Constitution d'une base de références phonétiques pour la reconnaissance de mots isolés pour un système multi-locuteurs

Christine Delia

THESE

présentée à

l' Université Scientifique et Médicale de Grenoble

pour obtenir le grade de
DOCTEUR-INGENIEUR
«Mathématiques appliquées»

par

Christine DELIA

CONSTITUTION D'UNE BASE DE REFERENCES PHONETIQUES POUR LA RECONNAISSANCE DE MOTS ISOLES POUR UN SYSTEME MULTI-LOCUTEURS.

Thèse soutenue le 28 septembre 1984 devant la commission d'examen.

B. VAN CUTSEM	Président
C. BELLISSANT	Examinateurs
G. BENBASSAT	
B. GROC	

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

Année universitaire 1982-1983

Président de l'Université : M. TANCHE

MEMBRES DU CORPS ENSEIGNANT DE L'U.S.M.G.

(RANG A)

SAUF ENSEIGNANTS EN MEDECINE ET PHARMACIE

PROFESSEURS DE 1ère CLASSE

ARNAUD Paul	Chimie organique
ARVIEU Robert	Physique nucléaire I.S.N.
AUBERT Guy	Physique C.N.R.S.
AYANT Yves	Physique approfondie
BARBIER Marie-Jeanne	Electrochimie
BARBIER Jean-Claude	Physique expérimentale C.N.R.S. (labo de magnétisme)
BARJON Robert	Physique nucléaire I.S.N.
BARNOUD Fernand	Biosynthèse de la cellulose-Biologie
BARRA Jean-René	Statistiques - Mathématiques appliquées
BELORISKY Elie	Physique
BENZAKEN Claude (M.)	Mathématiques pures
BERNARD Alain	Mathématiques pures
BERTRANDIAS Françoise	Mathématiques pures
BERTRANDIAS Jean-Paul	Mathématiques pures
BILLET Jean	Géographie
BONNIER Jean-Marie	Chimie générale
BOUCHEZ Robert	Physique nucléaire I.S.N.
BRAVARD Yves	Géographie
CARLIER Georges	Biologie végétale
CAUQUIS Georges	Chimie organique
CHIBON Pierre	Biologie animale
COLIN DE VERDIERE Yves	Mathématiques pures
CRABBE Pierre (détaché)	C.E.R.M.O.
CYROT Michel	Physique du solide
DAUMAS Max	Géographie
DEBELMAS Jacques	Géologie générale
DEGRANGE Charles	Zoologie
DELOBEL Claude (M.)	M.I.A.G. Mathématiques appliquées
DEPORTES Charles	Chimie minérale
DESRE Pierre	Electrochimie
DOLIQUE Jean-Michel	Physique des plasmas
DUCROS Pierre	Cristallographie
FONTAINE Jean-Marc	Mathématiques pures
GAGNAIRE Didier	Chimie physique

.../...

GASTINEL Noël	Analyse numérique - Mathématiques appliquées
GERBER Robert	Mathématiques pures
GERMAIN Jean-Pierre	Mécanique
GIRAUD Pierre	Géologie
IDELMAN Simon	Physiologie animale
JANIN Bernard	Géographie
JOLY Jean-René	Mathématiques pures
JULLIEN Pierre	Mathématiques appliquées
KAHANE André (détaché DAFCO)	Physique
KAHANE Josette	Physique
KOSZUL Jean-Louis	Mathématiques pures
KRAKOWIAK Sacha	Mathématiques appliquées
KUPTA Yvon	Mathématiques pures
LACAZE Albert	Thermodynamique
LAJZEROWICZ Jeannine	Physique
LAJZEROWICZ Joseph	Physique
LAURENT Pierre	Mathématiques appliquées
DE LEIRIS Joël	Biologie
LLIBOUTRY Louis	Géophysique
LOISEAUX Jean-Marie	Sciences nucléaires I.S.N.
LOUP Jean	Géographie
MACHE Régis	Physiologie végétale
MAYNARD Roger	Physique du solide
MICHEL Robert	Minéralogie et pétrographie (géologie)
MOZIERES Philippe	Spectrométrie - Physique
OMONT Alain	Astrophysique
OZENDA Paul	Botanique (biologie végétale)
PAYAN Jean-Jacques (détaché)	Mathématiques pures
PEBAY PEYROULA Jean-Claude	Physique
PERRIAUX Jacques	Géologie
PERRIER Guy	Géophysique
PIERRARD Jean-Marie	Mécanique
RASSAT André	Chimie systématique
RENARD Michel	Thermodynamique
RICHARD Lucien	Biologie végétale
RINAUDO Marguerite	Chimie CERMAV
SENGEL Philippe	Biologie animale
SERGERAERT Francis	Mathématiques pures
SOUTIF Michel	Physique
VAILLANT François	Zoologie
VALENTIN Jacques	Physique nucléaire I.S.N.
VAN CUTSEN Bernard	Mathématiques appliquées
VAUQUOIS Bernard	Mathématiques appliquées
VIALON Pierre	Géologie

PROFESSEURS DE 2ème CLASSE

ADIBA Michel	Mathématiques pures
ARMAND Gilbert	Géographie

.../...

AURIAULT Jean-Louis	Mécanique
BEGUIN Claude (M.)	Chimie organique
BOEHLER Jean-Paul	Mécanique
BOITET Christian	Mathématiques appliquées
BORNAREL Jean	Physique
BRUN Gilbert	Biologie
CASTAING Bernard	Physique
CHARDON Michel	Géographie
COHENADDAD Jean-Pierre	Physique
DENEUVILLE Alain	Physique
DEPASSEL Roger	Mécanique des fluides
DOUCE Roland	Physiologie végétale
DUFRESNOY Alain	Mathématiques pures
GASPARD François	Physique
GAUTRON René	Chimie
GIDON Maurice	Géologie
GIGNOUX Claude (M.)	Sciences nucléaires I.S.N.
GUITTON Jacques	Chimie
HACQUES Gérard	Mathématiques appliquées
HERBIN Jacky	Géographie
HICTER Pierre	Chimie
JOSELEAU Jean-Paul	Biochimie
KERCKOVE Claude (M.)	Géologie
LE BRETON Alain	Mathématiques appliquées
LONGEQUEUE Nicole	Sciences nucléaires I.S.N.
LUCAS Robert	Physiques
LUNA Domingo	Mathématiques pures
MASCLE Georges	Géologie
NEMOZ Alain	Thermodynamique (CNRS - CRTBT)
OUDET Bruno	Mathématiques appliquées
PELMONT Jean	Biochimie
PERRIN Claude (M.)	Sciences nucléaires I.S.N.
PFISTER Jean-Claude (détaché)	Physique du solide
PIBOULE Michel	Géologie
PIERRE Jean-Louis	Chimie organique
RAYNAUD Hervé	Mathématiques appliquées
ROBERT Gilles	Mathématiques pures
ROBERT Jean-Bernard	Chimie physique
ROSSI André	Physiologie végétale
SAKAROVITCH Michel	Mathématiques appliquées
SARROT REYNAUD Jean	Géologie
SAXOD Raymond	Biologie animale
SOUTIF Jeanne	Physique
SCHOOL Pierre-Claude	Mathématiques appliquées
STUTZ Pierre	Mécanique
SUBRA Robert	Chimie
VIDAL Michel	Chimie organique
VIVIAN Robert	Géographie

S O M M A I R E

Mon travail s'est effectué dans les laboratoires de traitement de la parole à TEXAS INSTRUMENT à Villeneuve Loubet.

Aussi j'exprime toute ma reconnaissance à M. Gérard BENBASSAT pour m'avoir accueillie dans l'équipe de Recherche et Développement de TEXAS INSTRUMENT.

Je remercie tout particulièrement M. Bernard VAN CUTSEM pour avoir consenti à présider le jury de cette thèse et pour les innombrables conseils qu'il m'a donnés.

M. Camille BELLISSANT m'a efficacement guidé dans mes recherches ; qu'il trouve ici mes plus sincères remerciements.

J'ai aussi été très sensible à l'aide apportée par les membres de l'équipe de traitement de la parole à TEXAS INSTRUMENT ; plus précisément, j'ai beaucoup profité des conseils que m'a prodigués M. Xavier DELON spécialiste en Phonétique.

Un grand merci pour M. Bernard GROC pour avoir accepté de participer au jury de cette thèse.

Enfin, j'exprime toute ma considération à Mme Paulette SOUILLARD et Mme Jocelyne PATRUNO pour avoir assuré la frappe de ce mémoire.

C H A P I T R E I

Introduction

1.1. AVANT PROPOS

L'ordinateur a largement contribué au développement de la science. Mais sa fonction de communication avec l'homme doit être améliorée; qu'il s'agisse de communication auditive visuelle ou orale. C'est pourquoi, le problème d'interface homme-machine a suscité dernièrement un grand nombre de recherches.

La reconnaissance automatique de la parole est un des points les plus délicats à traiter. Les difficultés rencontrées sont analysées au cours de ce mémoire, mais on peut dès à présent en comprendre les causes. Tout d'abord on doit admettre que la compréhension du langage n'est pas simplement liée à l'ouïe mais que certaines informations correspondant le plus souvent au contexte sont prises en compte par d'autres sens. Ainsi, la multiplicité des locuteurs et des accents, et la dégradation du message dans des conditions bruitées ajoutent des paramètres supplémentaires qu'il faut parvenir à écarter. Jusque là, ces conditions ont été vérifiées mais les moyens technologiques existant ne permettaient pas d'effectuer des études sérieuses sur ce sujet. La révolution informatique apparue de nos jours a donc fait accélérer les démarches.

Deux types de méthode sont exploitées : la reconnaissance analytique et la reconnaissance globale. La première cherche à isoler les éléments phonétiques caractérisant un message inconnu. Suivant les cas, les unités à identifier peuvent être des phonèmes des diphones ou des syllabes. Pour la reconnaissance globale, le vocabulaire de taille finie contient des mots ou des phrases à reconnaître. Ces derniers sont mémorisés au cours d'une phase d'apprentissage. Le mot inconnu est identifié au plus proche des éléments en mémoire.

Pour qu'un système de reconnaissance fonctionne, il faut donc lui apprendre préalablement les unités qu'il doit reconnaître par la suite. Là réside le principe de tout procédé: déterminer

la ressemblance entre un signal inconnu et un certain nombre d'unités acoustiques connues par la machine. On les appellera les "modèles". Leur nature est différente suivant le type de méthode choisie : ce sont des mots entiers (reconnaissance globale) ou des éléments atomiques comme les phonèmes, les diphones, les syllabes ... (méthode analytique).

L'analyse du message inconnu comprend différents niveaux. Le signal peut ainsi être considéré comme :

- une suite de phonèmes
- une suite de mots : étape lexicale
- une suite de phrases: étape syntaxique.

Le choix d'une analyse particulière détermine une stratégie. On en distingue trois : celle qui procède par reconnaissance de phonèmes, sans reconnaissance de phonèmes ou par l'utilisation des deux méthodes. Nous proposons ici le schéma général du processus de reconnaissance. (Klatt,1977).

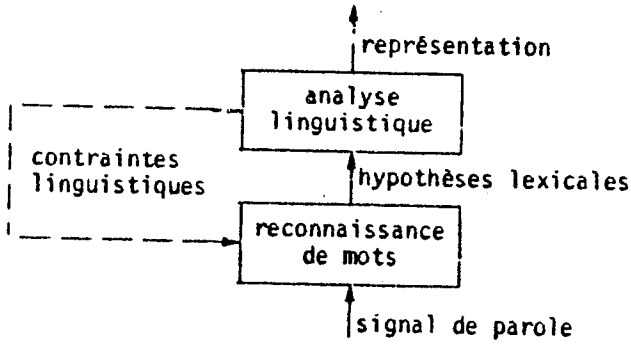


figure 1.1.1. Processus général de reconnaissance

Le signal de parole, après passage dans le module de reconnaissance est converti en hypothèses lexicales c'est-à-dire en une séquence de mots reconnus. Une analyse linguistique mettant en oeuvre des contraintes d'ordre syntaxique détermine alors la cohérence du dialogue proposé. Il peut y avoir retour à la solution proposée.

Nous nous intéressons, dans le cadre de ce projet, à la reconnaissance par phonème. Parmi les systèmes de reconnaissance de parole continue qui font appel à ce type de méthode, on peut citer KEAL et MYRTILLE 1 et 2. Le signal temporel est considéré comme une séquence de vecteurs correctement choisis. C'est le découpage du signal en segments de très courte durée. La phase de reconnaissance est l'identification des segments parmi une liste de phonèmes caractérisant la langue.

Le but est de reconnaître des mots isolés. Les systèmes de reconnaissance de parole continue agissent à un niveau supérieur, cela est dû à l'intégration de contraintes d'ordre syntaxique et linguistique. Le système KEAL, par exemple, utilise trois niveaux : un analyseur phonétique, un analyseur lexical et un analyseur syntaxique. Le dernier agit au niveau de la cohérence du discours. La reconnaissance de mots isolés est un problème moins complexe. Le type de stratégie adopté est différent dans la mesure où la solution doit être proposée à un premier niveau. Les systèmes existant consistent à reconnaître un mot parmi un vocabulaire fixé à l'avance. La reconnaissance globale apporte des solutions satisfaisantes. Mais les systèmes qui en résultent sont dépendants du locuteur et à vocabulaire figé. Ils répondent souvent à des applications précises. C'est pourquoi, notre objectif a été d'éliminer ces deux contraintes. Cela permet ainsi de pouvoir diversifier le type d'applications.

1.2. SYSTEME PROPOSE

le but est ici de concevoir pratiquement un système de reconnaissance de parole dont le principe a été proposé par la société TEXAS INSTRUMENT FRANCE. Le procédé d'analyse indépendant du locuteur a fait l'objet d'un brevet d'invention déposé par TEXAS INSTRUMENT FRANCE. Notre objectif s'oriente donc vers la réalisation pratique du système : recherche d'un méthodologie pour l'apprentissage puis programmation de la méthode de reconnaissance suivant le principe proposé. Dans un deuxième temps, nos efforts se porteront sur la réalisation d'un certain nombre de tests, afin de pouvoir juger de la validité de la méthode.

Les caractéristiques du système sont les suivantes :

- Indépendance du locuteur
- Reconnaissance de mots isolés parmi un vocabulaire donné ; ces mots sont représentés sous forme phonétique.
- Apprentissage du système réalisé pour l'ensemble des mots d'une langue donnée.

La description du système que nous exposons par la suite est inspirée de celle donnée dans le brevet d'invention intitulé : "Procédé d'analyse de la parole indépendant du locuteur" (TI, G. Benbassat, 8 Nov 1983).

1.2.1. Caractéristiques de base de la méthode

La méthode est de type analytique, c'est un point qui est à souligne En effet, les méthodes existantes donnant des résultats satisfaisants pour la reconnaissance de mots isolés sont des méthodes de reconnaissance globale. Les systèmes résultant fonctionnent en monolocuteur et l'utilisation de ce procédé pour une reconnaissance multilocuteur s'avère être une tâche difficile. On est alors amené à des solutions très lourdes à manipuler. En effet, pour résoudre le problème d'indépendance du locuteur il est nécessaire

[LEVINSON, AL 79] d'utiliser une dizaine de références par mot. Celles-ci sont obtenues à partir d'un grand nombre de locuteurs (environ 100) sélectionnés avec soin. Cela implique une phase d'apprentissage très importante compte tenu des moyens informatiques mis en oeuvre. De plus, l'apprentissage est effectué avec les mots du vocabulaire d'entrée ; il doit être réalisé à nouveau, pour un vocabulaire différent. Ce qui limite considérablement les applications du système.

Les solutions analytiques ont été jusqu'à présent, très peu exploitées. Le type de méthode a surtout été étudié pour la reconnaissance de parole continue [L.D. ERMAN] (Hearsay II) [MERCIER , AL] (keal)] où il fallair considérer un autre niveau de difficulté. L'utilisation d'une méthode analytique, pour reconnaître des mots isolés indépendamment du locuteur, est donc une idée originale : c'est un procédé très peu utilisé dans ce domaine et il éloigne le problème d'apprentissage posé pour une reconnaissance globale.

Néanmoins, certaines difficultés interviennent; la première réside dans la recherche de formes acoustiques caractéristiques des unités phonétiques de la langue. Pour résoudre ce problème, il est commode de considérer un espace acoustique divisé en sous espaces caractéristiques de chacune de ces unités. Cependant, et cela était prévisible pour un système multilocuteur, on observe un grand chevauchement de ces domaines. C'est pourquoi, il est difficile d'associer avec rigueur chaque sous espace à un phonème de la langue. La méthode proposée contourne cette difficulté par le fait qu'elle considère pour chacun de ces domaines une liste d'unités phonétiques susceptible de le représenter.

La détermination d'une distance liée à cet espace acoustique est un autre point délicat. La reconnaissance d'une forme inconnue consiste à trouver le meilleur représentant de celle-ci en calculant toutes les distances de cette forme à chacun des domaines acoustiques . C'est là qu'intervient le premier niveau de décision.

Il est nécessaire que les distances calculées ne soient pas toutes trop grandes. C'est pourquoi l'idée est de séparer l'espace acoustique considéré en un nombre de domaines aussi grand que souhaité. Cela permet d'améliorer la validité de ce premier niveau de décision, puisque la distance acceptable correspondant à l'association d'une forme inconnue au meilleur représentant peut être rendue aussi courte que nécessaire.

Nous retiendrons donc les points caractéristiques de cette méthode afin de bien comprendre la démarche qui va suivre :

1) C'est une méthode de type analytique.
2) Elle nécessite l'établissement de "formes acoustiques" dans un espace approprié ; chacune de ces formes est directement liée à une liste d'unités phonétiques et non à une unité phonétique précise.
3) La distance utilisée est une distance euclidienne.

1.2.2. Description schématique du système :

Un certain nombre de figures commentées permettront de mieux comprendre le principe de la méthode. La figure 1.2.1. décrit le principe de reconnaissance. Le mot inconnu présenté sous forme numérique est codé suivant l'ensemble des modèles constituant le dictionnaire : c'est le codage vectoriel des mots à reconnaître [LIND, BUZO, GRAY]. Chaque modèle correspond à une table contenant les phonèmes qu'il est susceptible de représenter avec telle ou telle probabilité. La phase de reconnaissance consiste donc à comparer la séquence des modèles résultant du codage vectoriel avec les mots du vocabulaire introduits sous forme phonétique.

Cette comparaison s'effectue par programmation dynamique. Le mot reconnu est celui qui correspond à la probabilité la plus grande. Le principe de cette méthode est schématisé sur la figure 1.2.2.

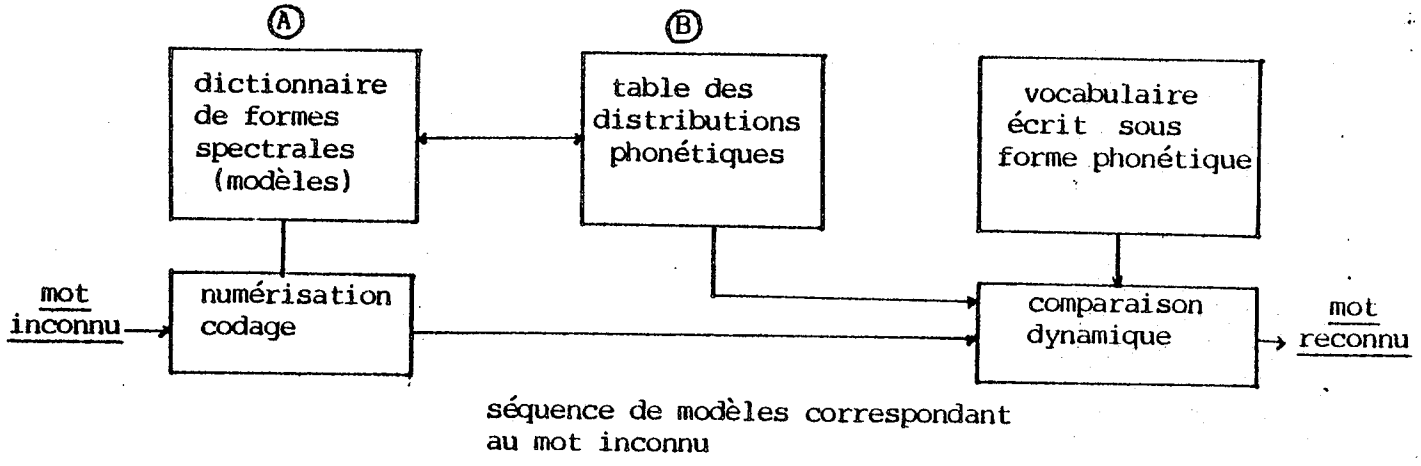


figure 1.2.1. Principe de reconnaissance

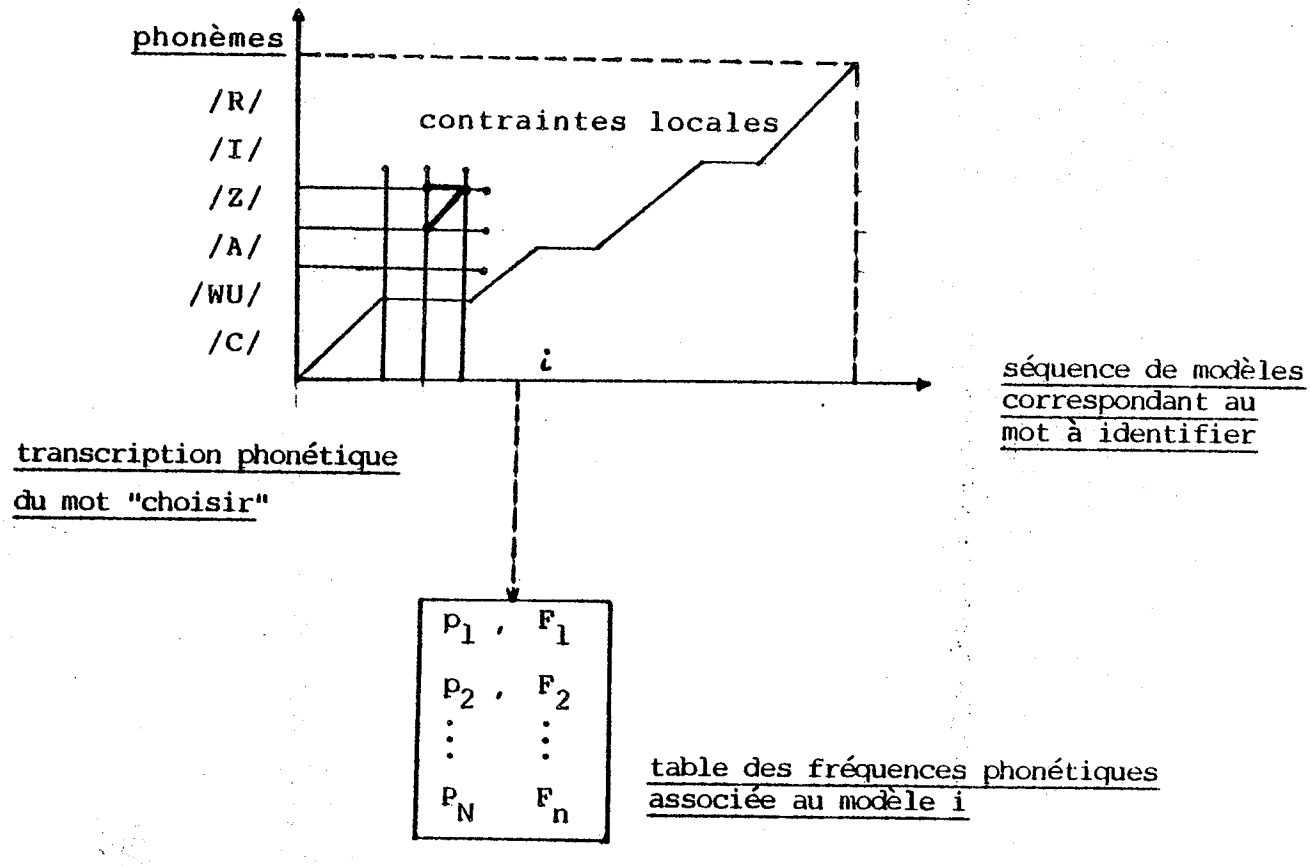


figure 1.2.2. Reconnaissance par programmation dynamique

Le dictionnaire des formes spectrales (figure 1.2.1. A) est issu d'une phase d'apprentissage au cours de laquelle des méthodes de classification adaptées agissent au niveau du corpus d'apprentissage de manière à extraire un ensemble de modèles jugé optimum. (figure 1.2.3.). La détermination d'une méthode de classification répondant au problème posé a été une des étapes les plus délicates de la phase d'apprentissage.

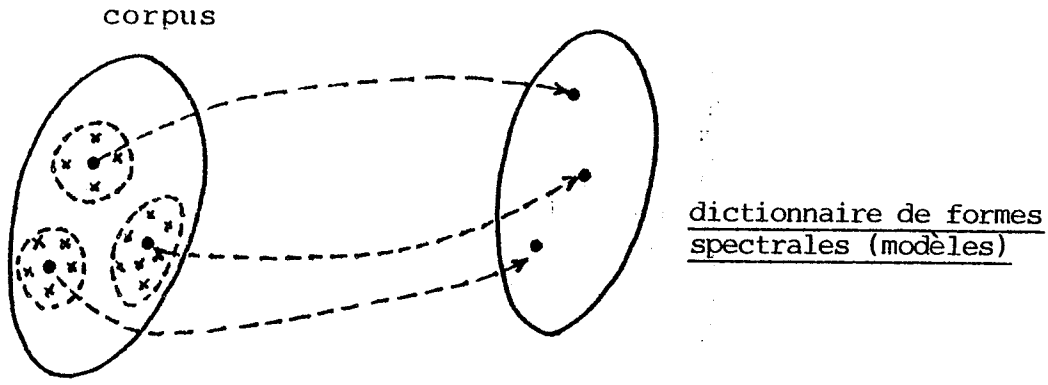


figure 1.2.3. Phase d'apprentissage à partir du corpus de données : extraction de formes spectrales représentant les unités phonétiques de la langue.

Les tables de fréquences phonétiques utilisées lors de la phase de reconnaissance (figure 1.2.1 B et 1.2.2.) donnent pour chacun des modèles sa probabilité de représenter tel ou tel phonème. Pour construire ces tables, il est nécessaire au préalable de segmenter tous les mots contenus dans le corpus de données. Ce dernier est introduit sous forme numérique (vecteurs LPC puis vecteurs cepstraux) chaque segment très court (12,5 ms) est donc étiqueté. Après le codage vectoriel sur les modèles du dictionnaire, la construction des tables consiste à identifier et à compter les phonèmes dans chacune des classes (figure 1.2.4).

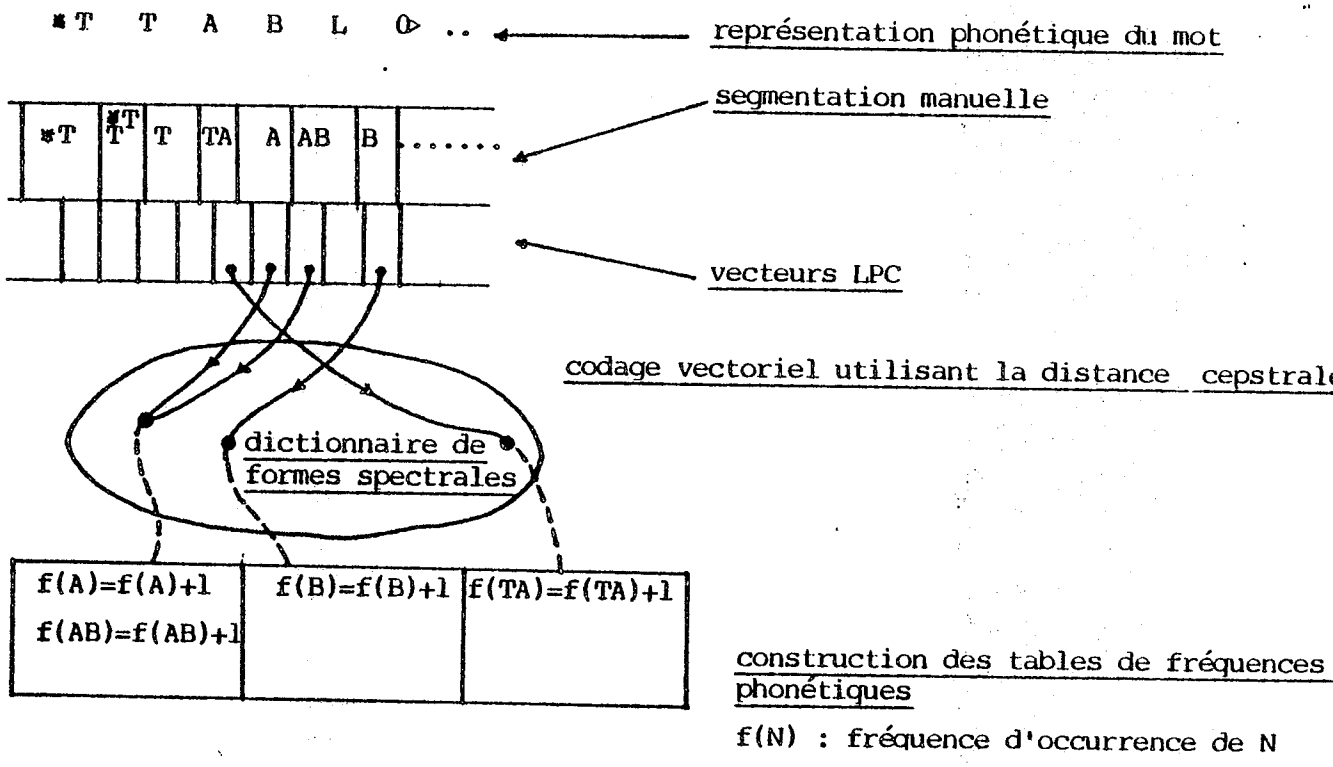


figure 1.2.4. Segmentation du corpus et construction des tables de fréquences phonétiques

1.3. NOTRE DEMARCHE

On distingue deux grandes étapes : l'apprentissage du système et la phase de reconnaissance. Pour la mise en oeuvre du procédé proposé, de nombreux points restaient à éclaircir : constituer un corpus d'apprentissage suffisamment représentatif, structurer les données à traiter, les classer de manière optimale, rechercher un traitement adapté au cas des transitions, mettre au point la méthode de reconnaissance. Tous les programmes ont été développés dans le cadre d'un logiciel existant et spécialement conçu pour le traitement du signal (ILS).

1.3.1. Le corpus d'apprentissage

La détermination des modèles s'effectue à partir du corpus d'apprentissage (figure 1.2.3.). Ce dernier doit transmettre au système toute l'information nécessaire pour la reconnaissance phonétique. La reconnaissance fonctionne pour tout vocabulaire de taille finie. C'est pourquoi, le corpus doit contenir toutes les unités phonétiques de la langue et cela, dans des contextes divers. De plus, pour une reconnaissance multilocuteur, des locuteurs différents doivent participer à l'apprentissage de manière à pouvoir introduire les phénomènes de variabilité inter-locuteur. Un bon corpus d'apprentissage doit donc avant tout être représentatif de la langue considérée et de ses variantes. La qualité de la reconnaissance en dépend fortement. L'apprentissage du système est une tâche très longue, surtout au niveau de la segmentation des données (figure 1.2.4.). C'est pourquoi il nous est apparu intéressant de mettre au point une méthode de segmentation automatique fondée sur le même principe que celui de la reconnaissance (chapitre 4). Ce procédé ne pouvait cependant fonctionner qu'à partir d'un corpus déjà segmenté manuellement.

1.3.2. Recherche d'une structure

Le corpus de données est préalablement enregistré. Dans un deuxième temps, il faut exploiter au mieux l'information transmise. Cela implique un choix judicieux de paramètres descriptifs ainsi qu'une bonne mesure de ressemblance (structure). Pour toute mesure statistique, les individus (ici les phonèmes) sont représentés par les caractères mesurés. L'espace acoustique que nous manipulon est l'espace vectoriel des coefficients cepstraux en dimension dix. La distance utilisée est la distance cepstrale (distance euclidienne). Ce premier choix (paramètres descriptifs, distance) est justifié dans le chapitre 2. Nous avons, à ce propos, essayé de faire une synthèse sur les méthodes numériques adaptées au traitement de la parole. Le choix d'un espace de mesures de grande dimension engendre une première difficulté : elle réside dans

l'observation des nuages de points. L'analyse des données est à ce niveau indispensable. Un autre point délicat se manifeste dans la recherche d'une méthodologie pour le traitement des transitions entre phonèmes. Une première question se pose : comment les représenter de manière pertinente ? Existe-t-il des classes caractéristiques permettant d'en réduire le nombre ? (environ six cents pour la langue française). Les recherches axées sur ce sujet (chapitre 3) ont eu pour objectif principal de pouvoir donner une représentation condensée des transitions. Cette démarche était nécessaire vu le type de reconnaissance choisie.

1.3.3. Analyse statistique des données

Les données à traiter sont donc assez complexes. L'observation des nuages de points est difficile. Une bonne méthode de classification est nécessaire pour la détermination des formes acoustiques (modèles) (figure 1.2.2.). Chacune d'elles doit représenter une liste d'unités phonétiques avec telle ou telle probabilité. Dans le chapitre 3, nous donnons un moyen simple de mesurer la représentativité de ces modèles par le tracé d'histogrammes. Ces modèles sont alors jugés plus ou moins caractéristiques d'un phonème ou d'une classe de phonèmes. La méthode de classification mise au point est décrite dans le même chapitre. Sa détermination est fondée sur le fait que les données à traiter étaient étiquetées au départ grâce à la segmentation du corpus. On étudiera plus spécialement le cas des phonèmes en effectuant une classification naturelle au préalable. La méthode utilisée est de type "nuées dynamiques" (DIDAY).

1.3.4. Reconnaissance

C'est la phase d'apprentissage qui est la plus longue à mettre en pratique : enregistrement du corpus, numérisation, choix d'une distance appropriée, classification ...

L'étape de reconnaissance n'a pu être mise au point (chapitre 4) qu'après la constitution des modèles de référence et du système

d'informations phonétiques associé. Le découpage du mot du vocabulaire en états phonétiques était alors à préciser. Notre idée a été de considérer ce dernier comme un processus temporel déterminé par une séquence d'états caractéristiques. (états stables, états transitoires). L'identification au cours du temps de ces états, est alors assurée par la programmation dynamique.

1.3.5. Environnement de travail

Le développement du projet s'est effectué avec l'aide préalable d'un ensemble de programmes spécialement conçus pour le traitement du signal de parole et d'une base de données adaptée (ILS) (annexe 3). La mise en oeuvre du procédé est schématisé par la figure 1.3.1. où toutes les étapes concernant la phase d'apprentissage sont décrites. Cette étape est préparatoire à la reconnaissanc Il en résulte un ensemble de modèles (A) associé à un système d'information B (tables des fréquences phonétiques). Ces données sont transmises au module de reconnaissance (figure 1.2.1). Les mots ou les phrases contenus dans le corpus de données sont enregistrés et le signal résultant est échantillonné (convertisseur A/N). Des fichiers de type ILS ("sample data file") permettent de stocker les échantillons. Un programme ILS procède alors à l'analyse du signal. On obtient alors des fichiers analysés ILS ("analysis data files"). La segmentation permet ensuite l'obtention des fichiers label ILS ("label files"). La sélection des modèles se fait après l'analyse cepstrale. Des fichiers à format variable ("record files") contiennent les coefficients cepstraux. Les programmes d'accès à ces fichiers étaient assez complexes mais l'utilisation de ces derniers était nécessaire dans la mesure où de nombreuses commandes ILS étaient applicables.

En ce qui concerne les moyens informatiques, les programmes ont été développés au départ sur un mini ordinateur DS 990 TEXAS INSTRUMENT puis ont dû être transférés, en cours de projet, sur un VAX DIGITAL 750 ainsi que toute la base de donnée correspondante.

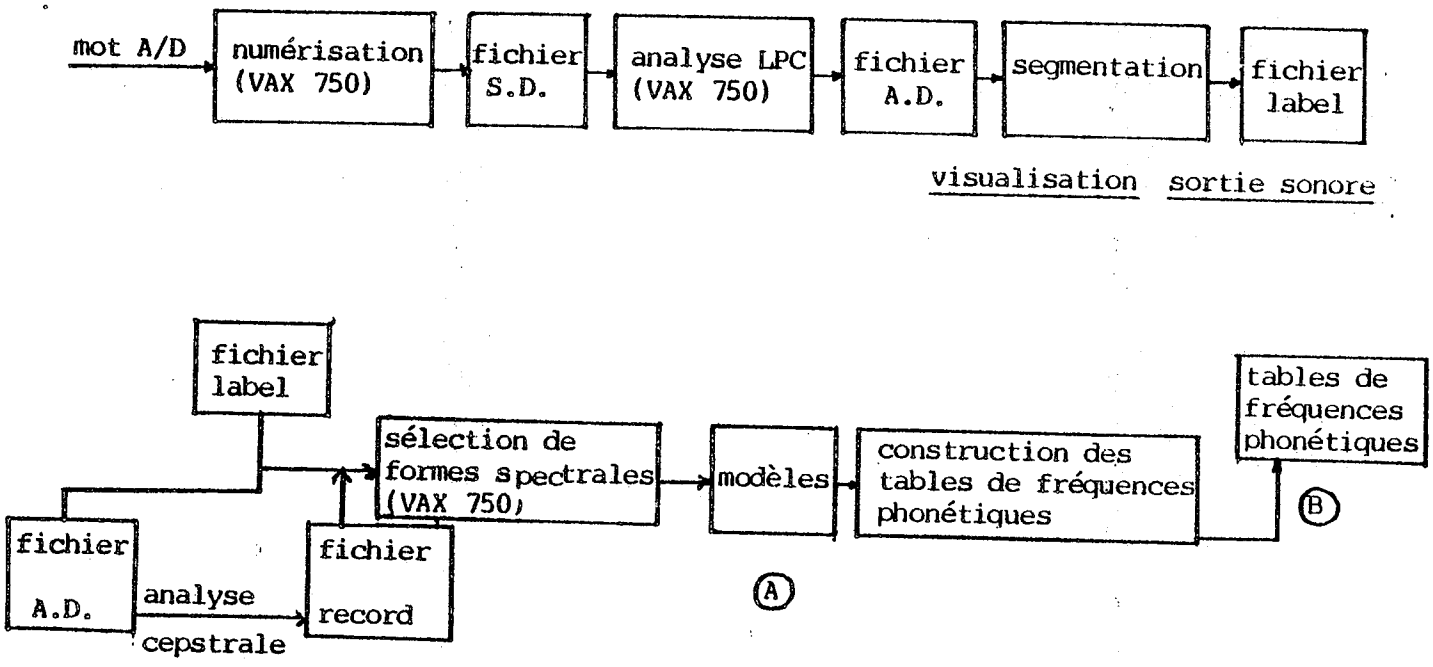


figure 1.3.1. Phase d'apprentissage

1.3.6. Plan de l'ouvrage

Notre ouvrage comporte par la suite trois chapitres principaux :

- Le traitement numérique de la parole : (recherche de paramètres descriptitfs, distance).
- La phase d'apprentissage (liste des phonèmes de la langue française, segmentation, classification).
- La phase de reconnaissance (segmentation automatique, principe et programmation de la reconnaissance, tests de reconnaissance).

Des annexes permettent au lecteur moins averti d'éclaircir quelques points particuliers de l'énoncé.

C H A P I T R E I I

Traitement numérique de la parole

TRAITEMENT NUMERIQUE DE LA PAROLE

La complexité de la nature même du signal de parole nécessite l'utilisation d'une bonne méthode d'analyse. Le traitement numérique du signal consiste à extraire un nombre de paramètres suffisants et caractérisant le signal de manière pertinente. Les méthodes classiques peuvent être divisées en deux grandes classes : les méthodes spectrales et les méthodes temporelles sans oublier celles qui font appel aux deux.

Dans un système multilocuteur, les phénomènes de variabilité dus au passage d'un locuteur à l'autre viennent s'ajouter en tant que paramètres non discriminants pour la reconnaissance (variabilité inter-locuteur). La recherche de paramètres numériques caractérisés par une variabilité interindividuelle peu importante est un problème délicat. Certaines des méthodes numériques exposées ici ont donné de bons résultats en reconnaissance monolocuteur. L'utilisation de distances bien adaptées (spectrale, Itakura, mel...) laisse supposer une faible variabilité intra-locuteur (articulation, vitesse d'élocution, phénomènes physiologiques, psychologiques, etc ...), mais les résultats restent encore discutables en ce qui concerne la variabilité inter-locuteur.
La représentation paramètrique du signal est donc certes nécessaire mais s'avère insuffisante en reconnaissance de parole et pour un système multilocuteur. Nous exposons ici les techniques de traitement numérique en soulignant celles que nous avons adoptées.

2.1. ANALYSE EN PREDICTION LINEAIRE (LPC)

L'analyse en prédiction linéaire a été introduite par B. ATAL et S. HANOWER en 1971. Depuis, c'est une méthode qui est largement exploitée en traitement de la parole. Fondée sur l'hypothèse d'un modèle autorégressif, elle a servi d'appui à de nombreuses techniques de calcul.

En ce qui concerne notre approche, l'analyse prédictive nous a permis d'obtenir facilement les coefficients cepstraux par résolution d'un système linéaire simple. Ce chapitre a pour but d'exposer le principe de l'analyse prédictive, les différentes méthodes de calcul, et l'intérêt de cette analyse en traitement de parole. Ensuite nous ferons un lien avec le traitement que nous avons adopté en montrant comment l'utilisation de l'analyse LPC a rendu facile l'extraction des coefficients cepstraux.

2.1.1. Principe :

L'analyse en prédiction linéaire a été utilisée pour le problème de production de parole. Le filtre qui en résulte est décrit par la figure 2.1.1. où U_n est soit un bruit aléatoire soit un signal quasi périodique.

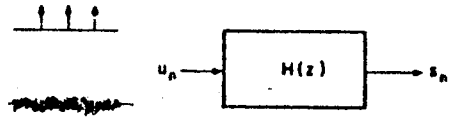


Fig. 2.1.1. d'après [J. MAKHOUL] :
Un modèle de production de parole échantillonnée.

Le principe de production de parole est le suivant :
Soit un signal s(t) échantillonné suivant ses valeurs $s(nt)=s_n$ toutes les "T" secondes.
On considère donc la suite $[s_n]$ des intervalles réguliers et contigus. On obtient alors une série d'intervalles appelée

"fenêtre d'échantillonnage". L'analyse en prédiction linéaire est effectuée à l'intérieur d'une de ces fenêtres.
Un échantillon de parole s_n peut être prédit en fonction des p précédents et de la séquence U_{n-l} (l=O, l=q) correspondant à la source inconnue :

$$Sn = \sum_{k=1}^{p} a_k \hat{S}n-k + G \sum_{l=0}^{q} b_l \, Un-l \qquad bo=1 \tag{1}$$

G est le gain, p et q sont les paramètres du système p est l'ordre de prédiction linéaire.

Dans le domaine fréquentiel l'équation (1) devient :

$$H(z) = \frac{S(z)}{U(z)} = \frac{G \left(1 + \sum_{l=1}^{q} b_l \, z^{-l}\right)}{1 + \sum_{k=1}^{p} akz^{-k}} \tag{2}$$

H(z) est la fonction de transfert du filtre de prédiction linéaire. Le modèle général décrit par l'équation (2) est appelé "pole-zéro model". Plus particulièrement, on appelle :
- "all-zero model" ce qui correspond à ak=o $1 \leqslant k \leqslant p$ (modèle MA)
- "all-pole model" ce qui correspond à bl=o $1 \leqslant l \leqslant q$ (modèle AR)

Le modèle "MA" est appelé modèle à moyenne glissante ("mowing average"), le modèle "AR" est le modèle autoregressif ("autoregressive model").
Nous nous intéressons au modèle AR car c'est celui qui est le plus utilisé par les spécialistes du traitement du signal [MAKHOUL, 1975] pour l'analyse des séries temporelles.

