REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE
DEPARTEMENT D'ELECTRONIQUE

# THESE



Présentée en vue de l'obtention du diplôme de **MAGISTER**
en : ELECTRONIQUE APPLIQUEE
OPTION : Acquisition et traitement de l'information

PAR :

**BOUCHEFRA KHELIFA**

## THEME

# CONTRIBUTION A LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE CONTINUE :

# ETUDE ET REALISATION D'UN SYSTEME DE RECONNAISSANCE ACOUSTICO-PHONETIQUE

Soutenue devant le jury composé de :

| | | | |
|---|---|---|---|
| Mr | D. BERKANI | Maître de conférences (ENP) | : PRESIDENT |
| Mr | B. BOUSSEKSOU | Chargé de cours (ENP) | : RAPPORTEUR |
| Melle | M. GUERTI | Maître de conférences (ENP) | : EXAMINATEUR |
| Mr | B. DERRAS | Maître de conférences (ENP) | : EXAMINATEUR |
| Mr | A. FARAH | Maître de conférences (ENP) | : EXAMINATEUR |

JUIN 1995

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE
DEPARTEMENT D ' ELECTRONIQUE

# THESE



Présentée en vue de l' obtention du diplôme de **MAGISTER**
en : **ELECTRONIQUE APPLIQUEE**
**OPTION : Acquisition et traitement de l' information**

**PAR :**

**BOUCHEFRA KHELIFA**

## THEME

# CONTRIBUTION A LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE CONTINUE :

# ETUDE ET REALISATION D' UN SYSTEME DE RECONNAISSANCE ACOUSTICO-PHONETIQUE

**Soutenue devant le jury composé de :**

| | | | |
|---|---|---|---|
| **Mr** | **D. BERKANI** Maître de conférences (ENP) | : | **PRESIDENT** |
| **Mr** | **B. BOUSSEKSOU** Chargé de cours (ENP) | : | **RAPPORTEUR** |
| **Melle** | **M. GUERTI** Maître de conférences (ENP) | : | **EXAMINATEUR** |
| **Mr** | **B. DERRAS** Maître de conférences (ENP) | : | **EXAMINATEUR** |
| **Mr** | **A. FARAH** Maître de conférences (ENP) | : | **EXAMINATEUR** |

JUIN 1995

# REMERCIEMENTS

Ce travail n' aurait sans doute jamais eu lieu sans le concours volontaire ou involontaire de tous ceux qui m' ont formé, soutenu, encouragé ou aidé... Parents et amis, professeurs et collégues qu' il m' est impossible ici de citer tous, mais à qui je suis sincèrement reconnaissant pour tout ce qu' ils ont fait pour moi.

Je tiens pourtant ici tout spécialement a :

exprimer ma profonde reconnaissance et mes vifs remerciements au professeur Mr B. BOUSSEKSOU pour avoir bien voulu me proposer un sujet et diriger ma thèse, et surtout pour sa précieuse collaboration, sa grande disponibilité et l' aide constante qu' il m'a apportée durant mes travaux. Pour moi, il est plus qu'un directeur de thèse, c' est un ami. Il ne m' est impossible ici de le remercier pour tout. J' espère sincèrement que cette thèse ne sera qu' une étape dans notre collaboration.

remercier vivement le professeur Mr D. BERKANI, d' avoir bien voulu accepter de donner un avis sur mon travail, je lui en suis très reconnaissant pour les conseils et les critiques fructueuses et aussi d' avoir bien voulu accepter de présider le jury de ma thèse.

exprimer ma reconnaissance au professeur Melle M. GUERTI pour l' attention qu' elle a manifestée à l' égard de mon travail. Les conseils et encouragements qu' elle m' a prodigués m' ont été très utiles. Elle a bien voulu juger mon travail et je la remercie vivement.

remercier le professeur Mr B. DERRAS pour toute l' attention qu' il m' a accordée et l' importante aide qu' il a voulu me consentir. Je lui suis très reconnaissant d' avoir bien voulu accepter de participer au jury.

adresser mes vifs remerciements au professeur Mr A. FARAH qui a bien voulu examiner mon travail et de participer au jury de ma thèse.

