation of Speech Communication Systems". Dept. ESD-TDR- 65-153. Electronic Systems Division, Air Force Systems Command, Hanscom Field, Mass.

WILLIAMS C.E. & STEVENS K.N. (1969)
On Determining the Emotional State of Pilots during Flight.
Aerospace Medicine 40 1369-1372

WILLIAMS C.E. & STEVENS K.N. (1972)
Emotions and Speech : some Acoustical Correlates.
J.A.S.A. 52. 1238-1250.

WOLF J.J. (1969)
Acoustic Measurement for Speaker Recognition
Q.P.R. M.I.T. 94, 216-222.

WOLF J.J. (1969)
Acoustic Measurements for Speaker Recognition.
PhD. Th. Depart. Elect. Eng. M.I.T.

WOLF J.J. (1970)
Simulation of the Measurement Phase of an Automatic Speaker Recognition System.
J.A.S.A. 47, 83 (A)

WOLF J.J. (1971)
Efficient Acoustic Parameters for Speaker Recognition.
J.A.S.A. 51, 2044-2055.

WOLF J.J. (1971)
Choice of Speaker Recognition Parameters.
Q.P.R. M.I.T. 97 125-133.

YEGNANARAYANA B., SARMA V.V.S. & VENUGOPAL D. (1977)
Studies on Speaker Recognition using a Fourier Analysis System.
IEEE Conf. ASSP. 776-779.

YOUNG M.A. & CAMPBELL R.A. (1967)
Effects of Context on Talker Identification
J.A.S.A. 42, 1250-1254.

ZALEWSKI J., MAJEWSKI W., HOLLIEN H. (1975)
Cross Correlation of Long-Term Speech Spectra as a Speaker Identifica Technique.
Acustica 34, 20-24.

NOTA BENE

Ne figurent pas les références citées suivantes :

ROSENBERG, 1977 ; HATON, 1977 ; PETER SON & LEHISTE, 1968 ; DEMICHELIS & DE MORI, 1978 ; CHEN, 1970 ; J.J. BAKER, 1975 ; STEVENS & HOUSE, 1961 ; STEVENS, 1972 ; SOTTER & STEINBERG, 1950 ; FUJIMURA, 1962 ; DICKSON,1962 ; KRAUSE, 197: LUCK, 1961 ; WILLIAMSON, 1961 ; SCHROEDER, 1968 ; HEIGHLEIRAN, 1962 ; GLICH, 1978 ; MARILL & GREEN, 1963 ; LEWIS, 1962 ; KAMENTSKY & LIU, 1963 ; HODGE & a 1978 ; BOË & RAKOTOFIRINGA, 1975 ; GRENIER, 1978 ; ABRY, BOË & ZURCHER, 1974.

dernière page de la thèse

AUTORISATION DE SOUTENANCE

VU les dispositions de l'article 3 de l'arrêté du 16 Avril 1974,

VU les rapports de présentation de Messieurs :

- C. BELLISSANT, Maître-Assistant à l'Université Scientifique et Médicale de GRENOBLE
- G. BENBASSAT, Ingénieur C.I.T. ALCATEL

Monsieur Patrick C O R S I

est autorisé à présenter une thèse en soutenance pour l'obtention du diplôme de DOCTEUR-INGENIEUR, spécialité "Génie Informatique".

Grenoble, le 22 Octobre 1979
Le Président de l'I.N.P.G.

Ph. TRAYNARD
Président
de l'Institut National Polytechnique