L'équation (1) pour un modèle AR devient :

$$Sn = -\sum_{k=1}^{p} ak\ Sn-k + G\ Un$$

L'équation (2) s'écrit :

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} akz^{-k}}$$

La figure 2.1.2. montre le modèle AR dans le domaine fréquentiel et dans le domaine temporel :

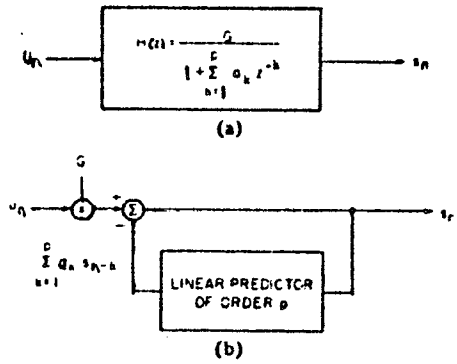


Figure 2.1.2. d'après [J. MAKHOUL]

2.1.2. Méthodes de calcul :

Une fois le signal échantillonné, l'analyse prédictive se fait à largeur de fenêtre fixe.
Pour un modèle AR :

$$s_n = \sum_{i=1}^{p} a'i\ s_{n-i} + G\ Un$$

On note $\hat{s}_n = \sum_{i=1}^{p} a'i\ S_{n-i}$ le signal "prédit" d'erreur de prédiction peut être exprimée par :

$$e_n = s_n - \hat{s}_n = G\ Un \qquad (3)$$

2.1.2.1. Minimisation quadratique : On minimise la quantité α

$$\alpha = \sum_n e_n^2$$

α est rendue minimale pour :

$$\frac{\partial \alpha}{\partial ai'} = o \qquad i = 1, p$$

La minimisation quadratique donne des équations linéaires. En effet :

$$C_{ij} = \sum_n s(n-i)\ s(n-j)$$

$$\alpha = \sum_n e_n^2 = \sum_n \left[\sum_{i=o}^{p} ai\ s(n-i) \right]^2 \quad \text{avec} \begin{cases} a_o = 1 \\ a_i' = -a_i \end{cases}$$

$$\alpha = \sum_n \sum_{i=o}^{p} ai\ cij\ aj$$

d'où :

$$\frac{\partial \alpha}{\partial ai} = o = \sum_{j=o}^{p} cij\ aj$$

d'où le système d'équation :

$$\sum_{i=1}^{p} ai\ cij = -c_oj \qquad j = 1, p \qquad (4)$$

système p équations, p inconnues

2.1.2.2. Méthodes de résolution :

2.1.2.2.1. Méthode de covariance [MARKEL, GRAY]

Soit une séquence de n échantillons s(o), s(1) ... s(n-1), on choisit l'intervalle compris entre les échantillons S_p et S_{n-1}. De ce fait, l'erreur quadratique α n'est minimisée que dans l'intervalle [P, n-1]

$$\alpha = \sum_{j=p}^{j=n-1} e_j^2$$

alors que les n échantillons sont utilisés pour calculer les terme de la matrice cik. Ce qui donne le système :

$$\sum_{i=i}^{p} aiCik = - c_o k \qquad k = 1 \,..\, p$$

$$Cik = \sum_{m=p}^{m=n-1} s(m-i)\, s(m-k)$$

$$en = \sum_{i=o}^{p} (ai\, s(n-i))$$

avec $a_o = 1$ et $n = p \,...\, n-1$

2.1.2.2.2. Méthode d'autocorrélation [MARKEL, GRAY]

On prend l'intervalle $]-\infty, +\infty]$ et on impose $s(m)=o$ pour $m<o$ et $m \geqslant n$, alors :

$$r(|i-j|) = \sum_{m=o}^{m=n-1+|i-j|} s(m)\, s(m+|i-j|)$$

ce qui donne le système :

$$\sum_{i=1}^{p} (a_i \, r|i-j|) = -\, r(j) \qquad \begin{matrix} i = 1 \,..\, p \\ j = 1 \,..\, p \end{matrix}$$

avec :

$$r(l) = \sum_{m=0}^{m=n-1-l} s(m)\, s(m-l) \qquad l = 1 \,..\, p$$

avec l'erreur :

$$e(n) = \sum_{i=0}^{i=p} a_i \, s(n-i) \quad (a_o = 1)$$

pour $n = o, \ldots n+p-1$

2.1.2.3. Discussion :

Une fois les coefficients de prédiction linéaire calculés se pose le problème de stabilité du filtre résultant. Il est montré [MAKHOUL], que la méthode d'autocorrélation donne une solution stationnaire contrairement à la méthode de covariance (Fig. 2.1.3.). Si cette stabilité est garantie en théorie, elle n'est pas assurée pour des mots de longueur finie.

COMPARISON BETWEEN DIFFERENT PROPERTIES OF VARIOUS LINEAR-PREDICTION METHODS

PROPERTY	LINEAR PREDICTION METHOD			
	AUTOCORRELATION	COVARIANCE	REGULAR LATTICE	COVARIANCE LATTICE
WINDOWING	NECESSARY	NONE	NOT NECESSARY	NONE
STABILITY	THEORETICALLY GUARANTEED	NOT GUARANTEED	CAN BE GUARANTEED	
STABILITY WITH FINITE WORDLENGTH COMPUTATIONS	NOT GUARANTEED		CAN BE GUARANTEED	
COMPUTATIONAL EFFICIENCY	EFFICIENT		EXPENSIVE	EFFICIENT
LEAST-SQUARES OPTIMALITY	OPTIMAL		POSSIBILITY ONLY SUBOPTIMAL	
QUANTIZATION OF REFLECTION COEFFICIENTS WITHIN RECURSION	NOT POSSIBLE		POSSIBLE	
NUMBER OF SAMPLES FOR ANALYSIS	"		CAN BE REDUCED TO 0.7N FOR THE SAME RESOLUTION	

Fig.2.1.3. - Comparaison de méthodes pour le calcul des coefficients de prédiction linéaire - d'après [J. MAKHOUL]

D'autres méthodes de calcul sont également exploitables ; celle proposée par KALMAN par exemple (filtre de KALMAN) [GUEGEN] qui fait appel à la récurrence du système.
C.J. GUEGEN propose une méthode de prédiction linéaire modifiée en imposant toutes les racines de A(2) sur le cercle unité.
Quoi qu'il en soit, la méthode la plus courante utilise le critère d'optimisation quadratique de l'énergie résiduelle du filtre LPC.

2.1.3. Applications pratiques :

Nous donnons ici quelques applications :

2.1.3.1. Estimation du spectre :

Soit un signal x(t) et son spectre x(ν), la relation de Parseval appliquée à x(ν) (annexe 1) donne :

$$||x||_t^2 = \langle x, x\rangle = \langle x, x\rangle = ||X||_\nu^2$$

En appliquant la relation à

$$||e||_t^2 = \sum_2 e_n^2 = ||E||_\nu = \frac{1}{2\pi}\int_{-\Pi}^{\Pi} |E\ (e^{jw})\ |^2\ dw$$

soit : $P(w) = |S(e^{jw})\ |^2$

d'après la relation : $E(z) = A(z)\ S(z)$ on a :

$$E = \frac{1}{2\pi}\int_{-\Pi}^{\Pi} P(w)\ A(e^{jw})\ A(e^{-jw})\ dw$$

soit : $\hat{P}(w)$ l'estimateur de p(w)

$$\hat{P}(w) = |H(e^{jw})\ |^2 = \frac{G2}{|A(e^{jw})|^2}$$

$$P(w) = |S(e^{jw})| = \frac{|E(e^{jw})|^2}{|A(e^{jw})|^2}$$

et donc

$$E = \frac{G2}{2\pi}\int_{-\Pi}^{\Pi} \frac{P(w)}{\hat{P}(w)}\ du$$

Il en résulte alors une autre formulation du problème de prédiction linéaire :

"Etant donné un spectre P(w), une estimation de ce spectre $\hat{p}(w)$ est telle que la quantité :

$$E = \frac{G2}{2\pi} \int_{-\pi}^{\pi} \frac{P(w)}{\hat{P}(w)} d\omega \text{ soit rendue minimale".}$$

Le problème de prédiction linéaire peut être formulé aussi bien dans le domaine temporel (autorégression du signal s(t) que dans le domaine fréquentiel (minimisation de la distance entre P(w) et $\hat{P}(w)$. La figure 2.1.4. donne une estimation du spectre (ou enveloppe spectrale) par la méthode LPC.

2.1.3.2. Calcul de distance [ITAKURA]

Le calcul des coefficients LPC a été utilisé dans le domaine de la reconnaissance de la parole [ITAKURA, 1975]. Le problème était ainsi formulé : " quelle est la mesure optimale qui puisse servir de test au problème de similarité entre un échantillon inconnu et un modèle connu par la détermination de ses coefficients LPC ?". Il a alors été montré que la distance exprimée par le log du rapport de vraisemblance (voir chapitre distance) donnait les meilleurs résultats.

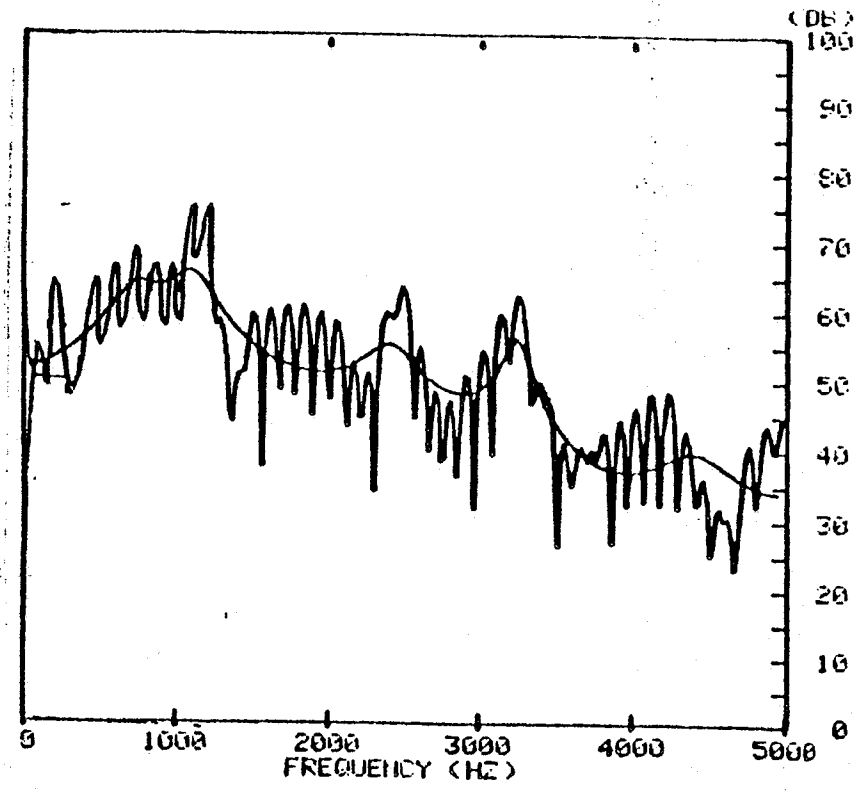


Fig. 2.1.4. - Estimation du spectre par la

2.1.4. Ordre de prédiction :

Soit V_p l'erreur normalisée : $V_p = \frac{Ep}{R(o)}$

où R(o) est l'énergie du signal (annexe 1). P est l'ordre de prédiction linéaire. Si P croît E_p décroît. Le problème est de déterminer un nombre P_o à partir duquel la variation de E_p est négligeable. Le test à effectuer s'exprime par la formule : $1 - \frac{V_p+1}{V_p} < S$ c'est à dire regarder l'instant où le rapport $\frac{V_p+1}{V_p}$ se rapproche de 1. Un autre moyen d'optimisation a été proposé par AKAIKE (1974). Il consiste à calculer $I(p) = \log (U_p) + \frac{2p}{Ne}$ où Ne est le nombre effectif de points dans la fenêtre d'étude (Ne ≈0.4N pour une fenêtre de Hamming, N étant le contexte d'échantillonnage).

La Figure 2.1.5. donne le tracé de U_p et I(p).

La valeur optimale est obtenue pour p = 10.

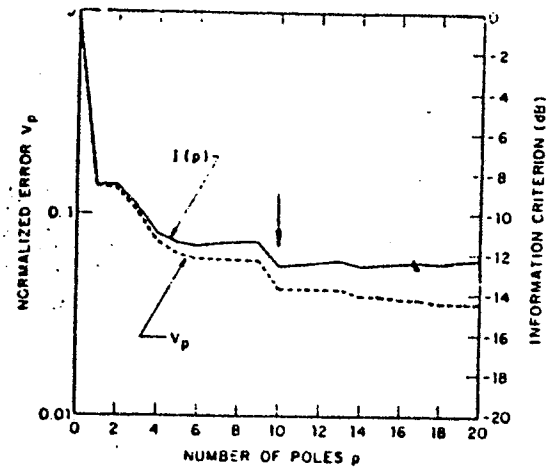


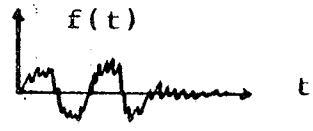
Fig. 2.1.5. [J. MAKHOUL]

2.1.5. Intérêt de l'analyse prédictive dans notre système de reconnaissance

2.1.5.1. Echantillinage du signal

(figure 2.1.6.)

enregistrement ⟶ signal analogique f(t)



filtrage comprenant 2 filtres en séquence

⟶ - 1 filtre passe bas F_M = 3,4 KHz
- 1 filtre passe haut F_m = 60 Hz

A/N
⟶ convertisseur analogique/numérique : échantillonnage
Fe = 8 KHz

fichiers échantillonnés type ILS

figure 2.1.6. Séquence d'acquisition du signal digitalisé

En première approximation, on admet qu'un signal n'ayant pas de composantes fréquentielles au delà d'une fréquence FM est entièrement décrit par la suite de ses échantillons sous la condition suivante : la période d'échantillonnage Te ($Te = \frac{1}{Fe}$) doit être strictement inférieure à la quantité $\frac{1}{2FM}$ (ou $Fe > 2\ FM$) (cf théorème d'échantillonnage d'un signal (WIENER et SHANNON)).

Pour les voix d'hommes, il n'y a pour ainsi dire pas de composantes fréquentielles significatives au delà de 3000 H_z. On choisit donc dans ce cas, une fréquence d'échantillonnage de 8000 H_z. Pour les voix de femmes cette fréquence est insuffisante, puisqu'il existe des composantes fréquentielles au delà de 4000 H_z. Il est bon [J.S. LIENARD] de choisir Fe = 2,5 FM afin de faciliter le passage numérique/analogique.
Le signal est donc capté en temps réel sous forme analogique f(t). Un filtrage résultant d'un filtre "passe haut" à 60 H_z et d'un filtre "passe bas" à 3400 H_z permet de récupérer les composantes fréquentielles entre ces deux fréquences. Un convertisseur A/N procède alors à un échantillonnage à 8000 H_z (supérieur à 2 x (3400) H_z). Une fois l'échantillonnage effectué, il faut déterminer le contexte d'échantillonnage, c'est-à-dire le nombre de points contenus dans chacune des fenêtres d'étude. Le choix de ce contexte est important par la suite, au moment de l'analyse LPC. En effet, le processus constitué par la suite des échantillons est considéré comme stationnaire à l'intérieur de chacune de ces fenêtres. Dans notre étude, le découpage est effectué toutes les 12,5 ms ce qui correspond à 100 points par fenêtre d'échantillonnage. Ces informations sont récupérées dans des fichiers d'échantillonnage type ILS (Fig. 2.1.7.).

SECTOR 1, FRAME 50

0	-1	0	0	1	0	-2	0	0	-1
-2	0	-1	-2	0	-1	-3	-4	-2	-3
-3	-1	-2	-4	-3	-2	-4	-2	-3	-4
-4	-3	-3	-5	-4	-3	-5	-3	-1	-1
-6	-3	-3	-4	-3	-3	-4	-5	-3	-2
-3	-3	-2	1	-1	0	-1	-1	-2	-1
-1	0	0	0	-1	1	4	2	2	2
2	4	4	3	2	3	4	4	4	5
5	3	6	6	5	2	6	9	6	7
4	3	4	5	5	5	5	6	5	6

Fig. 2.1.7. - Echantillonnage du signal à Fe = 8 K H_Z fenêtre d'échantillons à 100 points (fichier ILS)

2.1.5.2. Analyse LPC :

L'analyse LPC se fait directement sur les fichiers échantillonnés type ILS. On obtient des fichiers d'analyse type ILS. La première étape consiste à n'analyser que les segments de parole véritable, c'est-à-dire que l'on élimine préalablement le début et la fin des mots par des calculs d'énergie. Ensuite on calcule les coefficients de réflexion du filtre inverse A(Z) par la méthode de LEVINGSTON [MARKEL, GRAY, "Linear Prediction of Speech" p 46]. Les coefficients LPC sont ensuite calculés par une récurrence (Fig. 2.1.8., Fig. 2.1.9). L'ordre de calcul est choisi égal à 10. Dans les fichiers analysés sont stockés les 10 premiers coefficients de réflexion.

FORTRAN PROGRAM

```
      SUBROUTINE AUTOCR(X,N,M,ALPHA,RC)
      DIMENSION X(1),A(11),RC(1)
      DIMENSION R(11),B(10)
      NP=M+1
      DO 100 K=1,NP
      R(K)=0.
      NK=N-K+1
      DO 100 MP=1,NK
      MPK=MP+K-1
  100 R(K)=R(K)+X(MP)*X(MPK)
      ALPHA=R(1)
      RC(1)=-R(2)/R(1)
      A(1)=1.
      A(2)=RC(1)
      ALPHA=ALPHA-RC(1)*RC(1)*ALPHA
      NP=M
      DO 400 MINC=2,NP
      M=MINC-1
      DO 300 J=1,MINC
      JB=MINC-J+2
  300 B(J)=A(JB)
      M=M+1
      S=0.
      DO 350 I=1,M
      MI=M-I+2
  350 S=S+R(MI)*A(I)
      RC(M)=-S/ALPHA
      DO 360 I=1,M
  360 A(I)=A(I)+RC(M)*B(I)
      A(M+1)=RC(M)
      ALPHA=ALPHA-RC(M)*RC(M)*ALPHA
      IF(ALPHA.LE.0.)WRITE(6,500)
  500 FORMAT(' UNSUFFICIENT ACCURACY')
  400 CONTINUE
      RETURN
      END
```

Fig. 2.1.8. [MARKEL, GRAY] Calcul des coefficients de réflexion (RC) du filtre inverse A(z).

```
      SUBROUTINE RC2A(A,RC,M)
      DIMENSION A(2),RC(1)
      A(1)=1.
      A(2)=RC(1)
      MM1=M-1
      IF(M.LE.1) GO TO 130
      DO 120 I=1,MM1
      IP1=I+1
      I2=IP1/2
      RCIP1=RC(IP1)
      DO 110 J=1,I2
      IJ=IP1-J
      TA=A(J+1)+RCIP1*A (J+1)
  110 A(J+1)=TA
      A(IP1+1)=RCIP1
  120 CONTINUE
  130 RETURN
      END
```

Fig. 2.1.9 - Programme ILS permettant le calcul des coefficients LPC à partir des coefficients de réflexion.

Conditions d'analyse :

L'analyse se fait sur les segments de parole véritable. Le numéro de la première fenêtre analysée ainsi que le nombre de fenêtres analysées est enregistré dans l'entête du fichier analysé (Fig. 2.1.11). Pour une fenêtre d'échantillonnage de 100 points, on choisit une fenêtre d'analyse à 100 points (appelé contexte d'analyse). Une fenêtre d'analyse supérieure à la fenêtre d'échantillonnage aurait pour effet de tenir compte de la forme du signal avant et après le segment étudié ce qui n'est pas nécessaire dans notre cas. De toute façon, la fenêtre d'analyse qui détermine le nombre de points

sur lesquels est effectuée la minimisation quadratique doit au moins être égale à la longueur de la fenêtre d'échantillonnage. Pour éviter tout "effet de bord" on utilise une fenêtre de Hamming $W_H(n)$ exprimée par :

$$W_H(n) = \begin{cases} (\alpha+(1-\alpha)\cos\frac{2\Pi n}{N}) & si\ -\frac{(N-1)}{2} \leqslant n \leqslant \frac{(N-1)}{2} \\ 0.0 \text{ ailleurs} \end{cases}$$

avec $\alpha = 0.54$ (fenêtre de Hamming)

$\alpha = 0.5$ (fenêtre de Hanning)

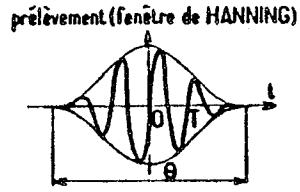


Fig. 2.1.10 - Fenêtre de Hanning appliquée à une Sinusoide [J.S. LIENARD].

```
                ANALYSIS DATA
SECTOR    1,FRAME     60
  7400     17809     1423     1181     6898     2877    -3163    -2368     7288    -6679
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0        0
     0         0        0        0        0        0        0        0        0       60
     0         0        0        0        0        0        0        0
 24683        10       14       93       10      100     5901        0
```

Fig. 2.1.12 - Fichier analyse ILS

Les dix premiers coefficients sont les coefficients de prédiction correspondant à la fenêtre numéro 60. Les coefficients LPC sont ensuite calculés par le programme RC2A (Fig. 9).

On utilise une préemphase à 6dB/Octave où le signal résultant de la préemphase s*(t) est égal à la différence du signal initial s(t) et de la quantité K×s(t-1) (où K=0.93). Cette transformation a l'intérêt de donner une allure à peu près plate à la fréquence à long terme de la parole.

100	10	93	100	Y	0	100	60	80	1
0	0				0	0	0	1	
0	0				0	0	0	0	
256	768	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
256	8000	-29000	32149	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
256	0	8000	0	-32000	-1	32149	0		

Fig. 2.1.11 - Fichier analyse ILS

Entête de fichier d'un fichier analyse

100 = contexte d'échantillonnage
10 = ordre de prédiction
93 = préemphase
100 = contexte d'analyse
y = "Hamming window"
60 = première fenêtre analysée
80 = nombre de fenêtres analysées

2.2. ANALYSE CEPSTRALE

Nous donnons dans ce chapître, un bref exposé de l'analyse cepstra
Nous nous pencherons plus particulièrement sur la définition des coefficients cepstraux, leurs propriétés, et leur intérêt dans notr étude.

2.2.1. Fonction cepstre complexe ; Fonction cepstre réel

Soit un signal $x(t)$ et son spectre $X(\nu)$.
La fonction "log" appliquée à une quantité complexe peut être auss définie [Rabiner] :
Pour c complexe : $\text{Log}(c) = \text{Log}|c| + j \arg(c)$
On note alors : $\bar{X}(e^{jw}) = \text{Log}|X(e^{jw})| + j \arg(X(e^{jw}))$.
Par Transformée de Fourier inverse appliquée à $\bar{X}(e^{jw})$, on obtient la quantité :

$$\hat{x}(n) = \frac{1}{2\Pi}\int_{-\Pi}^{\Pi} \bar{X}(e^{jw})e^{jwn}\,dw \qquad (1)$$

$\hat{x}(n)$ est appelé cepstre complexe de $x(t)$

La fonction cepstre réel est déterminée par l'équation :

$$c(n) = \frac{1}{2\Pi}\int_{-\Pi}^{\Pi} \text{Log}|X(e^{jw})|\, e^{jwn}\,dw \qquad (2)$$

L'équation (1) peut être exprimée [Rabiner] par une Transformée de Fourier discrète :

1) $$X_p(k) = \sum_{n=o}^{n=N-1} x(n)e^{-2i\Pi kn/N}$$

2) $$\bar{X}_p(k) = \text{Log}(X_p(k))$$

$$3)\ \hat{x}_p(k) = \frac{1}{N} \sum_{n=o}^{N-1} (\bar{X}(k) e^{2j\Pi kn/N})$$

$\hat{x}_p(k)$ est une approximation de $\hat{x}(n)$.

Parallèlement, une approximation du cepstre réel peut être faite par :

$$C_p'(n) = \frac{1}{N} \sum_{k=o}^{N-1} X_p(k) e^{2i\Pi kn/N}$$

Nous nous intéresserons désormais au cepstre réel :

Propriété de c(n)

Une transformation de Fourier appliquée à c(n) permet de retrouver

$$\log |X(e^{jw})| = \frac{w}{2\Pi} \int_{w/4\Pi}^{w/4\Pi} c(n) e^{jwn}\, dn$$

Interprétation :

Schéma de calcul de c(n) à partir de x(t) :

$$x(t) \overset{F}{\longleftrightarrow} X(e^{jw}) \longleftrightarrow \text{Log } |X(e^{jw})|$$

$$\overset{F-1}{\longleftrightarrow} c(n)$$

Propriétés :

- déphasage : soit le signal x(t) déphasé de t_0

$$x(t-t_0) \longleftrightarrow e^{-2\Pi N t_0} X(e^{jw})$$

si $x(t) \longleftrightarrow X(e^{jw})$

(notations de l'annexe 1)

on a alors : si $\bar{X}(e^{jw}) = \text{Log}\ X(e^{jw}) + j\ \arg(W(e^{jw}))$

$$\text{Log}\ (e^{-2\Pi N t_0}\ X(e^{jw})) = \text{Log}\ (|X(e^{jw})|) + j\ \arg(e^{-2\Pi N t_0}\ X(e^{jw}))$$

et donc on obtient le résultat suivant :

Le cepstre réel de deux signaux déphasés est le même.

Ce résultat est important en ce qui concerne notre étude. En effet (Fig. 2.2.1.) si le cepstre est déterminé pour chacun des segments correspondant au contexte d'échantillonnage et dans une zone de stabilité (phonème), le déphasage imposé par la détermination des bornes d'un segment n'a pas d'effet dans le résultat.

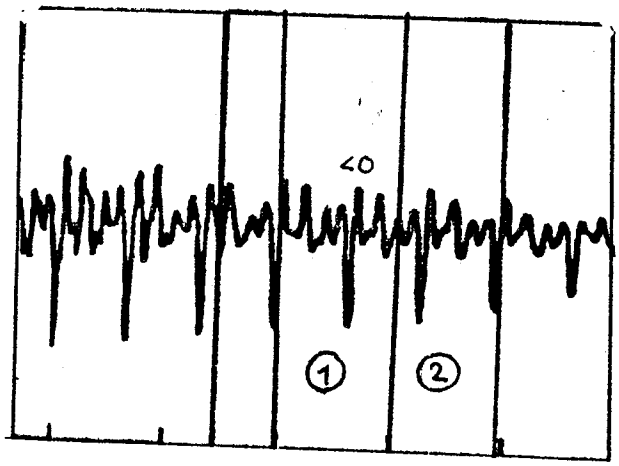


Fig. 2.2.1. : Calcul du cepstre sur les segments (1) et (2)
Le déphasage imposé par les borne des segments (nécessité de conte te d'échantillonnage) n'a pas d'effet sur le calcul du cepstre

- amplification :

Supposons maintenant que le signal x(t) est multiplié par a :

$$ax(t) \longleftrightarrow a\ X(e^{jw})$$

$$\text{Log}\ |a\ X(e^{jw})| = \text{Log}\ |a| + \text{Log}\ |X(e^{jw})|$$

Si on calculele cepstre du signal ax(t) :

$$\text{cepstre}\ (ax(t)) = \frac{\text{Log}\,|a|}{\Pi n} \sin \Pi n + \text{cepstre}\ (x(t))$$

pour $n \in N$ $\quad \frac{\log |a|}{\Pi n} \sin \Pi n = 0$

donc : l'amplification d'un signal ne change pas la valeur du cepstre.

2.2.2. Intérêt du cepstre dans le traitement de la parole :

- estimation de la période du fondamental :

Le spectre instantané d'un signal de parole peut être considéré comme la multiplication du spectre du signal source et de la fonction de transfert du conduit vocal. Ce produit devient une somme lorsque l'on passe à l'échelle logarithmique. Le spectre total devient donc la somme de deux fonctions : l'une représentant le signal source (oscillations rapides et périodiques) l'autre représentant le conduit vocal (oscillations lentes et généralement apériodiques).

Par transformation de Fourier inverse on obtient le CEPSTRE. Une transformation correspondant à un spectre donnant des oscillations rapides et périodiques donne une raie d'abscisse éloignée indiquant la période du signal source. Pour un spectre donnant des oscillations lentes et apériodiques on observe une courbe plus étalée et voisine de l'origine. La figure 2.2.3. montre une série de "Log" de spectres avec le cepstre correspondant. Pour les sons non voisés (les 9 premiers) le deuxième maximum du cepstre n'est pas observé. Pour les sons voisés l'abscisse correspondant au deuxième maximum correspond à $T_0 = \frac{1}{F_0}$. Le cepstre est donc aussi un moyen de distinguer les sons voisés des sons non voisés. Le début du cepstre (courbe plus étalée) correspond au cepstre du conduit vocal ; la courbe située de part et d'autre du deuxième maximum correspond au cepstre instantané du signal source (Fig. 2.2.2).

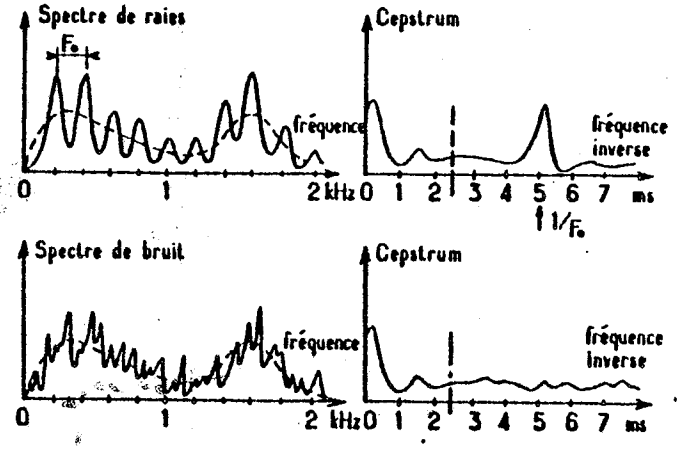


Figure 2.2.2. d'après [J.S. LIENARD]

Calcul du cepstre pour un spectre de raie et pour un spectre de

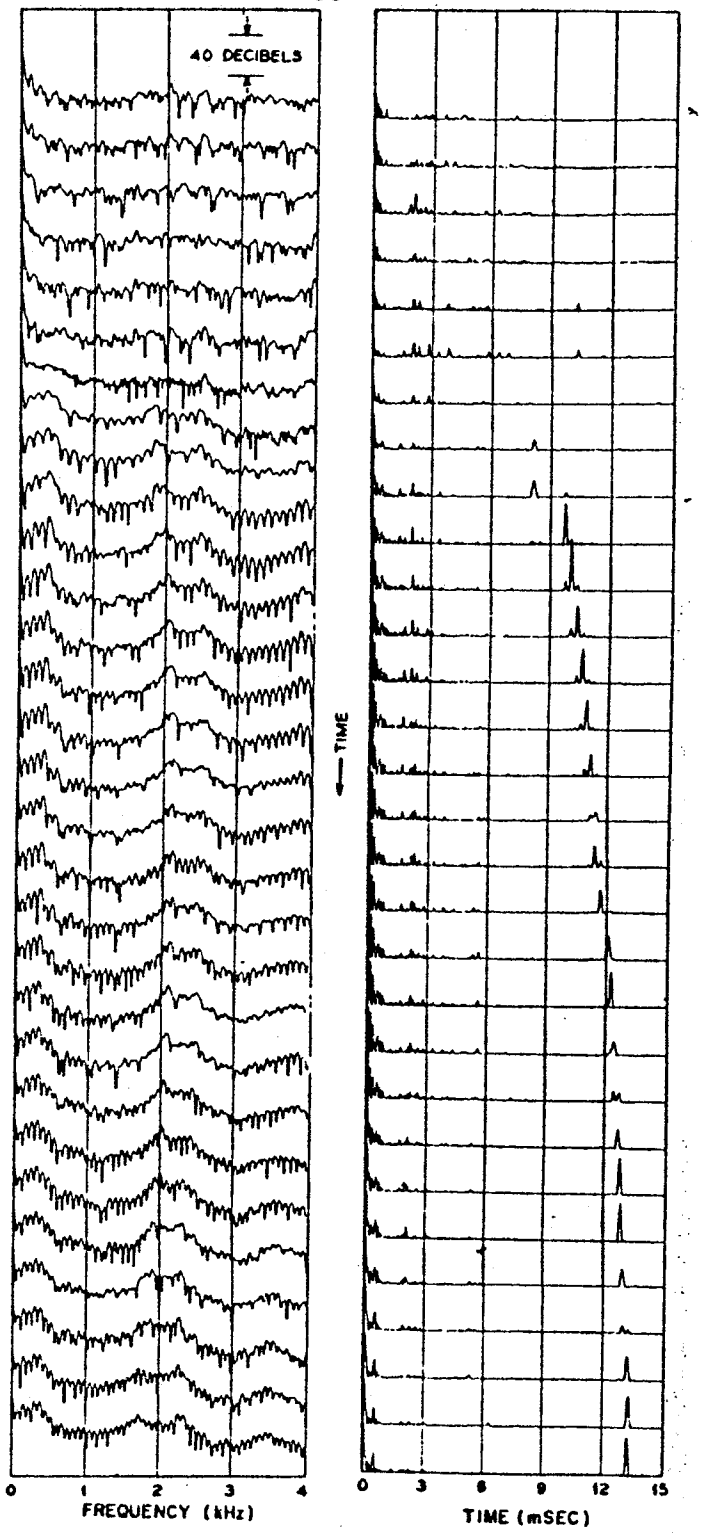


Fig. 2.2.3. Log du spectre et cepstre correspondant : distinction des sons voisés et non voisés d'après [L.R. RABINER - R.W. SHAF

2.2.3. Coefficients cepstraux :

2.2.3.1. Nature des coefficients :

Soit la fonction Log $|X(e^{jw})|$ que l'on développe suivant la série de fonctions orthogonale e^{jwn} ($n \in \mathbb{Z}$)

$$\text{Log } |X(e^{jw})| = \sum_{n=-\infty}^{n=+\infty} (c_n \, e^{jwn})$$

La suite e^{jwn} étant orthogonale et normée :

$c(n) = \langle e^{jw}, \text{ Log } |X(e^{jw})| \rangle$ (produit scalaire)

d'après la définition du produit scalaire dans le domaine fréquentiel :

$$c(n) = \frac{1}{2\Pi} \int_{\Pi}^{\Pi} \text{Log } |X(e^{jw})| \, e^{-jwn} \, du$$

La suite c(n) représente donc la suite des amplitudes fréquentielles de la fonction Log $|X(e^{jw})|$ suivant les fonctions e^{jwn} ($n \in \mathbb{Z}$)

2.2.3.2. Calcul des coefficients

Dans notre étude, les coefficients cepstraux sont calculés jusqu'à un certain ordre. Nous soulignons ici l'importance de l'analyse LPC du signal enregistré, car le calcul des coefficients de prédiction linéaire, nous a permis une obtention rapide des coefficients cepstraux ; le principe est le suivant :
On suppose avoir procédé à l'analyse LPC du signal et on considère le filtre de prédiction linéaire résultant :

$$H(z) = \frac{\alpha}{A(z)}$$ (α représente le gain du filtre)

H(z) est la fonction de transfert du filtre. On développe alors l'expression :

$$\text{Log } \left(\frac{\sqrt{\alpha}}{A(z)}\right) = \text{Log } (\sqrt{\alpha}) - \text{Log } (A(z))$$

$$= \text{Log } (\sqrt{\alpha}) - \sum_{k=1}^{\infty} c(k) z^{-k}$$

Par différenciation par rapport à $\frac{1}{z}$, puis par multiplication par z

$$-n\ c(n) \quad -n_{an} = \sum_{k=1}^{n-1} (n-k)\ c(n-k)\ a_k \qquad (3) \qquad \text{pour } n > 1$$

Décomposons maintenant :

$$\text{Log } \frac{\alpha}{|A(z)|^2} = \text{Log } \alpha - \text{Log } (A(z)) - \text{Log } (A(z^{-1}) = \sum_{k=-\infty}^{+\infty} c(k) z^{-k}$$

en posant $c_o = \text{Log } (\alpha) \quad c_1 = - a_1$

$c_n = c(-n)$

On obtient la formule globale :

$$\text{Log } \left(\frac{\alpha}{|A(e^{jw})|^2}\right) = \sum_{k=-\infty}^{k=+\infty} c(k) e^{jwk}$$

Pratiquement : Pour une fenêtre d'échantillonnage on connaît les coefficients LPC (ak, k = 1, p) à l'ordre p. Le calcul des coefficients cepstraux se déduit de (c_0 = Log (α), c_1 = - a_1) et de la formule de récurrence (3) (figure 2.2.4.).

```
C...            INTERACTIVE LABORATORY SYSTEM
C...
C...            ILS SUBROUTINE  ** A2CP **
C...
C...            COPYRIGHT - OCTOBER 1981
C...            SIGNAL TECHNOLOGY, INC.
C...            SANTA BARBARA, CA.
C...
      SUBROUTINE A2CP(A,C,M,NC)
C...
C...  SUBROUTINE FOR EFFICIENT COMPUTATION
C...  OF CEPSTRAL COEFFS
C...
C...
C...  A       AUTOREGRESSIVE COEFFICIENTS (INPUT)
C...  C       CEPSTRAL COEFFICIENTS (OUTPUT)
C...  M       ORDER OF AUTOREGRESSIVE POLYNOMIAL (INPUT)
C...  NC      ORDER OF CEPSTRAL POLYNOMIAL (MAY BE.GT.M)(INPUT)
C...
      DIMENSION A(2),C(2)
C...
C...  INITIAL CONDITIONS
C...
      C(1)=0.
      C(2)=-A(2)
C...
C...  COMPUTE FIRST M TERMS
C...
      DO 120 I=2,M
      IP=I+1
      AP=FLOAT(I)
      SUMA=AP*A(IP)
      DO 110 J=2,I
      JJ=I-J+2
110   SUMA=SUMA+A(J)*C(JJ)
120   C(IP)=-SUMA
C...
C...  COMPUTE TERMS FROM M+1 OUT TO NC
C...
      MP1=M+1
      NP=NC+1
      IF(NC.LE.M) GO TO 150
      DO 140 I=MP1,NC
      IP=I+1
      SUMA=0.
      DO 130 J=2,MP1
      JJ=I-J+2
130   SUMA=SUMA+A(J)*C(JJ)
140   C(IP)=-SUMA
150   DO 160 J=3,NP
160   C(J)=C(J)/FLOAT(J-1)
      RETURN
      END
```

Figure 2.2.4. - Programme ILS-A2CP

Calcul des coefficients cepstraux en fonction des coefficients LPC

Stockage : Les coefficients cepstraux sont stockés dans des fichiers type record ILS (figure 2.2.5.). L'ordre des coefficients est choisi égal à 10 (égal à l'ordre des coefficients LPC).

Ces coefficients nous sont utiles pour le calcul de distance (distance cepstrale). L'ordre de calcul (égal à celui des coefficients LPC) assure le critère de positivité de la distance (voir chapitre distance).

```
**************************************************************************
* DUA1:[DELIA.IDATA]WD3001.                    3001,       80  RECORDS *
**************************************************************************

------------------------------------------------------------------------
  REC    STFR     NFR    SC    CTX    EX    FIELD-1    FIELD-2     FIELD-6
   10      60      80     1    100  8000    <

   SOURCE:DATA              N     M  PR  NP  W     SF    FIELD-4    FLD3  FLD5
DUA1:[DELIA.IDATA]WD10    100    10  93 100  Y      0                  0     0

  5.9904E-01    -2.4108E-01    -4.2464E-01    -1.9649E-01     8.5720E-02
 -4.6285E-01     1.6811E-01    -3.1236E-01    -1.7006E-01     1.1206E-01
  1.5800E+02     4.7344E-01
```

Figure 2.2.5. - Fichiers type record ILS - Stockage des coefficients cepstraux

Calcul à l'ordre 10 - Les numéros 11 et 12 correspondent à des informations supplémentaires.

2.3. DISTANCES UTILISEES EN TRAITEMENT DU SIGNAL

2.3.1. Introduction

Pour mesurer la ressemblance entre deux échantillons de parole, il faut en premier lieu se situer dans un espace de mesure bien adapté au problème (choix des paramètres numériques). Ensuite, il est indispensable de choisir une métrique dans cet espace. Dans la plupart des cas, en effet, la mesure de similarité entre deux signaux s'effectue sur un certain nombre d'informations de nature numérique et extraites du signal digitalisé. D'autres distances plus directes utilisent les filtres de prédiction linéaire qui minimisent l'énergie résiduelle (rapport de vraisemblance). Enfin, il existe une méthode plus classique qui fait appel à la norme L_2 dans l'espace caractérisé par les "Log" des modules du spectre (mesure spectrale). Nous nous pencherons plus particulièrement sur cette mesure même si elle est difficilement calculable sur ordinateur (à cause de l'intégrale). On verra, par ailleurs, qu'il existe des moyens astucieux de l'approximer de manière satisfaisante.

Avant tout, voici les principes de base vérifiés par toute distance en traitement du signal :

1/ $d(x,y) = d(y,x)$
2/ $d(x,x) = 0$
3/ $d(x,y) > 0$ pour $x \neq y$
4/ $d(x,y) \leqslant d(x,z) + d(z,y)$
5/ $D(x,y)$ doit être physiquement interprétable et son évaluation doit être facile numériquement.

Dans la suite, on appellera "test" le signal inconnu (ou la portion de signal) auquel il faudra associer la "meilleure référence" par le calcul de leur distance (d(test, référence)).

2.3.2. Mesure spectrale

Soit deux modèles spectraux $\frac{\sigma}{A(z)}$ et $\frac{\sigma'}{A'(z)}$, la différence entre ces deux modèles peut être exprimée par :

$$V(w) = \text{Log}\ [\sigma^2 \oint |A(z)|^2\ ] - \text{Log}\ [\sigma'^2 \oint |A'(z)|^2\ ] \qquad (1)$$

avec $z = e^{jw}$.

On définit alors la norme L_p :

$$L_p(V(w)) = d_p\ (S,S') = \int_{-\Pi}^{\Pi} |V(w)|^p\, dw \qquad (2)$$

Pour p = 1, on obtient la moyenne de V(w) sur $[-\Pi, \Pi]$. Pour p = 2, d_2 exprime la moyenne quadratique de V(w) sur le même intervalle.

Cette mesure, pour tout p positif vérifie les relations mathématiques énoncées au chapitre précédent.

Bien sûr, l'augmentation de l'ordre de p, favorise les variations les plus grandes des deux spectres à comparer (figure 2.3.1.).
La distance d_p lorsque p tend vers l'infini est une mesure des plus grandes différences observées (figure 1)

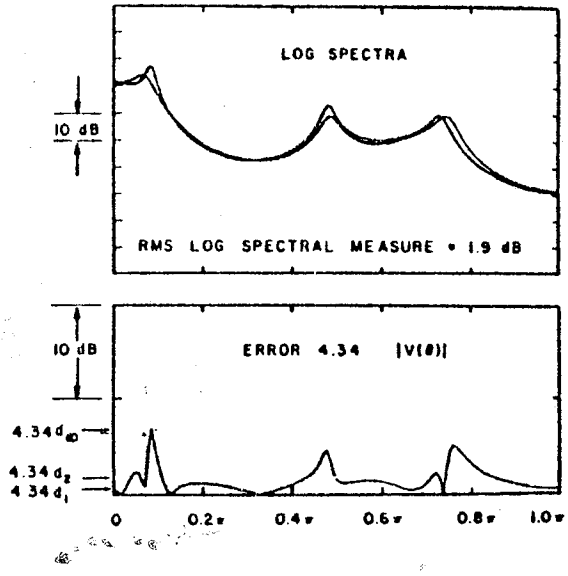


figure 2.3.1. d'après [Markel, Gray]

A première vue, il n'existe pas de relations linéaires évidentes entre les distances d_p lorsque p varie. La figure 2.3.2. nous donne une mesure de la corrélation entre d_2 et d_8.

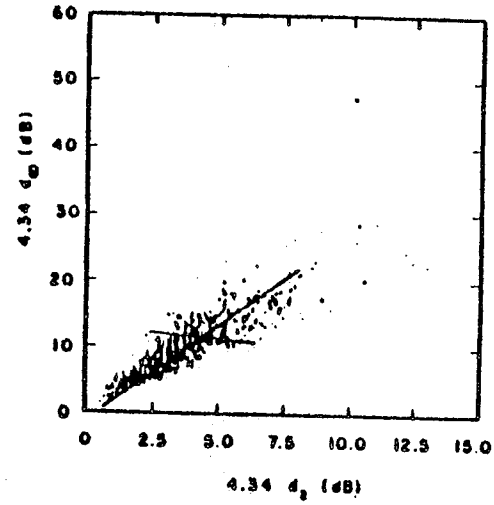


Figure 2.3.2. d'après [Markel, Gray]
Corrélation entre les deux distances d_2 et d_8.

2.3.3. Une bonne approximation de la distance spectrale : la distance cepstrale

D'après la définition des coefficients cepstraux (voir chapitre s'y rapportant) on peut écrire :

$$\text{Log}\,[A(z)] = -\sum_{k=0}^{k=+\infty} C_k\, z^{-k}$$

$$\text{Log}[\sigma^2 / A(z)] = \sum_{k=-\infty}^{k=+\infty} C_k\, z^{-k}$$

avec $C_0 = 2\text{Log}\,(\sigma)$ et $C_k = C_{-k}$

D'après (1) :

$$|V(w)|^2 = \sum_{k=-\infty}^{k=+\infty} (C_k - C'_k)^2 + \sum_{k_1=-\infty}^{k_1=+\infty} (C_{k_1} - C'_{k_1}) + \sum_{\substack{k_2=-\infty \\ k_2 \neq k_1}}^{k_2=+\infty} (C_{k_2} - C'_{k_2})\, e^{jw(k_2-k_1)}$$

D'après (2) :

$$d_2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_{k=-\infty}^{k=+\infty} (C_k - C'_k)^2 + \sum_{k_1=-\infty}^{k_1=+\infty} \sum_{\substack{k_2=-\infty \\ k_2 \neq k_1}}^{k_2=+\infty} (C_{k_1} - C'_{k_1})(C_{k_2} - C'_{k_2}) e^{jw(k_2-k_1)} \right]$$

d'où :

$$d_2 = \sum_{k=-\infty}^{k=+\infty} (C_k - C'_k)^2$$

Si on tronque la série à l'ordre L :

$$d_2 = \sum_{k=-L}^{k=+L} (C_k - C'_k)^2 = 2 \sum_{k=0}^{k=L} (C_k - C'_k)^2$$

avec $C_0 = \text{Log}(\sigma)$

L = ordre des coefficients cepstraux

En pratique

Sur le signal échantilloné, et pour une fenêtre déterminée on commencera par calculer les coefficients de prédiction linéaire à l'ordre M. Le calcul des coefficients cepstraux s'en déduit par résolution d'un système linéaire. Il est nécessaire, pour conserver le critère de positivité de la distance, que L soit supérieur ou égal à M. Des mesures de corrélation [Markel, Gray] ont permis de conclure que le choix L = M suffit [figure 2.3.3.] puisqu'il donne une corrélation de 0,98 contre 0,997 pour L = 2M et 0,999 pour L = 3M.