Enfin, je ne saurais manquer de remercier tous les collègues et amis qui m' ont toujours soutenu et encouragé et, en particulier, Mr A. BOULARES Directeur de LTTS USTHB, Messieurs S. OUZNADJI et A. NACER, enseignants à l' ITS USTHB ainsi que Mr R. ZERGUI enseignant à l' ENP.

# S O M M A I R E

# *PARTIE A*

# INTRODUCTION A LA RECONNAISSANCE ET A LA COMPREHENSION DE LA PAROLE

## *CHAPITRE A1* : INTRODUCTION A LA PAROLE ET SA RECONNAISSANCE



## *CHAPITRE A2* : COMPREHENSION DE LA PAROLE

## *PARTIE B*

## ETUDE ET REALISATION DU SYSTEME DE RECONNAISSANCE ACOUSTICO-PHONETIQUE PROPOSE

## *CHAPITRE B1* : ETUDE DU TRAITEMENT ACOUSTICO-PHONETIQUE



## *CHAPITRE B2* : MISE EN OEUVRE DU DECODEUR ACOUSTICO-PHONETIQUE

## *CHAPITRE B3* : EXPERIENCES, TESTS ET RESULTATS



## *CHAPITRE B4* : PRESENTATION ET UTILISATION DU SYSTEME SRAPH REALISE

# INTRODUCTION

Depuis plus d'une quarantaine d'année, le développement des machines informatiques a été considérable, principalement sur le plan du matériel. Du premier ordinateur de 1946 contenant 18 000 tubes à celui de nos jours occupant un volume 10 000 fois plus faible pour une puissance de calcul nettement supérieure, l'évolution semble remarquable. Malheureusement, le géant informatique a un talon d'Achille dont la fragilité assombrit les perspectives d'utilisation généralisée de l'ordinateur par le grand public : la machine doit être programmée suivant un code très précis où le moindre écart est fatal. Certes, des langages évolués, comme le FORTRAN ou le PL1, ont vu le jour, qui facilitent largement son utilisation et sont à l'origine du fantastique développement des calculs scientifiques et de gestion sur ordinateur, mais celle-ci reste l'épanage de quelques initiés.

Dés le début de l'informatique, la nécessité d'une communication Homme-Machine directe à l'aide du langage naturel a vu le jour. En 1950, TURING espérait que le développement des méthodes de programmation s'effectuerait à un tel rythme qu'en l'an 2 000 il serait pratiquement impossible, en dialoguant avec une machine de se rendre compte de sa nature mécanique. On sait que, pour l'instant, ces espoirs sont irréalistes et que l'ordinateur reste un outil dont la mise en oeuvre est astreignante. L'informatique n'a pas, à notre avis, atteint l'âge adulte à partir duquel son utilisation serait simple et accessible à tous. La réalisation de terminaux à l'échelle humaine nous semble impérative et la parole est certainement le vecteur le mieux adapté à ce but.

C'est dans cette optique fortement orientée vers le dialogue Homme-Machine que se situent les travaux qui font l'objet de cette thèse. Ces recherches concernent plus spécialement l'étude de la parole comme l'un des moyens de communication entre l'homme et la machine, complément souhaitable, pour ne pas dire indispensable, aux moyens graphiques et visuels, et aux moyens actuellement existants sur les ordinateurs.

Mais ce domaine d'étude de la parole est encore beaucoup trop vaste pour être pris en charge entièrement dans un seul projet de recherche. Nous nous sommes donc limités volontairement à la reconnaissance et à la compréhension de la parole, à l'exclusion de la synthèse, de l'identification du locuteur, etc. Et dans ce cadre, nous avons surtout centré nos recherches, sur une partie de la chaîne de communication parlée, à savoir essentiellement l'étape du décodage acoustico-phonétique qui constitue le premier maillon de cette chaîne.

Le principal objectif de ce travail est donc l'étude et la réalisation d'un système expert pour le décodage acoustico-phonétique en vue d'une reconnaissance automatique de la parole continue (R. A. P. C). Ce système peut être considéré à la fois comme un module autonome réalisant la reconnaissance d'une séquence de parole, ou bien comme une étape importante d'un système plus général intégrant d'autres informations linguistiques (syntaxe, sémantique, pragmatique, prosodie).

L'organisation pratique de ce mémoire fait apparaître essentiellement deux parties A et B rassemblant une large documentation sur le sujet.

La partie A est une introduction à la reconnaissance et à la compréhension de la parole. Cette partie sera décomposée en deux chapitres.

Dans le premier, on évoquera tout d'abord l'importance que constitue le dialogue oral "Homme-machine", ensuite, après avoir mis l'accent sur la nature complexe du signal vocal, nous proposerons une classification des différentes tâches de la reconnaissance automatique de la parole, conduisant ainsi aux différentes méthodes utilisées en reconnaissance.

Le second chapitre A2 sera consacré à la compréhension de la parole. Après avoir montré clairement la différence entre la reconnaissance et la compréhension, nous nous intéresserons essentiellement aux diverses sources de connaissances et leurs stratégies d'utilisation en compréhension de la parole. Puis nous consacrerons un paragraphe à la présentation des différents modèles proposés en reconnaissance automatique. On présentera aussi l'état actuel des travaux de recherche où nous donnerons les objectifs, les résultats ainsi que les enseignements tirés des principaux systèmes existants. Nous terminerons enfin ce chapitre par une simple schématisation du processus de la compréhension de la parole en spécifiant le rôle de chaque module qui intervient.

La partie B quand à elle, sera toute entière consacrée à l'étude et la réalisation du Système de Reconnaissance Acoustico-Phonétique **S R A P H** que nous avons proposé. Cette partie regroupe quatre chapitres.

Le premier chapitre B1 expose d'une manière générale le module du décodage acoustico-phonétique. On décrira les différentes étapes qui constituent ce module ainsi que les approches associées à chacune d'elles. On étudiera dans ce chapitre, en particulier, l'approche système expert qui sera adoptée à l'étape d'identification lors de la mise en oeuvre de notre système.

Le chapitre B2 aborde la mise en oeuvre du décodeur acoustico-phonétique à travers les explications qui seront données sur chacune des étapes constituantes : acquisition des paramètres, segmentation et identification.

Dans le chapitre B3, Nous retrouverons un certain nombre d' expériences qui nous seront très utiles pour l' établissement de notre base de connaissances. Cette base sera ensuite testée sur notre système et un recueil de résultats sera donné sous une forme appropriée aidant ainsi l' expert à une meilleure appréciation. Nous terminerons ce chapitre par une discussion des résultats obtenus.

Le dernier chapitre B4 est une présentation du système **S R A P H** que nous avons réalisé ainsi que son utilisation.

Nous terminerons cette étude par des conclusions et des perspectives.

# *PARTIE A*

## INTRODUCTION A LA RECONNAISSANCE ET A LA COMPREHENSION DE LA PAROLE

A1 : INTRODUCTION A LA PAROLE ET SE RECONNAISSANCE

A2 : COMPREHENSION DE LA PAROLE

# *CHAPITRE A1*

# INTRODUCTION A LA PAROLE ET SA RECONNAISSANCE

## 1. UTILITE ET IMPORTANCE DU DIALOGUE HOMME-MACHINE AU MOYEN DE LA PAROLE

La parole est le moyen de communication le plus naturel et familier, le plus facilement utilisable et le plus rapide. La parole ne demande pas un apprentissage particulier comme, par exemple, l'utilisation d'une machine à écrire. La parole présente aussi d'autres avantages :

- l'utilisation de la parole libère complètement l'usage de la vue et des mains (contrairement à l'écran et au clavier) et laisse l'interlocuteur humain libre de ses mouvements ;

- la possibilité de transmettre la parole à distance par radiodiffusion ou par réseau téléphonique ;

- la parole peut nous informer sur l'idendité du locuteur.

En conséquence, si l'on admet qu'il est possible de construire des systèmes de reconnaissance, ou plus exactement de compréhension, de la parole suffisamment complexes et généraux (c'est-à-dire avec un lexique de grande taille, une syntaxe peu contraignante et aucune restriction sur les locuteurs), les applications envisageables d'une entrée-sortie parlée sont les suivantes :

- la saisie de données ;
- la gestion et la consultation de bases de données ;
- l'enseignement assisté ;
- la commande de processus comme par exemple de machines-outils, de systèmes d'aide aux handicapés ou robots travaillant en ambiance polluée ou hostile.

On peut ajouter à ceci les domaines des transports, banques, ventes et bien d'autres secteurs de l'activité humaine.

Remarquons cependant que les possibilités offertes par les programmes actuels sont encore très éloignées de ce but.