2.3.4. Echelle de mel
Coefficients de mel
distance de mel

Principe :

L'utilisation de l'échelle de mel est fondée sur les caractéristiques de perception de l'oreille humaine (bande critique de l'oreille humaine). Un échelonnement linéaire est effectué au niveau des basses fréquences et un échelonnement logarithmique au niveau des hautes fréquences. On obtient donc une dilatation dans la zone des hautes fréquences.

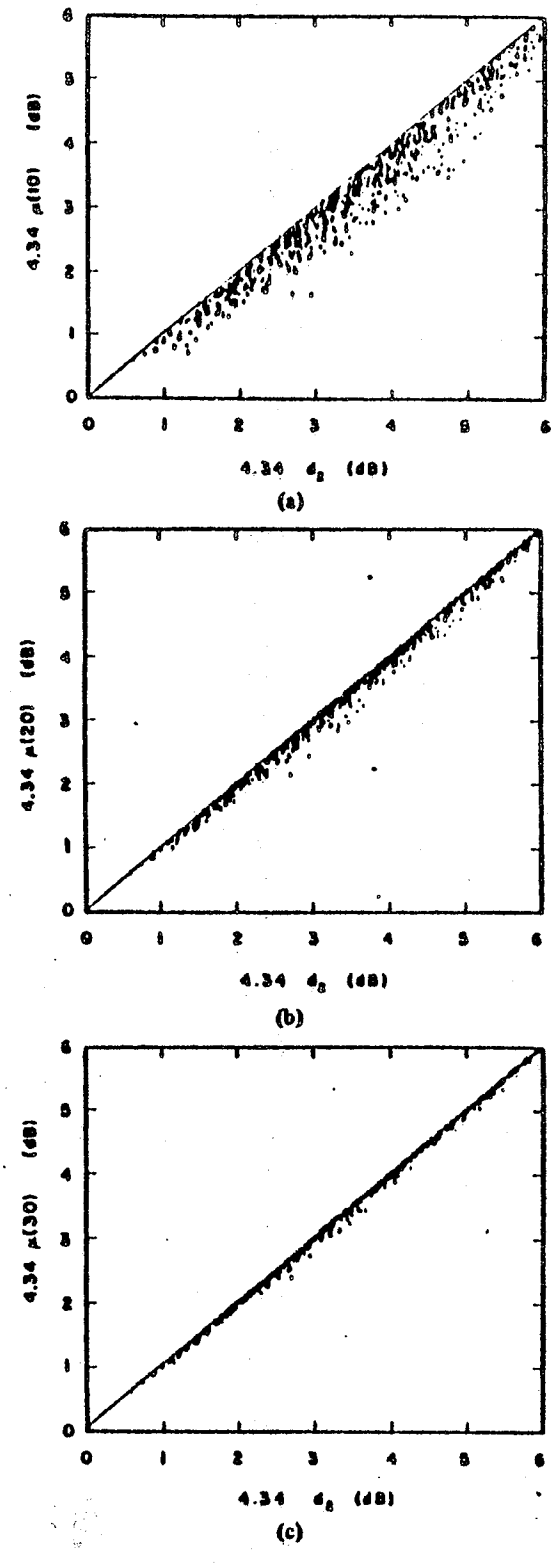


Figure 2.3.3, [d'après Markel, Gray]
Mesures de corrélation entre d_2 et d_L
(pour L = M, L = 2M, L = 3M)

En pratique, on choisit un ensemble de 20 filtres triangulaires (figure 2.3.4) Le passage du signal ou de la portion de signal à travers ces filtres particuliers permet d'obtenir une suite X_k {k=1,20} de valeurs correspondant au logarithmique de l'énergie résultant du filtre k.
Avec ce type de filtrage, on récupère 50% des valeurs entre 0 et 1000 Hz (de X_1 à X_{10}) la fréquence maximale étant de 4600 Hertz.

Coefficients de mel : Par leur définition (transformée de Fourier réelle sur le module du spectre) les coefficients de mel sont de même nature que les coefficients cepstraux. Ils sont obtenus par une projection des X_k {k=1,20} sur la partie réelle de la base $(e^{i(k-\frac{1}{2})\frac{\Pi}{20}})$ i=1,M où M est l'ordre des coefficients. Le choix des axes de projection permet de récupérer les caractéristiques du signal en tenant compte de l'échelle adoptée.
La formule devient :

$$C_i(\text{mel}) = \sum_{k=1}^{20} X_k \cos\left[i\ (k - \tfrac{1}{2})\ \frac{\Pi}{20}\right] \quad k=1,2,\ldots,\ M.$$

Distance de mel :
La distance euclidienne entre les coefficients de mel est appelée distance de mel.
De nombreux tests [STEVEN B. DAVIS ; 1980] ont été faits en ce qui concerne la reconnaissance et les résultats paraissent bons (figure 2.3.5.). Les performances obtenues sont séduisantes mais les calculs qui en découlent le sont moins (filtrage).

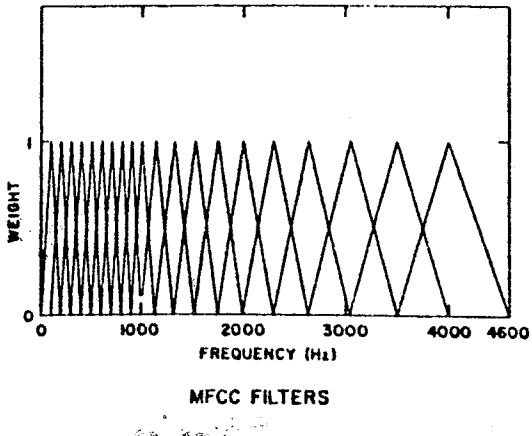
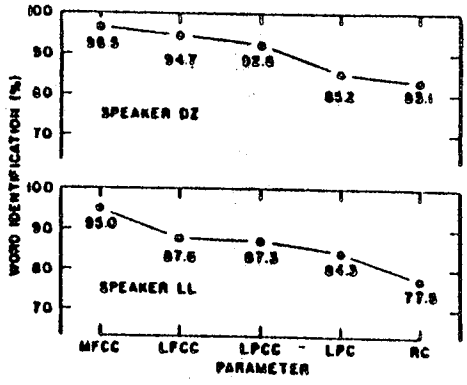


Figure 2.3.4. d'après DAVIS AND MERMELSTEIN 1980
Système permettant de générer les coefficients de mel.



Locuteur D2 = ayant participé à l'apprentissage.
Locuteur LL = n'ayant pas participé à l'apprentissage.

Figure 2.3.5. d'après DAVIS AND MERMELSTEIN 1980

Comparaison de diverses représentations paramétriques pour la reconnaissance.

MFCC coefficients de mel

LFCC coefficients cepstraux obtenus par calcul direct FFT

LPCC coefficients cepstraux après calcul LPC

RC coefficients de réflexion

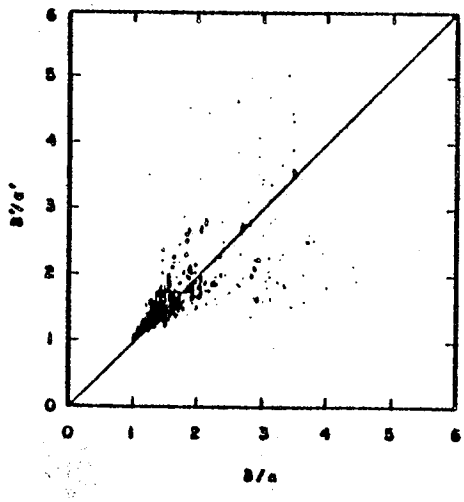


Figure 2.3.6. d'après [Markel, Gray]

Corrélation entre les deux mesures $\frac{\delta}{\alpha}$ et $\frac{\delta'}{\alpha'}$

2.3.5. Distance utilisant les filtres de prédiction linéaire

Principe : Si le signal s(t) échantilloné suivant ses valeurs s(n) est considéré comme un processus temporel autorégressif alors :

$$\hat{s}(n) = - \sum_{i=1}^{p} a_i \, s(n-i)$$

L'erreur de prédiction e(n) à l'instant n s'exprime par la différence $(s(n) - \hat{s}(n))$, $\hat{s}(n)$ étant le signal prédit. Donc :

$$e(n) = \sum_{i=0}^{p} a_i s(n-i) \text{ avec } a_0 = 1$$

Soit donc le filtre A qui minimise l'énergie α (minimisation quadratique)

$\alpha = \sum_{n=-\infty}^{n=+\infty} [e(n)]^2$. Soit A' le filtre qui minimise l'énergie α' :

$\alpha' = \sum_{n=-\infty}^{n=+\infty} [e'(n)]^2$ pour un autre signal s'(t). Alors si s(t) passe à travers A' l'erreur en résultant notée δ est :

$$\delta = \sum_{n=-\infty}^{n=+\infty} \left[\sum_{i=0}^{p} a'_i \, s(n-i) \right]^2$$

D'après le critère de minimisation on a nécessairement : $\delta > \alpha$.

$\frac{\delta}{\alpha}$ mesure donc la similarité entre les deux signaux s(t) et s'(t). C'est le rapport de vraisemblance entre s(t) et s'(t).

Le procédé inverse qui consiste à faire passer s'(t) à travers A nous donne une erreur δ' vérifiant : $\delta' > \alpha'$

$$d(s,s') = \frac{\delta}{\alpha}$$

$$d(s',s) = \frac{\delta'}{\alpha'}$$

Cette distance n'est pas symétrique. Une corrélation entre les deux est presque visible lorsqu'on effectue le tracé des couples $(\frac{\delta}{\alpha} , \frac{\delta'}{\alpha'})$.

Cette corrélation est plus forte lorsque les deux signaux à comparer sont peu différents ($\frac{\delta}{\alpha} \simeq \frac{\delta'}{\alpha'} \simeq 1$) mais elle ne peut être observée pour deux échantillons assez différents (figure 2.3.6).

Une autre corrélation est mesurée entre $\frac{\delta}{\alpha}$ et d_2. Celle-ci n'étant pas satisfaisante (figure 2.3.7) la transformation :

$\frac{\delta}{\alpha} \rightarrow \sqrt{2}\ \sqrt{\frac{\delta}{\alpha} - 1}$ donne de meilleurs résultats. (figure 2.3.8.)

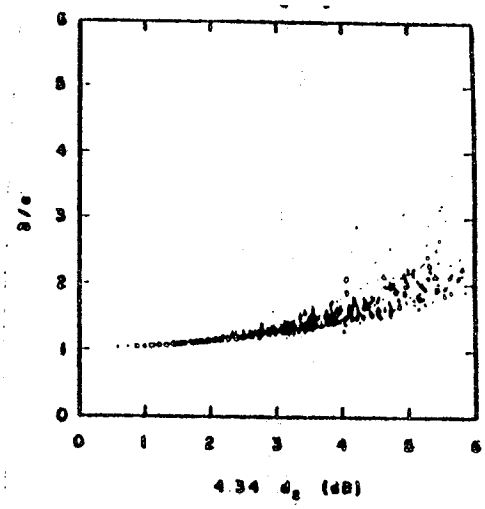


Figure 2.3.7. d'après [Markel, Gray] 1976
tracé de $(d_2, \frac{\delta}{\alpha})$

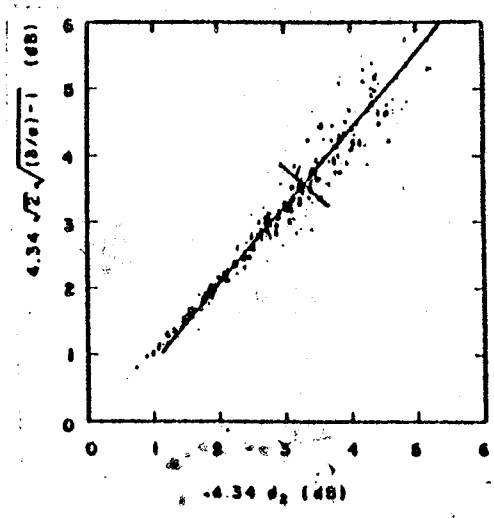


figure 2.3.8. : corrélation entre d_2 et $\sqrt{2}\ \sqrt{\frac{\delta}{\alpha} - 1}$

Distance d'Itakura :

La distance d'Itakura est dérivée du rapport de vraisemblance $\frac{\varepsilon}{\alpha}$. Son expression est :

$$d_{ITAKURA}(\text{référence, test}) = \text{Log}\ \frac{[LPC\ \hat{R}\ LPC^T]}{[\hat{LPC}\ \hat{R}\ \hat{LPC}]}$$

où : LPC = vecteur de coefficients LPC de la référence

$\hat{LPC}$ = vecteur de coefficients LPC du test

$\hat{R}$ = matrice de corrélation obtenue à partir du signal test.

2.3.6. Autres distances

- distance entre les coefficients d'autocorrélation [GUPTA, GOWDY, BRYAN] (1977)

Pour deux fenêtres d'échantillonnage n et m

$$d(n,m) = \sum_{i=1}^{k} [r_i'(n) - r_i(m)]^{\frac{1}{2}}$$

Soit deux signaux (référence, test) comprenant respectivement N et M fenêtres d'échantillonnage. La distance entre les deux signaux peut être calculée dynamiquement fenêtre par fenêtre :

$$D = \sum_{m=1}^{M} d(n(m), m) \text{ où n est une fonction de m}$$

autre possibilité en prenant :

$$d(n,m) = \text{Log}\ [\sum_{i=1}^{k} (r_i'(n) - r_i(m)^2]$$

- simple distance eudidienne entre coefficients LPC

$$d(\text{test, référence}) = \sum_{i=1}^{P} (a_i' - a_i)^2$$

2.3.7. Conclusion

L'utilisation de la distance cepstrale permet d'avoir une approximation de la distance L_2. De ce fait, il en découle une bonne interprétation physique des deux spectres à comparer. L'obtention des coefficients cepstraux à partir des coefficients LPC est très rapide. On peut donc espérer une adaptation en temps réel, condition nécessaire en reconnaissance automatique de la parole. L'ordre L égal à M suffit pour donner une bonne corrélation de d_L avec d_2 en même temps qu'il assure le critère de positivité. Les résultats issus de quelques essais [STEVEN B. DAVIS] sont encouraqeants. En règle générale, les coefficients cepstraux (ou coefficients de mel) sont beaucoup plus porteurs d'information que les coefficients LPC ou les coefficients de réflexion. Une simple distance eudidienne paraît suffire alors qu'elle ne paraît pas satisfaisante avec les coefficients LPC. [STEVEN B. DAVIS]. En ce qui concerne la distance d'Itakura, il est difficile d'interpréter les différences phonétiques résultant d'une grande distance.

Distance utilisée :

Nos essais ont été effectués avec la distance cepstrale. Une fois les coefficients cepstraux calculés, cette distance (simple distance eudidienne) ne nécessite pas le calcul d'une matrice de pondération (Itakura). Son calcul est donc rapide et la place en mémoire est peu importante. L'ordre des coefficients de prédiction linéaire est choisi égal à 10. C'est l'ordre à partir duquel la minimisation de l'erreur résultant du modèle d'autoreqression peut être observée [J. MAKHOUL]. L'ordre de calcul des coefficients cepstraux est donc aussi égal à 10 : la corrélation de la distance cepstrale avec la distance d_2 est satisfaisante (MARKEL, GRAY]. La variabilité intralocuteur est assez faible (STEVEN B. DAVIS], mais il est difficile d'assurer la variabilité inter-locuteur.

Nécessité d'une approche statistique

Un traitement statistique s'impose. En effet, la recherche de paramètres caractérisés par une faible variabilité inter-locuteur est difficile. Des essais ont pourtant été faits [Mario ROSSI, YUKIDIRO NISHINUMA, G. MERCIER]198 avec l'utilisation d'indices acoustiques fondés sur la distribution spectrale et temporelle de l'énergie, donnant de bons résultats dans le traitement des voyelles. La détermination de certains traits phonétiques (type graveaiqu, ouvert-fermé, occlusif, constrictif, nasal) a aussi été assez efficace [P. FONSALE. 1983].

Cependant, la plupart des systèmes de reconnaissance multilocuteur [LAWRENCE R. RABINER 1978] font appel à des méthodes générales de reconnaissance des formes. L'analyse des données est un outil puissant dans ce domaine. Dans le chapitre III, nous décrivons l'apprentissage phonétique du système par des méthodes statistiques.

C H A P I T R E I I I

Phase d'Apprentissage

3.1. CORPUS D'APPRENTISSAGE

Le corpus de données que nous avons utilisé a été constitué en plusieurs étapes. Au départ, nous avons enregistré un ensemble de phrases phonétiquement équilibré (au sens des proportions de chacun des phonèmes trouvées dans un discours naturel). Ensuite, une liste de mots a dû être rajoutée afin de faire apparaître des phonèmes peu fréquents.

3.1.1. Liste des phrases et mots contenus dans le corpus :

a) Liste des phrases prononcées par Locuteur 1 (Guy)

1 "On guinche ensemble demain, j'aimerais, du pain, du vin chaud, du boursin, du gateau".

2 "Tu peux causer encore, on ne peut rien faire pour toi, c'est la loi vois-tu".

3 "Cuicui, fit un oiseau épuisé tombé dans un puits"

4 "La bière est moins forte que le rhum".

5 "Ici, il fait toujours très froid en hiver".

6 "J'aime Sylvie quand elle est mignonne".

7 "Diane ne reviendra pas avant lundi".

8 "Aimez-vous ce dessin ?".

9 "J'ai déjà lu la réponse qu'il m'a envoyée par la poste ".

10 "Mes gants sont usés".

b) Mots prononcés par Locuteur 2 (Ellias) (80 mots)

abnégation
voiture
chouanerie
yoyo
chaumière
fraîcheur
la soeur
vendangeur
il geint
du raffut
vison
bovidé
un élégant
embrassa
bonze
bourbe
fondeur
rendez-vous
garenne
il ronronne
champagne
avalanche
maréchal
oblongue
lapin
le pont neuf
autobus
barbouse
rempailleux
décevoir
villégiature
pompadour
valvule
remugle
savoir-vivre
immonde
blanc bec
les limbes
chair de poule
le déluge
débridé
synode
cochon d'Inde
les vagues
les intrigues
son intime
alarme
éborgné
les algues
physiologiste
miasme
le bridge
s'incurve
tilleul
récipiendaire
juillet
la sauge
rimaille
ensommeillé
mésange
frimousse
soixante et onze
insigne
felouque
glauque
la fougue
renfourné
engouffré
revêche
mauve
il s'engonse
exégèse
exige
chemise
ratache
la bache
effiloché
embauche

c) Phrases prononcées par Locuteur 3 (Xavier) (10 phrases)

1 "Mes gants sont usés"
2 "Est-ce que le conducteur arrête l'auto ?
3 "C'est toujours comme ça depuis dix ans, tu sais"
4 "Le cheval ne peut pas marcher au pas"
5 "La bière est moins forte que le rhum"
6 "Ici, il fait toujours très froid en hiver"
7 "J'aime Sylvie quand elle est mignonne"
8 "Diane ne reviendra pas avant lundi"
9 "Aimez-vous ce dessin ?"
10 "J'ai déjà lu la réponse qu'il m'a envoyée par la poste".

d) Mots prononcés par 3 locuteurs

- locuteur 3 (Xavier)
- locuteur 4 (Yves)
- locuteur 5 (Daniel)

TONGUE
TINTIN
COUTEUX
VIN
UN VOLTAIRIEN
VONT, VIENNENT
VULGAIRE
CANCAN
ZESTE
GOND
UN VOEUX
ZINC
UN GAIN
JE VEUX
UN FEU
LE LOT
UN CODE
UN JEU
FIOLE
JE TE JAUGE
UN VOL
CREOLE
TAUTOLOGIE
VEUVE

UN BOEUF
UN GUEUX
UN COEUR
DU BEURRE
QUELQU'UN DE CHOUETTE
DES OEUVRES
D'HEURE EN HEURE
SUIVEUR
PUITS
SUISSE
HUIS CLOS

PION
PIONNIER
UN PONT
UN PAON
PINPON
POLYGONE
DES POIDS
PUSTULE
PUROTIN
PURULENCE
PUANTEUR
PROMENOIR
PONANT
PONCER
UN PUNK
POMPEUX
POUPE
PUNAISE
PUISSANTE
PUISQUE
PUINE
PONEY
PUCE
PUCHE
PUFFIN
PUBERE
TOUTOU
JE TOUILLE
TOUCHETTE
TOTEM
TOUCHANT
TOUCAN
TOSCAN
TOMME
TOAST
TETE-BECHE
OEUF COQUE
COQUIN

Cinq locuteurs ayant des accents assez différents ont donc participé à l'apprentissage. Certains (locuteur 1 et locuteur 3) ont prononcé une liste de phrases phonétiquement équilibrée . Un ensemble de mots (locuteur 2) avait été constitué pour sa variété allophonique, et pour une étude sur la synthèse de la parole. Nous nous sommes servi de cette liste car tous les mots avaient été segmentés préalablement. Des mesures statistiques sur le corpus constitué par a b c ont mis en évidence des phonèmes peu fréquents (plosives, semi-voyelles, certaines voyelles nasales etc ...). Des mots ont alors été ajoutés (liste d) afin d'effectuer un équilibrage au niveau phonétique. Cette dernière liste a été prononcée par 3 locuteurs, dont deux d'entre eux étaient de nouveaux locuteurs (locuteurs 4 et 5).

Le temps de parole enregistrée est environ de 6mn, ce qui est certainement insuffisant pour la conception d'un système de reconnaissance multilocuteur. Il est nécessaire dans ce cas de tenir compte de la variabilité inter-individuelle très importante. Pour cela, le corpus doit être assez consistant : il devrait tenir compte de toutes les variétés rencontrées. De plus, notre système n'est pas à vocabulaire fixe. Pour un système utilisant un vocabulaire déterminé, il suffit d'enregistrer les mots constituant ce dernier par une grande variété de locuteurs, et cela plusieurs fois. Bien sûr, ce type de système est moins pratique d'utilisation.

Pour résoudre notre problème, l'idée serait d'utiliser une liste de mots ou de phrases contenant tous les phonèmes dans des proportions satisfaisantes, en les faisant intervenir dans des contextes divers, de manière à résoudre le problème de variabilité engendré. Ensuite cette liste serait prononcée plusieurs fois par des locuteurs d'origines géographiques différentes.

La constitution d'un corpus pour l'apprentissage est une manipulation très longue. C'est pourquoi, nous nous sommes contentés, dans un premier temps des listes présentées ci-dessus. La stratégie de construction d'un "bon corpus" exposée dans le paragraphe précédent n'a pu être exploitée faute de temps. En revanche, nous avons mis au point un programme d'exploitation du corpus donnant un certain nombre de résultats statistiques. Cette étude fait l'objet du prochain chapitre.

3.1.2. Statistiques établies à partir du corpus

Le programme que nous avons conçu, procède à l'exploitation de tous les fichiers "label" caractérisant les mots du corpus (cf. annexe 3). En particulier ces fichiers donnent l'étiquette phonétique des phonèmes d'un mot ainsi que sa durée. Les résultats qui nous intéressaient sont de deux types : d'une part connaître exactement le nombre d'apparitions de chaque phonème dans le corpus, d'autre part obtenir certaines informations sur la durée de ces phonèmes.

La fréquence de chacun des phonèmes dans le corpus nous a servi pour la caractérisation phonétique des modèles (chapitre 3.5.). Nous rappelons ici la figure 3.5.4. du chapitre 3.5. Le premier nombre indique le pourcentage des parties stables dans le corpus, le deuxième donne le pourcentage des transitions correspondantes (transition notée A?, cf. chapitre 3.5.)

P .	0.48.	2.61
T .	0.63.	2.87
K .	0.48.	1.20
#P.	1.24.	3.66
#T.	1.29.	1.84
#K.	0.87.	1.23
B .	0.54.	0.65
D .	0.61.	2.04
G .	0.51.	0.89
F .	0.76.	1.14
S .	1.54.	2.05
C .	1.15.	0.61
V .	0.78.	1.99
Z .	0.46.	1.20
J .	0.66.	2.08
N .	0.79.	0.42
M .	0.60.	0.59
L .	0.89.	2.07
R .	0.89.	1.65
A .	0.80.	0.92
E<.	0.94.	0.93
E>.	0.78.	0.97
O<.	0.80.	0.82
O>.	0.56.	0.49
I .	0.77.	1.03
Y .	0.66.	0.53
U .	0.48.	0.41
^<.	1.00.	0.90
^>.	0.90.	0.62
^ .	0.51.	0.23
<A.	0.86.	0.87
<O.	0.72.	0.75
<^.	0.74.	3.72
<E.	0.80.	0.73
WI.	0.44.	0.45
WU.	0.20.	0.22

figure 3.1.1.

Proportions des phonèmes "A" et des transitions notées "A?" correspondantes, dans le corpus de données;

A titre d'information, nous donnons ci-dessous, le nombre d'apparitions des phonèmes dans le corpus. (figure 3.1.2.).

P	115
T	111
K	67
B	35
D	53
G	36
F	36
S	74
C	33
V	65
Z	34
J,	41
N	49
M	41
L	100
R	115
A	75
E<	64
E>	76
O<	66
O>	37
I	79
Y	60
U	45
^<	35
^>	80
^	51
<A	63
<O	46
<^	44
<E	47
WI	41
WU	24
WY	26

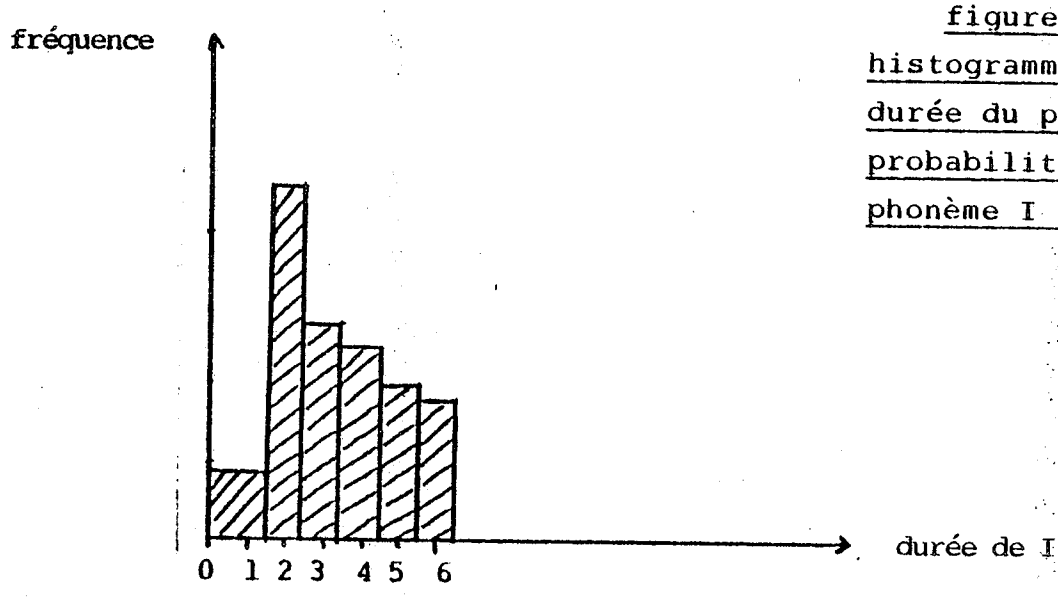
Figure 3.1.2.

Fréquences d'apparition des phonèmes
Certains phonèmes relativement peu fréquents (C par exemple) sont des phonèmes dont la durée moyenne est grande.

Les statistiques établies sur les durées des phonèmes ont été utiles en particulier pour le programme de segmentation automatique (cf. chapitres 4.1., 4.2.) pour la détermination des durées de vie. Les durées moyennes de chacun des phonèmes ont préalablement été calculées (figure 3.1.3.). Ensuite des histogrammes sur ces durées ont été tracés pour tous les phonèmes (figure 3.1.4.). Puis des probabilités exprimant la rupture ou la continuité du phonème ont été déterminées à partir de là (cf. chapitre 4.2.).

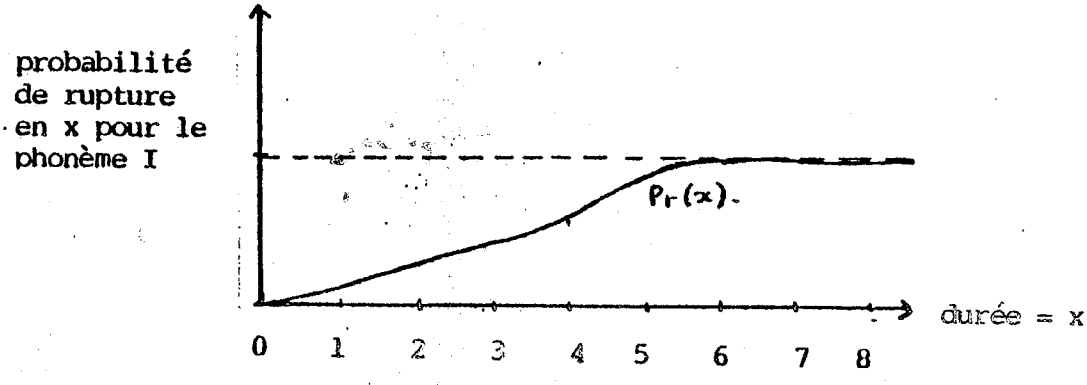
P	1.48
T	1.97
K	2.53
B	5.42
D	4.08
G	4.99
F	7.4
S	7.33
C	12.31
V	4.17
Z	4.72
J	5.64
N	5.68
M	5.15
L	3.12
R	2.70
A	3.74
E<	5.18
E>	3.57
O<	4.20
O>	5.38
I	3.42
Y	3.91
U	3.75
^<	10.04
^>	3.99
^	3.52
<A	4.81
<O	5.50
<^	5.80
<E	5.69
WI	3.58
WU	2.88
WY	1.81

figure 3.1.3.
durées moyennes



3.1.4. a

figure 3.1.4.
histogramme relatif à la durée du phonème I (3.1.4.a)
probabilité de rupture du phonème I (3.1.4.b)



3.1.3. Conclusion

Le corpus de donnée que nous avons présenté doit être segmenté de manière à mettre en évidence toutes les unités phonétiques de la langue ou, du moins, celles contenues dans le corpus. La tâche de segmentation est très longue surtout pour un système multilocuteur où la variabilité interlocuteur intervient en augmentant le nombre de représentants par phonème. C'est pourquoi, nous avons essayé de mettre au point un programme de segmentation automatique qui permettrait d'accroître considérablement et rapidement la taille du corpus. Le principe de la méthode est expliqué dans le chapitre 4.2. Cependant, dans un premier temps, la segmentation s'est faite manuellement notamment pour les listes a b et c , afin d'avoir un nombre suffisant d'unités phonétiques apprises par la machine pour faire démarrer le programme de segmentation automatique (testé sur la liste d). Les résultats obtenus sont commentés dans le chapitre 4.

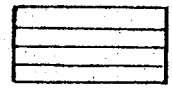
La suite du chapitre 3 est ainsi constituée :

- on définit tout d'abord les unités phonétiques pour la langue française.
- on décrit ensuite toute la procédure d'extraction des formes acoustiques représentant ces unités, à partir du corpus des données.
- enfin on associe à chacune de ces formes la liste des unités phonétiques qu'elle est susceptible de représenter.

3.2. LES PHONEMES : leur utilisation pour la reconnaissance analytique

3.2.1. Introduction :

Rappelons le principe de la reconnaissance phonétique :

signal inconnu → découpage → fichier numérisé  P T K

→ identification phonétique des segments

La reconnaissance par phonème implique donc deux grandes étapes :
- le découpage du signal digitalisé en intervalles (Fig. 3.2.1)
- l'identification de ces intervalles aux unités phonétiques de la langue.

Dans la première étape, le signal est considéré comme une suite d'évènements successifs tous de même durée en général très courte (de l'ordre du centiseconde). La seconde étape est l'étiquetage phonétique de ces segments. Elle peut se faire par une approche directe, en mesurant certaines caractéristiques propres au signal correspondant à tel ou tel phonème par application d'un ensemble de règles (indices et traits par exemple). Elle peut aussi résulter d'un traitement statistique élaboré, à partir d'un corpus de données correctement construit. C'est alors au niveau de la décision statistique que s'effectue l'identification phonétique.

En reconnaissance phonétique, il faut choisir, en rapport avec ses propres besoins, la liste de unités phonétiques de la langue. Nous exposons ici celles de la langue française et nous expliquerons l'utilisation d'unités phonétiques définies plus globalement. En effet, certaines nuances résultant de phénomènes articulatoires ("a" antérieur et postérieur par exemple) ou contextuels (cas du "r"), utiles pour les phonéticiens, ne sont pas nécéssaires dans notre étude.

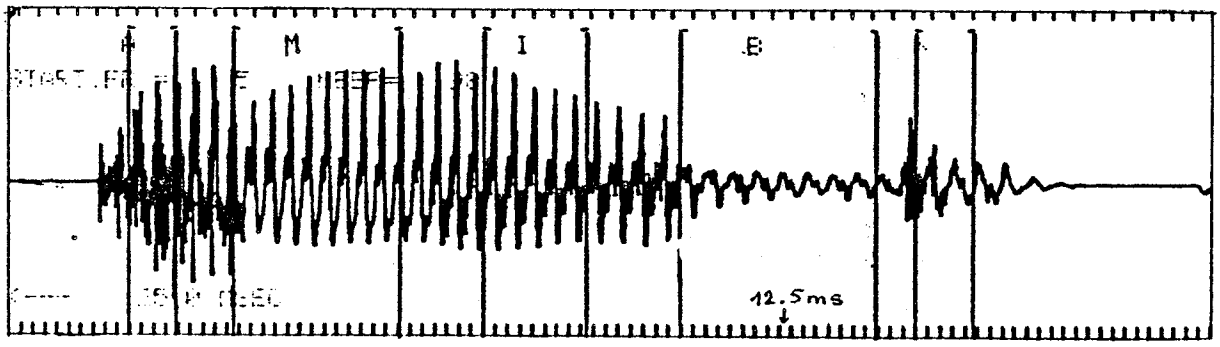


Figure 3.2.1. - découpage du signal ("amibe")
intervalles de 12.5 ms

Quelle définition donner du mot "phonème" ? Pour J.S. LIENARD (1972), c'est "la plus petite unité capable de changer un mot en un autre". Bien sûr, tout être doué de facultés d'audition normale distingue le mot "table" du mot "cable". Cette définition nous paraît satisfaisante si l'on s'intéresse au problème de décodage phonétique par le récepteur humain. Elle peut donc être comprise dans le domaine perceptif. Cependant, l'unité phonétique est plus difficile à définir lorsqu'il s'agit de décodage automatique de parole sur ordinateur. Déterminer les parties stables d'un signal est un problème délicat. Ce dernier est lié au choix d'une méthode d'analyse souvent difficile (mesure de paramètres discriminants, stratégie d'identification) La définition de J.S. LIENARD est cependant à la base de tout système de reconnaissance. En effet, la différence entre les deux mots cités doit aussi être apparente dans la phase de reconnaissance et cela, avec un taux de discrimination suffisamment grand.

Quoiqu'il en soit, la reconnaissance phonétique nécessite la connaissance des phonèmes de la langue. La liste que nous décrivons doit être considérée comme un outil dans notre chaîne de traitement. Son utilisation joue un rôle dans l'interface homme-machine et par là tout mot décodé par l'ordinateur peut être compris par l'homme grâce à la transcription phonétique de ce mot.

3.2.2. Transcription phonétique d'un mot :

On utilise 34 symboles pour désigner les sons, divisés en deux grandes classes, les consonnes et les voyelles. L'alphabet phonétique constitué par la suite de ces symboles est en apparence peu différent de l'alphabet courant mais inversement, il sert à désigner les sons élémentaires de la langue et non les éléments nécessaires pour constituer ces sons. Ces symboles phonétiques sont appelés phonèmes. Chacun d'eux correspond à un élément de l'alphabet phonétique international (Fig. 3.2.2).

Phoneme

consonnes	*	voyelles
fricatives	*	orales
F S C V Z J	*	A I Λ Λ> Λ< E< E> Y U O> O<
nasales	*	semi-voyelles
M N	*	WI WU WY
plosives	*	nasales
P T K B D G	*	<O <E <A <Λ
liquides	*	
L R	*	

< (>) ouverture(fermeture) signe postérieur
< nasalite signe antérieur

Alphabet phonétique International correspondant

consonnes	voyelles
f s ʃ v z ʒ m n p t k b d g l ʁ	a ɪ ə ø œ ɛ e y u o ɔ j w ɥ ɔ̃ ɛ̃ ɑ̃ œ̃

Figure 3.2.2. - Liste des phonèmes et correspondance avec l'Alphabet Phonétique International

Dans notre étude, d'autres éléments nécessaires seront ajoutés à cette liste standard. Ces éléments ne devront pas être considérés comme des phonèmes à part entière dans le sens phonétique du terme. Ils seront nécessaires à la segmentation des mots ; il s'agit de :

* pour la représentation du silence
*T *P *K pour les silences situés avant les plosives.

Cette liste de phonèmes pourrait paraître insuffisante si on oublie le but poursuivi : c'est-à-dire non pas produire des sons (synthèse) mais les reconnaître. C'est pourquoi certaines nuances ne sont pas prises en compte : celle de la voyelle orale a par exemple dont la réalisation est différente dans les mots "pas" et "arbre" respectivement antérieure et postérieure. De plus, pour un système multilocuteur, de nombreux paramètres (vitesse d'élocution, accent, sexe ...) diffèrent lors du passage d'un locuteur à l'autre. Tous ces paramètres apparaissent en tant que facteurs non discriminants. De plus, dire qu'un phonème correspond à une réalité acoustique bien définie n'est qu'une première approximation. De manière plus juste, on peut considérer que chacun des symboles phonétiques de notre liste est le représentant d'un ensemble de classes de même nature phonétique mais exprimant les différentes variantes rencontrées (phénomènes articulatoires, perceptifs, acoustiques ... etc. Ces classes résultent d'une méthode statistique dont l'exposé fait l'objet d'un des chapitres suivants.
Notre dictionnaire de phonèmes, en donnant un répertoire des sons, doit donc être considéré comme outil servant à représenter canoniquement les classes de phonèmes.

Voici des exemples de transcription phonétique de mots :

TABLE	*TTABL(Λ)
CHIEN	CWI<E
ROQUEFORT	RO *KK(Λ)FO R
ODIEUX	O>DWI
LUI	LWYI
LOI	LWUA

.
.
etc

3.2.3. Notions de traits phonétiques

Il est difficile de définir un phonème avec rigueur à partir de phénomènes acoustiques articulatoires ou perceptifs. C'est pourquoi, les phonétitiens préfèrent travailler dans un ensemble de mesures plus universel, celui constitué par les traits phonétiques. Les traits sont une représentation de la réalité acoustique et articulatoire et par là ils jouent le rôle de classificateurs. Ils sont représentés de manière binaire. L'utilisation de 8 traits binaires (JACKOBSON, FANT et HALL) basés sur les caractéristiques acoustiques a permis de construire un système de description suffisant et jouant le rôle de classificateur. A chaque phonème correspond une chaîne binaire fondée sur les propriétés de ce phonème.
Nous donnons ici un aperçu des traits phonétiques utilisés en français :

Trait voisé/non voisé (Figure 3.2.3)

Un signal est dit voisé lorsque l'on observe une périodicité de ce signal. C'est le cas de toutes les voyelles. Pour les consonnes le voisement peut dépendre du contexte (R, L) ; il n'est en aucun cas observé pour les fricatives F S C par opposition aux fricatives voisées VZJ ou pour les plosives PTK par rapport à BDG. Le trait de non-voisement se traduit par une force d'articulation plus forte.

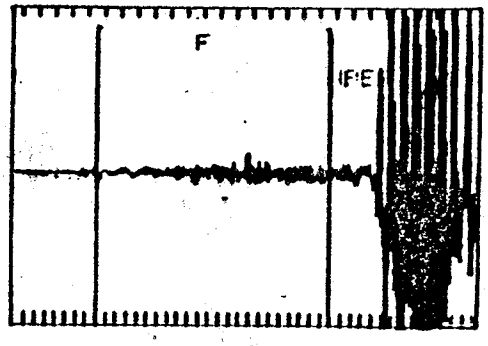


Figure 3.2.3. - Trait voisé/non voisé D/F dans "felouque"

Trait consonantique/non consonantique :
Le trait "consonantique" s'exprime par un caractère bruité (largeur de bande en un point). On observe aussi une faiblesse d'énergie globale par rapport aux voyelles adjacentes.

Trait continu/non continu :
Ce trait s'observe lorsque le débit d'air est continu c'est-à-dire sans blocage. Les plosives ont la caractéristique d'être non continues (présence d'une tenue) (Fig. 3.2.4).

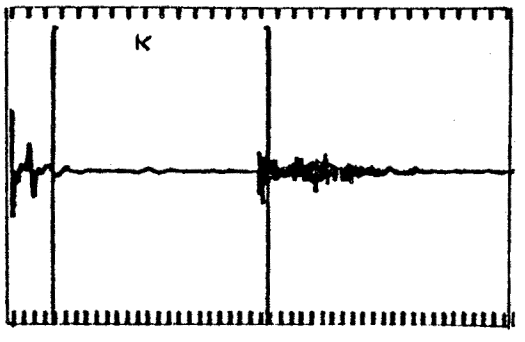


Figure 3.2.4. - non continuité du débit d'air (cas de la plosive K)

Trait interrompu/non interrompu :
Le trait "interrompu" se traduit par la présence d'une explosion La réalisation des deux traits interrompu-continu est possible dans la mesure où l'intervention d'une explosion n'implique pas l'interruption du débit d'air (nasales M, N).
C'est lors de la présence d'une tenue (non continu) et d'une explosion (P, T, K, B, D, G) que les traits non continu-interrompu pourraient être confondus (Fig. 3.2.4).

Trait vocalique/non vocalique :
Ce trait s'exprime par la présence d'une structure de formant et d'une excitation périodique à amplitude progressive. Toutes les voyelles sont vocaliques. Pour les consonnes, ce trait s'observe chez les nasales M et N (Fig. 3.2.5).

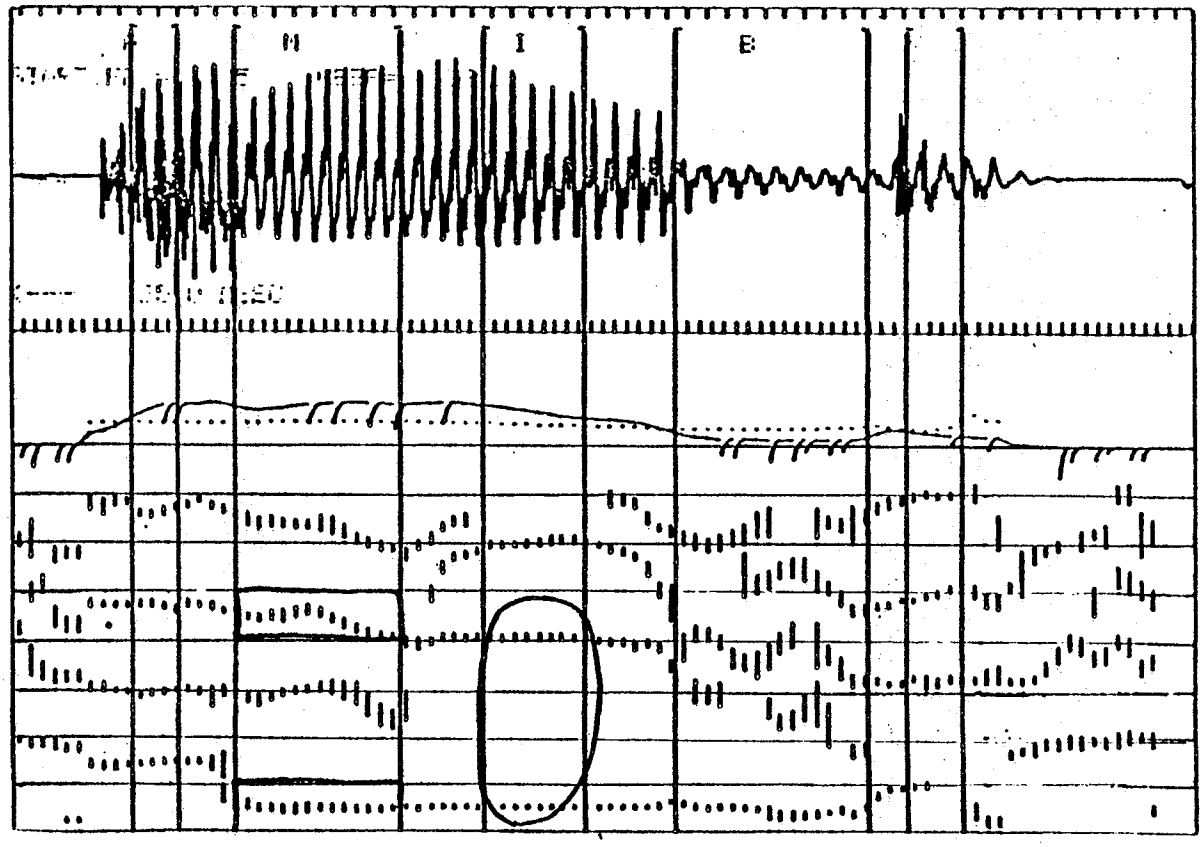


Figure 3.2.5.