## 2. PROPRIETES SPECIFIQUES DU SIGNAL VOCAL

La grande difficulté de la reconnaissance automatique de la parole provient du caractère même du processus de la communication parlée et des propriétés intrinsèques du signal vocal. Les messages vocaux subissent une série de transformations depuis l'idée à émettre jusqu'au signal acoustique, ce qui correspond à un codage très complexe. Le décodage d'un tel message est à l'évidence particulièrement difficile. La figure A1.1 montre un modèle hiérarchique du processus de perception de la parole chez l'homme [ 1 ].

Le signal vocal, tel qu'il apparaît sur la figure A1.2 possède des propriétés très spécifiques et qui se résument par :

a - **la continuité** : Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont l'occurrence est aléatoire ; il peut très bien y avoir des intervalles de silence au milieu d'un mot et aucun intervalle entre deux mots successifs. Il est donc très difficile de déterminer le début et la fin des mots composant la phrase ;

b - **la variabilité** : La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que se soit pour un même locuteur ou plusieurs. Pour un même locuteur, des différences importantes de prononciations peuvent apparaitre suivant l'état émotionnel du sujet et l'intensité de sa voix ; celui - ci peut crier ou murmurer, être enroué ou enrhumé. De même, des contrastes considérables peuvent se manifester entre plusieurs locuteurs suivant l'âge, le sexe, l'origine géographique et le milieu social. On peut ajouter aussi les perturbations apportées par le microphonne (selon le type, la distance, l'orientation) et l'environnement (bruit, réverbération), etc ;

c - **la redondance** : Le signal de la parole est très redondant. Son traitement automatique nécessite, en effet, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter les durées du traitement, lequel doit se faire en temps réel. A l'inverse, le débit ne doit pas être trop faible pour conserver un bon rapport signal / bruit. Une valeur de 100 ou 50 bits/s paraît convenir à la reconnaissance ;

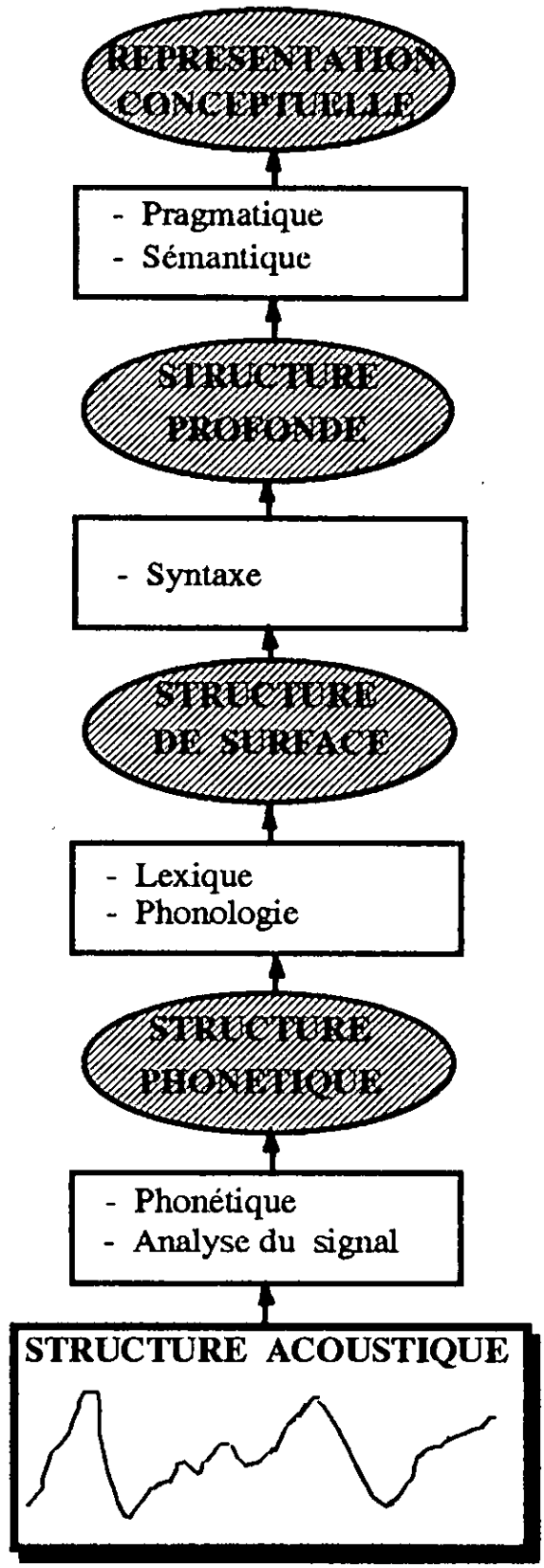



**Fig. A1.1 :** Un modèle hiérarchique de la perception humaine de la parole chez l' homme d' après [ 1 ]

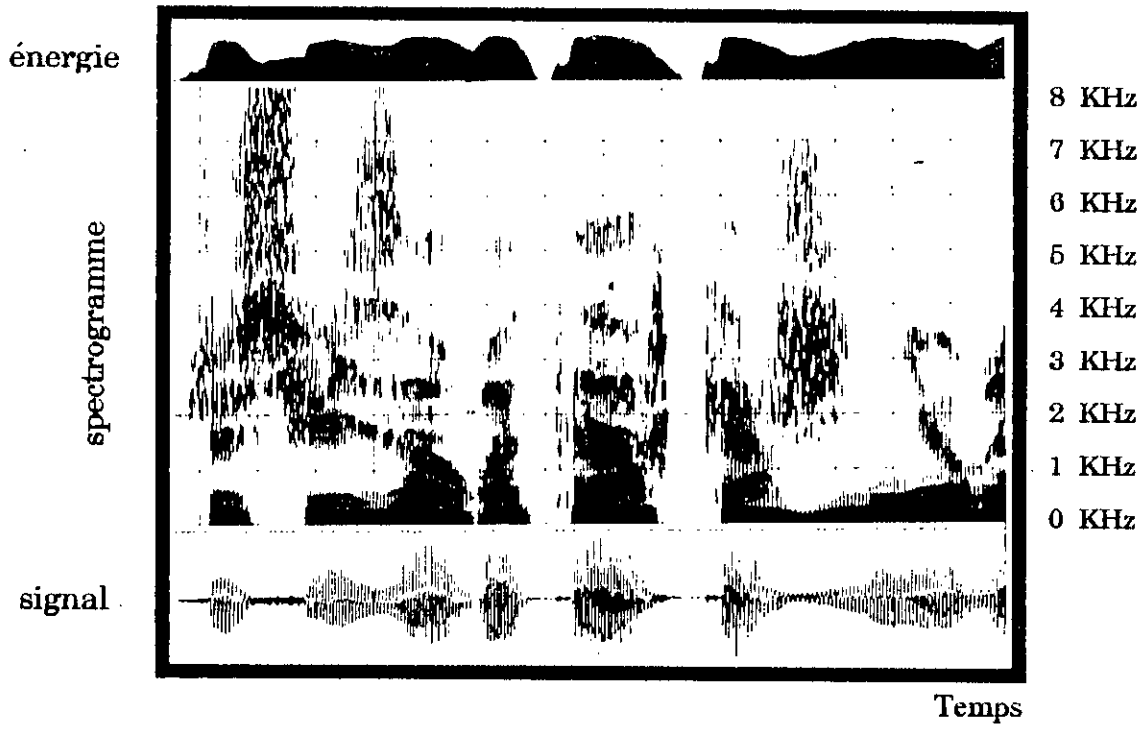



**Fig. A1.2 :** Exemple de signal vocal temporel, spectrogramme et courbe d'énergie de la phrase "je suis en retard car..." [39]

d - **la grande liberté du langage parlé** : la syntaxe du langage parlé est généralement moins stricte que celle du langage écrit. Les programmes de reconnaissances évolués doivent obligatoirement en tenir compte si l'on veut qu'ils soient utilisables en pratique.

## 3. CLASSEMENT DES TACHES DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Compte tenu des caractéristiques de la parole, on distingue plusieurs démarches pour aborder les problèmes de reconnaissance. Ces démarches portent sur les points suivants :

- **mode d'élocution** : mots isolés ou parole continue ;

- **indépendance du système vis-à-vis du locuteur** : systèmes monolocuteurs ou multilocuteurs ;

- **complexité du langage autorisé** : qui porte à la fois sur la taille du vocabulaire et sur la difficulté de la grammaire autorisée ;

- **reconnaissance ou compréhension** : qui distingue entre une simple reconnaissance au niveau des mots ou l'interprétation de tout le message émis.