- M vocalique (structure de formants)
  nasal (formants 250, 2500 H_z)
- opposition A compact/I diffus
- I aigu écartement des formants

Trait aigu/grave

Ce trait s'observe lors de la prédominance de la partie basse du spectre (F_1, F_2). Pour les voyelles F_2 1200 H_z (Fig. 3.2.6).

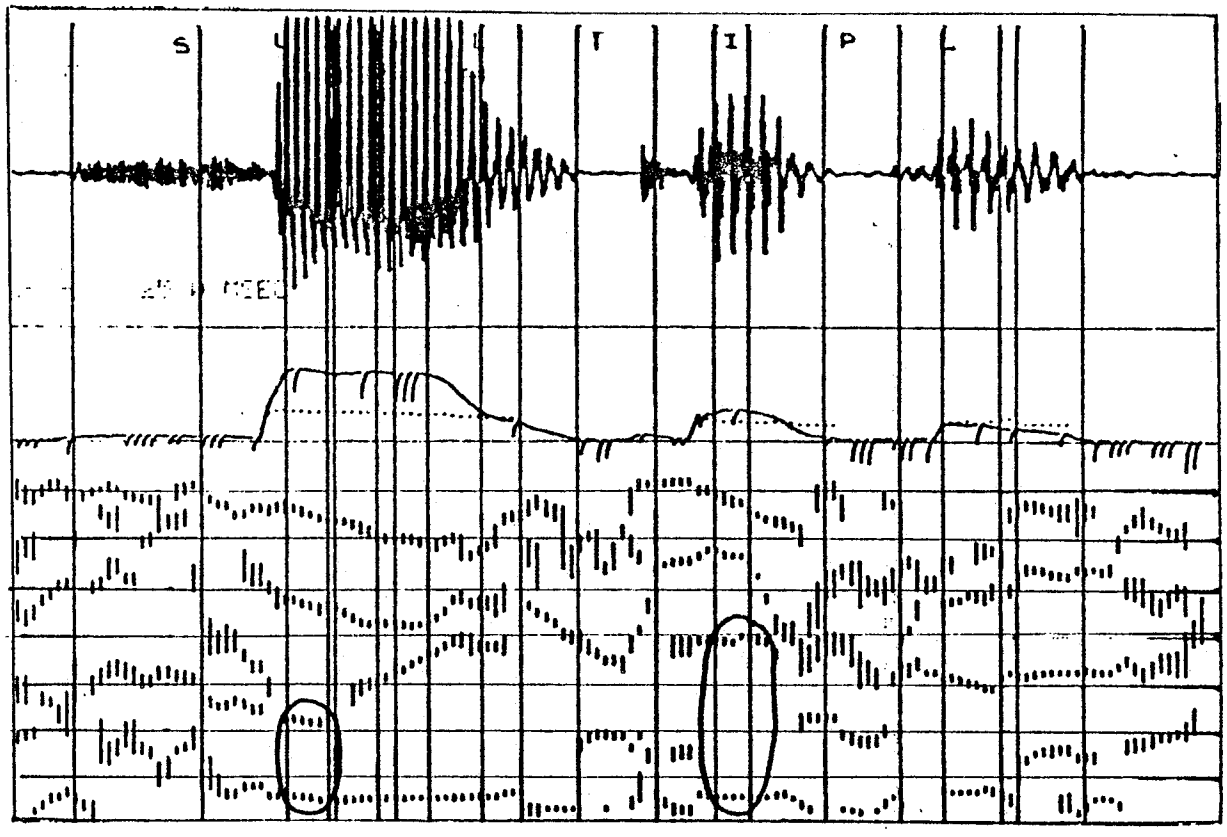


Figure 3.2.6. - Opposition grave/aigu U/I dans "sous multiple"

- Nasalité du M (formants de nasalité)

Trait bémolisé/diésé

Le trait "bémolisé" se traduit par un abaissement en énergie dans la zone du 2ème et 3ème formant. Ce trait ne concerne que les voyelles (U, U, O>, O<, Λ<, Λ>).

Trait tendu/relâché

La tension est exprimée par l'écartement des formants. Pour les voyelles le trait "tendu" est caractérisé par l'ouverture. ((O>, O<) (E>, E<) (Λ>, Λ<) ..)

Trait nasal/non nasal

Un son dit nasal est dû à la cavité nasale qui reste fixe. Pour les consonnes, cela se traduit par la présence de formants, le 1er vers 250 H_z, le 2ème vers 2500 H_z(Fig. 3.2.5). Pour les voyelles, il y a élargissement des largeurs de bande.

Trait compact/diffus (Figure 3.2.5)

Ce trait s'observe lorsque la zone centrale du spectre est remplie (2000-3500 H_z, pour les consonnes). Cette zone est caractéristique de la partie basse du spectre pour les graves, de la partie haute pour les aigus.

La figure 3.2.7 donne la table des traits phonétiques pour la liste de phonèmes utilisés. La représentation binaire 1,0 exprime la réalisation ou la non réalisation du trait. Cette table sert de référence. La définition des traits n'est pas figée et certains sont très dépendants du contexte (voisement de L, R par exemple).

3.2.4. Un aperçu de l'utilisation des traits en reconnaissance par phonème

Un ensemble de traits traduit des phénomènes acoustiques ce qui laisserait supposer qu'un phonème correspond à une réalité acoustique bien définie. Cette hypothèse a déjà été discutée. C'est pourquoi la notion de traits est difficilement exploitable pour le décodage de parole. Le problème est encore plus complexe lorsqu'il s'agit de décodage adapté à un système multilocuteur. Certes, la définition des traits éloigne les variations de type contextuel. Mais on conçoit bien que les variations dépendantes du locuteur jouent un rôle important dans la recherche des traits. Pourtant, des études sur ce sujet ont déjà été faites [P FONSALE 1983] [ROSSI NISHIMA, TREVARIN, MERCIER] [M. ESKENAZI, JS LIENARD]. Des essais [P FONSALE] ont permis de dire que les traits d'ouverture, d'acuité et de nasalité étaient des traits multilocuteurs. Mario ROSSI, suite à la critique due au binarisme, propose un système plus compliqué introduisant plusieurs niveaux fondés sur des caractéristiques acoustiques à l'intérieur d'un même trait. De même, au lieu d'utiliser une structure arborescente conduisant à la définition binaire d'un évènement à identifier, la méthode qu'il propose consiste à mesurer tous les traits . L'analyse est faite échantillon par échantillon. Les essais résultant de ces remarques ont été effectués sur les voyelles avec des taux de reconnaissance de 50 à 80 %.

	N	VC	I	CT	CP	A	V	CS	BM	TD
N	1	1	1	1	0	1	1	?	?	?
M	1	1	1	1	0	0	1	?	?	?
G	0	0	1	0	1	?	1	1	?	?
K	0	0	1	0	1	?	1	1	?	?
B	0	0	1	0	0	0	1	1	?	?
P	0	0	1	0	0	0	0	1	?	?
D	0	0	1	0	0	1	1	1	?	?
T	0	0	1	0	0	1	0	1	?	?
J	0	0	0	1	1	1	1	1	?	?
C	0	0	0	1	1	1	0	1	?	?
V	0	0	0	1	0	0	1	1	?	?
F	0	0	0	1	0	0	0	1	?	?
Z	0	0	0	1	0	1	1	1	?	?
S	0	0	0	1	0	1	0	1	?	?
L	0	1	1	1	0	1	?	1	?	?
R	0	1	?	1	1	0	1	1	?	?
WU	0	1	0	1	0	0	?	1	1	0
WI	0	1	0	1	0	1	?	1	?	?
WY	0	1	0	1	0	1	?	1	1	0
<A	1	1	0	1	1	0	1	0	0	?
<E	1	1	0	1	0	1	1	0	0	?
<O	1	1	0	1	0	0	1	0	1	?
<^	1	1	0	1	0	1	1	0	1	?
A	0	1	0	1	1	?	1	0	0	1
U	0	1	0	1	0	0	1	0	1	1
I	0	1	0	1	0	1	1	0	0	1
Y	0	1	0	1	0	1	1	0	1	1
O<	0	1	0	1	1	0	1	0	1	1
O>	0	1	0	1	1	0	1	0	1	0
E<	0	1	0	1	0	1	1	0	0	1
E>	0	1	0	1	0	1	1	0	0	0
^<	0	1	0	1	0	1	1	0	1	1
^>	0	1	0	1	0	1	1	0	1	0

N: NASAL
VC: VOCALIQUE
I: INTERROMPU
CT: CONTINU
CP: COMPACT
A: AIGU
V: VOISE
CS: CONSONANTIQUE
BM: BEMOLISE
TD: TENDU

Figure 3.2.7. Traits phonétiques

3.2.5. Utilisation des traits dans notre étude : segmentation du corpus de données en événements phonétiques

Suite à la méthode que nous avons mise au point, nous ne faisons pas une reconnaissance de phonème par identification des traits phonétiques. La connaissance des traits phonétiques utiles pour la langue française nous a cependant servi dans la phase d'apprentissage notamment dans la segmentation manuelle du corpus de données. Pour chacun des événements à identifier, l'analyse des traits (fondée sur la courbe d'énergie, la représentation spectrale (formants) ou le signal temporel) nous a permis de déterminer de manière assez fine les limites des parties stables des sons. La segmentation manuelle du corpus de donnée fait l'objet du paragraphe suivant.

3.3. APPRENTISSAGE AU NIVEAU PHONETIQUE : segmentation manuelle du corpus

3.3.1. Introduction

Pour reconnaître les unités phonétiques de la langue lors de la phase de reconnaissance, il est nécessaire de procéder à un apprentissage préalable du système. Comme nous l'avons déjà dit, cet apprentissage s'effectue à deux niveaux :

- l'apprentissage phonétique indépendant de la méthode numérique choisie : il consiste à étiqueter phonétiquement les segments contenus dans le corpus de données ;

- l'extraction d'un ensemble de références caractéristiques de ces phonèmes, à l'aide de méthodes de type statistique (traitement de l'information contenu dans le corpus de données).

Nous nous interessons ici à l'apprentissage phonétique du système. Pour segmenter les mots, nous utilisons un programme de segmentation manuelle interne à Texas Instrument France. Ce programme étiquette le signal suivant des bornes de segmentation déterminées préalablement par l'utilisateur. Il est muni d'une sortie numérique/analogique ce qui permet d'obtenir l'écoute d'une partie de signal déterminé par l'utilisateur. La précision de la segmentation se fait à l'échantillon près.

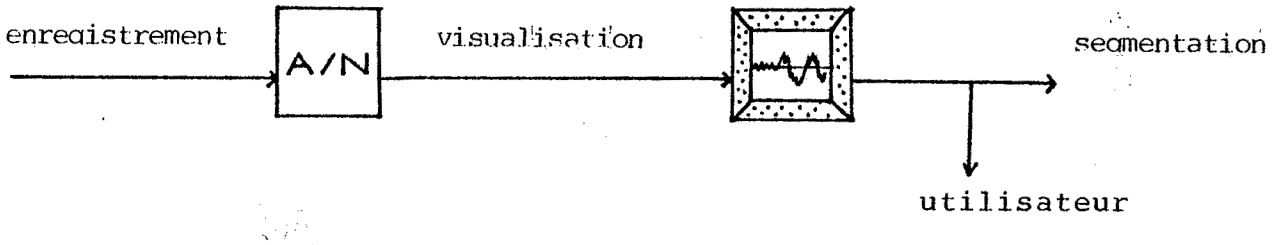


Fig. 3.3.1. - Segmentation manuelle

Le programme accepte en entrée les fichiers de type ILS (fichier échantillonné et fichier analysé) correspondant au mot à segmenter. En sortie, l'utilisateur récupère le fichier étiqueté de type ILS. Dans ce dernier, est stocké un certain nombre d'informations et en particulier les bornes des segments et le phonème correspondant (Fig. 3.3.2).

```
LB50
1000010000100          ;00;7          ;      ;**DANIEL                ;
   4799 ;    1000.; 8000;ECD300.ILSEDT.CHRIS2.WD50        ;16  MAY 1;
1001100000101          ;00;F          ;      ;**DANIEL                ;
   8050 ;     705.; 8000;ECD300.ILSEDT.CHRIS2.WD50        ;16  MAY 1;
0000000000000          ;00;H          ;      ;**DANIEL                ;
   9903 ;     601.; 8000;ECD300.ILSEDT.CHRIS2.WD50        ;16  MAY 1;
0101000101000          ;00;           ;      ;**DANIEL                ;
  10576 ;     277.; 8000;ECD300.ILSEDT.CHRIS2.WD50        ;16  MAY 1;
```

Figure 3.3.2. - Fichier de type label ILS

3.3.2. Détermination des limites des segments

La détermination des segments est basée sur :

- l'écoute donnée par la sortie N/A
- l'observation du signal temporel (t, x(t))
- le tracé de la courbe d'énergie correspondant
- la connaissance du spectre à chaque instant sur un plan de coupe (pour les sons vocaliques ; variation des formants bande de fréquence) (figure 3.3.3).

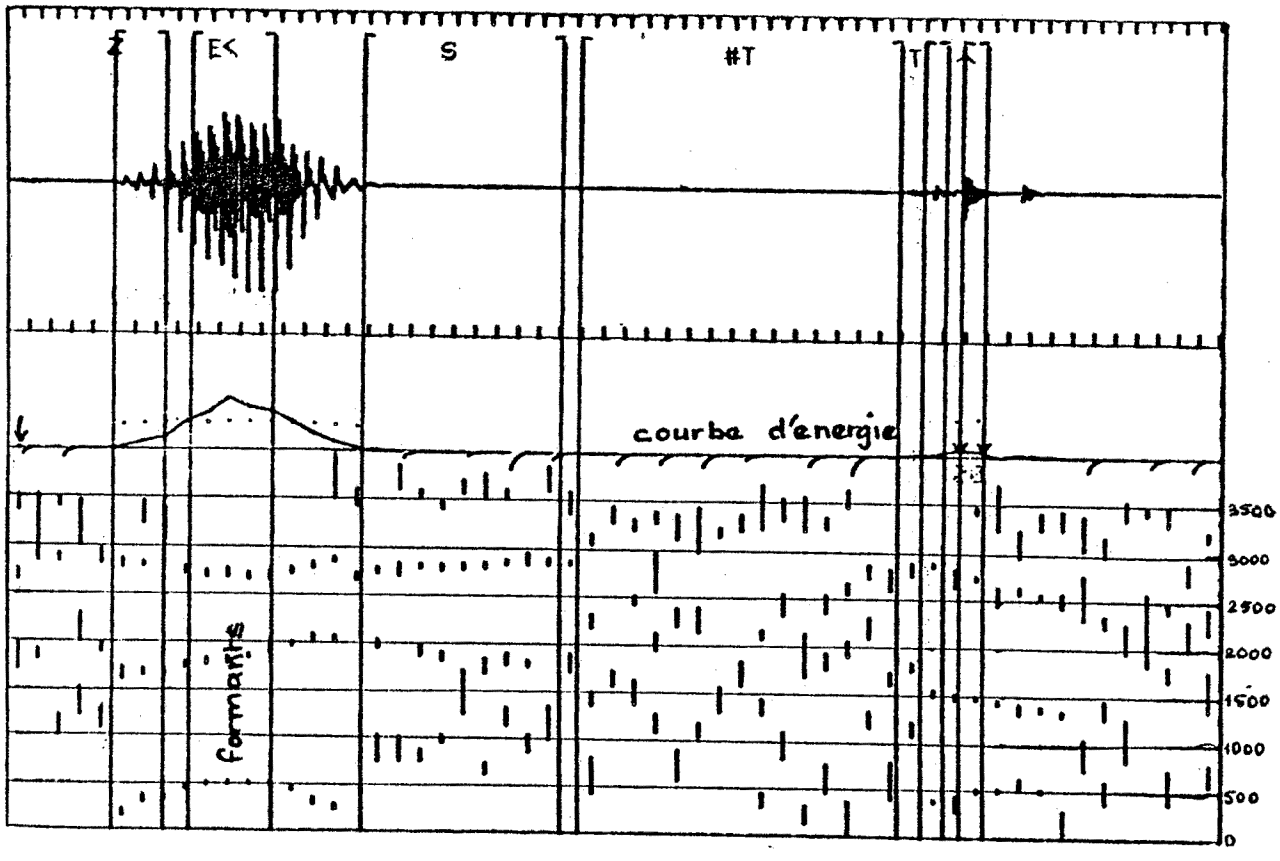


Figure 3.3.3. - Segmentation de "zeste"

Courbe d'énergie : énergie plus forte dans les zones correspondant aux voyelles (E<, A)

Formants : bandes élargies en certains points (caractère bruité) dans les zones correspondant aux consonnes (Z, S, T)

Signal : présence d'une périodicité pour les voyelles

Pour les sons vocaliques, (présentant une structure de formants) nous utilisons une table standard donnant les valeurs formantiques. Cette table donne des mesures de référence et pour chacune

d'elles il existe une zone d'incertitude (due aux variations contextuelles, individuelles ou autres) (Fig. 3.3.4).

3.3.3. Conclusion

La nécessité d'une bonne segmentation est primordiale. En effet, à partir du corpus de donnée, cette dernière permet de récolter tous les segments de parole caractéristiques de chacun des phonèmes. Il faut donc agir avec précision. Une fois la segmentation effectuée, l'utilisation de méthodes de classification appropriées, sur l'ensemble des points étiquetés, a pour but de condenser l'information recueillie (recherche de modèles). Dans le chapitre suivant nous décrivons le procédé de classification que nous avons adopté.

FORMANTS VOCALIQUES

***	F1	F2	F3	(FN1 FN2) (NASALITE)
<A	550	1000		
<E	550	<1800		
<O	550	750		
<^	550	1400		
A	750	1350		
^	500	1500	2500	
^>	500	<1500	2400	
^>	400	1600	2400	
I	250	2300-2500		
Y	250	1800		
U	250	750		
O>	375	800		
O<	500-550	950		
E>	350	2200	3000	
E<	350/500	1800	2500	
M				250 2500
N				250 2500
L	400	1700	2700	
R	700	1200		
WU	AS U			
WY	AS Y			
WI	AS I			

Figure 3.3.4. - Valeurs des formants pour les sons vocaliques

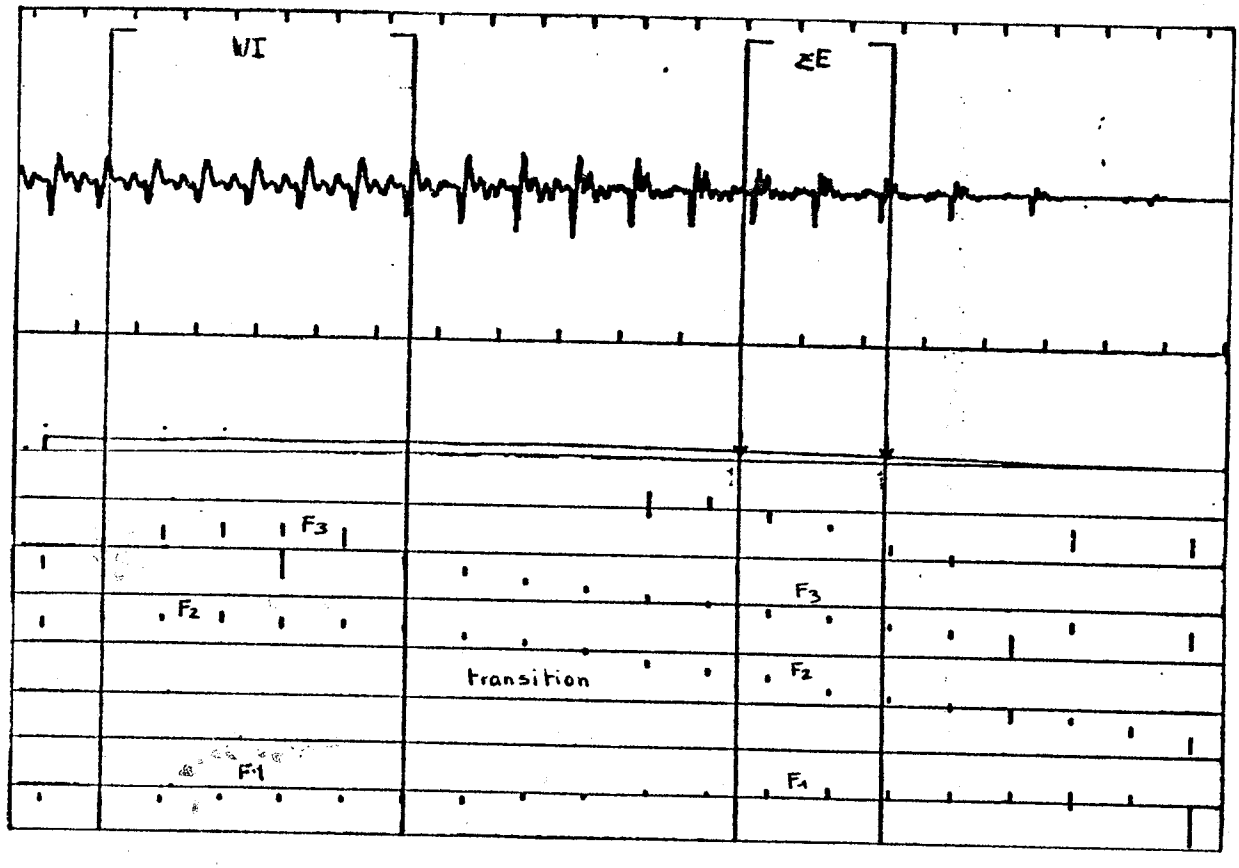


Figure 3.3.4. - Utilisation des formants pour la segmentation WI/<E dans le mot "voltairien"

3.4. CLASSIFICATION DU CORPUS DE DONNEES ET CONSTITUTION D'UNE BASE DE REFERENCES PHONETIQUES ("MODELES")

3.4.1. Introduction

On rappelle les trois principaux aspects concernant toute approche statistique pour la reconnaissance de parole (Fig. 3.3.1) :

1. - En premier lieu, sélectionner les paramètres ;
2. - Dans un deuxième temps, choisir une mesure de ressemblance adéquate ;
3. - Ensuite mettre au point une méthode pour la construction d'une base de références phonétiques.

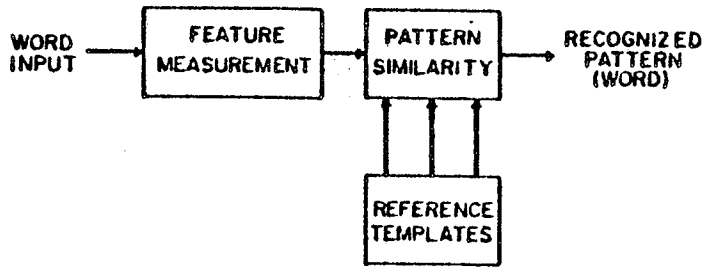


Figure 3.4.1. - d'après [RABINER] (1978)
schéma de reconnaissance d'un mot

Pour la conception d'un système multilocuteur le troisième aspect est des plus importants. La différence avec un système dépendant du locuteur est que de nouveaux phénomènes de variabilité interviennent lorsque l'on change de locuteur. Les variations intra-locuteur influent peu lorsque les unités à reconnaître (que ce soit des mots ou des phonèmes) sont répétées plusieurs fois par le même locuteur et dans des contextes variés. Mais la variabilité inter-locuteur agit de manière prépondérante dans la classification. On peut citer les essais effectués par Laurence L. RABINEI (1978) dans le cas où les unités apprises sont des mots (reconnaissance globale par mot). Trois classes ont été obtenues lors d'une classification entre plusieurs références prononcées par différents locuteurs. Cela signifie qu'une représentation des données doit se faire au moins par 3 modèles. L'introduction de deux nouveaux locuteurs (A,B) laisse penser qu'une partition en 3 classes est insuffisante (Fig. 3.4.2). Là se pose le problème fondamental de la classification pour un système multilocuteur : à quel moment peut-on considérer que la partition est stable ? Est-ce que l'introduction de nouveaux locuteurs va faire apparaître de nouvelles classes Si cette question se pose au niveau de la classification des données,

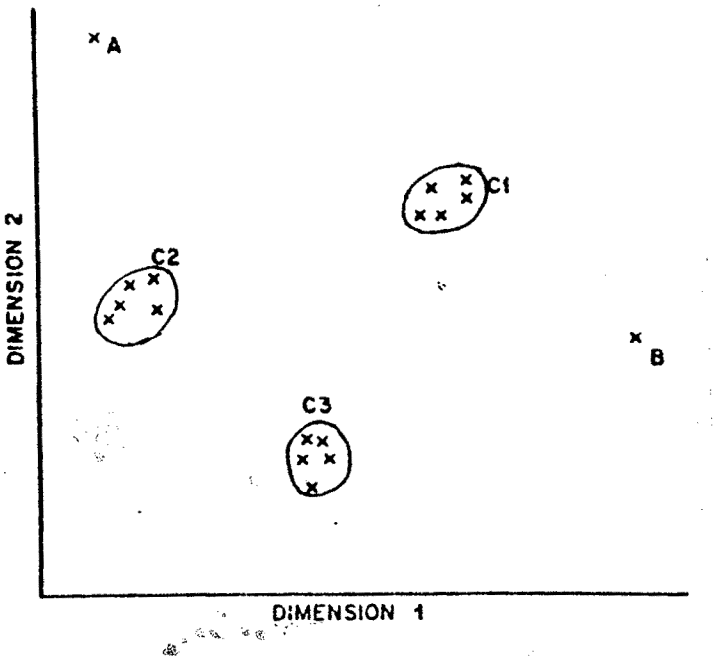


Figure 3.4.2. - d'après [RABINER] (1978)
Problème de la variabilité inter-locuteur lors de la phase d'apprentissage : 2 locuteurs n'ayant pas participé à l'apprentissa (A,B) introduisent deux nouvelles classes.

il va de soi qu'il faut essayer d'y répondreavant : choix de paramètres discriminants, recherche d'une distance. Au niveau de la classification, il s'agit d'avoir à traiter des données assez consistantes (diversité des locuteurs, des mots, des phonèmes, des contextes ... etc). Ensuite, il faut évaluer le nombre suffisant de modèles pour représenter ces données. Et puis, bien sûr rechercher une méthode de classification donnant les meilleurs résultats au niveau de la séparabilité des classes.

3.4.2. Recherche de la meilleure partition

La recherche d'une classification à l'intérieur d'un ensemble de données peut être schématisée par la figure 3.4.3. Dans un premier temps, l'acquisition des données permet d'avoir un système de description des formes à étudier. Ce système, après l'obtention des données brutes est déterminé à l'aide de paramètres aux caractéristiques choisies préalablement. Ensuite, le but est d'exploiter l'information transmise afin de trouver la meilleure partition adaptée à l'ensemble à traiter. L'important est de réduire la taille des données tout en considérant le maximum d'information.

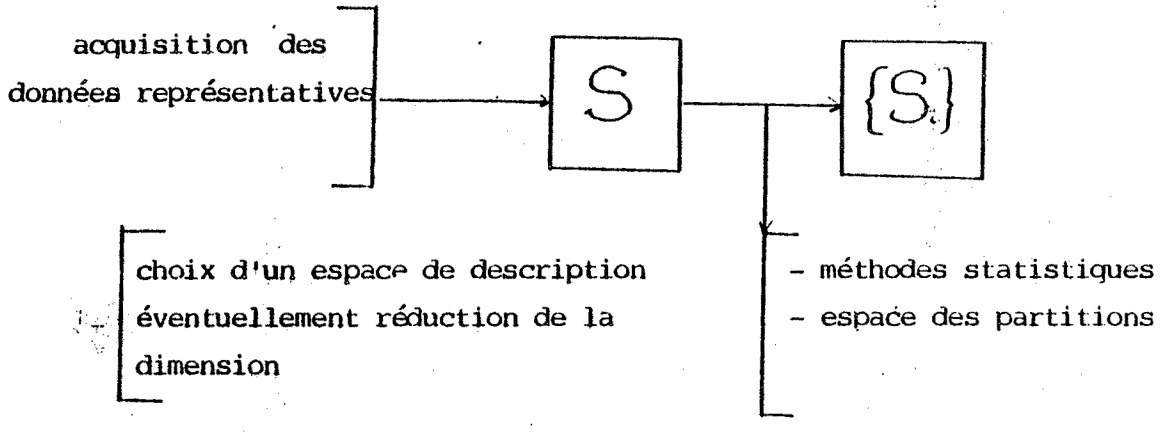
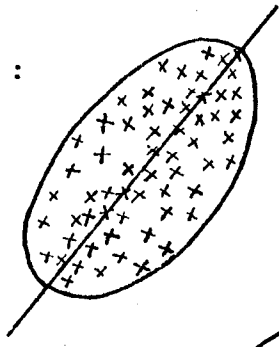


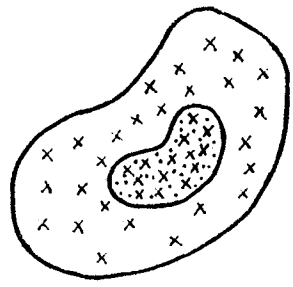
Figure 3.4.3. - Schéma de classification de données

Une fois l'espace de description déterminé, on peut en réduire la dimension (analyse en composantes principales). Mais il est clair que au fur et à mesure des étapes du traitement, la représentativité des données acquises doit être conservée dans la mesure du possible. Il en est donc de même dans la phase de séparation des formes. Pour chacune d'elles, ensuite, on peut déterminer un ensemble de représentants. Ces derniers doivent s'adapter au mieux à la forme des nuages des partitions. Ils résultent par exemple de l'optimisation d'un critère (minimisation quadratique). On les appellera "les noyaux des partitions construites. Ces noyaux peuvent être :

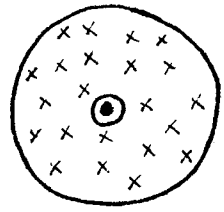
a) une droite :



b) un groupe de points :



c) un centre de gravité :



Quelles que soient les unités à reconnaître, que ce soit des phonèmes, des mots entiers ou des syllables, l'important est de procéder à un apprentissage suffisant de manière à faire connaître à la machine les modèles représentant au mieux l'ensemble de ces unités. Pour toute reconnaissance de type statistique, une bonne classification s'impose. Les méthodes proposées rejoignent souvent

celles des nuées dynamiques [DIDAY]. D'autres sont plus originales et consistent à isoler successivement les boules de plus forte densité [SUGAMA, SHIKANO, FURUI].

Dans un cadre plus général, on distingue deux grands types de méthodes : les méthodes hiérarchiques et les méthodes non hiérarchiques. Pour les premières, une analyse descendante des données permet d'obtenir à chaque itération une description des parties et sous parties intéressantes. C'est le cas, par exemple, de la méthode MNV (Mutual Neiborhood Value) fondée sur les proximités relatives. Ces méthodes nécessitent une grande place mémoire. Leur grand avantage est que l'on peut visualiser à chaque pas l'évolution des partitions. La méthode des nuées dynamiques fait partie du deuxième type de méthodes. Ce type de classification est bien adapté dans le cas d'un grand nombre de données. L'idée est la suivante : on suppose avoir à classer un ensemble de points pour lequel il existe à priori k classes. On choisit alors N représentants de cet ensemble (N > K). Si on applique la méthode des nuées dynamiques, ces représentants s'organisent en fonction des formes les plus caractéristiques et s'agrègent en leur centre. Les familles de représentants résultant de la classification se trouvent alors dans les zones de plus forte densité (Fig. 3.4.4).

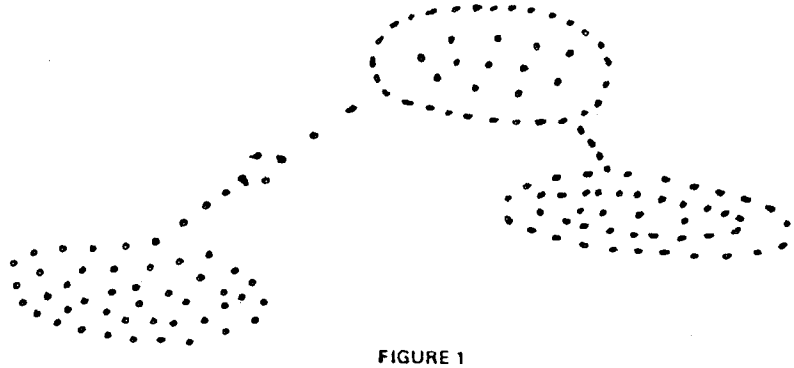


FIGURE 1

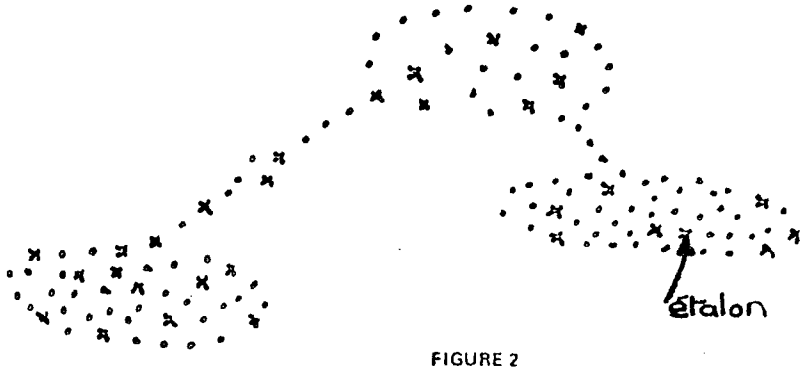


FIGURE 2

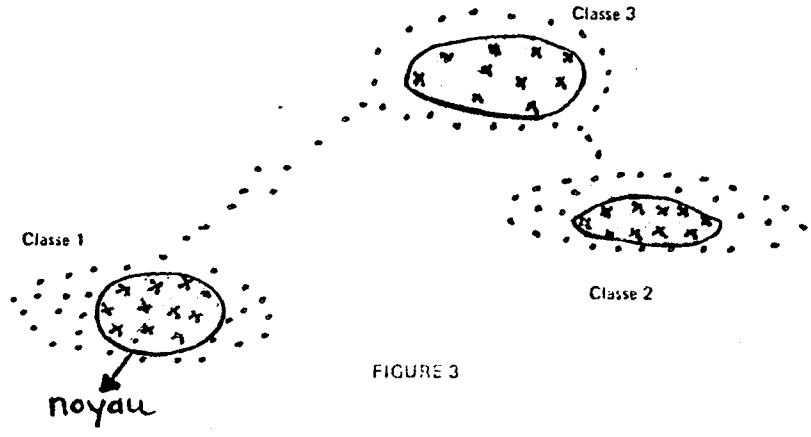


FIGURE 3

Figure 3.4.4. - d'après [DIDAY]

On choisit aléatoirement sur un ensemble de points se répartissant en trois classes, un certain nombre de représentants (étalons)(X). Ces derniers se retrouvent regroupés autour du centre des zones de plus forte densité (noyau). Les familles formées par les noyaux sont appelées nuées dynamiques.

3.4.3. Méthode des nuées dynamiques

L'avantage de la méthode des nuées dynamiques est qu'il n'est pas nécessaire d'avoir une grande place mémoire. En revanche, pour une analyse hiérarchique de N données on a besoin du tableau des distances prises deux à deux : soit une mémorisation d'un tableau de dimension $\frac{(N-1)\ N}{2}$. Le choix d'un test d'arrêt est très simple et la convergence est assez rapide (au maximum 5 itérations). La convergence étant obtenue, on définit alors l'optimalité de la solution dans un sens défini préalablement (minimisation de la distorsion globale, invariance des classes ...etc).

3.4.3.1. Description de la méthode [DIDAY]

Soit un ensemble noté E à classer et pour lequel on a défini :

- la distance "d" entre deux points :

$d : E \times E \rightarrow + \ / \ (x,y) \rightarrow d(x,y)$

- l'ensemble P des parties de E muni de la distance "D"

$$P \times P \rightarrow R^{+} \ / \ D(P_1, P_2) = \frac{1}{\text{card}(P_1) \times \text{card}(P_2)} \left[\sum_{x \in P_1} \sum_{y \in P_2} d(x,y) \right]$$

On suppose que le nombre de classes est fixé à l'avance : soit K classes.
On choisit alors dans E une famille de parties (A_i) $1 < i < k$ appelées noyaux telle que l'on ait :

$$\sum_{i=1,k} \text{card}\ (A_i) \ll \text{card}\ (E)$$

On utilise alors deux règles définies comme suit :

- En premier lieu, on détermine pour chaque élément de E le noyau le plus proche. A chaque noyau A_i $(i=1,\ldots,k)$, correspond donc une classe P_i. Cette classe est constituée des éléments x tels que : $\forall x \in P_i \ d(x,A_i) = \underset{j}{\text{Min}}\ d(x,A_j)$

- Dans un deuxième temps, k nouveaux noyaux sont calculés à partir des classes P_i (i=1,...,k) construites. Ces noyaux sont considérés comme plus représentatifs dans un sens à préciser. (par exemple, centre de gravité de la classe).

La classification automatique consiste à appliquer simultanément les deux règles. En même temps, est définie une fonction critère ω qui fournit un indice de qualité de la partition. ω s'exprime ainsi :

$$\omega = \sum_{i=1}^{k} \sum_{x \in P_i} d(x, A_i)$$

ω est décroissante. En effet :

Soit $N = (A_1, \ldots, A_K)$ $P = (P_1, \ldots, P_K)$

Par application de la première règle, on obtient $P = f(N)$.
En appliquant la deuxième, il vient $N = g(P)$.

ω est décroissante c'est-à-dire :

$\omega(N,P) > (N, f(N))$.

$\omega(N,P) > (g(P), P)$. En effet :

* $f(N) = Q = (a_1 \ldots 0_k)$

$$\omega(L,P) = \sum_{i=1}^{k} \sum_{x \in P_i} d(A_i, x)$$

$$\omega(L,Q) = \sum_{i=1}^{k} \sum_{x \in Q_i} d(A_i, x)$$

Soit $z \in Z$, $d(A_j, z) = \min_{1 \leqslant n \leqslant k} d(A_n, z) \leqslant d(A_i, z)$

avec $A_j \in f(N)$ et $A_i \in N$

donc $\omega(N,P) \geqslant \omega(N, f(N))$

* $g(P) = (B_1, B_2 \ldots B_k)$

$$\omega(N,P) = \sum_{1}^{k} \sum_{x \in P_i} d(A_i, x)$$

$$\omega(g(P), P) = \sum_{1}^{k} \sum_{x \in P_i} d(B_i, x)$$

Or : $\sum_{x \in P_i} d(B_i,x) \leqslant \sum_{x \in P_i} d(A_i,x)$.

d'où $\omega(N,P) \geqslant \omega(g(P),P)$

3.4.3.2. Discussion

Bien sûr, la partition obtenue peut différer suivant les noyaux choisis au départ. De là, apparaît la notion de formes fortes (formées des éléments qui restent dans une même classe quelle que soit la partition initiale) ou de formes faibles (parties contenant des éléments qui sont au moins une fois dans une même classe).

Comme beaucoup de méthodes, il faut effectuer des choix préalables (étalons initiaux, nombre de classes, métrique). Pour savoir dans quel esprit il faut faire ces choix il est préférable de connaître à l'avance la nature des données à traiter. Dans le cas où on ne connaît rien a priori, les étalons de départ peuvent être déterminés de manière aléatoire. Si le nombre de partitions est trop grand, on risque de trouver un certain nombre de classes vides. Il faut aussi savoir combien d'étalons prendre pour chacun des noyaux. Enfin, la convergence dépend du test d'arrêt choisi. Si on utilise la fonction ω décroissante, il suffit de se fixer un seuil "α" en deça duquel l'optimisation est suffisante.

Les variantes de la méthode proposée par E. Diday jouent sur le critère de convergence (Forgy). La convergence est obtenue lors de l'invariance des classes d'une itération à l'autre. D'autres

différent par le choix des noyaux initiaux (Mac Queen). Dans toutes ces méthodes, la détermination de nouveaux noyaux à partir des partitions formées résulte du calcul du centre de gravité (minimisation quadratique de la distance). Souvent ce centre de gravité est recalculé à chaque attribution (Mac Queen). Il est possible aussi de fusionner deux classes trop proches l'une de l'autre. (Mac Queen - "Méthode K-meano avec paramètre C et R).

Notre classification utilise la méthode des nuées dynamiques. Grâce à la segmentation du corps de donnée effectuée préalablement, il nous a été possible de connaître la nature des données à traiter l'initialisation des noyaux ne s'est pas faite de manière aléatoire pour ce qui concerne les nuages phonétiques (parties stables). Pour les transitions, cependant, il a été difficile de trouver un nombre limité de classes représentatives. Dans ce cas, la démarche est beaucoup plus classique.

3.4.4. Adaptation de la méthode à notre problème :

Etant donnée la complexité des données à traiter, il était préférable de rechercher une certaine organisation parmi elles. Les points que nous avons à classer sont des vecteurs de dimension 10 correspondant aux 10 premiers coefficients cepstraux. Comme nous l'avons vu auparavant ces points représentent une portion de signal de 12,5 ms, celle-ci ayant été analysée puis numérisée. La segmentation du corpus des données enregistrées a permis un étiquetage phonétique de chacun de ces points (figure 3.4.5.).

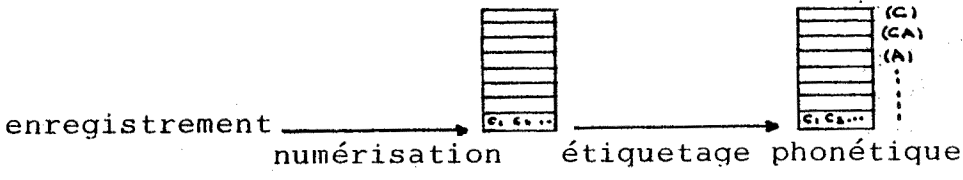


Figure 3.4.5. Phase de segmentation du corpus enregistré.

L'organisation que nous proposons est la suivante :

- A partir de l'ensemble des vecteurs constituant le corps de données, isoler ceux qui correspondent à un phonème de la liste adoptée. Ce sont les parties stables ou événements phonétiques proprement dits.
- Regrouper les vecteurs restant c'est-à-dire ceux qui correspondent aux états de passage entre deux états stables. Ce sont les transitions.

La première de ces deux opérations nous permet de construire naturellement les classes phonétiques. Si l'on se réfère à la table des unités phonétiques utilisées on effectue un pré-classement suivant 38 groupes phonétiques. En première approximation nous obtenons donc 38 partitions parmi les événements correspondant aux états stables. Mais comme nous l'avons déjà souligné, il est difficile d'associer à une unité phonétique donnée, une réalisation acoustique bien définie. Pour un système monolocuteur on parlera de variante de type contextuelle. Pour un système multilocuteur viennent s'ajouter les variantes inter-individuelles. C'est pourquoi, chacune des 38 partitions est elle-même constituée de sous-classes. Le nombre de ces sous-classes diffère suivant la sensibilité du phonème aux paramètres extérieurs (contexte, accent, ... etc). La sensibilité du phonème est plus ou moins forte suivant le phonème étudié. On remarque par exemple une grande influence des voyelles avoisinant les consomnes L et R.

Dans la deuxième opération, la recherche de classes caractéristiques parmi les transitions a été vaine. La langue française comporte à peu près 600 transitions distinctes et donc une méthode similaire à la première était impossible. La stratégie a donc consisté à ne pas différencier au départ les événements correspondant aux transitions et d'effectuer une classification globale.