Tout système de reconnaissance de la parole résulte d'un compromis entre ces différentes démarches. Il est fonction du but et des objectifs à atteindre.

## 4. DIFFERENTES APPROCHES DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Le problème de la reconnaissance de la parole peut être abordé suivant deux approches très différentes :

### 4.1. Approche globale

Elle consiste à considérer le mot ou le groupe de mots comme une entité que l'on cherche à reconnaître. Au cours d'une phase préalable d'apprentissage, les formes de référence correspondant aux différents mots du vocabulaire sont stockées dans un dictionnaire. La reconnaissance d'un mot inconnu se fait alors par comparaison de sa forme aux différentes formes de référence.

Le traitement acoustique préliminaire est assez simplifié ; il n' y a pas de problèmes de segmentation (tous les mots sont séparés par environ 30 ms, ce qui permet une segmentation triviale en mots), par contre il est nécessaire de conserver en mémoire les différentes représentations possibles de tous les mots. Il n' est pas possible de conserver toute l' information car il nous faudrait une très grande capacité mémoire. De plus, cette approche nécessite une normalisation en fréquence et en temps ; en effet, lorsqu ' une même personne prononce deux fois de suite le même mot, on constate l' existance de différences sensibles, qui portent à la fois sur le rythme d' élocution (certaines parties du mot sont prononcées plus rapidement, d ' autres plus lentement), sur l' intensité et sur le timbre. Il sera donc impossible de faire coïncider exactement les deux images acoustiques du mot référence et du mot inconnu.

Pour résoudre ce problème, on a recours à une technique appelée *programmation dynamique* qui a été mise au point par Richard BELLMAN et qui fut ensuite reprise par les chercheurs SAKOE et CHIBA en 1971 [ 2 ]. Celle-ci consiste à déterminer la façon optimale de mettre en correspondance les deux mots à comparer, et l' écart entre leurs images acoustiques.

Des études faites sur cette méthode [ 3 ] ont montré sa validité sur des vocabulaires de faible taille, dont la limite semble se situer aux alentours d' une centaine de mots avec un taux de reconnaissance assez élevé et son indépendance vis-à-vis des particularités de la langue à reconnaitre (du fait de la phase d' apprentissage).

Son principal inconvénient, outre qu' elle ne permet pas de traiter de gros vocabulaires, est son inadaptation au traitement des phrases. En effet, elle nécessite un énoncé mot à mot des phrases (avec un silence entre chaque mot de l' ordre de quelques dizaines de ms) et cette contrainte, difficilement mise en place dans le cadre de langages pseudo-naturels, donne une impression très artificielle au dialogue ainsi obtenu.

## 4.2. Approche analytique

Les limites des méthodes globales ont conduit les chercheurs vers un second type de méthodes, dites analytiques dont l' objectif est de déterminer dans un premier temps des éléments minimaux ou sons élémentaires (phonèmes, syllabes, ..) de la langue, toujours en nombre restreint, pour ensuite reconstruire la phrase de départ comme séquence de ces sons élémentaires. Les difficultés rencontrées sont importantes et loin d' être résolues : il faut en effet obtenir cette suite d' éléments phonétiques minimaux se rapprochant le plus possible des

phonèmes dont personne à notre connaissance n'est capable de fournir la liste des éléments acoustiques caractéristiques.

Cette méthode est beaucoup plus générale et peut seule permettre de traiter la parole continue. Mais elle est aussi beaucoup plus délicate à mettre en oeuvre, en particulier au niveau des algorithmes de segmentation du message.

Le tableau A1.1 résume quelques considérations sur les qualités et les défauts de ces deux méthodes :

| CRITERES | Méthode globale | Méthode analytique |
|---|---|---|
| Taille du vocabulaire | limitée | indépendante |
| Taux de reconnaissance actuel | très élevé (> 95 %) | faible |
| Indépendance vis-à-vis de la langue | oui | non |
| Traitement de la parole continue | impossible | possible |
| Exploitation et mise en oeuvre | facile | difficile |
| Problèmes de segmentation | simple | très difficile |
| Adaptation au locuteur | difficile | relativement facile |
| Domaine d'application | spécialisé | vaste |

**Tableau A1.1 :** Avantages et inconvénients des méthodes globales et analytiques.

## 5. CHOIX DES UNITES MINIMALES DE RECONNAISSANCE

Le message sonore a comme caractéristique principale sa continuité. Il sera donc nécessaire pour parvenir à une meilleure représentation de retrouver les diverses unités linguistiques minimales composant ce message. Sans cette étape indispensable on ne peut espérer mener à bien un tel décodage.

Les systèmes de reconnaissance de la parole diffèrent non seulement par les algorithmes et les stratégies employés mais aussi par les éléments qu'ils cherchent à identifier.

L'élément minimal à reconnaître peut être le mot, la syllabe, le phonème ou le phone. Pour la reconnaissance de vocabulaires limités, le mot est l'unité de reconnaissance la plus utilisée, mais pour des vocabulaires importants, il faut créer un grand ensemble de références d'où la nécéssité d'introduire des unités plus petites que le mot.

Les avantages et les inconvénients liés à l'utilisation de l'une de ces unités comme unité de reconnaissance sont résumés dans le tableau A1.2 :

| UNITE DE BASE | Avantages | Inconvénients |
|---|---|---|
| MOT | - coarticulation incluse<br>- indépendance de la langue<br>- reconnaissance simple | - grands vocabulaires<br>- adaptation au locuteur |
| SYLLABE | - facile à localiser (voyelle)<br>- coarticulation incluse | - nombre total élevé<br>- frontières difficiles à localiser |
| DIPHONE | - contient une partie de la coarticulation<br>- nombre total peu élevé | - problèmes de segmentation<br>-très dépendante du contexte<br>- nombre total très élevé (>1000) |
| PHONEME | - codage aisé des mots dans le lexique | - pas facle à localiser<br>- algorithmes et règles complexes pour segmentation et reconnaissance |
| PHONE | - peut correspondre à un segment acoustique<br>- facile à localiser et à reconnaitre | - nombre élevé<br>- peut dépendre du contexte |

**Tableau A1.2 :** Unités de base de la reconnaissance de la parole

# CHAPITRE A2

# COMPREHENSION DE LA PAROLE

## 1. INTRODUCTION

Avec ce second chapitre, nous abordons l'essentiel de nos recherches : l'étude des diverses sources de connaissances en compréhension de la parole, leurs représentations et leurs stratégies d'utilisation.

Mais avant d'aborder ce chapitre, nous estimons qu'il est nécessaire de rappeler quelques remarques qui ont guidé notre réflexion à ce sujet :

Dans le premier chapitre, nous avons vu que les signaux du discours continu ne peuvent pas être traités en appliquant simplement chaque mot du signal contre les schémas enregistrés pour les mots du vocabulaire. La prononciation des mots individuels change quand les mots sont juxtaposés pour former une phrase. En fait, trouver les limites entre les mots dans un discours continu est une tâche très difficile du processus. En bref, le signal acoustique ne ressemble pas du tout à la concaténation des signaux des mots individuels.

Les difficultés introduites en tentant de reconnaître un discours continu nécessitent donc une nouvelle vision dans la méthodologie. Les chercheurs ont fait la spéculation qu'il y avait plus d'informations introduites par l'utilisateur que simplement le signal acoustique.

De nombreuses experiences ont montré que l'homme avait recours de façon banale et souvent inconscientes à une source de connaissances relevant de domaines très différents (acoustiques, linguistiques, psychologiques, etc.) [4] et de traiter un grand volume de données, pour dégager les informations pertinentes à la reconnaissance de la parole. Une analyse de ce type lui permet de lever des ambiguïtés acoustiques et, dans bien des cas, de prévoir la fin d'une phrase avant que celle-ci ne soit terminée.

Ces changements de perspective dans la recherche du discours continu est souvent considérée comme la différence entre la **reconnaissance** et la **compréhension.** Actuellement la plupart des travaux sont du second type. Essayons néanmoins de fournir une définition de ces termes.