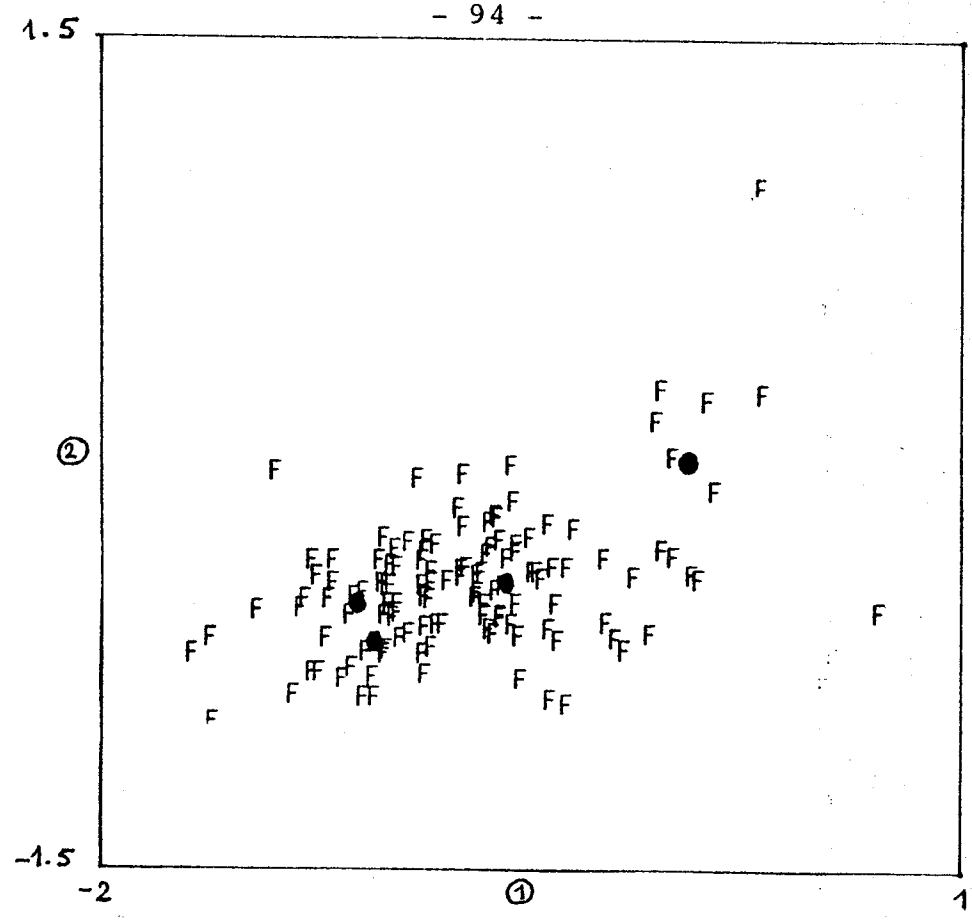
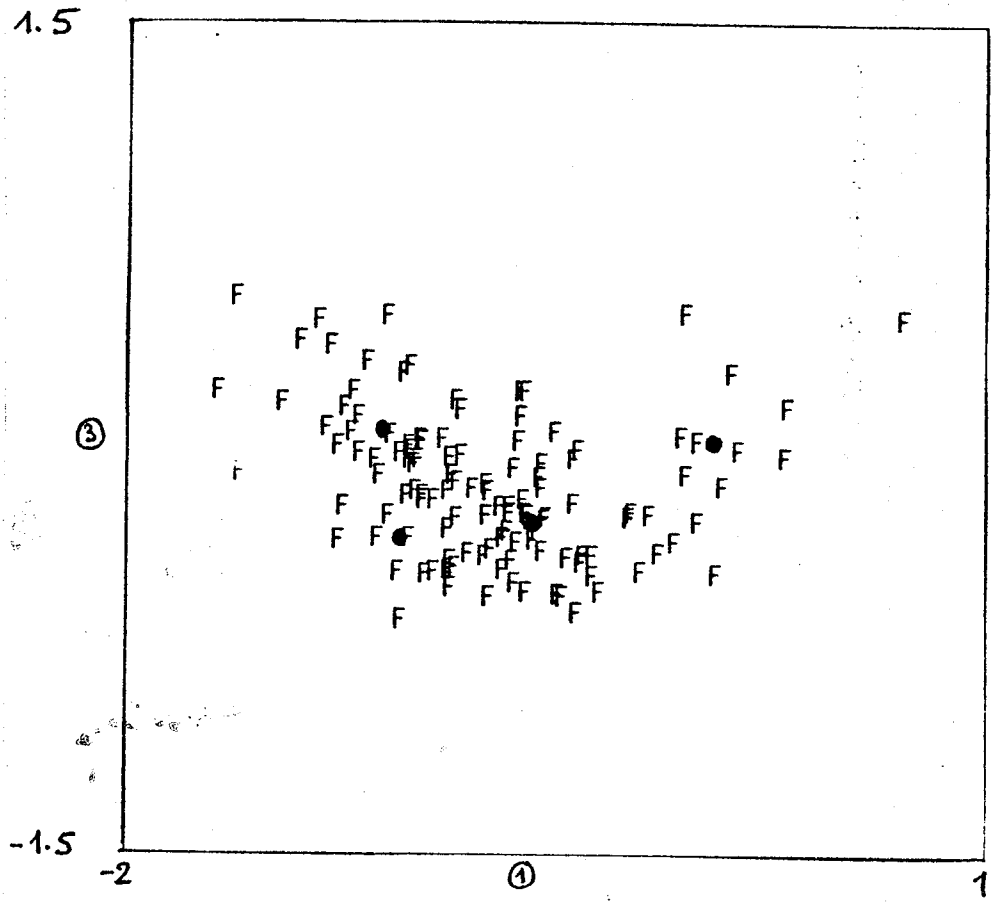


figure 3.4.6. Phonème F dans les plans cepstraux (C_1, C_2) (C_1, C_3



3.4.4.1. Classification des phonèmes

a) Recherche de noyaux à l'intérieur des partitions phonétiques :

En premier lieu, on isole tous les vecteurs correspondant à un phonème déterminé. On obtient alors des nuages phonétiques. La figure 3.4.6. nous montre l'ensemble des vecteurs représentant le phonème F, ainsi que le noyau du nuage (.).

Le problème est donc :

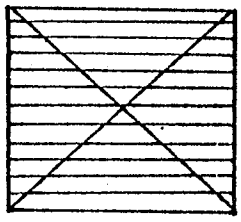
Pour chacun des nuages phonétiques :

- déterminer le nombre d'étalons suffisantspour décrire le nuage.
- initialiser et optimiser chacun des noyaux.

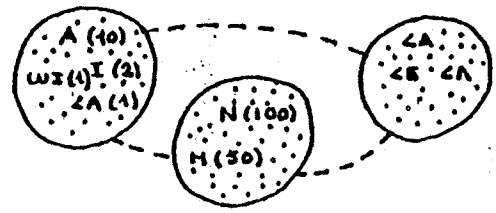
b) Effectif des noyaux :

Voici la procédure effectuée pour chacun des phonèmes :

a prendre tous les vecteurs correspondant au 38 parties stables.
b sur l'ensemble de ces vecteurs appliquer la méthode des nuées dynamiques avec initialisation des étalons (100 à 150 étalons).
c Pour chacun des phonèmes, examiner leur répartition sur chacun des étalons (histogramme des répartitions phonétiques). Pour cela, il est nécessaire préalablement d'étiqueter les classes résultant de la classification (figure 3.4.7.).

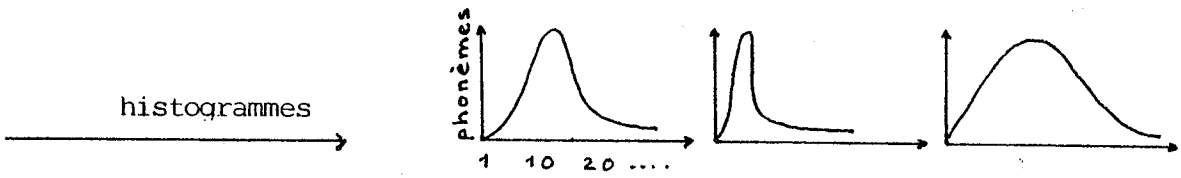


nuées dynamiques →



corpus des parties stables segmentées

100 à 150 familles résultant de la classification. Etiquetage des éléments avec fréquence d'apparition.



1 histogramme par famille

répartitions phonétiques

Figure 3.4.7.

L'effectif suffisant pour chacun des noyaux peut être déterminé à partir de cette étude. En général, on choisit de 3 à 5 étalons suivant les phonèmes. On s'assure après ce choix d'une représentation à 60% ou plus du phonème. On peut citer en exemple le cas du phonème "F" que l'on représente par 4 références (figure 3.4.6.) La figure 3.4.8. donne la répartition phonétique du "F".

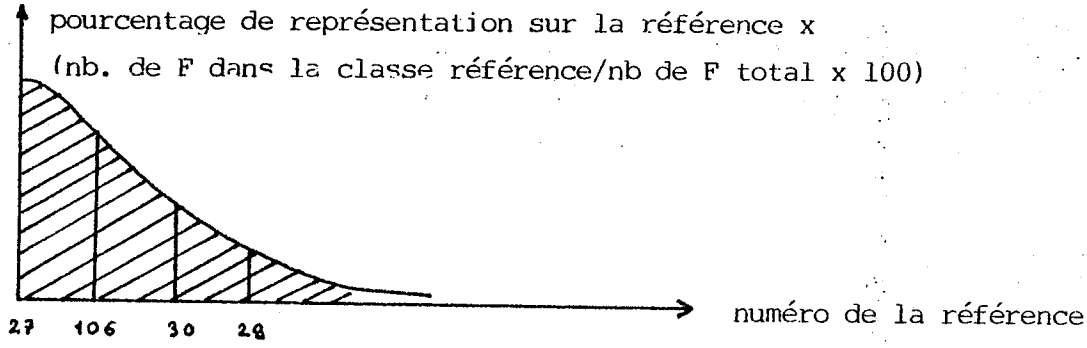


Figure 3.4.8.

c) Construction des noyaux

Pour les 38 groupes phonétiques, on a fixé le nombre des étalons constituant le noyau. L'initialisation des étalons consiste à :

- calculer le centre de gravité du nuage phonétique.
- compléter en déterminant des points supplémentaires à l'aide de l'algorithme décrit par la figure 3.4.9.

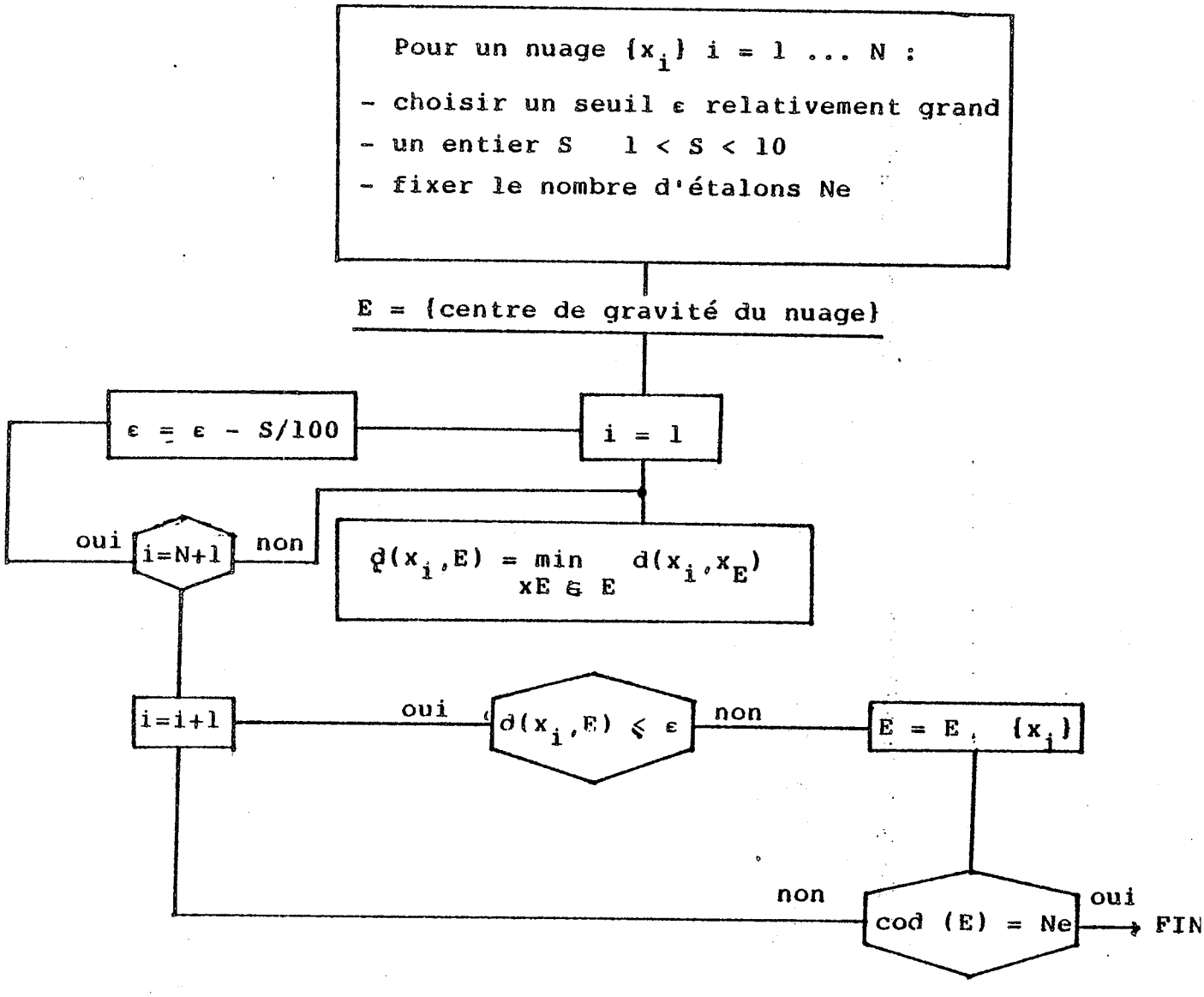


figure 3.4.9.

Initialisation des étalons

L'ensemble des étalons définitifs est construit par la méthode des nuées dynamiques. Il est alors visible que le centre de gravité est presque inchangé puisqu'il correspond déjà à la minimisation quadratique des distances du nuage à ce point. En revanche, on constate un net déplacement des étalons supplémentaires résultant de l'initialisation. (figure 3.4.10.)

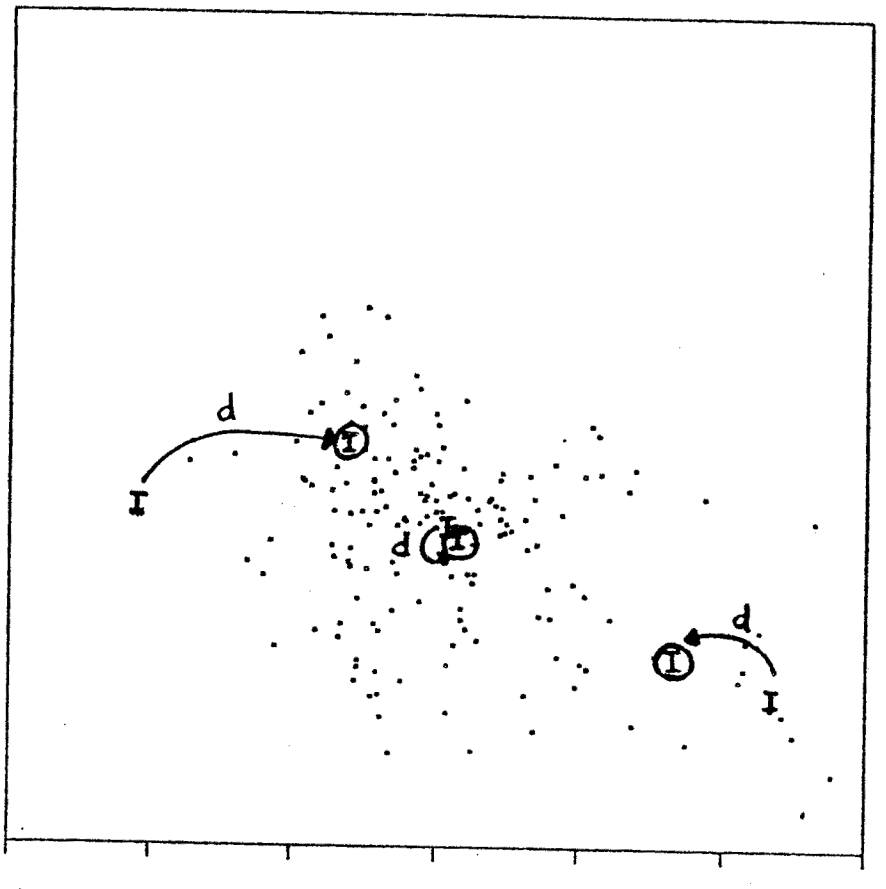


figure 3.4.10. Nuage du phonème I

On a choisi un nombre de 3 étalons :
- initialisés ("I")
- puis optimisés (" I ").

Le critère de convergence de notre algorithme est fondé sur la minimisation de la fonction ω. L'entrée du programme accepte donc un seuil de distance à partir duquel l'optimisation est supposée être suffisante. Cependant, pour éviter un trop grand nombre d'itérations pour une décroissance relative faible de ω, le programme s'arrête lorsque la variation de la distance ω est négligeable. Cela se résume par le schéma suivant : (figure 3.4.11.)

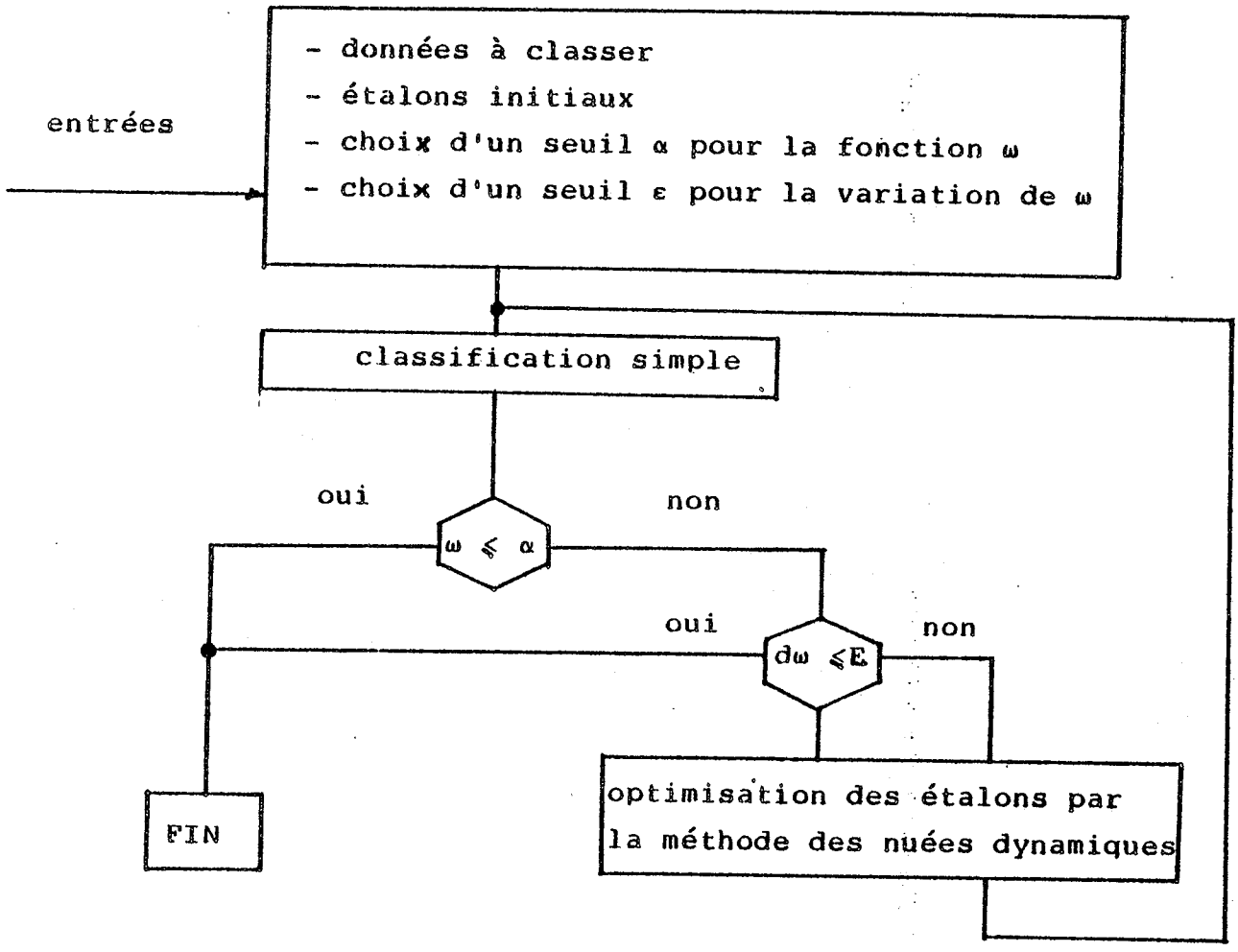


figure 3.4.11. optimisation du noyau

3.4.4.2. Classification des transitions

La méthode de classification pour les transitions est beaucoup plus simple. Après un regroupement parmi le corpus enregistré, on obtient environ 18000 segments de 12,5 ms correspondant à des états transitoires. On choisit un nombre de 200 étalons. En premier lieu, on calcule le centre de gravité de toutes les transitions, ensuite on initialise les étalons suivant l'algorithme décrit par la figure 3.4.9. ; puis on procède à l'optimisation de l'ensemble des étalons suivant la méthode exposée par la figure 3.4.10. On obtient alors 200 modèles pour les états correspondant aux transitions.

L'optimisation est effectuée en 2 itérations : la première donnant une distorsion globale de 0.45 (valeur de la fonction critère ω). Cette valeur correspond à la distance du corpus des transitions avec l'ensemble initial des étalons. A la deuxième itération la distorsion globale est optimisée à 0.23.

3.4.5. Conclusion

Notre démarche est une approche classique de reconnaissance de formes :

- trouver un ensemble assez représentatif des formes à étudier (corpus).
- partitionner cet ensemble en un certain nombre de classes pertinentes.
- trouver des représentants à l'intérieur de chacune de ces classes. Ces derniers représentent au mieux les formes à étudier.

Nous avons utilisé deux méthodes différentes pour les phonèmes et pour les transitions. Une approche plus élaborée pour les premières provient de la connaissance plus précise que nous avons des états stables. Le nombre de classes "a priori" étant fixé à 38 il a été possible d'effectuer une analyse pas à pas. A l'inverse, le cas des transitions est beaucoup plus complexe. D'abord les états

sont plus difficiles à définir acoustiquement. Ensuite leur nombre trop important dans la langue (environ 600) ne pouvait pas induire une approche similaire à celle des états stables. En revanche, dans la phase d'identification phonétique des modèles, (voir chapitre suivant) nous avons considéré toute transition d'un phonème A vers un phonème B comme un événement particulier ; cet événement est caractérisé à chaque instant par deux sous-états. Le premier détermine l'éloignement du phonème A, le deuxième l'approche du phonème B. Cette idée a l'avantage de réduire le nombre d'états transitoires (en nombre de 38) tout en conservant l'information qui est exprimée par le passage progressif d'un phonème à un autre.

Après la classification, le regroupement des étalons relatifs aux transitions avec les étalons correspondant aux états stables constitue l'ensemble des modèles phonétiques utilisés pour la reconnaissance. Pour cela, d'après le principe de la méthode de reconnaissance phonétique (voir chapitre s'y rapportant) il est nécessaire de caractériser phonétiquement chacun de ces modèles.

3.5. NATURE PHONETIQUE DES MODELES

3.5.1. Introduction

Une reconnaissance de type statistique se fait de manière indirecte lorsqu'un mot inconnu est prononcé on procède tout d'abord à un découpage de ce mot en segments très courts analysés numériquement. Ensuite chacun des segments est identifié au modèle le plus proche, suivant une proximité établie. Pour connaître la nature phonétique du segment inconnu, il faut donc posséder certaines informations phonétiques sur le modèle auquel il est associé. L'identification phonétique du segment en découlera (figure 3.5.1)

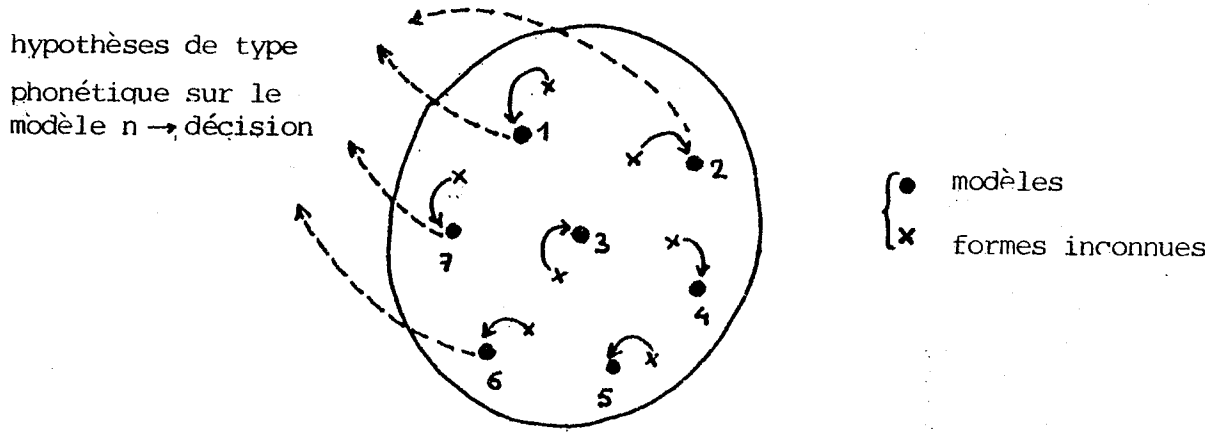


figure 3.5.1. De l'affectation des formes inconnues (segment de 12,5 ms) à l'identification de ces formes (décision)

Les hypothèses que l'on peut émettre sur la nature phonétique des modèles résulte de la constitution des classes correspondantes. On peut identifier phonétiquement chaque élément dans une classe donnée grâce à l'étiquetage préalable du corpus (segmentation). Un certain nombre de mesures effectuées sur les classes permet de déduire la nature phonétique de leur représentant.

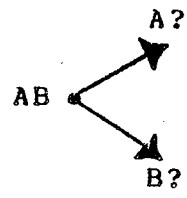
3.5.2. Méthodologie

3.5.2.1. Méthode relative aux transitions

Soit la transition AB du phonème A vers le phonème B. L'événement qui lui correspond est considéré comme une suite de deux états :

- le premier exprime le passage de A à B par l'éloignement de A. On le note "A" ?
- le deuxième traduit la transition AB par le fait que l'on se rapproche du phonème B. On l'appellera B ?

De manière plus schématique on effectue une séparation de la transition AB :



L'important est de pouvoir quantifier ces deux états dans la transition AB. Supposons ainsi que la transition à étudier soit constituée de n segments de 12,5 ms. (figure 3.5.2)

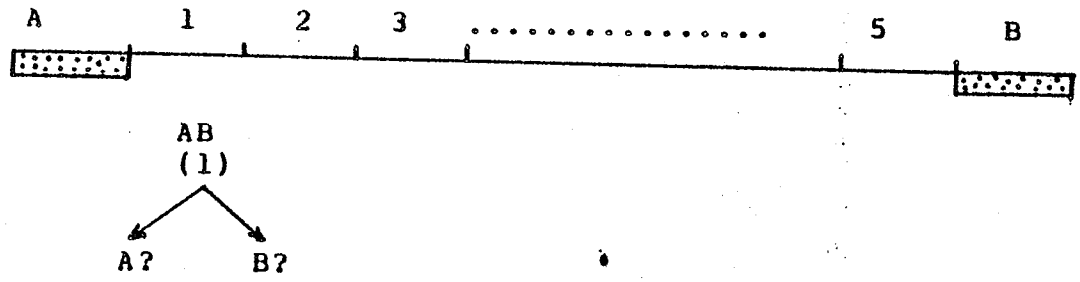


figure 3.5.2.

Lorsqu'un de ces n segments se trouve dans une classe C on augmente sa fréquence de 1 dans C. Si on ne considère plus l'étiquette "AB" mais l'ensemble des deux étiquettes A? et B?, La fréquence qui leur correspondra sera fonction de la position du segment dans toute la transition. Intuitivement, si on prend le premier segment de AB on conçoit qu'il se situe davantage dans l'état A?. C'est l'inverse pour le dernier segment.

En première approximation, on quantifiera l'évolution des deux états A? et B? par une évolution linéaire (figure 3.5.3)

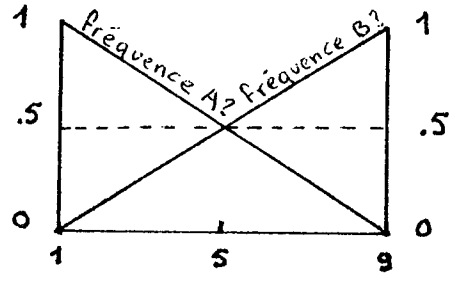


figure 3.5.3. a cas n impair

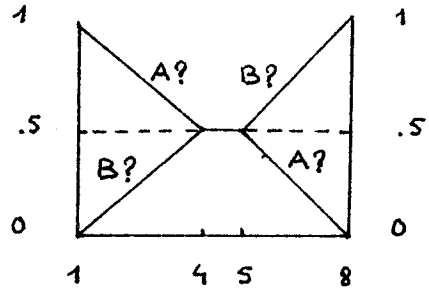


figure 3.5.3. b cas n pair

figure 3.5.3. : Evolution des états A? et B?

Par exemple, supposons que la transition AB ait une durée de 5 x 12,5 ms. Soit une classe C dans laquelle on trouve les états A? et B? avec les fréquences respectives $f_{A?}$ et $f_{B?}$. Si le troisième segment de AB est classé dans C, alors, d'après le découpage ci-dessous :

1	2	3	4	5
A?(1)	A?(3/4)	A?(½)	A?(¼)	A?(0)
B?(0)	B?(¼)	B?(½)	B?(3/4)	B?(1)

$f_{A?} \Rightarrow f_{A?} + \frac{1}{2}$

$f_{B?} \Rightarrow f_{B?} + \frac{1}{2}$

Discussion de la méthode

1 L'homogénéité est conservée. En effet, pour un segment donné, la somme des fréquences $f_{A?}$ et $f_{B?}$ reste égale à 1.

2 L'avantage de ce traitement est que le nombre d'états à considérer pour les transitions est le même que pour les phonèmes (38). Les classes obtenues sont donc plus faciles à caractériser.

3 Pour simplifier la démarche, on a fait un certain nombre d'approximations :

- On suppose que les évolutions des états A? et B? sont symétriques ce qui n'est pas vrai dans certains cas : explosion suivie d'une voyelle, par exemple. Des études sur ce sujet pourraient être faites sur les différents cas, de manière à fournir une solution plus juste.

- Par conséquent, on émet aussi l'hypothèse de la symétrie (BA, AB), puisque le symbole A? exprime un état allant vers A ou provenant de A ce qui n'est pas justifié dans certains cas. Une approche plus fine pourrait être faite en distinguant les transitions venant de A de celles allant vers A.

3.5.2.2. Effectif des classes

La reconnaissance statistique dépend du corpus enregistré. Pour obtenir des mesures cohérentes, il faut avoir à traiter un grand nombre de données et cela, pour chacun des phonèmes et des transitions correspondantes. Sous cette condition, il est

intéressant aussi que le corpus soit équilibré phonétiquement c'est-à-dire avoir sensiblement le même nombre de segments pour tous les états. Ce qui est difficile à obtenir. C'est pourquoi, avant toute mesure statistique, les effectifs réels des phonèmes dans une classe donnée ont été affectés d'un poids inversement proportionnel à leur pourcentage de représentation dans le corpus initial.

La table suivante donne les proportions des phonèmes type A et des transitions A? correspondantes. Pour le cas des transitions A? le calcul a consisté à compter toutes les transitions contenant A et à diviser l'effectif obtenu par deux. Ce calcul résulte des considérations exposées dans le chapitre précédent.

P ,	0.48,	2.61
T ,	0.63,	2.87
K ,	0.48,	1.20
P,	1.24,	3.66
T,	1.29,	1.84
K,	0.87,	1.23
B ,	0.54,	0.65
D ,	0.61,	2.04
G ,	0.51,	0.89
F ,	0.76,	1.14
S ,	1.54,	2.05
C ,	1.15,	0.61
V ,	0.78,	1.99
Z ,	0.46,	1.20
J ,	0.66,	2.08
N ,	0.79,	0.42
M ,	0.60,	0.59
L ,	0.89,	2.07
R ,	0.89,	1.65
A ,	0.80,	0.92
E< ,	0.94,	0.93
E> ,	0.78,	0.97
O< ,	0.80,	0.82
O> ,	0.56,	0.49
I ,	0.77,	1.03
Y ,	0.66,	0.53
U ,	0.48,	0.41
^< ,	1.00,	0.90
^> ,	0.90,	0.62
^ ,	0.51,	0.23
<A ,	0.86,	0.87
<O ,	0.72,	0.75
<^ ,	0.74,	3.72
<E ,	0.80,	0.73
WI ,	0.44,	0.45
WU ,	0.20,	0.22
WY ,	0.13,	0.47
* ,	0.01,	26.89

figure 3.5.4. Proportions des phonèmes A et des transitions A?

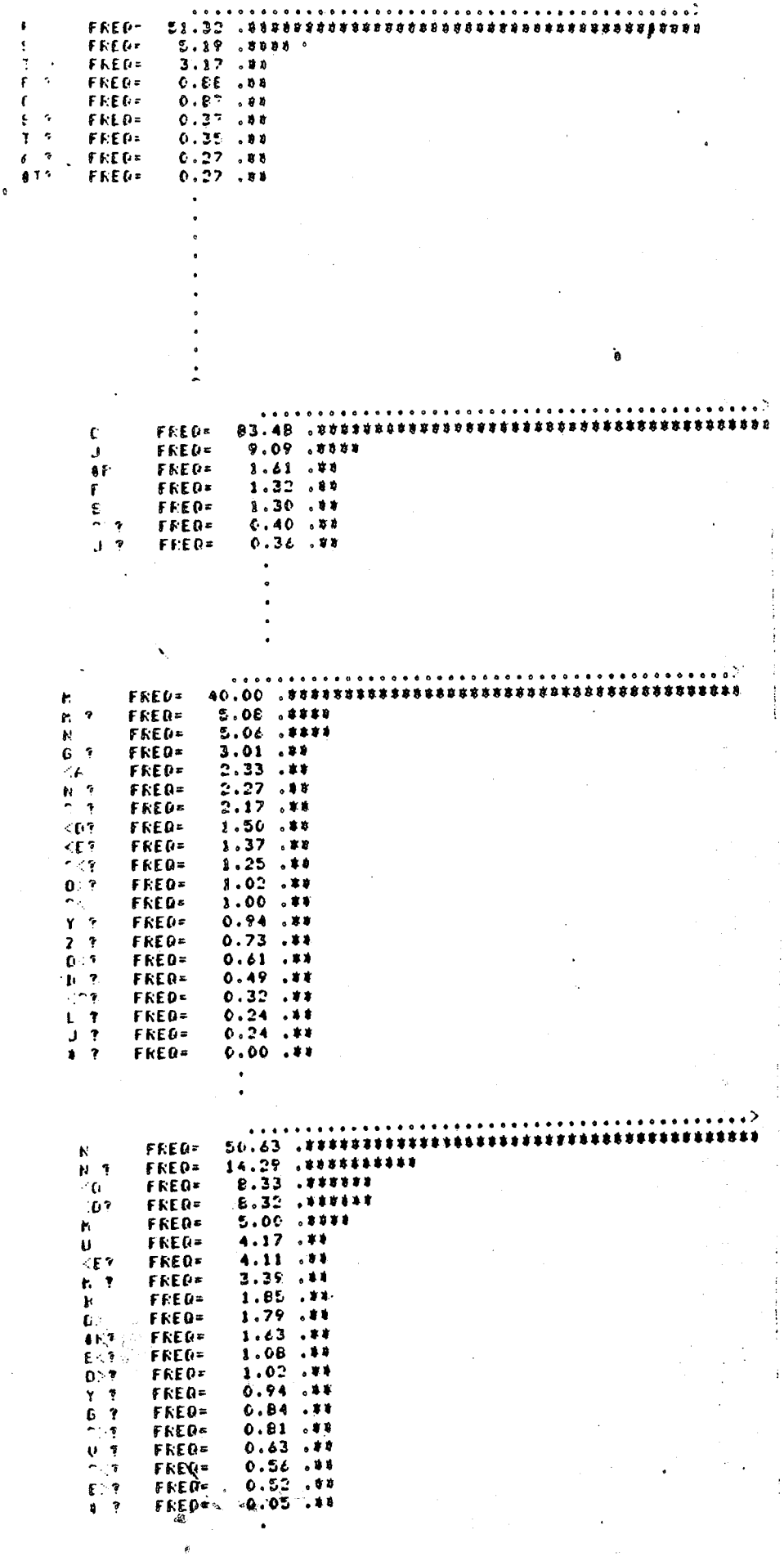


figure 3.5.5. Histogrammes montrant la représentativité des modèles

Soit maintenant une classe C, de centre de gravité le point G. G est un des modèles résultant de l'apprentissage. L'important est de connaître la liste des phonèmes ou des transitions qu'il est sensé représenter avec telle ou telle probabilité.

La méthode est très simple :
Soit $(\varphi_1, \ldots, \varphi_n)$ les éléments de la classe C représentés avec les fréquences respectives $(f_n, \ldots, f_n)$. Pour les transitions A?, ces fréquences sont calculées suivant la méthode exposée dans le chapitre précédent. Pour les parties stables A, ce sont des valeurs entières incrémentées de la valeur "un" chaque fois que le phonème se trouve dans la classe. On pondère $(f_1, \ldots, f_n)$ par les proportions phonétiques correspondantes (figure 3.5.4.)

$$f'_i = f_i / p_i$$

Alors la probabilité pour que G représente le phonème φ_i est égale à la fréquence effective f'_i sur la somme de toutes les fréquences. On note :

$$P(G = \varphi_i) = \frac{f'_i}{\sum_{i=1}^{n} f'_i} .$$

3.4.3. Réalisations pratiques

La classification des données du corpus permet d'obtenir un ensemble de modèles jugé optimum. Pour chacun d'eux une table appelée "table de fréquences phonétiques", donne la liste de tous les éléments contenus dans la classe représentée par le modèle en question. La lecture de ces tables mesure la qualité du modèle.

Pratiquement, on obtient deux types de fichier de type "record ILS" (voir annexes) formatés différemment. Ces fichiers agissent en parallèle, l'un contenant la liste des modèles, l'autre la succession des tables des fréquences phonétiques. Ces deux fichiers, résultant de l'apprentissage, seront les deux seuls outils nécessaires dans la phase de reconnaissance.

Parallèlement, on construit pour chaque modèle un histogramme donnant la liste des éléments qu'il peut représenter. Nous en donnons ici quelques exemples. Dans le cas de la figure 3.5.5. on reconnaît de "bons modèles" des consomnes F, C, N, M avec une bonne discrimination d'une part entre les deux fricatives non voisée d'autre part entre les deux consonnes nasales.

On constate, sur les figures 3.5.5.c et 3.5.5.d l'apparition de la transition M? dans la classe M et celle de la transition N? dans la classe N. Comment interpréter ce résultat ? Il ne faut pas oublier à ce niveau, que l'étiquette est donnée aux segments au moment de la segmentation manuelle du corpus (voir chapitre s'y rapportant) et qu'il est parfois difficile, même manuellement de déterminer les frontières exactes d'un état stable. C'est pourquoi, il est assez fréquent d'observer les regroupements (phonème, transition) dans une même classe, comme le montre la figure 3.5.6. par exemple.

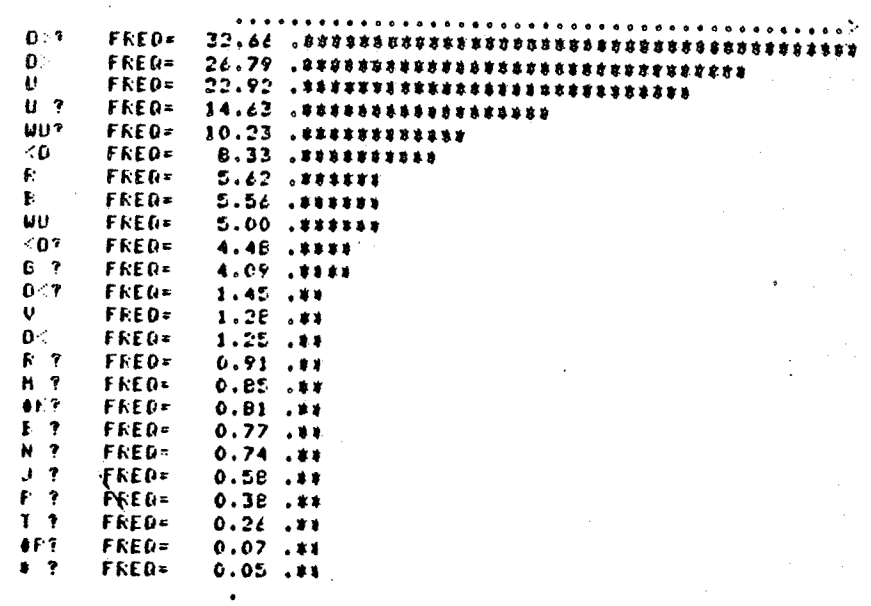


figure 3.5.6. effets d'une segmentation peu précise

En règle générale, on obtient de bons histogrammes pour les états stables, souvent fortement unimodaux. De bonnes discriminations sont observées à l'intérieur de groupes phonétiquement voisins. La figure 3.5.5. en est déjà un exemple. Citons encore :

- la bonne séparation du "euh" ouvert et fermé. (figure 3.5.8)
- celle des deux voyelles nasales <0 et <A (figure 3.5.9.)
- ou encore la bonne représentation de la voyelle orale A (figure 3.5.7.)

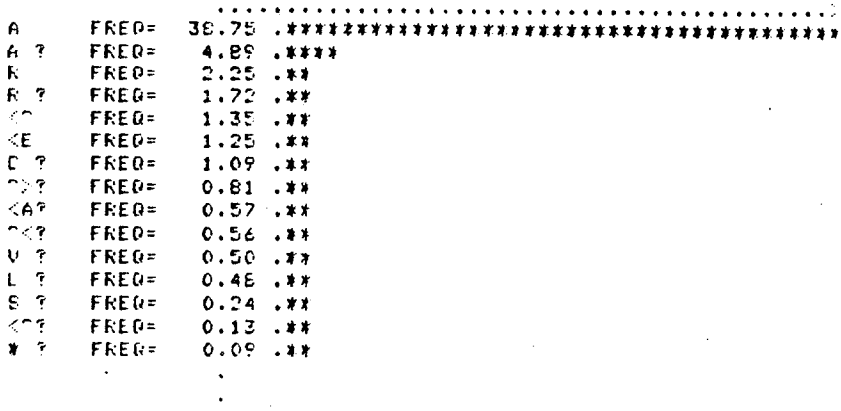


figure 3.5.7. : un bon modèle du "A"

13

```
                    .....................................................
^       FREQ=  79.00 .*****************************************************
^^?     FREQ=  17.11 .*********
O       FREQ=  15.00 .*******
<E      FREQ=  10.00 .*****
A       FREQ=   6.25 .***
E<?     FREQ=   5.97 .***
O^?     FREQ=   5.49 .***
^>?     FREQ=   4.49 .***
^ ?     FREQ=   3.99 .***
^       FREQ=   3.32 .***
F ?     FREQ=   2.91 .***
<^      FREQ=   2.70 .***
T ?     FREQ=   2.26 .***
F       FREQ=   2.25 .***
E       FREQ=   2.13 .***
<E?     FREQ=   1.71 .***
U ?     FREQ=   1.35 .***
B ?     FREQ=   1.15 .***
W1?     FREQ=   1.11 .***
A ?     FREQ=   1.09 .***
^^?     FREQ=   0.99 .***
N ?     FREQ=   0.83 .***
L ?     FREQ=   0.76 .***
<O?     FREQ=   0.67 .***
<A?     FREQ=   0.57 .***
Z ?     FREQ=   0.42 .***
J ?     FREQ=   0.16 .***
# ?     FREQ=   0.01 .***
```

```
                    .....................................................
^       FREQ=  52.22 .*****************************************************
^^?     FREQ=  33.76 .*********************************
O^?     FREQ=  15.48 .***************
O ?     FREQ=  11.74 .***********
^^?     FREQ=   7.82 .*******
^<      FREQ=   7.00 .*******
L ?     FREQ=   6.34 .******
O       FREQ=   5.00 .*****
N ?     FREQ=   4.76 .*****
^       FREQ=   3.92 .****
F ?     FREQ=   2.31 .***
W1      FREQ=   2.27 .***
W1?     FREQ=   2.22 .***
^ ?     FREQ=   2.17 .***
U       FREQ=   2.08 .***
G       FREQ=   1.96 .***
Y       FREQ=   1.52 .***
E       FREQ=   1.28 .***
N       FREQ=   1.27 .***
<E      FREQ=   1.25 .***
L       FREQ=   1.12 .***
G ?     FREQ=   0.98 .***
Y ?     FREQ=   0.94 .***
T ?     FREQ=   0.93 .***
<A?     FREQ=   0.86 .***
<O?     FREQ=   0.83 .***
F ?     FREQ=   0.66 .***
E ?     FREQ=   0.54 .***
U ?     FREQ=   0.50 .***
B ?     FREQ=   0.49 .***
F ?     FREQ=   0.44 .***
J ?     FREQ=   0.36 .***
Z ?     FREQ=   0.31 .***
# ?     FREQ=   0.19 .***
```

figure 3.5.8. Séparation ($\Lambda^<, \Lambda^>$)

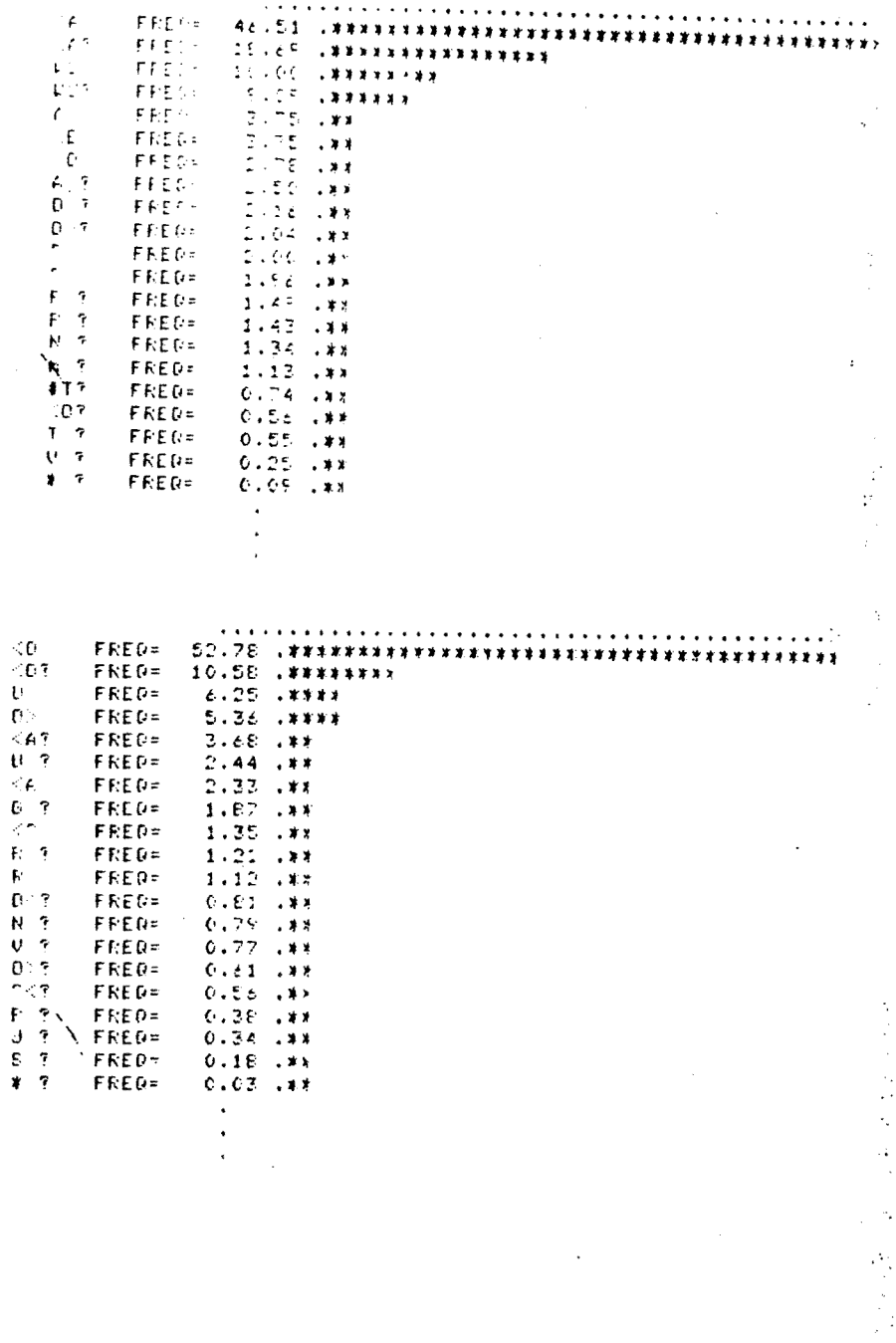


figure 3.5.9. séparation <0, <A

Les résultats obtenus pour les transitions sont beaucoup moins exploitables. Cela était d'ailleurs prévisible. Les histogrammes que nous obtenons sont assez pointus mais ce qui est plus frappant c'est le petit nombre de représentants contenus dans les classes. La transition, telleque nous l'avons définie dépend fortement du contexte. Aussi, il est évident qu'un <E? après une nasale

n'a pas les mêmes caractéristiques q'un <E? après une plosive. Nous donnons ici deux exemples d'histogrammes pour les transitions (figure 3.5.&0)

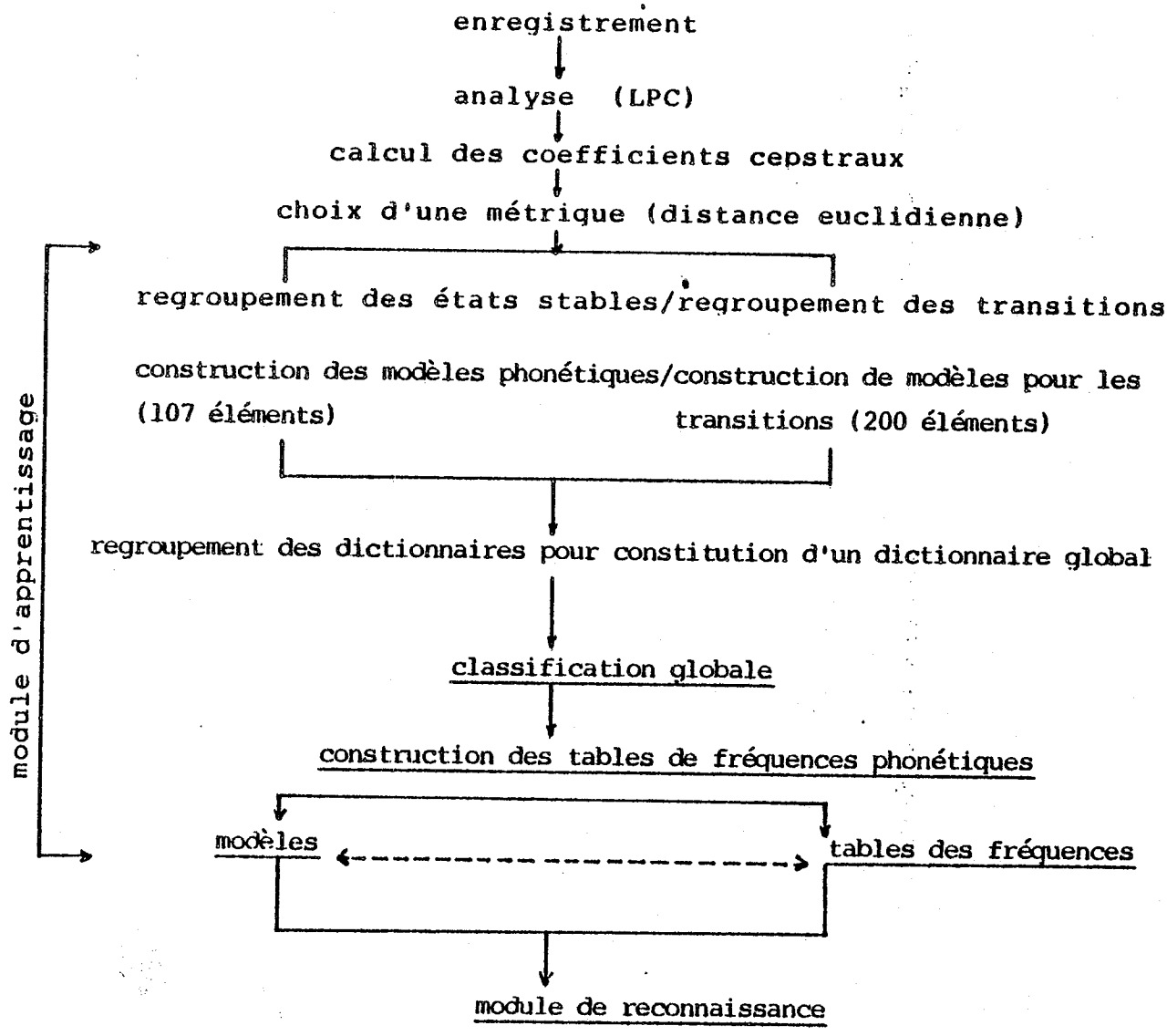
```
                    .....................................................
I ?     FREQ=  11.17 .#####################################################
WI?     FREQ=   6.76 .################################
J       FREQ=   6.06 .############################
Z       FREQ=   4.35 .####################
E       FREQ=   3.85 .##################
Z ?     FREQ=   2.92 .#############
I\      FREQ=   2.60 .############
E:?     FREQ=   2.51 .###########
WI      FREQ=   2.27 .##########
V       FREQ=   1.28 .#####
J ?     FREQ=   1.20 .#####
L       FREQ=   1.12 .#####
E:?     FREQ=   1.03 .##
F ?     FREQ=   0.88 .##
S ?     FREQ=   0.49 .##
# ?     FREQ=   0.00 .##
                    .
                    .
                    .
HERE IS THE HISTOGRAM NUMBER 244
                    .....................................................
<E?     FREQ=  15.45 .#####################################################
<       FREQ=   4.05 .###################
M ?     FREQ=   3.56 .#################
E<?     FREQ=   2.15 .##########
D:?     FREQ=   1.83 .#########
J       FREQ=   1.52 .#######
F ?     FREQ=   1.40 .##
<?      FREQ=   1.26 .##
<E      FREQ=   1.25 .##
L       FREQ=   1.12 .##
R       FREQ=   1.12 .##
?>      FREQ=   1.11 .##
E       FREQ=   1.06 .##
P ?     FREQ=   0.91 .##
A ?     FREQ=   0.82 .##
V ?     FREQ=   0.35 .##
T ?     FREQ=   0.35 .##
G ?     FREQ=   0.28 .##
L ?     FREQ=   0.24 .##
# ?     FREQ=   0.13 .##
J ?     FREQ=   0.05 .##
                    .
                    .
                    .
```

figure 3.5.10 histogrammes moins représentatifs pour les transitions

Il est évident que la qualité des modèles agira fortement sur le module de reconnaissance. Cette qualité est mesurée par les histogrammes de représentation relatifs à chacun des modèles. Comme nous l'avons vu, il est dans certains cas, difficile d'assurer l'unimodalité des courbes obtenues. Le cas des transitions est des plus délicats, mais le fait d'avoir pu les caractériser en un petit nombre d'états nous a permis de mettre au point la méthode de reconnaissance. Le principe de la méthode est la recherche d'un chemin optimal caractérisant l'identification d'un mot inconnu avec un mot présent sous forme phonétique dans un vocabulaire donné. La méthode a pour avantage de considérer le mot globalement même si chaque segment est identifié à chaque pas avec telle ou telle probabilité. C'est le but de la programmation dynamique.

3.5.4. Conclusion

Rappelons le but de la phase d'apprentissage : fournir au module de reconnaissance une base de références phonétiques associée à un système d'informations permettant de mesurer la qualité des éléments de cet ensemble de références. L'apprentissage agit depuis l'acquisition du signal jusqu'à la construction des modèles phonétiques. La méthodologie adoptée est résumée par le schéma qui suit :



C H A P I T R E IV

RECONNAISSANCE

4.1. LA PROGRAMMATION DYNAMIQUE

4.1.1. Introduction

La programmation dynamique donne en général des résultats satisfaisants pour la résolution de problèmes concernant la recherche d'optimum. Dans le domaine du traitement de la parole, ce type de méthode a souvent été utilisé pour la reconnaissance de mots isolés ou de discours continu, et aussi pour la segmentation automatique de mots.

Les algorithmes les plus fréquemment présentés sont fondés sur le principe proposé par Sakoe et Chiba. Cette méthode a en particulier été mise au point afin de pouvoir effectuer une normalisation temporelle entre deux mots. En effet, les variations dues à la différence de débit de parole causent des fluctuations non négligeables et gênantes pour résoudre les problèmes de reconnaissance.

A chaque point du plan déterminé par le couple C = (i,j) est associée une distance :

$$d(C) = a(i,j) = ||a_i - b_j||$$

qui exprime la "proximité" entre les deux vecteurs a_i et b_j.

A la forme F correspond alors la mesure :

$$E(F) = \sum_{k=1}^{K} d(C(k))\ \omega(k)$$

où $\omega(k)$ est une pondération.

Lorsque E(F) atteint son minimum F est considérée comme la meilleure fonction déterminant l'ajustement temporel entre les deux formes.

$$D(A,B) = \min_{F} \frac{\sum_{k=1}^{K} d(C(k))\ \omega(k)}{\sum_{k=1}^{K} \omega(k)}$$

où le dénominateur a pour effet d'atténuer l'influence dûe au nombre de points (K) de F.

Cette méthode a beaucoup été utilisée pour la reconnaissance globale de mots isolés [Myers, Rabiner, Rosenberg], [GUAGNOULET JOUVET] où les formes A et B à comparer sont décrite numériquement (coefficients représentatifs du signal). C'est aussi un outil puissant par la segmentation automatique. [Jordan R. Cohen]. Nous en parlerons plus en détail dans un des chapitres suivants.

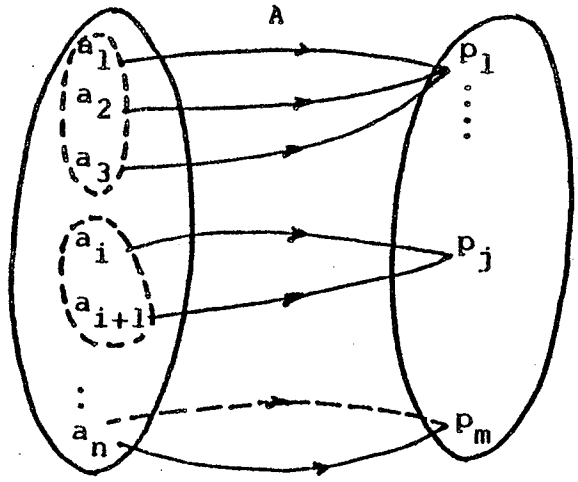
4.1.2. Description de notre méthode

- Principe

Notre méthode diffère un peu des méthodes de programmation dynamique classiques. En effet, dans notre cas, les deux formes à comparer ne sont pas de même nature. L'une est décrite qualitativement : c'est la séquence phonétique correspondant à un mot L'autre est représentée de manière quantitative au moyen d'une suite de paramètres numériques : suite des vecteurs contenant les coefficients cepstraux. Pour ce type de méthode, le problème d'alignement temporel est exclu puisque l'idée consiste à trouver la "meilleure" fonction liant l'ensemble des vecteurs caractérisant le signal inconnu et l'ensemble des états phonémiques de la chaine d'entrée. On désigne ainsi par :

- $T = (a_1, \ldots, a_n)$ la suite des vecteurs caractérisant le mot à identifier.
- $P = (p_1, \ldots, p_m)$ la séquence des états phonémiques à préciser à l'entrée.

De manière plus schématique, on représente ci-dessous l'application A entre T et P.



A chaque vecteur a_i (i = 1 ... n) correspond un état p_j (j=1...m) déterminé de manière optimale, avec les contraintes :

- limites aux bornes $A(a_1) = p_1$ et $A(a_n) = p_m$

- monotonie : $\forall\ i_1 < i_2$ avec $A(a_{i1}) = p_{j1}$

et $A(a_{i2}) = p_{j2}$ $j_1 < j_2$

- surjectivité de l'application

Chacune des classes C_j (j = 1...m) déterminée par $C_j = \{A^{-1}(p_j)\}$ représente les segments de parole correspondant au phonème p_j.

Voici le principe général de la méthode :
Soient deux formes A et B à comparer :

$A = a_1 \ldots a_i \ldots a_I$

$B = b_1 \ldots b_j \ldots b_J$ (où a_i, b_j sont des vecteurs).

I étant différent de J, la durée des deux formes n'est pas forcément identique.

On considère alors l'ensemble des couples :

$$C = (i,j)$$

et F, une fonction, dite "fonction de déformation"

$$F = C(1), C(2) \ldots C(K)$$

avec $C(K) = (i(K), j(K))$.

Si aucune différence n'est observée alors F décrit la diagonale, pour tout K, $i(K) = j(K)$.
Une représentation de F peut être visible sur la figure 1.

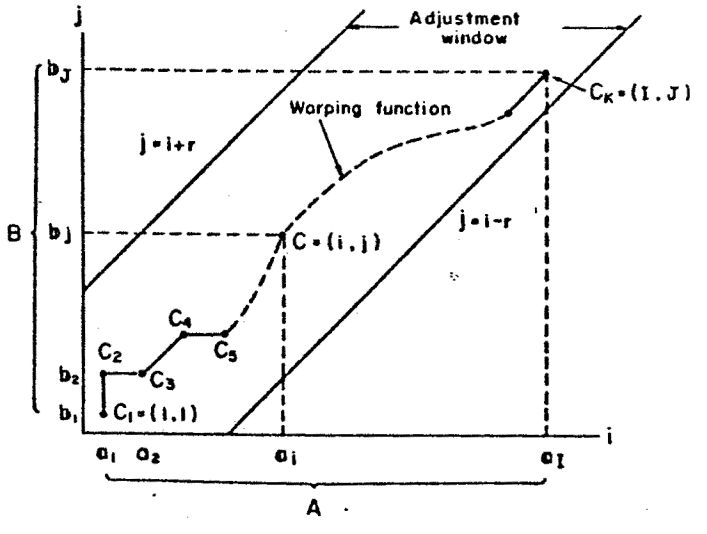


figure 4.1.1. d'après [Sakoe, Siba]
Fonction de déformation

- en pratique :

Le mot inconnu est échantillonné segmenté puis numérisé (coefficients cepstraux). A chacun des segments (12.5ms) est

associé le modèle phonétique le plus proche et la table des fréquences phonétiques associée. L'identification du mot consiste non pas à caractériser le mot lui-même mais la suite des modèles phonétiques qui lui correspond. (figure 4.1.2.)

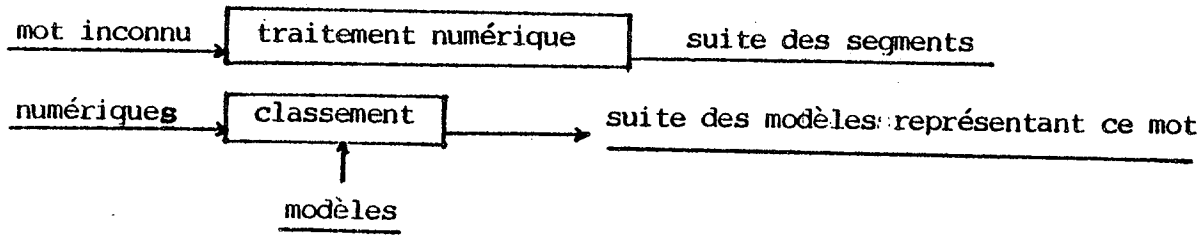


figure 4.1.2. Prétraitement pour l'identification d'un mot

La reconnaissance fait alors appel à la programmation dynamique afin de rechercher le meilleur chemin entre la suite des modèles phonétiques représentant le mot inconnu et la séquence phonétique d'un mot du vocabulaire. (figure 4.1.3.).

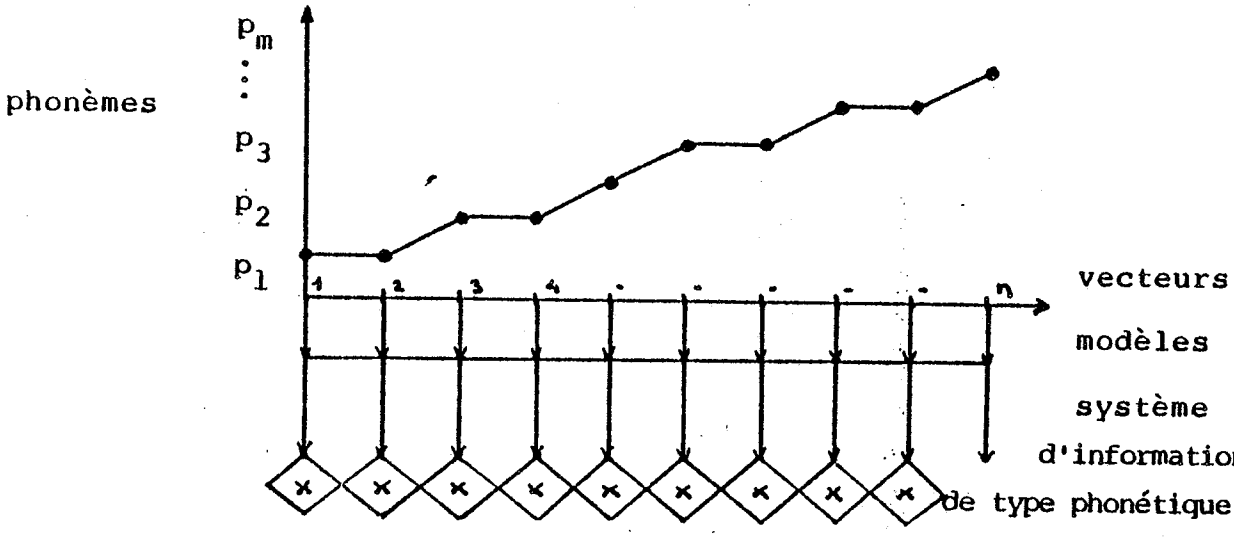


figure 4.1.3. : Recherche du meilleur chemin pour l'identification d'un mot

- Transcription phonétique du vocabulaire

La transcription phonétique d'un mot distingue deux états :

- les états stables : phonèmes
- les transitions.

Rappelons que la transition entre le phonème A et le phonème B est considérée comme la juxtaposition de deux états notés A? et B?. (voir chapitre s'y rapportant). En conséquence, la transcription phonétique d'un mot prend en compte deux types de symbole :

- symbole type phonème : A
- symbole type transition : A? (contenant le symbole caractérisant le début et la fin d'un mot : * ?)

Exemple : le mot "chat" est représenté par la séquence
*? C? C C? A? A A? *?

La suite de ces symboles permet de caractériser chacun des mots du vocabulaire. Certains mots nécessitent de plus d'une transcription phonétique, ces derniers pouvant être prononcés différemment suivant les locuteurs. Nous en donnons quelques exemples :

UN : <Æ, <E
JAUNE : JO < N, JO > N
PEINTURE : P < ETYR, P <ΛTYR
⋮

- Contraintes relatives au chemin

L'application A définie précédemment étant surjective tout chemin doit se situer dans la zone C (figure 4.1.4.)

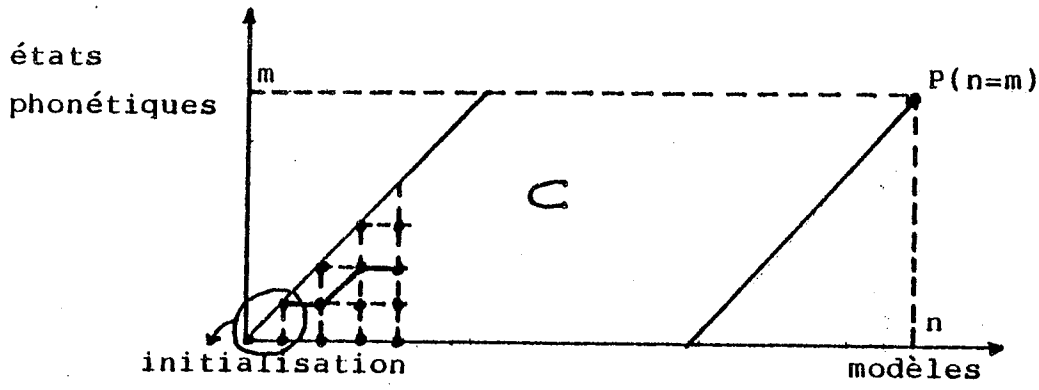


figure 4.1.4. zone de cheminement possible

De même, considérons un point M(i,j) correspondant à la réalisation : modèle i = état j. L'identification du modèle précédent i-1 doit se faire soit dans l'état j (dans ce cas il y a séjour dans cet état) soit dans l'état j-1 (dans ce cas il y a changement d'état) (figure 4.1.5.)

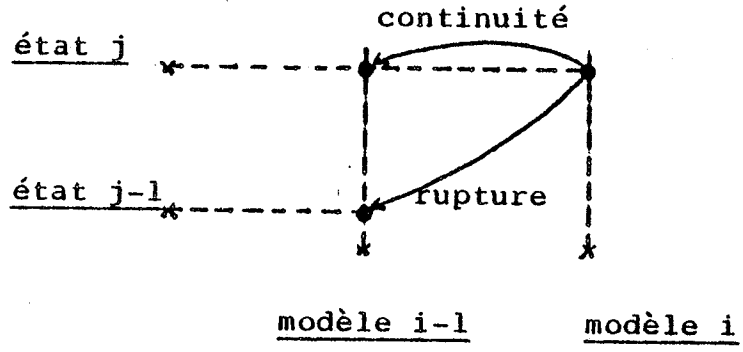


figure 4.1.5. contraintes locales

- Equations de la programmation dynamique

Compte tenu des considérations précédentes, on en déduit les équations donnant le chemin optimal.
On suppose connaître pour le modèle i-1 tous les chemins optimaux tels que l'on ait :

modèle (i-1) = état(j) $\quad (j \in C)$

A chacun de ces chemins est affecté une fonction P(i-1,j) exprimant la meilleure probabilité du chemin sachant que modèle i-1 = état j ($j \in C$). Soit le modèle i, alors :

$P(i,j) = \max (P(i-1,j)\ \omega(j),\ P(i-1,\ j-1)\ \omega(j-1))$

si $i \neq j$ x $p(i,j)$

$P(i,j) = [P(i-1,\ j-1)\ \omega(j-1)]\ P(i,j)$

si $i = j$

ω étant une pondération liée à l'état
P étant une pondération liée au modèle.

- <u>détermination de P, p,ω</u> .

A chacun des modèles correspond une table de probabilités donnant un certain nombre d'informations sur sa nature phonétique. (voir chapitre s'y rapportant). On peut donc connaître la probabilité pour que ce modèle représente tel ou tel phonème. Ainsi, à la réalisation "modèle i = phonème j" on fait correspondre le nombre p(i,j) tel que : P(i,j) = probabilité (modèle i = phonème j).

ω est une fonction dépendante de l'état précédent. Elle exprime soit la possibilité de continuité dans un état soit la probabilité de rupture de cet état (passage à l'état suivant).
P est déterminé par récurrence à partir de p et ω. La connaissance des deux premières colonnes (initialisation) permet de faire démarrer l'algorithme (figure 4). P(n,m) donnera la probabilité du chemin optimal connu grâce à un marquage pas à pas

- <u>Calcul pratique de P ; probabilité du chemin</u>

On rappelle l'expressionde P :
pour $i \neq j$ $P(i,j) = \max(P(i-1,j)\omega(j),\ P(i-1,j-1)\omega(j-1)p(i,j)$

pour $i = j$ $P(i,j) = [P(i-1,j-1)\ \omega(j-1)]\ p(i,j)$

Pour des commodités de calcul (multiplications successives de petits nombres) on effectue la transformation suivante :

- p(i,j) probabilité relative au modèle est exprimé en "pour-cent". Même chose pour la quantité ω_j.

- on note $p_1(i,j) = 50 \log_{10} (p_{\%}(i,j))$
  (si $p_{\%}(i,j) = 100$, $p_1(i,j) = 100$)

A chaque étape des probabilités p_1 sont ajoutées :

$$p_1(i,j) = \max (p_1(i-1,j) + 50 \log_{10}(\omega(j)), p_1(i-1,j-1) + 50 \log_{10}(\omega(j-1)) + p_1(i,j)$$

La probabilité finale est obtenue par : $p_1(n,m)$ on la transforme par :

$$p_1(n,m) \rightarrow (\frac{1}{2\,n} p_1(n,m) \times 100)\%$$

4.2. APPLICATION A LA SEGMENTATION AUTOMATIQUE

4.2.1. Procédé

Notre procédé de segmentation automatique a été inspiré de la méthode proposée par JORDAN R. COHEN. En voici l'idée : le mot à segementer est considéré comme une succession d'états binaires assimilée à une chaîne de Markov. La description de cette suite exprime la continuité ou la discontinuité d'un phonème (transition). Ainsi tout mot I est considéré comme une séquence booléenne $B_I(t)$ $0 < t < N-1$. L'ensemble $C = \{t / B_I(t) = 0\}$ détermine les états de continuité. L'ensemble $D = \{t / B_I(t) = 1\}$ désigne les états de discontinuité (figure 4.2.6.). La segmentation consiste alors à trouver le meilleur ensemble B_I correspondant à une observation O (mot prononcé). Cette expression [COHEN] s'évalue par programmation dynamique en considérant une machine à M états (figure 4.2.7.)

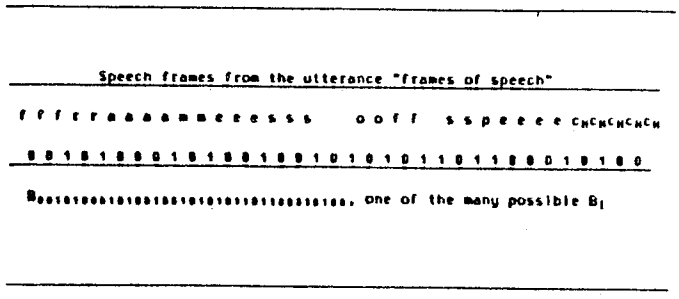


figure 4.2.6. vecteur booléen représentant le mot : "frames of speech"

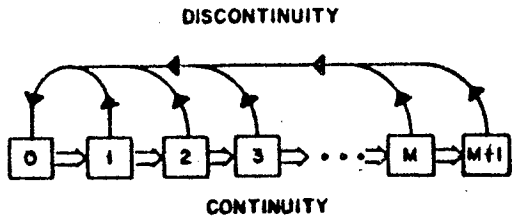


figure 4.2.7. machine à M états

L'état désigne la longueur du phonème considéré. Le retour en "0" exprime la présence d'une transition. En général la longueur d'un phonème varie entre 300 et 600 ms.

Pour en revenir à notre étude, la séquence booléenne que nous utilisons est constituée :

- d'une part de sous-suites de "0" exprimant la continuité dans un état stable (A) ou dans un état transitoire (A?).
- d'autre part de la présence de "1" signifiant la rupture d'un état (stable ou transitoire).

La séquence correspondant à l'exemple donné par la figure 4.2.8. est donc :

"000100100001001001000010010"

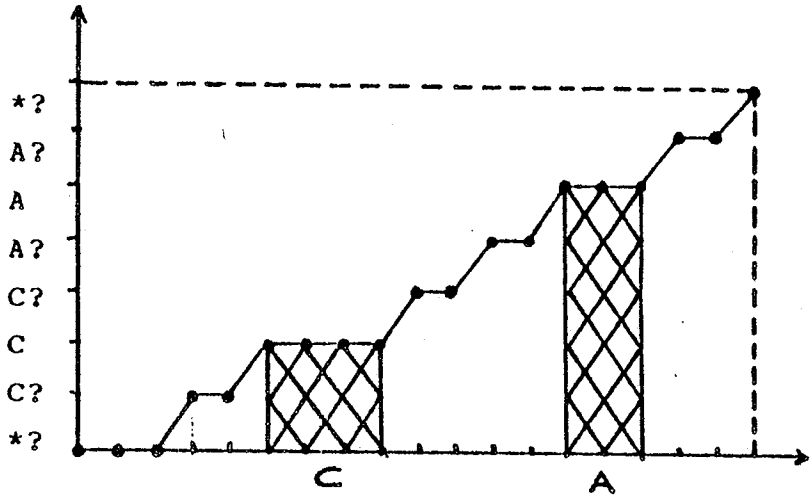


figure 4.2.8. chemin donnant la meilleure segmentation du mot "chat"

L'intérêt de la segmentation automatique réside dans sa rapidité et donc dans sa capacité d'augmenter facilement le corpus de données servant à l'apprentissage (auto-apprentissage). Celle que nous proposons est moins précise que la segmentation manuelle puisqu'elle travaille au niveau des segments de 12.5ms (et non de 12.5 10^{-2}ms). De même, le procédé d'apprentissage automatique n'a pu être exploité qu'à partir d'un corpus assez consistant puisque les statistiques calculées en découlaient. C'est pourquoi, la segmentation manuelle était nécessaire au préalable. Nous avons donc utilisé la segmentation automatique seulement dans un deuxième temps sur une liste de mots prononcés par trois locuteurs.

Le calcul de la pondération ω découle des observations faites à partir du corpus de données. Pour chacun des phonèmes, on a tracé des histogrammes relatifs à la durée en nombre de segments de 12.5ms. Pour les transitions, le traitement a été fait globalement. Pour des résultats plus cohérents, un lissage des histogrammes a été effectué.

* calcul de la probabilité de rupture d'un état (P_r) :

Soit l'état j, de durée déjà égale à d à un certain stade de la segementation. Alors :

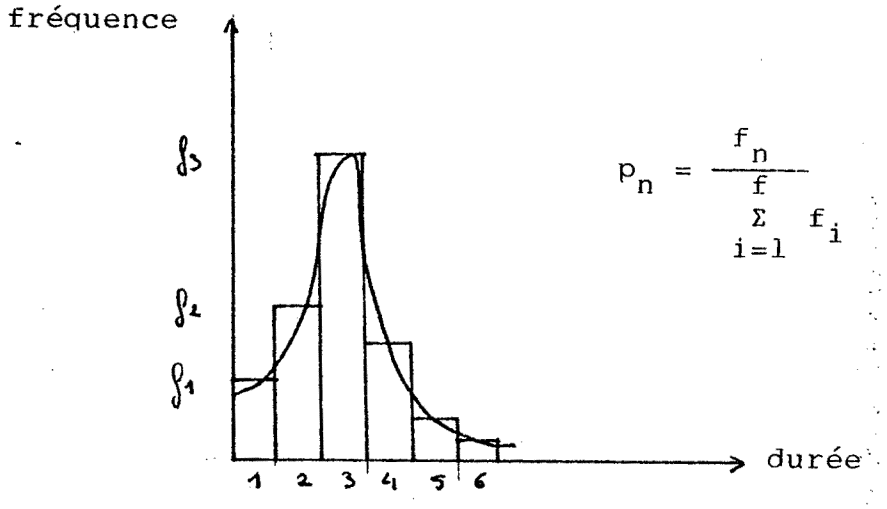
$$P_r(d) = \text{Prob (durée = d/durée} \leqslant d) = \frac{p_d}{p_1+p_2+..p_d}$$ (figure 4.2.9.

* calcul de la probabilité de continuité d'un état (p_c) :

$$p_c(d) = \text{Prob (durée} \geqslant d+1/\text{durée} \leqslant d) = \frac{p_{d+1} + \ldots p_f}{p_1 + p_2 + \ldots p_d}$$

(figure 4.2.9)

Bien entendu, $p_c + p_r = 1$



$$p_n = \frac{f_n}{\sum_{i=1}^{f} f_i}$$

figure 4.2.9. : histogramme des durées après lissage

4.2.2. Exemples de résultats

On présente ici quelques résultats de segmentations automatiques de mots prononcés par deux types de locuteur :

LOC1 : a participé à l'apprentissage préalable.

LOC2 : n'a pas participé à l'apprentissage .

Les mots enregistrés ne sont pas contenus dans le corpus initial. Ils y ont été introduits afin d'y ajouter des phonèmes peu fréquents dans le corpus présent.
(cf. chapitre 3.1.).

4.2.2.1. Résultats sur LOC1

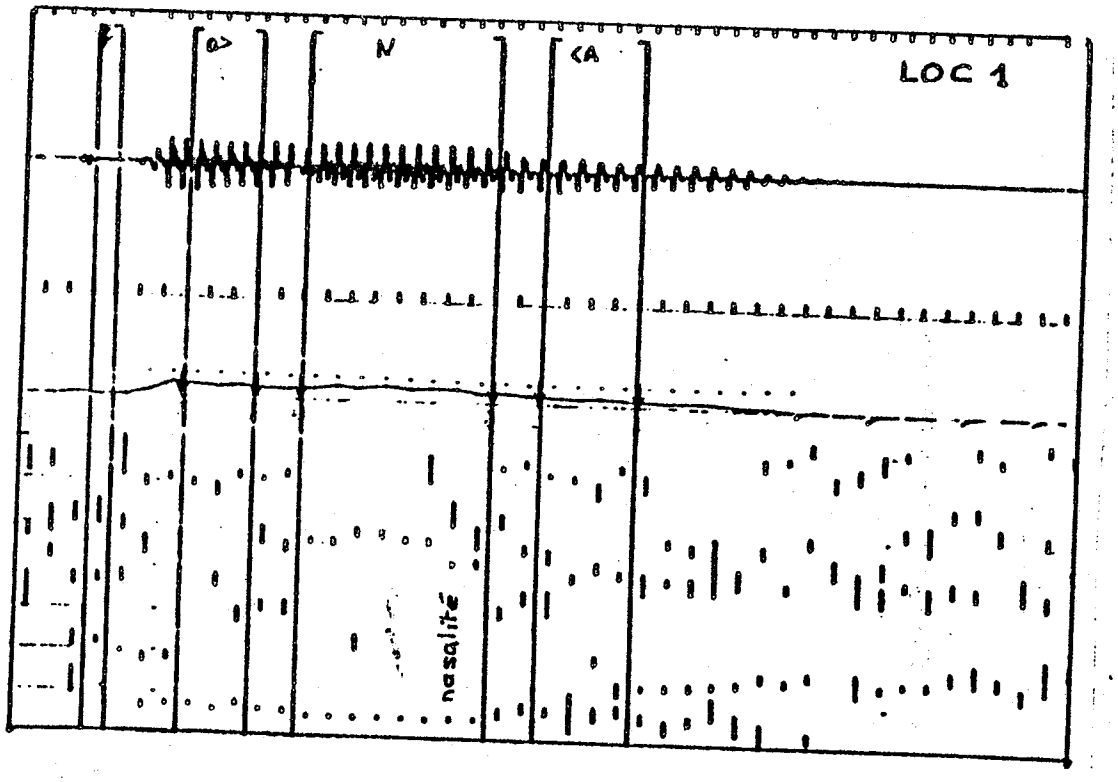
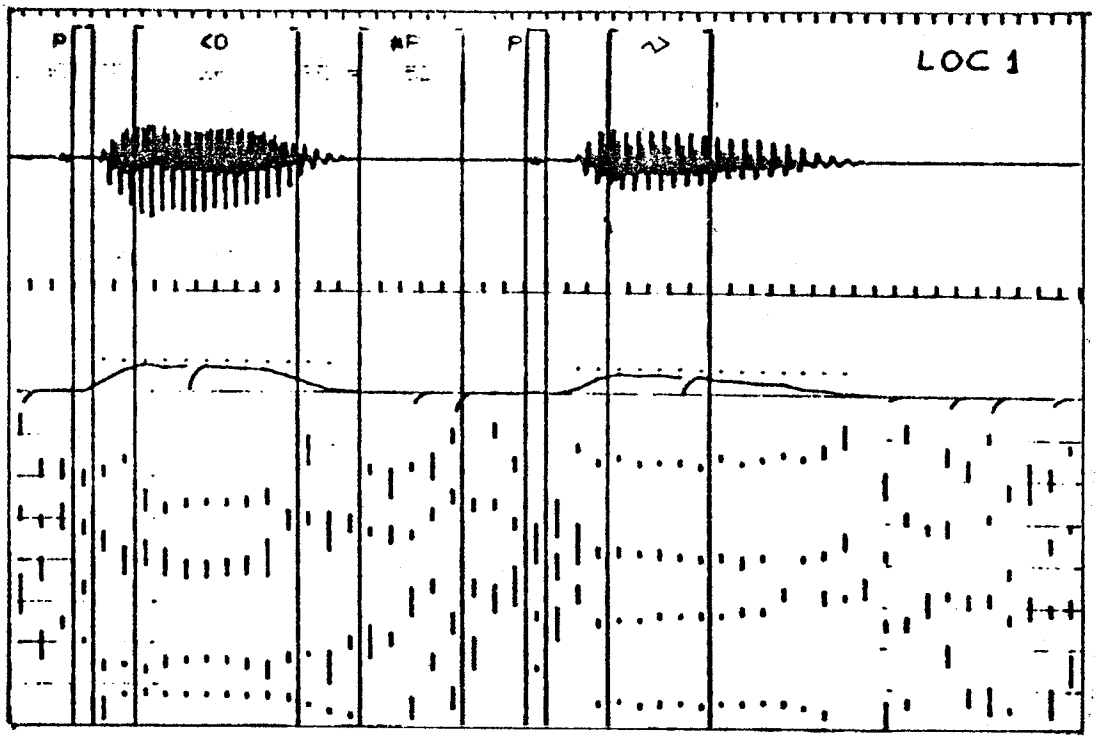
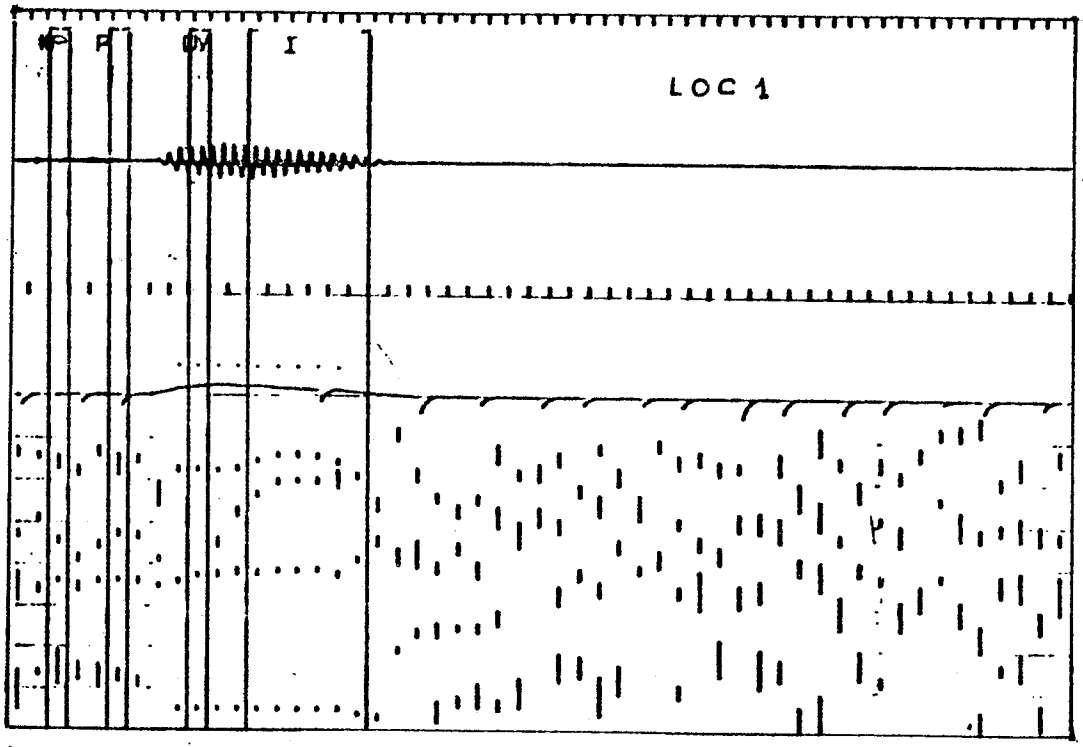


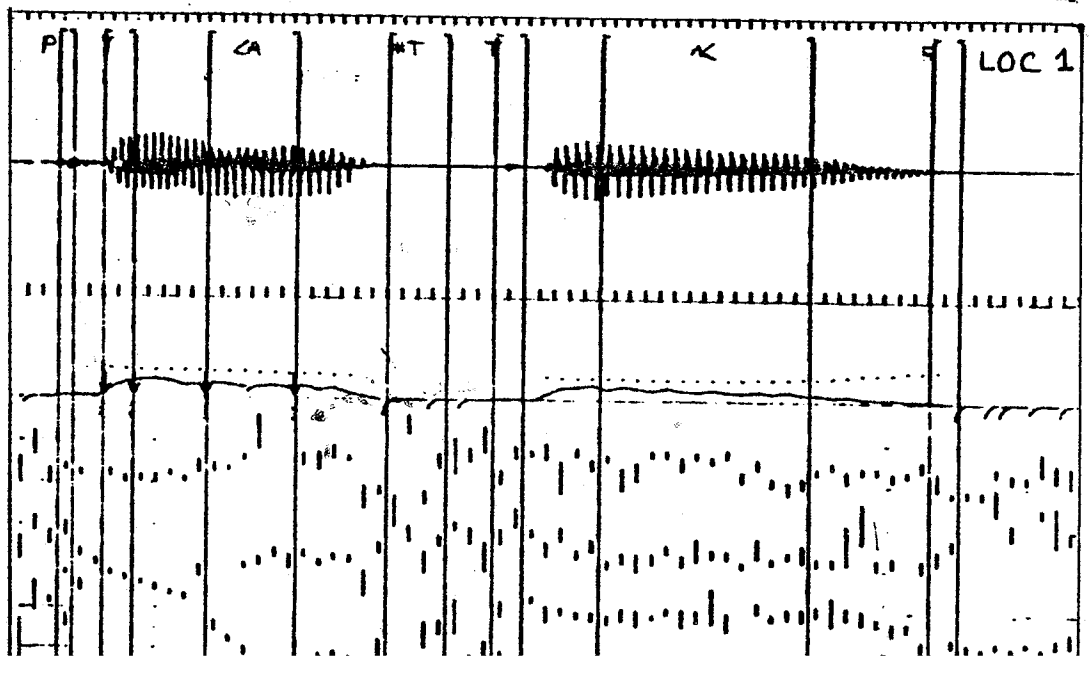
figure 4.2.10.a "ponant"



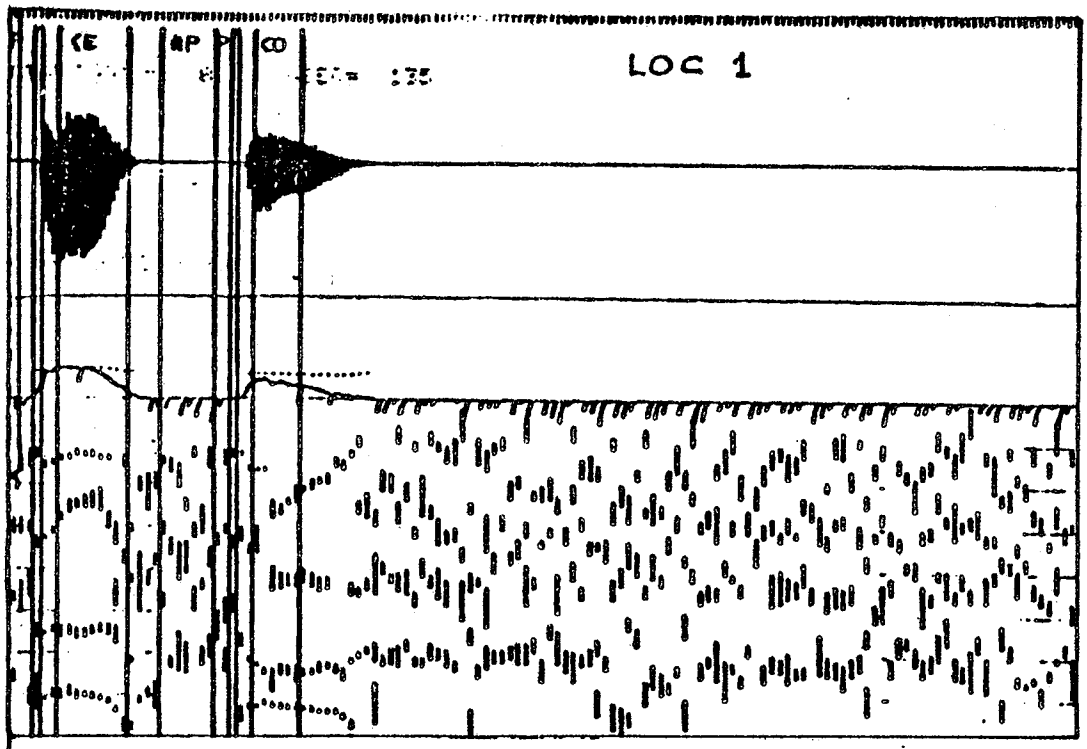
4.2.10.b
"pompeux"



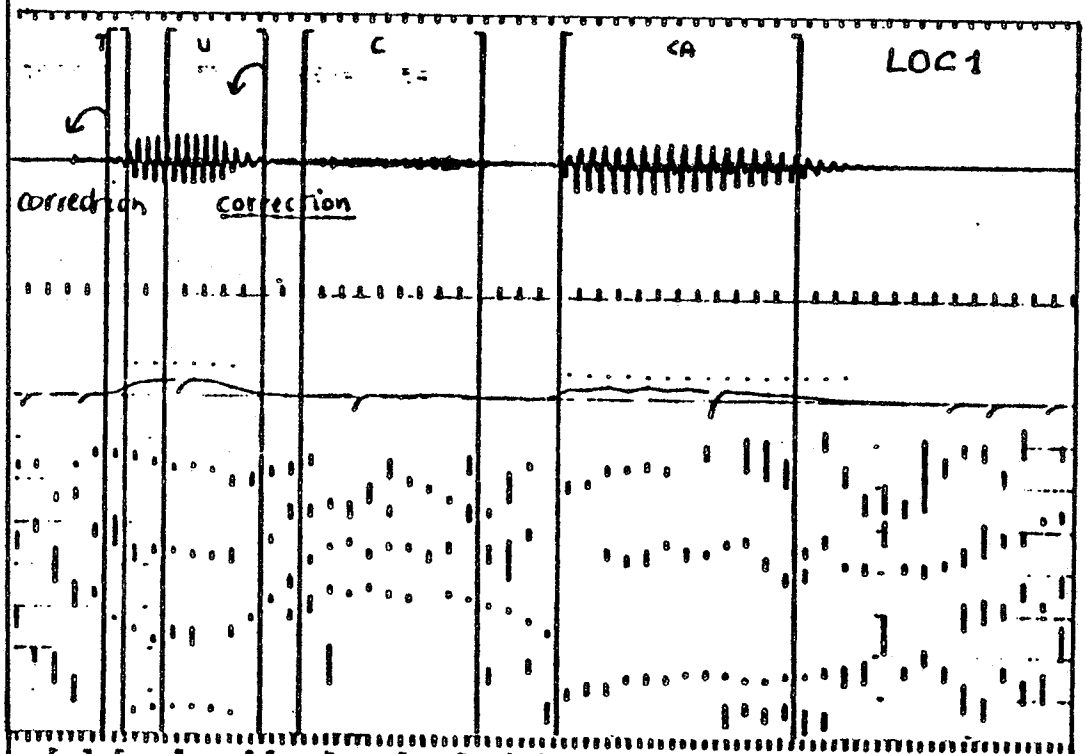
4.2.10. c
"puits"



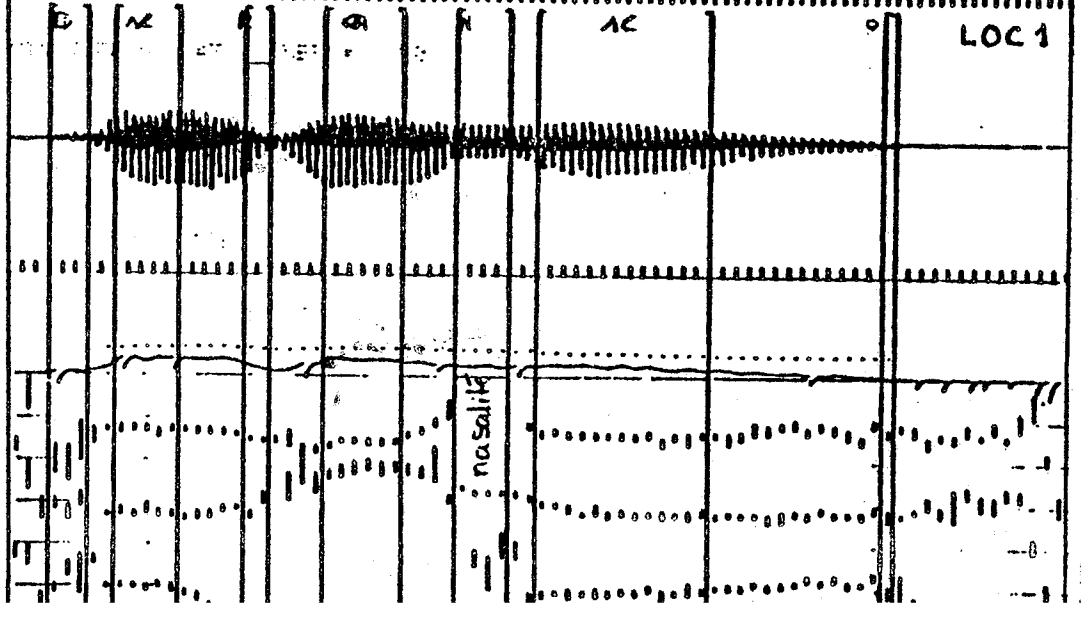
4.2.10.d
"puanteur"



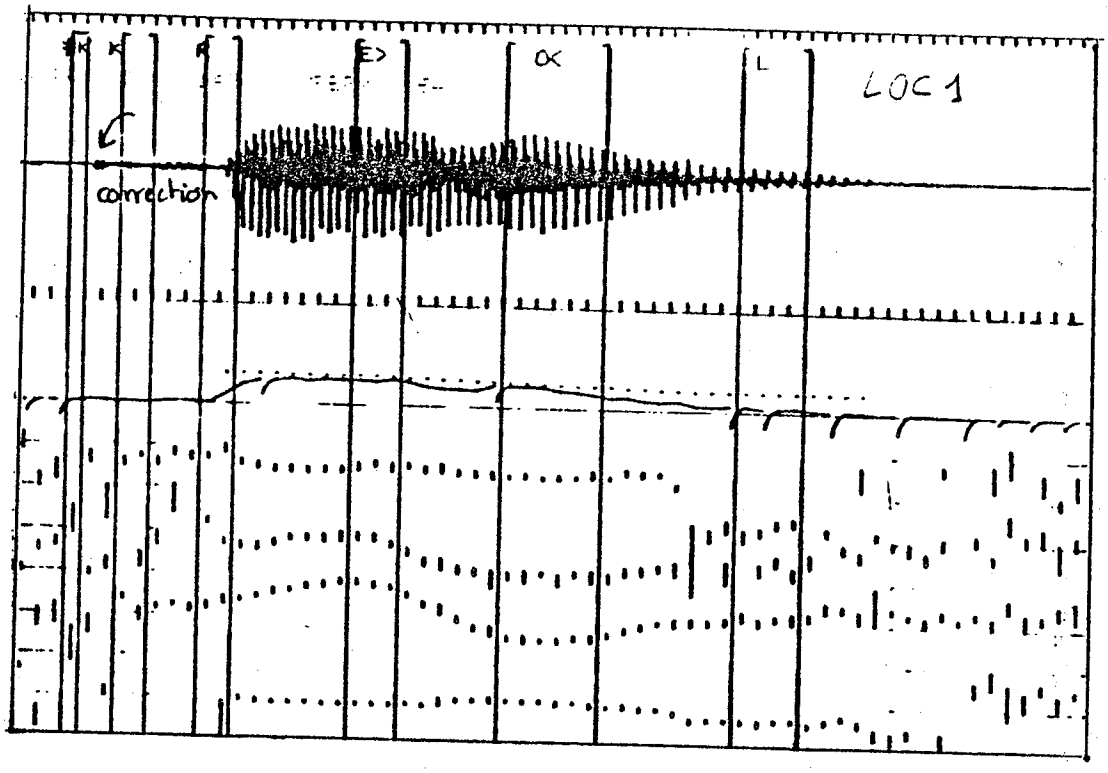
4.2.10.e
"pinpon"



4.2.10.f
"touchant"



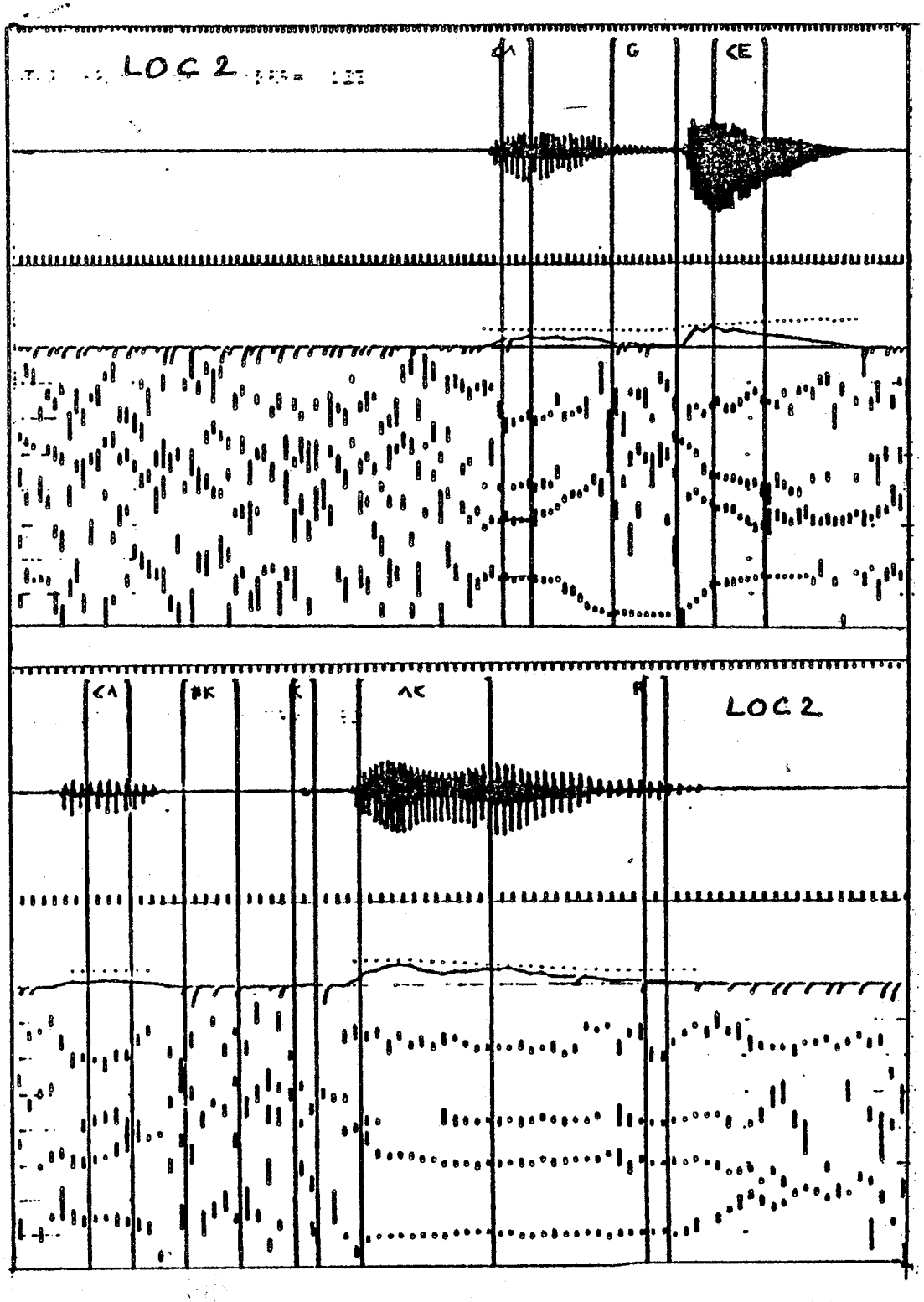
4.2.10.g
"d'heure en heure"



4.2.10. h
créole

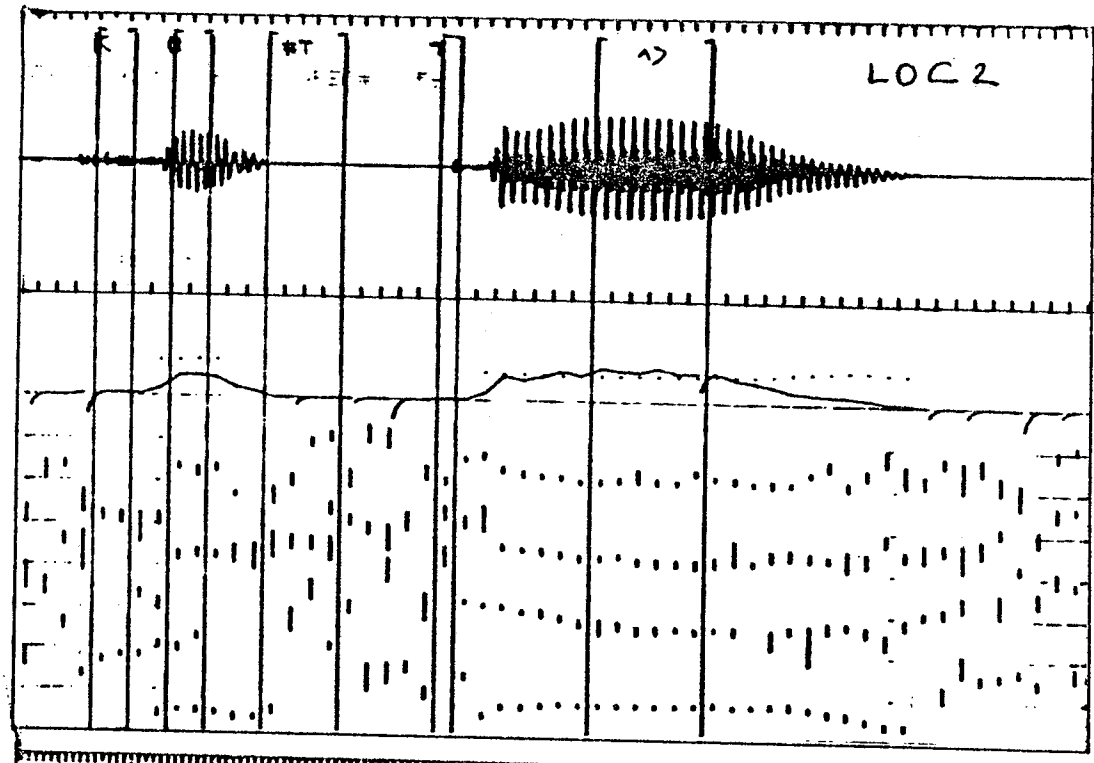
figure 4.2.10. Exemples de segmentation sur LOC1

Les résultats sont satisfaisants sur des mots courts à syllables bien séparées (4.2.10a, 4.2.10b). Sur un mot mal cadré (mauvaise détection de parole lors de la phase d'analyse, 4.2.10b), la segmentation effectue un recadrage du mot. En règle générale les phonèmes bien segmentés sont les fricatives sourdes (C dans "touchant"), les voyelles ainsi que les consomnes nasales avec une bonne prise en compte des formants de nasalité (4.2.10g, 4.2.10a). Les plosives sourdes de très courte durée sont souvent mal cadrées (4.2.10.f, 4.2.10.h) ce qui nécessite dans la majeure partie des cas une correction manuelle (voir chapitre "segmentation manuelle"). Les voyelles successives sont assez bien séparées (4.2.10.a, 4.2.10.c).

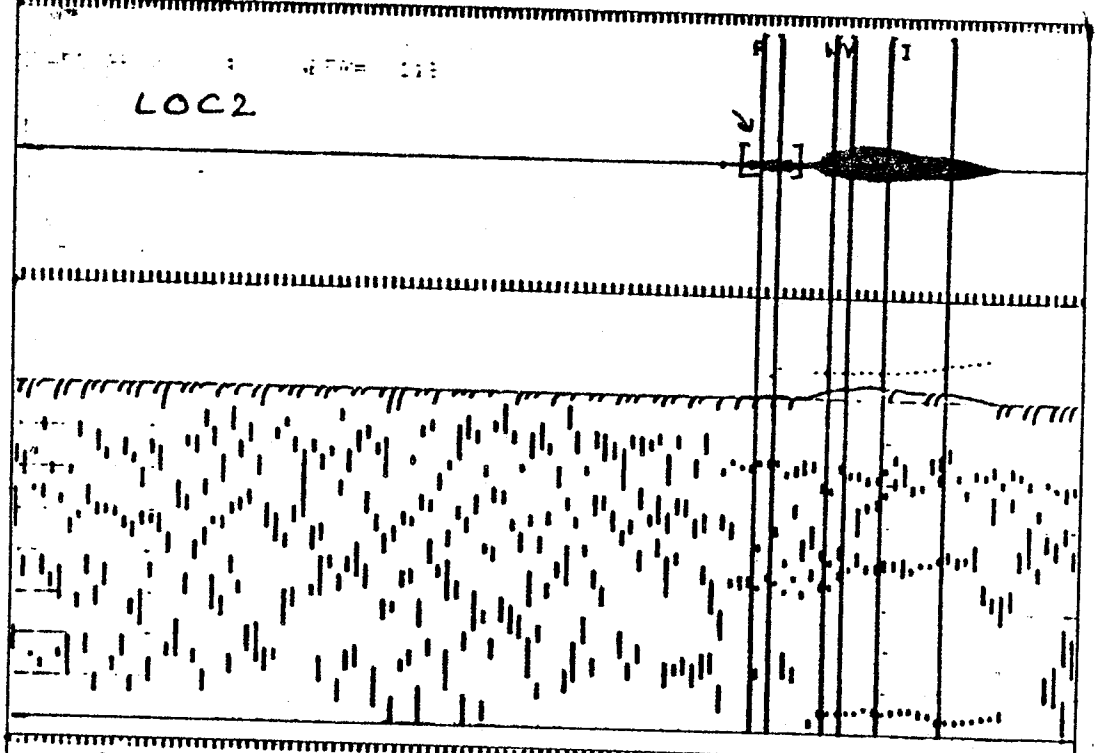


4.2.11.a
"un gain"

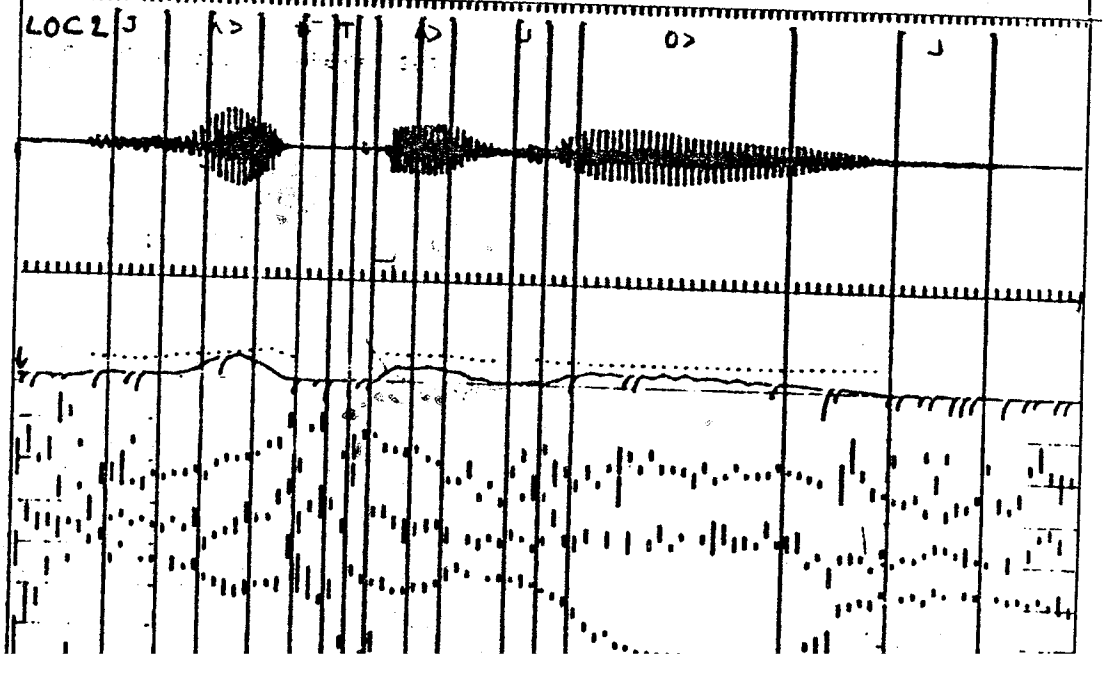
4.2.11.b
"un coeur"



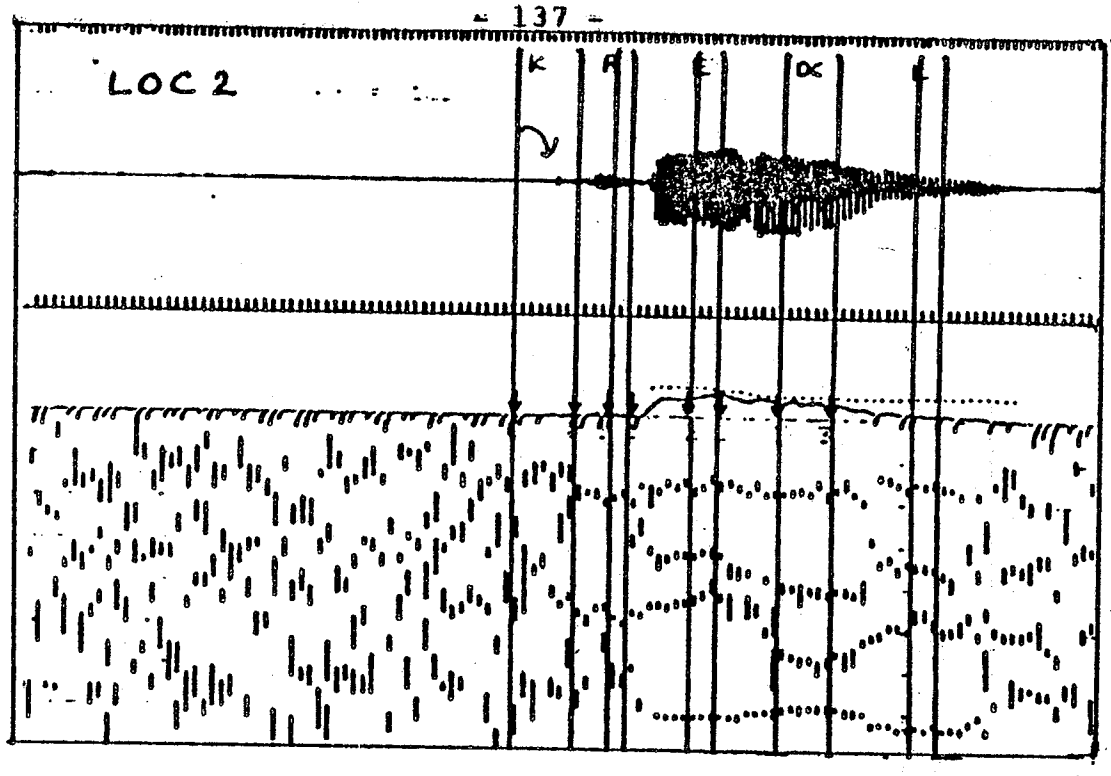
4.2.11. c
"couteux"



4.2.11. d
"puits"



4.2.11.e
"je te jauge"



4.2.11. f "créole"

figure 4.2.11 : résultats de segmentations sur LOC2

Les remarques sont à peu près les mêmes que celles faites sur LOC1 c'est-à-dire :

- bon codage du mot (4.2.11a, 4.2.11.d)
- stabilité des formants pour les voyelles (<A, <E dans 4.3.11.a, 1> , 0> dans 4.2.11.e)
- bonne séparation des voyelles adjacentes (4.2.11.f)
- difficulté de segmentation des plosives sourdes (4.2.11.f) avec souvent corrections nécessaires.

Cependant, le fait que LOC1 soit un locuteur connu (ayant participé à l'apprentissage) n'a pas d'effet sur la qualité des résultats.

II.3. CONCLUSION

La segmentation automatique de mots est un outil puissant pour l'apprentissage. Comme nous l'avons déjà souligné, elle permet une augmentation rapide du corpus de donnée de taille nécessairement importante pour un système multilocuteur Cependant, son utilisation ne peut se faire que lorsque le corpus est jugé assez consistant pour donner des statistiques cohérentes (étiquette phonétique des modèles, statistiques sur les durées). C'est pourquoi, nous n'avons exploité le procédé que durant la deuxième phase de notre apprentissage et de plus sur des phonèmes assez rares dans le corpus présent (puisque l'opération consistait à fournir au corpus des phonèmes peu fréquents : plosives sourdes,semi-voyelles etc ...).

4.3. RECONNAISSANCE PAR PHONEME

4.3.1. Rappel du principe : (figure 4.3.12)

On rappelle les étapes de notre méthode :

- le mot inconnu est enregistré, échantillonné puis numérisé. Les dix premiers coefficients cepstraux représentent chaque intervalle de 12.5ms. La suite constituée par ces vecteurs donne une représentation numérique du mot inconnu.

- On recherche pour chacun de ces vecteurs le modèle le plus proche dans le dictionnaire construit lors de la phase d'apprentissage. Pour cela on utilise la distance cepstrale (chapitre "distances"). Chaque élément de la suite des vecteurs est donc codé suivant le meilleur modèle et une table de probabilités (système d'information de type phonétique) lui est associé.

- la phase de reconnaissance proprement dite commence alors; le vocabulaire des mots à reconnaître est écrit sous forme phonétique avec une ou plusieurs transcriptions possibles par mot (variantes dues à la prononciation suivant les locuteurs). La suite des modèles est identifiée à chacun des mots du vocabulaire par programmation dynamique. Le mot

reconnu est celui qui correspond au mot le plus probable.

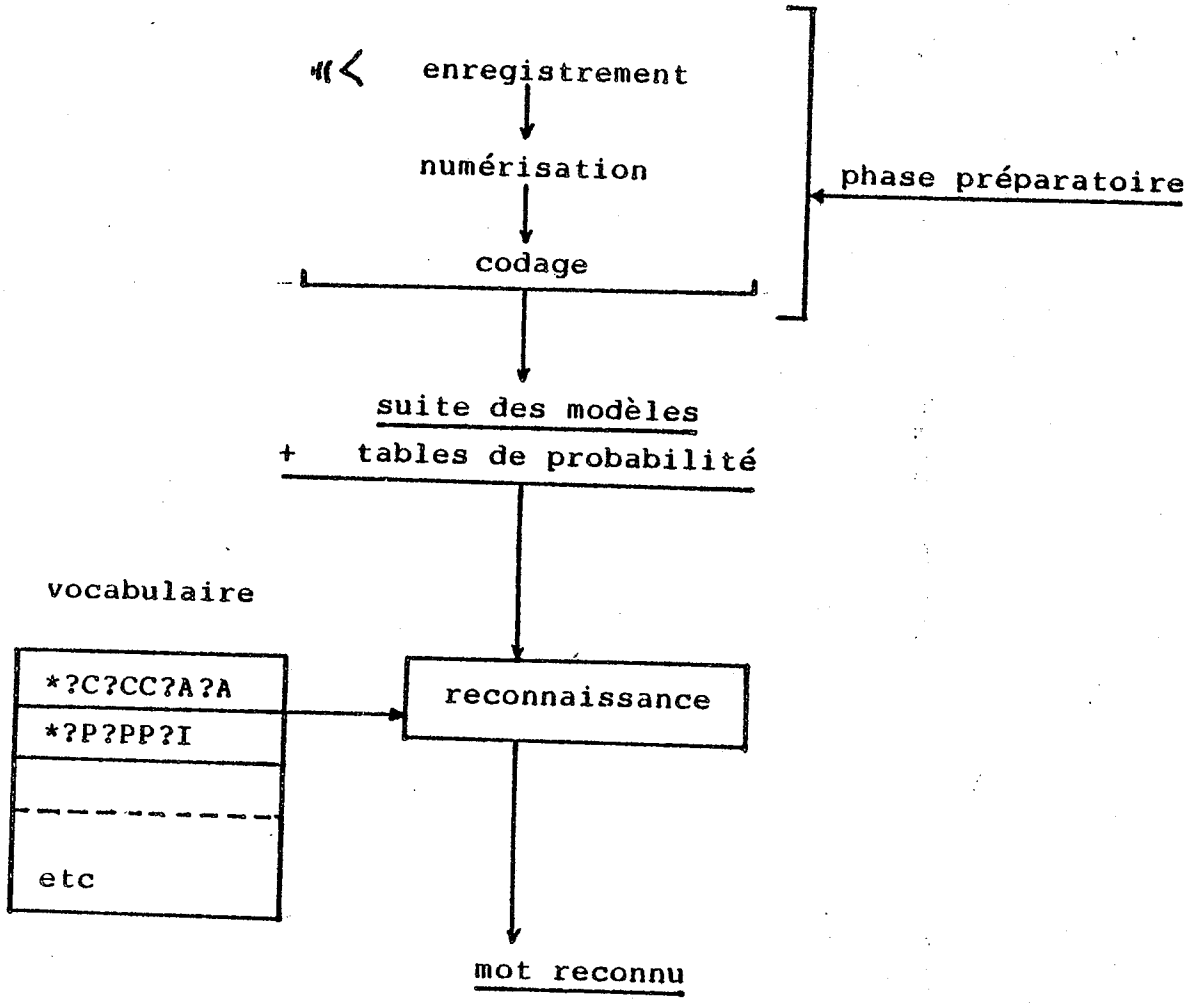


figure 4.3.1.2. : méthode de reconnaissance par phonème

4.3.2. Quelques tests de reconnaissance :

Les performances d'un système de reconnaissance sont difficiles à évaluer ; et pourtant, de nombreux fabricants annoncent des taux variant entre 95 et 100%. De tels pourcentages paraissent satisfaisants mais souvent il est difficile d'interpréter les résultats obtenus en détail. En effet, les conditions relatives aux mesures sont rarement précisées (type et nombre de locuteurs utilisés, conditions d'enregistrement, vocabulaire).
C'est pourquoi, nous avons effectué nos tests sur deux types de vocabulaires : un ensemble de séries minimales et quelques vocabulaires choisis pour d'éventuelles applications. De ce fait, l'utilisation des séries minimales (ensemble de mots de signifiés différents se distingant par une unité phonétique) nous a permis de mesurer l'aptitude du système à discriminer des sons très voisins, et cela dans la mesure où l'oreille humaine peut être amenée à le faire. Les tests réalisés sur des vocabulaires courants (contenant des mots phonétiquement assez différents) peuvent être exploités dans des cas plus précis pour des systèmes conçus pour certaines applications (ensemble des chiffres par exemple).

Quatre locuteurs ont participé à ces tests. On les notera LOC1, LOC2, LOC3, LOC4. Deux d'entre eux (LOC1 et LOC2) ont contribué à l'apprentissage préalable du système. Pour des raisons de temps, seulement LOC1 et LOC3 ont été enregistrés pour les tests relatifs aux paires minimales. Chacun des vocabulaires a été prononcé trois fois pour éliminer la variabilité intra-individuelle. Ce nombre nous paraît insuffisant si l'on se réfère aux test de reconnaissance réalisés par G. CHOLLET, où chacun des mots a été répété sept fois. C'est pourquoi, il est nécessaire de souligner que les résultats présentés ici ne sont ni assez nombreux ni assez consistants pour pouvoir donner des statistiques cohérentes. Chacun sait que les tests et les comparaisons de systèmes de reconnaissance représentent un lourd travail et qu'il est bon d'automatiser les méthodes d'exploitation afin de pouvoir diversifier facilement les tests. Aussi, le type de stratégie proposée ici pourra être pris en compte pour une étude ultérieure destinée à l'évaluation proprement dite du système.

4.3.2.1. Tests sur les paires minimales

Nos expériences ont été fortement inspirées de celles réalisées par G. CHOLLET. Elles consistent à faire répéter un certain nombre de séries minimales par différents locuteurs. Le choix de ces dernières essaie de tenir compte des variétés intra-phonémiques dépendantes du contexte. En voici la description :

- Tests sur les voyelles :

1) Série P Ⓥ R : voyelle en position syllable finale recouverte

2) Série L Ⓥ : voyelle en position syllabe finale absolue.

- Tests sur les consonnes

1) Série Ⓒ E > : consonne devant voyelle.

2) Série A Ⓒ O> : consonne inter-vocalique

3) Série A R Ⓒ : consonne en position syllabe finale absolue.

D'autres contextes peuvent être envisagés [CHOLLET] en particulier, pour les consonnes (K A Ⓒ O >, Ⓒ W U A ... etc).

Le but de ces tests n'est pas de fournir des résultats en terme de reconnaissance de vocabulaires. G. CHOLLET propose une analyse des confusions rencontrées au niveau des traits phonétiques. Il distingue quatre classes de confusions chez les voyelles (ouvert/fermé, grave/aigu, nasal/non nasal, diésé/bémolisé) et six chez les consonnes (vocalique/non vocalique, voisé/non voisé, nasal/non nasal, interrompu/non interrompu, grave/aigu, compact/diffus). Les résultats ont été comparés avec ceux obtenus avec des auditeurs humains. Nous les regroupons dans les figures qui suivent :

feature \ system	Human	A	B
open/close	09%	22%	06%
plain/flat	15%	22%	06%
nasal/non nasal	33%	50%	67%
grave/acute	42%	07%	20%

Table I: Analysis of confusions with vowel test; A and B are two A.S.R. systems.

figure 4.3.13. Résultats pour les voyelles d'après [CHOLLET]

feature \ system	Human M.N.	Human C.A.R.	C	D	E
vocalic/non vocalic	00%	00%	00%	04%	15%
voiced/unvoiced	00%	00%	00%	27%	13%
nasal/non nasal	00%	00%	00%	19%	12%
interrupted/non int.	20%	35%	7,9%	08%	30%
acute/Grave	23%	34%	06%	13%	05%
compact/diffuse	57%	27%	14%	29%	24%

Table II: Analysis of confusions with consonant test; M.N.=Miller and Nicely <6>, C.A.R.=Chollet Astier Rossi, C D E three A.S.R. systems.

figure 4.3.14. Résultats pour les consonnes d'après [CHOLLET]

Les expériences ont donc été réalisées sur cinq séries minimales. Dans la série L Ⓥ, la voyelle se situe en position finale absolue. Dans ce cas, on n'observe pratiquement que des réalisations fermées. On dit, [Henriette WALTER] que l'opposition ouvert/fermé est neutralisée dans cette position. Cependant, on signale que l'opposition E>/E< est encore présente chez un grand nombre de locuteurs. Rappelons aussi que le système phonologique de la langue française oppose les deux réalisations du phonème A en position antérieure et postérieure mais que cette différence

n'est pas présente dans la liste que nous utilisons (chapitre 3). La série L V fait aussi intervenir les voyelles nasales. Elle comporte donc douze éléments :

LU, LI, LÆ>, LA, LO>, LY, LE>, LE<, L<A, L<E, L<Æ, L<O

En position en syllabe finale couverte, les réalisations les plus fréquentes sont ouvertes. Mais c'est surtout la position devant le R final qui a une influence courante sur les voyelles. L'opposition O</O> trouvée dans les mots "saule" et "sole" est aussi neutralisée devant un R final. Il en est de même pour l'opposition Æ</Æ>. La série P Ⓥ R comporte donc 7 éléments :

PUR, PIR, PÆ<R, PAR, PO<R, PYR, PE<R.

En ce qui concerne les consonnes, les trois séries présentées font intervenir les plosives non voisées (P,T,K), les plosives voisées (B ,D, G), les fricatives voisées (V, Z, J) les fricatives non voisées (F,S,C) les nasales (N,M) les liquides (L,R). On trouve donc seize éléments dans les séries Ⓒ E> et Æ Ⓒ O> et quinze dans la série A R Ⓒ. (sans R).

RESULTATS :

Le but de cette étude n'est pas de donner des résultats en terme de vocabulaire. Nous avons néanmoins calculé les pourcentages de reconnaissance obtenus pour l'ensemble des deux locuteurs (LOC1 et LOC3) : on a observé 34.2% de confusions pour les voyelles et 46% pour les consonnes. Nous avons détaillé les confusions rencontrées dans les tables ci-dessous :

	LOC1	LOC2	Total
U	5/6	2/6	7/12
I	0/6	0/6	0/12
1	0/3	1/3	1/6
1	2/3	2/3	4/6
A	6/6	5/6	11/12
0	1/3	1/3	2/6
0	0/3	0/3	0/6
Y	0/6	0/6	0/12
E	1/3	0/3	1/6
E	0/6	2/6	2/12
A	1/3	1/3	2/6
E	1/3	0/3	1/3
1	2/3	3/3	5/6
0	0/3	3/3	3/6

figure 4.3.15. Nombre de confusions pour les voyelles.

	LOC1	LOC3	Total
P	4/9	4/9	8/18
T	5/9	7/9	12/18
K	3/9	2/9	5/18
B	5/9	6/9	11/18
D	5/9	6/9	11/18
G	3/9	6/9	9/18
F	4/9	7/9	11/18
S	4/8	6/9	10/17
C	0/8	0/9	0/17
V	3/8	7/9	10/17
Z	2/9	5/9	7/18
J	2/9	3/9	5/18
M	8/9	5/9	13/18
N	2/9	1/9	3/18
L	3/9	6/9	9/18
R	3/6	1/6	4/12

figure 4.3.16. Nombre de confusions pour les consonnes

Ces tables permettent de classer les phonèmes. Pour les voyelles, on peut dire que les résultats sont très satisfaisants pour I, Y, O< puisque aucune confusion n'est rencontrée. Les phonèmes E>, Λ>, E<, <A, <E, O> sont aussi assez bien reconnus. En revanche les phonèmes <O, U, Λ <, <Λ présentent une instabilité notable. Mais le cas le plus critique est celui du A pour lequel le taux de confusion est très important. En ce qui concerne les consonnes, on remarquera indiscutablement une très bonne reconnaissance de C ainsi que des résultats acceptables pour les consonnes N, K, J, R, Z, P L, G. Par contre les confusions de B, D, F, S, V, T, M sont rencontrées très fréquemment.

A quel niveau peut-on interpréter ces résultats ? Tout d'abord on doit noter que les phonèmes confondus arrivent souvent en deuxième ou troisième position par rapport aux autres phonèmes candidats. En d'autres termes, on pourrait examiner toutes les paires minimales constituées à partir de ces séries et le pourcentage des paires confondues est assez faible puisqu'il est en moyenne de 7.5% pour les voyelles et de 10% pour les consonnes. Cela permet donc de mieux cerner le problème, et cela pour chacun des cas. Plus précisément, en particulier pour les consonnes, les confusions se manifestent à l'intérieur d'une même classe. On en distingue 6 : les nasales, les liquides, les fricatives voisées ou non voisées, les plosives voisées ou non voisées. Pour les voyelles, ce phénomène est visible dans les nasales (instabilité <Λ/<E, <O/<A). Et donc, ce sont donc des sons très voisins (souvent qui ne diffèrent que par un trait phonologique) qui sont amenés à être confondus. Pour fixer les idées, on présente ci-dessous le système phonologique de la langue française en séparant les voyelles (-consonnantique) des consonnes (+ consonnantique) où la détermination de certains traits caractéristiques suffit pour décrire un phonème. Notre méthode de reconnaissance ne procède pas à l'identification des traits. C'est pourquoi il est difficile d'affirmer que la confusion d'un phonème implique la non reconnaissance d'un trait le caractérisant de manière pertinente.

Cependant, et cela est dû à la répétition des observations on peut donner un aperçu de l'aptitude du système à reconnaître tel ou tel caractéristique d'ordre phonologique (acuité, voisement, ... etc) en précisant éventuellement le contexte (parcours de l'arbre).

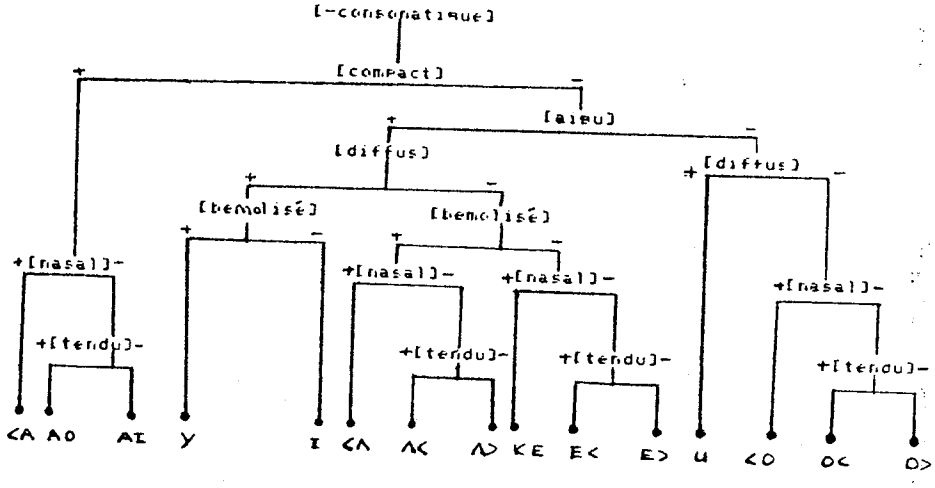


figure 4.3.15. Système phonologique pour les voyelles [-consonantiques] et les consonnes [+consonantiques]

Cas des voyelles

On étudie les traits suivants sur une liste de paires caractéristiques.

- tendu/relaché : E</E>
- nasal/non nasal : <A/A <Λ/Λ> <Λ/Λ< <0/0> <0/0< <E/E< <E/E>
- bémolisé/non bémolisé : Y/I <Λ/< E Λ</E< Λ>/E>
- diffus/non diffus : Y/<Λ, Y/Λ< Y/Λ> I/<E I/E< I/E> U/<0 U/0< U/0>.
- compact/non compact : comparaison de A avec toutes les autres voyelles sauf <A.
- aigu/grave : Y/U I/U <Λ/<0 <E/<0 Λ>/0> Λ</0< E>/0> E</0<

Pour certains traits, on effectue une comparaison asymétrique (compact/non compact, nasal/non nasal) car les confusions rencontrées sont beaucoup plus fréquentes dans ce sens. Nous regroupons les résultats dans le tableau ci-dessous :

tendu/relache relache/tendu	12	1	8.3 %
nasal/non nasal	30	4	13.3 %
bemolise/non bemolise non bemolise/bemolise	60	7	11.6 %
diffus/non diffus non diffus/diffus	108	8	7.4 %
compact/non compact	96	21	21.8 %
aigu/grave grave/aigu	120	6	5.0 %

figure 4.3.16. nombre de tests effectués et confusions rencontrées pour les voyelles

La détection de la nasalité se fait assez mal dans le cas du <Λ puisque sur six comparaisons effectuées, 3 phonèmes ont été confondus avec Λ>. Le trait de bémolisation présente aussi une certaine fragilité lors de l'analyse de la paire <Λ/<E et cela tout particulièrement chez le locuteur LOC2 d'origine méridionale (réalisation de <Λ avec absence de l'arrondissement des lèvres) En ce qui concerne la diffusion, on remarque une certaine instabilité de l'opposition U/0< notamment dans la position finale recouverte. Mais le cas le plus pertinent est celui de la voyelle A dont le trait caractéristique est la compacité. Les confusions observées proviennent de mauvaises discriminations des paires A/<Λ, A/<E A/U A/0< A/Y. Il est donc légitime de penser qu'un problème se pose au niveau de la compacité.

Cas des consonnes :
On mesure les oppositions suivantes : (figure 4.3.17)

voisé/non voisé : J/C Z/S V/F D/T B/P K/G

aigu/grave : Z/V S/F D/B T/P M/N M/L

compact/non compact : J/Z J/V C/S C/F G/D G/B K/T K/P

continu/non continu : J/G K/C Z/D S/T V/B P/F

vocalique/non vocalique : R/J L/Z M/V L/D N/D N/Z M/B R/G.

voise/non voise	216	5	2.3%
aigu/grave	216	30	13.8%
compact/non compact	285	35	12.2%
continu/non continu	216	15	6.9%
vocalique/non vocalique	252	6	2.3%

figure 4.3.17. Cas des consonnes - nombre des tests et confusions observées

On peut admettre qu'aucun problème ne se pose au niveau des traits voisé/non voisé, continu/non continu, vocalique/non vocalique. Par contre le trait d'acuité n'est pas pertinent pour les paires M/N D/B puisque sur 36 tests on observe respectivement 14 et 9 mauvaises discriminations. Pour les nasales, elle est pratiquement due à la confusion de M par rapport à N (13 confusions sur 18 tests). Pour la compacité on remarque la confusion des paires K/T (9 mauvaises discriminations sur 18) G/D, G/B, C/S, C/F, et K/P. Mais ce qui est plus caractéristique c'est qu'il semble que ces confusions n'agissent que dans le sens. (non compact → compact), les phonèmes G, C, K étant assez bien reconnus (figure 4.3.16). On peut donc admettre (et c'est le cas pour les voyelles) que la compacité ne donne pas des résultats satisfaisants. Aussi, des mesures supplémentaires portant directement sur le spectre permettraient d'améliorer les pourcentages. Ces indications pourraient porter sur le positionnement des formants (trait tendu/relaché, grave/aigu), l'énergie ou la variation d'énergie en certains points (trait bémolisé/diésé) ou la variation du spectre dans la zone centrale (trait compact/non compact).

4.3.2.2. Tests sur les vocabulaires

On donne des résultats sur quatre vocabulaires :

- VOCABULAIRE 1 : "oui, non"
- VOCABULAIRE 2 : "vrai, faux"
- VOCABULAIRE 3 : "zéro, un, deux, trois, quatre, cinq, six, sept, huit, neuf"
- VOCABULAIRE 4 : "blanc, noir, rouge, bleu, jaune"

a) nécessité de plusieurs transcriptions phonétiques par mot

Des premières mesures ont été obtenues à partir des mots du vocabulaire avec une seule représentation phonétique. (celle qui nous paraissait être la plus courante). En fait, on s'est aperçu au cours des essais que les réalisations étaient différentes suivant les locuteurs. Le cas le plus typique est celui

"un" à qui on peut associer soit le phonème <Æ, soit le phonème <E. On notera aussi la réalisation du "0" de "Jaune" donnant un "0" ouvert ou un "0" fermé. Il en est de même pour le "e" muet fréquemment positionné en fin de mot surtout pour les locuteurs d'origine méridionale.

Bien sûr, l'utilisation de plusieurs transcriptions phonétiques alourdit considérablement la méthode qui consiste à effectuer une comparaison avec chacun des mots du vocabulaire. Mais cela permet aussi de résoudre le problème de variabilité posé lors du passage d'un locuteur à l'autre (ou du moins les variantes de type articulatoire).
Nous donnons ci-dessous la liste des vocabulaires avec les transcriptions phonétiques correspondantes :

oui	WUI	vrai	VRE>, VRE<
non	N < 0	faux	F0>
zero	ZE>RO>		
un	<Æ,<E		
deux	DÆ>		
trois	TRWUA		
quatre	KA *TTR, KA *TTRÆ		
cinq	S<E *KK, S<E *KKÆ		
six	SIS		
sept	SE< *TT		
huit	WUI *TT, WYI *TTÆ		
neuf	NÆ<F		
blanc	BL<A		
noir	NWUAR		
rouge	RUJ		
bleu	BLÆ>		
jaune	J0<N, J0<N.		

b) résultats

Sur les quatre locuteurs utilisés pour ces tests, on distingue ceux qui ont contribué à l'apprentissage (LOCA) de ceux qui n'ont pas participé à l'apprentissage (LOCB) et nous organisons nos résultats en fonction de cela (figure 4.3.). Chacun des mots a été prononcé 3 fois.

	LOCA	LOCB	TOTAL	
OUI/NON	12/12	11/11	33/33	100%
VRAI/FAUX	12/12	11/11	33/33	100%
CHIFFRES	43/59	47/60	90/119	75.6%
COULEURS	28/30	28/30	56/60	93.3%

figure 4.3. Taux de reconnaissance

4.3.3. Conclusion

Nous nous sommes efforcés de poursuivre les objectifs fixés au départ, c'est-à-dire :

- assurer l'indépendance du locuteur.
- reconnaître des vocabulaires variés et de longueur finie.

Dans un premier temps, les quelques tests que nous avons réalisés témoignent de la validité du procédé, qui consiste à associer à un segment de parole inconnu le meilleur modèle dans la base des références phonétiques. En effet, on constate que la séquence des modèles caractérisant un message inconnu après codage vectoriel, présente un certain nombre de paliers. La suite de ces derniers permet le découpage phonétique du mot. De plus les modèles donnent dans la majeure partie des cas, une bonne caractérisation du phonème leur correspondant (par l'intermédiaire des tables des fréquences phonétiques). Un autre point à souligner est la faible variabilité intra-locuteur lorsque l'on passe d'une occurrence à l'autre : le profil du mot déterminé après codage vectoriel est pratiquement inchangé. Le même phénomène s'observe lorsque l'on change de locuteur et l'identification des formes inconnues ne s'avère pas plus difficile pour un locuteur n'ayant pas participé à l'apprentissage. La stabilité du système paraît donc assurée.

La méthode fonctionne très bien pour des vocabulaires de petite taille et contenant des mots assez différents. Bien sûr, les tests effectués sur un ensemble de mots phonétiquement voisins (séries minimales) fournissent des performances beaucoup moins élevées. Mais c'est bien sûr dans ce contexte de mesures qu'il faut se situer pour améliorer la discrimination entre deux réalisations très voisines : par exemple, par la recherche de paramètres supplémentaires (énergie, variation de l'énergie, du spectre, écartement des formants, ... etc) ou par l'établissement d'une distance plus appropriée.

Les résultats obtenus sont de toute façon très encourageants. Ils prouvent la cohérence de la méthode, et ils valident les hypothèses énoncées (indépendance du locuteur, variétés des vocabulaires). En ce sens, nous espérons que ce mémoire pourra servir d'ouvrage de référence au lecteur intéressé ou simplement curieux ... à moins qu'il ne contribue de manière efficace à l'avancement des recherches dans le domaine du traitement de la parole.

A N N E X E S

A N N E X E I

GENERALITES SUR LE TRAITEMENT DU SIGNAL

1. GRANDEURS REPRESENTATIVES DES SIGNAUX

1.1. Représentation directe

La représentation directe d'un signal résulte de la connaissance du couple (t, x(t)) où x(t) est une fonction calculable ou npn calculable de t. En pratique, la détermination de x(t) est impossible. La représentation graphique donnée par le tracé des couples (t, x(t)) permet de visualiser le signal à étudier.

1.1.1. Un exemple de représentation directe pour un signal sinusoïdal :

Un signal sinusoïdal peut être commodément représenté par le nombre complexe :

$$z(t) = Ae^{j(\omega t + \varphi)}$$

ω est la pulsation du signal

$\nu = \frac{\omega}{2\pi}$ est la fréquence

φ est la phase.

1.1.2. quelques définitions :

périodicité d'un signal : un signal est dit périodique si il existe T tel que : $x(t + T) = x(t)$. Un signal périodique est donc entièrement représenté sur un intervalle de temps égal à sa période.

énergie d'un signal sur $[T_0, T_1]$

un signal est dit d'énergie finie si :

Il existe k tel que :

$$\int_{T_0}^{T_1} |x(t)|^2 \, dt < K$$

Dans ce cas, l'énergie e d'un signal est :

$$e = \int_{T_2}^{T_1} |x(t)|^2 \, dt$$

La courbe d'énergie d'un signal est utile dans certaines études. On a par exemple remarqué que les paliers d'énergie correspondaient aux parties stables du signal (phonèmes) (figure 1). Les zones correspondant au silence sont des zones d'énergie plus faible.

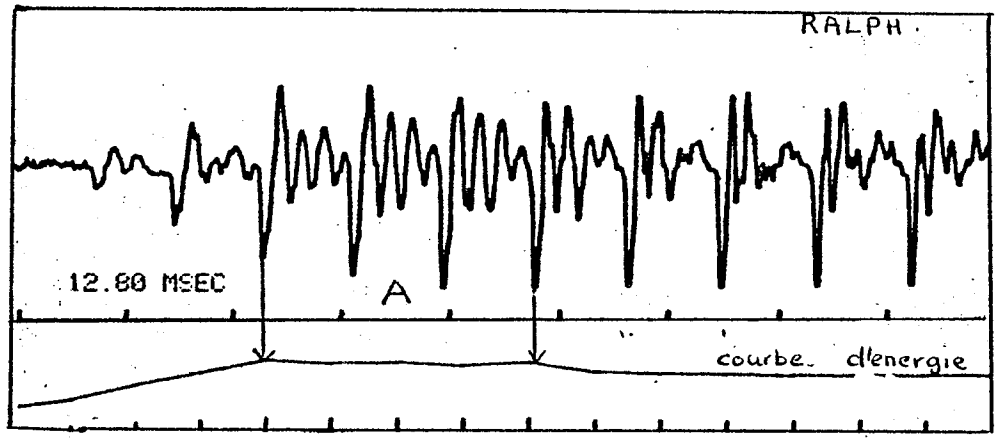


figure 1 : courbe d'énergie d'un signal.
Ici, la partie stable de la courbe correspond au phonème A de RALPH

1.1.3. Conclusion

La représentation (t, x(t)) est difficilement exploitable. Des méthodes plus sophistiquées permettent d'avoir davantage de renseignements sur la nature physique du signal.

1.2. Représentation codée

Soit $h_1, h_2, \ldots, h_n$ une série de n fonctions linéairement indépendantes. C'est-à-dire :

$$\sum_{m=1}^{n} (a_m h_m) = 0 \qquad a_m = 0 \quad \text{pour } m = 1 \ldots n$$

Sous cette condition, le signal x(t) peut être représenté par une décomposition unique sur chacune des fonctions h_m, $m = 1 \ldots n$.

$$x(t) = \sum_{m=1}^{n} (x_m h_m(t))$$

La suite x_m, m=1...n donne une représentation numérique du signal
De plus si la base $\{h_m\}$ est orthogonale c'est-à-dire si :

$$\langle h_i, h_j \rangle = \int h_i(t) \overline{h_j}(t)\, dt = \delta_{ij}$$

où $\delta_{ij} = 1$ si $i=j$ 0 sinon.

alors : $x_i = \langle x, h_i \rangle \quad i=1,n$

Exemple : Série de Fourier

On prend : $h_n(t) = e^{2i\pi nt/T}$

d'où
$x(t) = \Sigma x_n e^{2i\pi nt/T}$

et : $x_n = \frac{1}{T} \int_{[-\frac{T}{2}, \frac{T}{2}]} x(t)\, e^{-2i\pi t/T}\, dt$

1.3. Echantillonnage du signal

Au signal x(t) on fait correspondre les valeurs x_i du signal tel que :

$x_i = x\ (iTe)$ (figure 2)

Te étant la période d'échantillonnage.

$Ne = \frac{1}{T_e}$ est la fréquence d'échantillonnage.

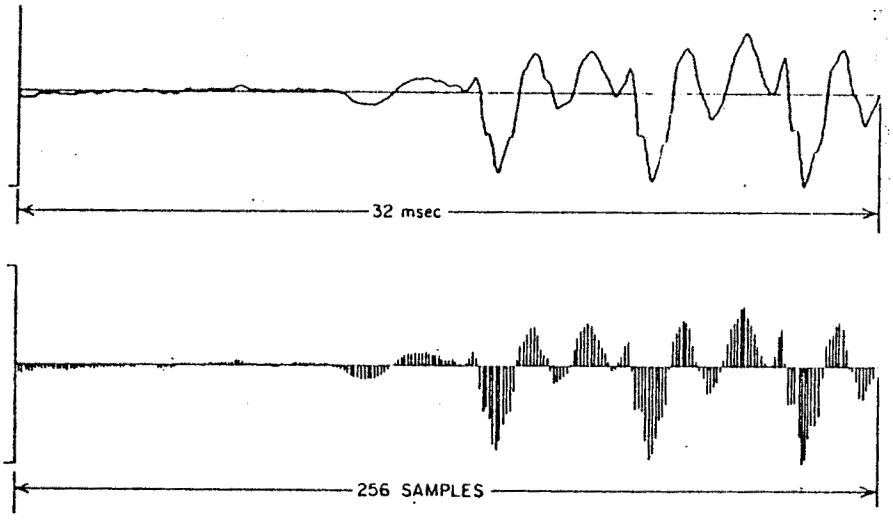


figure 2 : d'après [L.R. Rabiner, R.W. Schafer]
Représentation échantillonnée d'un segment de parole suivant une fréquence d'échantillonnage de 8k Hz.

Il est donc possible de représenter un signal par une séquence de nombres. Les techniques mathématiques concernant le traitement numérique du signal relèvent de cette hypothèse. Bien sûr la fréquence d'échantillonnage est importante car il est nécessaire de garder le maximum d'informations.

1.4. Représentation intégrale

1.4.1. Généralités

On reprend la représentation codée du signal :

$$x(t) = \sum_{m=0}^{\infty} x_m \, h_m(t), \; m \text{ variable discrète}$$

Par analogie, on peut écrire :

$$x(t) = \int_S (\hat{x}(s) \, h(t,s) \, ds)$$

où s est une variable continue décrivant un espace S. De même, les coefficients $\hat{x}(s)$ peuvent être définis par la transformation inverse :

$$\hat{x}(s) = \int_T q(t,s) \, x(t) \, dt$$

où t est une variable continue (temps) décrivant un espace T. On en déduit donc :

$$x(t) = \iint_{S,T} [([x(j) \, q(s,j)]dj)h(0,t)]ds$$

on note $i(t,j) = \iint_{S,T} q(s,j)h(s,t)ds$

$$x(t) = \int_T i(t,j) \, x(j) \, dj.$$

Cette expression donne x(t) en fonction des valeurs discrétisées x(j) projections de x(t) sur l'espace engendré par i(t,j).

Soit $\delta(t)$ la distribution de Dirac souvent utilisée pour représenter des impulsions très brèves et de grande amplitude. Soit $i(t,j) = \delta(t-j)$; $x(t) = \int_T \delta(t-j) \, x(j) \, dj$.

1.4.2. Transformée de Fourier

Soit V la variable fréquence
t la variable temps

La transformée de Fourier est obtenue pour le cas particulier où :

$$h(V,t) = e^{2i\pi Vt}$$
$$g(V,t) = e^{-2i\pi Vt}$$

$$c(t) = \int_{-\infty}^{+\infty} X(V)\ e^{2i\pi Vt}\ dV$$

$$X(V) = \int_{-\infty}^{+\infty} x(t)\ e^{-2i\pi Vt}\ dt$$

X(V) est la transformée de Fourier de x(t).

1.4.3. Transformée de Fourier de signaux périodiques

Un signal périodique x(t) peut s'écrire sous la forme :

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} \left(a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin \frac{2\pi nt}{T}\right)$$

avec : $a_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos (2\pi V_0 nt)\ dt$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin (2\pi V_0 nt)\ dt$$

avec $V_0 = \frac{1}{T}$

Si on pose :

$$X(nV_0) = \frac{1}{2}(a_n - ib_n)$$

Alors :

$$X(nV_0) = \frac{1}{T} \int_{-T/2}^{T/2} x(t)\ e^{-2i\pi V_0 nt}\ dt$$

$X(nV_0)$ est appelé spectre de fréquence du signal x(t).
Cette grandeur est en général complexe :

$$X(nV_0) = |X(nV_0 \text{à}| + j\ \arg(X\ nV_0)$$

On parle alors de spectre d'amplitude et de spectre de phase.

Le spectre de fonctions périodiques est constitué de raies dont l'écart minimum est égal à n_0.

1.4.4. Transformée de Fourier de signaux non périodiques

On considère que la "période" de fonctions non périodiques est extensible jusqu'à l'infini. L'écart entre les raies du spectre tend donc vers 0 rendant le spectre continu. Les expressions deviennent alors :

$$x(t) = \int_{-\infty}^{+\infty} X(V)\ e^{(2i\pi Vt)}\ dV$$

$$X(V) = \int_{-\infty}^{+\infty} x(t)\ e^{(-2i\pi Vt)} dt$$

on note $x(t) \leftrightarrow X(V)$

interprétation :

x(t) s'écrit comme somme linéaire de terme périodique $e^{2i\pi Vt}$,
X(V) dω étant la composante affectée à la bande de fréquence V, V + dV. D'une manière plus générale, on écrit :

$X(V) = c(V)\ e^{i\ a(V)}$

C(V) étant le spectre d'amplitude.

a(V) étant le spectre de phase.

X(V) et x(t) représentent la même grandeur qui est l'amplitude, l'un dans l'espace des temps, l'autre dans l'espace des fréquences. L'utilisation de spectre de fréquence X(V) est à retenir en traitement du signal notamment lorsqu'il s'agit de comparer deux signaux.

propriétés de la transformée de Fourier

$ax(t) \longleftrightarrow aX(V)$

$x(t) + g(t) \longleftrightarrow X(V) + Y(V)$

$x(t-t_0) \longleftrightarrow e^{-2i\pi Vt_0} X(V)$

$e^{2i\pi Vt_0} x(t) \longleftrightarrow X(V-V_0)$

$x(at) \longleftrightarrow \frac{1}{a} X(\frac{V}{a})$

si $x(t) \longleftrightarrow X(V)$

alors $\overline{x(t)} \longleftrightarrow \overline{X(-V)}$

relations de plancherel :

Transformée de Fourier d'une convolution.

$x(t) * y(t) \longleftrightarrow X(V)\ Y(V)$

$x(t)\ y(t) \longleftrightarrow X(V) * Y(V).$

relation de Parseval :

Le produit scalaire est invariant par transformée de Fourier.

$\langle x,y \rangle = \langle X,Y \rangle$

avec : $\langle x,y \rangle \ / \ x(t)\ \overline{y(t)}\ dt$

Les propriétés de symétrie sont résumées dans le tableau suivant :

x(t)	X(V)
Réelle et paire	Réelle et paire
Réelle et impaire	Imaginaire et impaire
Imaginaire et paire	Imaginaire et paire
Imaginaire et impaire	Réelle et impaire
Complexe et paire	Complexe et paire
Complexe et impaire	Complexe et impaire
Réelle quelconque	Partie réelle paire Partie imaginaire impaire
Imaginaire quelconque	Partie imaginaire paire Partie imaginaire impaire
Partie réelle paire Partie imaginaire impaire	Réelle
Partie réelle impaire Partie imaginaire paire	Imaginaire

ANNEXE 2

GENERALITES SUR LES FILTRES ET LES SYSTEMES LINEAIRES

Un filtre réalise une transformation de l'espace X des signaux d'entrée vers l'espace U des signaux de sortie.

Exemple : amplificateur :

$x(t) \rightarrow y(t) = ax(t)$

SCHEMA GENERAL D'UN FILTRE

signal d'entrée → [S] → signal de sortie

$y(t) = S(x(t))$

2.1. Opérateur filtre linéaire

Soit x(t) un signal d'entrée. On appelle filtre linéaire l'opérateur H défini par :

$y(t) = H(x(H))$ où H vérifie les relations de linéarité.

$H(x_1(t) + x_2(t) = H(x_1(t)) + H(y_1(t))$

$H(ax_1(t)) = a\ H(x_1(t))$

L'opérateur H est stationnaire si :
$y(t - t_0) = H(x(t-t_0)$

2.2. filtre linéaire

Un filtre est dit linéaire si il correspond à un opérateur linéaire continu et stationnaire.

Le système est dit de convolution si y(t) est obtenu à partir de x(t) au moyen d'un produit de convolution.

C'est-à-dire : $y(t) = x(t) * h(t)$

Plus explicitement :

$$y(t) = \int_0^t h(t_0)\ x(t-t_0)\ dt_0 = \int_0^t h(t-t_0)\ x(t_0)\ dt_0$$

Supposons appliquée l'impulsion de dirac à l'entrée du système.

$y(t) = x(t) * h(t) = h(t)$ (1)

h(t) est la réponse du filtre à l'impulsion de dirac ; on l'appelle réponse impulsionnelle du filtre.

La relation (1) se démontre facilement en passant par la transformée de Fourier.
En effet, si δ est l'impulsion de Dirac : en utilisant les mêmes notations que dans l'annexe 1 :

$\delta \longleftrightarrow 1$

$x(t) * \delta(t) \longleftrightarrow X(V)\ \delta(V) = X(V)$

$x(t) \longleftrightarrow X(V)$

d'où l'égalité $x(t) * h(t) = h(t)$.

* Transformation "Z" d'un signal :
On appelle transformation "Z" l'application de C dans l'espace des polynômes complexes qui à Z fait correspondre le polynôme en $\frac{1}{z}$ à partir des valeurs échantillonnées de x(t).

$$X(Z) = \sum_{n=0}^{\infty} x(n)\ (Z)^{-n}$$

Application aux filtres :

si $y(n) = x(n) * C(n)$ alors $y(Z) = X(Z)\ C(Z)$

2.3. Filtres physiquement réalisables

Un filtre est dit physiquement réalisable s'il vérifie le principe de causalité c'est-à-dire si la réponse impulsionnelle est nulle pour des temps de valeur négative.

Echelon de Heaviside :

$u(t) = 0$ pour $t < 0$

$u(t) = 1$ pour $t > 0$

Un filtre est dit physiquement réalisable si on sait trouver une fonction f telle que :

$u(t)\ f(t) = h(t)$

2.4. Fonction de transfert

Soit un filtre physiquement réalisable.
Soit H(V) la transformée de Fourier de la réponse impulsionnelle h(t).

$H(V) \longleftrightarrow h(t)$

H(V) est la fonction de transfert du filtre.

ANNEXE 3

ILS

ILS est un ensemble de programmes interactifs conçu spécialement pour le traitement du signal et agissant sur une base de donnée adaptée. Les types de fichiers utilisés sont au nombre de quatre :

1. les fichiers des échantillons : "sample data file"
2. les fichiers analysés : "analysis file"
3. les fichiers permettant des stockages d'informations diverses : "record file".
4. les fichiers dans lesquels sont rentrées des informations d'ordre phonétique avec possibilité de pointage : "label file".

Les trois premiers sont des fichiers binaires ; les fichiers label sont des fichiers de type ASCII.

Chacun des trois premiers types répond à un format standard défini comme suit : au départ une en-tête de fichier est spécifique du fichier correspondant ; elle contient un certain nombre d'informations propres. Ensuite sont rangées des zones de données différentes suivant le type considéré (figure 1).

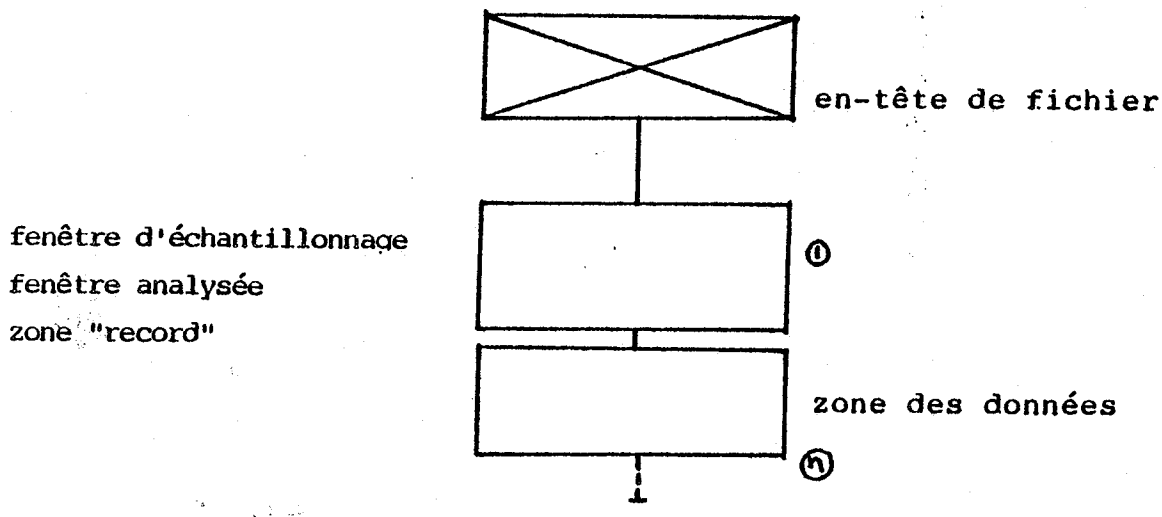


figure 1 : structure des fichiers type 1 à 3

L'en-tête de fichier est rangéedans une zone de 128 mots de 32 bits. Après traitement ("unpackage") les données à rentrer sont stockées dans 256 mots de 16 bits. Les 64 premiers mots ont une signification bien précise comme le montre la figure 2. Le type de fichier est reconnu par lecture du numéro 63.

LOC	NAME	DESCRIPTION
1	N	number of points in analysis window
2	M	number of autoregressive coefficients
3	ICON	preemphasis constant (0-100)
4	NSHFT	shift interval per sampled data frame (= NDPF)
5	IHAM	Hamming window ('Y ' OR 'N ')
6	NSPBK	number of sampled data blocks used in file
7	NP	number of spectral resonance peaks
8	ISTAN	starting frame for analysis
9	NAN	number of frames analyzed
10	NSCA	starting sector for analysis
11	MU	number of autoegressive coefficients for aperiodic frame (PAN,PNS)
12	NT	down-sampling factor (FLT, SIF)
13	IFLD(1)	FIELD-1 - 2 alphabetics
14	IFLD(2)	FIELD-1 - 2 alphabetics
15	IFLD(3)	FIELD-1 - 2 alphabetics
16	IFLD(4)	FIELD-1 - 2 alphabetics
17	NFR	number of variable size frames analyzed (PAN)
18	IAFIX	flag for autoregressive coefficients fixed (PAN)
19	IDK	disk number of sampled data file analyzed
20	NFL	file number of sampled data file analyzed
21	KP	number of analysis mnemonics (IAE)
23	ID(1)	identification - 2 alphabetics
24	ID(2)	identification - 2 alphabetics
25	ID(3)	identification - 2 alphabetics
26	ID(4)	identification - 2 alphabetics
27	ID(5)	identification - numeric
28	NASC	next available sector (TTL)
29	NAPT	next available point (TTL)
30	NZERO	number of zeros (TTL)
31	(FLAG)	= 1111 if secondary file initialized (TTL)
28-57	PARAM	mnemonic names (IAE)
58	ICHAN	starting A/D channel
59	NCHAN	number of channels
60	MULAW	flag set to 50 if 8-bit log quantization (REC)
61	IPWR	power of multiplier for sampling frequency
62	ISF	sampling frequency
63	(FLAG)	= -29000 if analysis data = -30000 if record data = -32000 if sampled data
64	(FLAG)	= 32149 if sampled file initialized

figure 2 : 64 premiers mots d'une en-tête de fichier ILS

fichiers des échantillons : Ils permettent l'acquisition du signal échantillonné après conversion A/N. (figure 3). Le contexte d'échantillonnage est à préciser. Il désigne le nombre de points obtenus dans une fenêtre d'échantillonnage.

fichiers analysés : C'est là que sont stockés les coefficients de reflexion ainsi qu'un certain nombre de mesures effectuées sur la fenêtre considérée (énergie par exemple) (figure 4). Les données sont mémorisées suivant les positions montrées par la figure 5.

```
SECTOR  1. FRAME     1
     0     0     0     1     0    -1     1     2     0     0
     1     0     1     1    -1    -2     0     1    -1    -1
     2     0    -1     1     0    -1    -1     0    -1    -1
     1    -2    -3    -1    -1     0     2     1     0     0
     1    -1    -1     0     1     1     0     1     0    -1
     0     2     1     1     1    -1     0     1    -1    -1
     0     1     0     0     1     0    -1     1     1     0
     0     0     1     0     0    -1     0     0     0     1
     1    -1    -1     0    -1     2    -1     0     0    -1
    -1     2     1     0     1     2     1     0     2     2
```

figure 3 a : 100 points d'une fenêtre d'échantillonnage (correspondant à une fréquence d'échantillonnage de 8000 Hz sur un segment de 12.5ms.

```
100    10    93   100    Y     0   100     1     1     1
  0     0          <          0     0     0     1
  0     0                     0     0     0     1
256   768     0     0     0     0     0     0     1     0
  0     0     0     0     0     0     0     0     0     0
  0     0     0     0     0     0     0     0     0     0
256  8000 -32000 32149     0     0     0     0     0     0
  0     0     0     0     0     0     0     0     0     0
  0     0     0     0     0     0     0     0     0     0
  0     0     0     0     0     0     0     0     0     0
  0     0     0     0     0     0     0     0     0     0
256     0  8000     0 -32000    -1 32149     0
```

figure 3b : en-tête de fichier. Les 64 premiers mots sont utilisés.

figure 3 : fichier des échantillons

```
                ANALYSIS DATA
SECTOR  1. FRAME   100
-18711  -11867   -3943    2186   15449    8023    3939    8763  -12408   -1343
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0
 19291     130     901      93      10     100    9901       0       0      60
```

figure 4a : 10 coefficients de réflexion + autres données

```
100      10      93     100       Y       0     100      60      80       1
  0       0                <              0       0       0       1
  0       0                               0       0       0       0
256     768       0       0       0       0       0       0       1       0
  0       0       0       0       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
256    8000  -29000   32149       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
  0       0       0       0       0       0       0       0       0       0
256       0    8000       0  -32000      -1   32149       0       0       0
```

figure 4b : en-tête de fichier

figure 4 : fichier analysé

LOC	NAME	DESCRIPTION
1-30	RC	reflection coefficients from autoregressive analysis, stored as RC * 32767. (ANA, API, PAN, SIF)
31-54	F,B,A	raw resonance frequencies, bandwidths and amplitudes from autoregressive analysis (maximum 8 sets), stored as F,B,A; A in dB * 100 + 10000 (i.e., stored value A=12574 represents 25.74 dB in the autoregressive spectrum) (SGM, RSO)
55-78	F,B,A	modified values from locations 31-54 (FTR)
79-98	PARAM	user-specified parameters for interactive analysis editting (IAE)
99	NZ	number of bandwidths at zero sampling frequency (RSO)
100	NH	number of bandwidths at half sampling frequency (RSO)
101-104	BZ	bandwidths at zero sampling frequency (RSO)
105-108	BH	bandwidths at half sampling frequency (RSO)
109	RC1	first reflection coefficient before preemphasis (API)
110	RMSS	(same as location 122) (API)
111	IZ	zero crossings per second (API)
112	ICPK	cepstral peak / 10 (API)
113	IPOINT	linear combination for voicing decision (API)
114	RMSUV	smoothed unvoiced signal energy (API)
115	RMSAVE	smoothed input signal energy (API)
116		not used
117	VTL	acoustic tube length, stored as centimeters X 100 (RSO N)
118	NP	number of sets of resonances in locations 31-54 and 55-78 (SGM, FTR, RSO)
119	F0	fundamental frequency, stored as Hz + .5
120	IP	pitch period, stored as (sampling frequency)/F0 + .5
121	RMSN	normalized residual signal energy RMSN=SQRT(AL(M+1)/AL(1)), stored as RMSN*32767
122	RMSS	unnormalized residual signal energy RMSS=SQRT(AL(M+1)*RD0*SCL/AL(1)), stored as RMSS+.5
123	R0	input signal energy R0=SQRT(SCL), stored as R0+.5
124	IPR	preemphasis, stored as percent
125	M	autoregressive filter order
126	N	window size in number of samples
127	ISPT	starting sampled data point in the analysis window (low order 15 bits)
128	IHSPT	starting sampled data point in the analysis window (high order 15 bits)

Nomenclature: AL is the 5th argument of AUTO, RD0=(1+PR*PR)-2*PR*R(2), PR=IPR/100., R is the 4th argument of (FACOR, RACOR), and SCL is the 5th argument of (FACOR, RACOR).

figure 5 : description d'une fenêtre analysée

fichiers d'enregistrement

Le type d'information stocké est indifférent. Ces fichiers ont l'avantage d'avoir un format par zone d'information, paramétrable. Chacune des zones est appelée "record". La structure d'un record est décrite dans des en-têtes de records (figure 6). Un record est structuré en "item" formés par des éléments. Dans la figure 7a on montre un record à 40 items de 3 éléments. Dans la figure 7b, on a un record de 1 items à 12 éléments.

RECORD HEADER FORMAT FOR FEATURE RECORDS

LOCATION	NAME	DESCRIPTION
1	LRH	length of record header = 40
2	NCELLS	length of record in computer words
3	NITEM	number of items in the record
4	NELE	number of elements per item
5	ISTAN	starting frame analyzed
6	NAN	number of frames analyzed
7	NSC	sector number analyzed
8	NDPF	context (number of data points per frame)
9	NFL	file number
10	IDK	disk number
11	IPWR	power of the 10 multiplier for sampling frequency
12	ISF	integer sampling frequency
13	N	number of data points per analysis frame
14	M	order of autoregressive analysis
15	ICON	pre-emphasis (0-100)
16	NP	number of spectral resonances
17	IHAM	hamming window flag ('Y' or 'N')
18	IEX	feature extraction code
19-22	IFLD1	FIELD-1 (8 characters)
23-26	IFLD4	FIELD-4 (8 characters)
27	IFLD3	FIELD-3 (integer)
28	IFLD5	FIELD-5 (integer)
29-33	IFLD2	FIELD-2 (10 characters)
34-38	IFLD6	FIELD-6 (10 characters)
39	ITYPE	record type
40		unused

figure 6 : Structure d'une en-tête de record à 40 mots

```
--------------------------------------------------------------------------------
 REC    STFR     NFR    SC    CTX    EX    FIELD-1    FIELD-2       FIELD-6
 19      0        0      0     0      0

  SOURCE DATA                 N     M   PR   NF   W      SF    FIELD-4     FLD3  FLD5
DUA1:[DELIA.IDATA]WDO.        0     0    0    0          0        0          0

          D            9.1803E+01    5.6000E+01
          G            3.9216E+01    2.0000E+01
          B            3.3333E+01    1.8000E+01
          Z            1.3043E+01    6.0000E+00
          N            1.0127E+01    8.0000E+00
          L            7.8652E+00    7.0000E+00
          WY           7.6923E+00    1.0000E+00
          V            2.5641E+00    2.0000E+00
          #K           1.1494E+00    1.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
     0.0000E+00        0.0000E+00    0.0000E+00
```

<u>figure 7a</u> : <u>fichier des fréquences phonétiques</u>

```
****************************************************************************
* DUA1:[DELIA.IDATA]WD3001.                           3001.        80  RECORDS *
****************************************************************************

----------------------------------------------------------------------------
 REC    STFR     NFR    SC    CTX    EX    FIELD-1    FIELD-2      FIELD-6
 50      60       80     1    100  8000    <

  SOURCE DATA              N     M   PR   NP   W      SF    FIELD-4     FLD3  FLD5
DUA1:[DELIA.IDATA]WD10   100    10   93  100   Y       0    #            0     0

 -2.0225E+00     2.4884E-01    -1.6534E-01    -1.6772E-01    -1.2708E-01
  2.2922E-01    -1.4757E-01     2.6231E-02     1.6108E-01    -2.2895E-01
  1.0000E+02     6.8846E-01
```

figure 7b : 10 premiers coefficients cepstraux - classe d'appartenance - distance à la classe.

<u>figure 7</u> : <u>Structure paramétrable pour les fichiers 'record' ILS</u>

fichiers label : Ils sont utilisés dans la phase de segmentation manuelle. Ils contiennent l'étiquette phonétique des segments successifs d'un mot prononcé (figure 8).

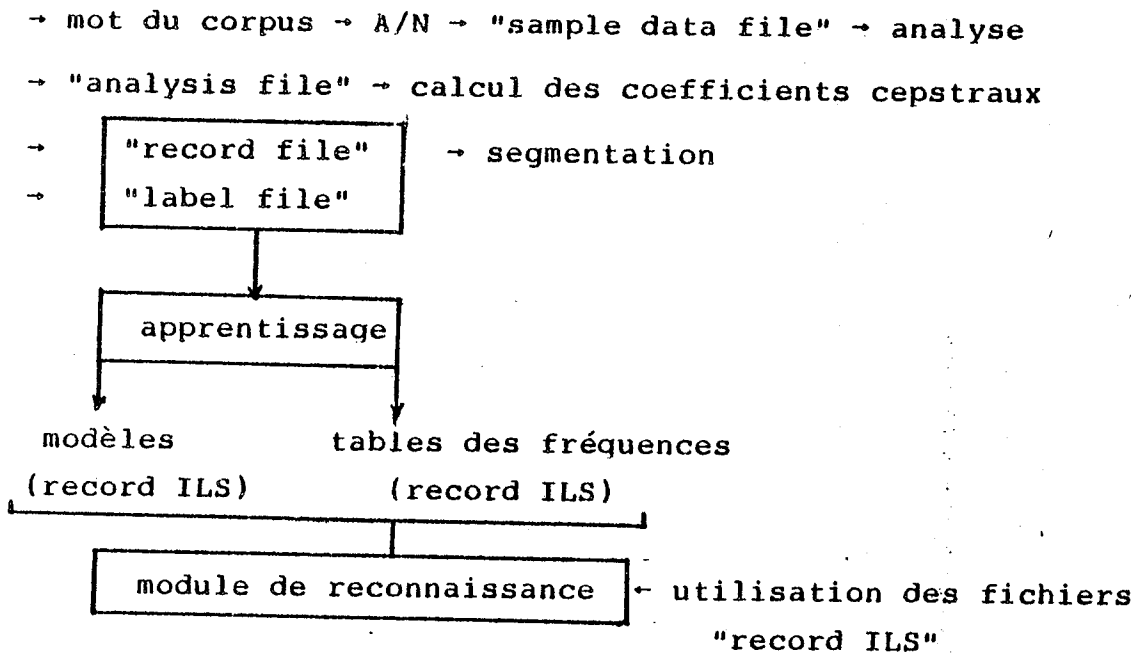
```
T -> DS03.ALL03.LB490
[1000000001111      ;10;A       ;     ;**ELIAS                ;
   6733.;    185.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1000100100000      ;00;B       ;     ;**ELIAS                ;
   7119.;    415.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1000100100101      ;00;N       ;     ;**ELIAS                ;
   7595.;    623.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1000100000101      ;00;E>      ;     ;**ELIAS                ;
   8415.;    142.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1000000101000      ;00;G       ;     ;**ELIAS                ;
   8937.;    304.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1000000001111      ;00;A       ;     ;**ELIAS                ;
   9574.;    153.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[0101001000100      ;00;S       ;     ;**ELIAS                ;
  10226.;    605.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[0000000010101      ;00;WI      ;     ;**ELIAS                ;
  11086.;    200.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
[1001101000011      ;00;<O      ;     ;**ELIAS                ;
  11748.;    306.; 8000;DS03.ALL03.WD490                ;31  MAY 1;
```

figure 8 fichier label

Un label est constitué de 2 lignes elles-mêmes structurées en un certain nombre de zones d'informations. Dans l'en-tête du fichier label (T) est rangé le nom du fichier. Les différents phonèmes du mot "abnégation" (de transcription phonétique ABNE>GASWI<0) sont repérés dans les zones marquées par le premier échantillon (au point) (6733 pour A) et le nombre d'échantillons correspondant à ce phonème (185 pour A). En nombre de segments de 12.5ms A est présent du 68ème segment au 69ème segment (il est à plus de 50% en nombre d'échantillons dans le 68ème et le 69ème à moins de 50% dans le 70ème). Les informations nous ont été utiles pour l'étiquetage des segments dans les fichiers de type "record ILS".

La base de donnée ILS nous a été très utile. Sa manipulation très pratique (pour une utilisateur averti !) permet une correspondance logique entre un fichier échantilloné, le fichier analysé correspondant, le fichier des coefficients cepstraux calculés (record ILS) et les zones phonétiques permettant le pointage sur les segments (coefficients cepstraux).
De plus, notamment par les fichiers d'enregistrement il existe des programmes d'accès à ces fichiers effectuant la lecture ou l'écriture suivant différentes options. Certaines commandes peuvent directement être utilisées. (calcul du centre de gravité par exempl

Nos programmes s'intègrent dans cet environnement et utilisent certains programmes ou fonctions ILS. Le schéma de notre traitement est tracé ci-dessous :



R E F E R E N C E S

[AKAIKE, H.]
"A new look at statistical model identification"
IEEE Decembre 1974.

[CHOLLET, ASTIER, ROSSI]
"Evaluating the performance of speech recognizers at the acoustic phonetic level".
IEEE 1981

[CORSI, P.]
"Reconnaissance des voyelles par les indices et les traits".
Polycopié INPG.

[ERMAN, L.D.]
"A fonctional description of the Hearsay II System"
IEEE 1977 conférence ASFP Mai 1977.

[ESKENAZI, M. ; LIENARD, J.S.]
"Sur les notions de traits et d'indices pour les parties stables du français émis par plusieurs locuteurs".
Processus d'encodage et de décodage phonétique.
24-25 septembre 1981 - Toulouse.

[FONSALE, P.]
"Traits multilocuteurs et leur utilisation dans un système de reconnaissance de mots isolés".
Congrès I CASSP 1983.

[FUMITADA, I]
"Minimum prediction residual principe applied to speech recognition"
IEEE 1975 (Février).

[GROC, B.]
"Conception et réalisation d'un système de reconnaissance de parole continue"
Application à l'interrogation orale de bases de données.
Description et exploitation de contraintes linguistiques.
Thèse INPG Grenoble 1981.

[GAGNOULET, C. ; JOUVET D.]
Séraphine :"Système de reconnaissance de courtes phrases".
Bruxelle 1984. GALF Conférence.

[GUEGEN, C.J.]
"Analyse de la parole par filtrage optimal de Kalman".
Automatisme TOME XVIII Mars 1973

[GUEGEN, C.J.]
"La prédiction linéaire modifiée et son application à la modélisation du signal de parole".
Colloque national sur le traitement du signal et ses applications (16-21 Juin).

[GUILLEMOT, DELIA]
"Etude d'une nouvelle méthode de classification".
Projet 3ème année - ENSIMAG- Juin 1982.

[GUPTA, GOWDI, BRYAN]
"Evaluation of some distance measures for computerized speech recognition".
6 april Clemson University. Report (1977).

[HIROAKI, CHIBA]
"Dynamic Programming Algorithm Optimization for Spoken word recognition".
IEEE 1978, February.

[JORDAN, COHEN]
"Segmenting speech using dynamic programming".
J. Acoust. Soc. 69 (5) May 1981

[KLATT]
"Review of the ARPA Speech understanding project".
J. Acous. Soc. am, 62, 1345-1366.

[LEVINSON, AL]
"Interactive clustering technics for selecting speaker independant reference templates for isolated word recognition".
IEEE ASFP Vol. 27 n°2 April 1979.

[LIENARD, J.S]
"Les processus de la communication parlée".
Introduction à l'analyse et à la synthèse de la parole.
Edition MASSON.

[LINDE, BUZO, GRAY]
"An algorithm for Vector Quantizer Design".
IEEE January 1980.

[MAKHOUL, J.]
"Linear Prediction, a Tutorial Review".
IEEE 1975 APril.

[MAKHOUL, J.]
"Stable and Efficient lattice methods of linear Prediction".
IEEE 1977 Octobre.

[MARKEL, GRAY]
"Linear Prediction of speech".
Springer Verlag Editor.

[MARKEL, GRAY]
"Distance mesures for speech processing ".
IEEE octobre 1980.

[MERCIER,AL]
Keal :"Un système pour le dialogue oral".
Congrès AFCET 13-15 Nov. Tome 2

[MYERS, RABINER, ROSENBERG]
"On the use of dynamic Time Warping for Word Spotting and
connected Word recognition ".
Bell System, technical Journal, vol. 60 March 1984.

[RABINER L.R. ; SCHAFER, R.W.]
"Digital processing of speech signal".
Prentice-Hall Signal Processing Series. Alan V. Oppenheim Editor.

[RABINER, L.R.]
"On creating reference templates for speaker independant recognition
of isolated words".
IEEE February 1978

[ROSSI, M. ; NISHINUMA, Y. ; MERCIER, G.]
"Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole".
Congrès ICA 1983.

[ROSSI, M. ; NISHINUMA, Y. ; TREVARAIN, O.; MERCIER, G.]
"Reconnaissance des voyelles par les indices et les traits".
Processus d'encodage et de décodage phonétique.
24-25 septembre 1981 Toulouse.

[STEVEN, DAVIS, MERMELSTEIN]
"Comparison of parametric representation for monosyllabique Word Recognition on continuously spoken sentences".
IEEE August 1980.

[SUGANA, SKIKANO, FURUI]
"Isolated word recognition using phoneme-like templates".
Conf. ICASSP 1983.

[TUFELLI, D.]
"Conception et réalisation d'un système de reconnaissance de parole continue".
Application à l'interrogation orale de bases de données.
Niveau acoustique et lexical : apprentissage et reconnaissance.
Thèse INPG Grenoble 1981.

[WALTER H.]
"La phonologie du français".
Edition PUF Le Linguiste.

DERNIERE PAGE D'UNE THESE

3E CYCLE, DOCTEUR INGÉNIEUR OU UNIVERSITÉ

Vu les dispositions de l'arrêté du 16 avril 1974,

Vu les rapports de M. BELLISSANT.....

M. .....

Mme Christine DELIA..... est autorisée

à présenter une thèse en vue de l'obtention du grade de DOCTEUR INGENIEUR...

.....

Grenoble, le 6 SEP. 1984

Le Président de l'Université Scientifique et Médicale

M. TANCHE

UNIVERSITÉ SCIENTIFIQUE ET MÉDICALE DE GRENOBLE