**a - système à reconnaitre la parole :** son objectif est d'essayer de reconstituer le message prononcé élément par élément ;

**b - système à comprendre la parole :** dans cette optique, on s'attache à reconnaitre globalement (grosso-modo) le message sans trop s'arrêter aux détails manquants et même si des erreurs subsistent au niveau de la reconnaissance d'un ou plusieurs mots de la phrase. De tels systèmes, comme nous le verrons, utilisent souvent des techniques de type intelligence artificielle, spécialement en ce qui concerne l'usage fréquent de retours en arrière ("back-tracking"), et surtout de connaissance sur le domaine d'application, pour déterminer la solution optimale.

## 2. SOURCES DE CONNAISSANCES ET LEURS STRATEGIES D'UTILISATION EN COMPREHENSION DE LA PAROLE

Les processus mentaux mis en oeuvre par l'être humain pour l'émission et la compréhension de la parole sont extrêmement complexes, comme on a pu le constater. Pour exprimer une idée, par exemple, il faut la formuler à l'aide de mots sous la forme d'une phrase, transformer les mots en commandes articulatoires pour émettre une suite de phonèmes. On retrouve cette hiérarchie de niveaux pour la reconnaissance et la compréhension de la parole. En effet, dans l'état actuel des recherches, les modèles de reconnaissance sont ainsi structurés. Les correspondances sont les suivantes (Tableau A2.1) :

| | *Parole naturelle* | *Compréhension automatique* |
|---|---|---|
| *complexité croissante* ↑ | idée | sémantique, compréhension |
| | phrase | syntaxique |
| | suite de mots | lexique, reconnaissance des mots |
| | suite de "phonèmes" | phonétique, règles phonologiques |
| | résolution de "phonèmes" | acoustique |
| | signal de parole | signal de parole |

**Tableau A2.1 :** Comparaison entre le processus de compréhension chez l'homme et la compréhension automatique de la parole

Le but de ce chapitre est de décrire les différents niveaux (ou sources de connaissances) nécessaires pour réaliser un système de compréhension de la parole en partant du niveau inférieur car c'est, actuellement, l'aspect acoustique du signal de la parole qui est le plus connu [ 5 ].

## 2.1. Différentes sources de connaissances

Le processus humain de compréhension de la parole met en jeu plusieurs sources de connaissances qu'il est nécessaire d'analyser, d'intégrer et de contrôler dans un système de reconnaissance de la parole. Parmi ces connaissances celles liées à la définition du langage, auxquelles s'ajoutent d'autres qui sont propres à la parole. Ces informations sont du type : acoustique, phonétique, phonologique, lexicale, syntaxique, sémantique, pragmatique et prosodique.

On a coutume généralement de classer ces différentes sources en deux sortes d'informations :

- ***Les sources de connaissances indépendantes du contexte*** :
  (acoustique, phonétique, phonologique, lexicale et prosodique)

- ***Les sources de connaissances contextuelles*** :
  (syntaxique, sémantique et pragmatique)

Tout en respectant cette classification, nous allons essayer d'expliciter un peu ce que recouvre chacun de ces mots. Cela n'est pas chose facile, car nous nous situons hors du domaine de l'informatique classique, sans pour autant pouvoir prétendre entrer dans le domaine de la linguistique. Or, tous ces termes ont un sens très précis en linguistique, même si tous les spécialistes du domaine n'en fournissent pas la même définition.

### 2.1.1. Sources de connaissances indépendantes du contexte

On distingue ici les niveaux: acoustique, phonétique, phonologique, lexical et prosodique. Les techniques associées à ces sources dépendent généralement plus de la langue que du domaine linguistique.

On distingue donc les niveaux :

**a - acoustique :** qui correspond à l'étape de la saisie du signal vocal, son traitement et l'extraction de paramètres pertinents par des techniques du traitement du signal ;

**b - phonétique :** régissant le passage du signal paramétré à sa description en termes d'unités phonétiques ;

**c - phonologique :** c'est la prise en considération de l'ensemble des règles décrivant les différentes variations individuelles (accent, . .) et des phénomènes d'altération contextuelle des sons (coarticulation, liaison, etc.) ;

**d - lexical :** ce niveau doit contenir l'ensemble des informations relatives aux mots, éléments de structuration essentiels d'une phrase.

**e - prosodique :** recouvrant le rythme, l'intensité et la mélodie de la voix. Elles sont spécifiques de la communication parlée et jouent un rôle important dans la compréhension d'un message.

### 2.1.2. Sources de connaissances contextuelles

Pour spécifier un domaine linguistique, il faut définir trois niveaux qui sont :

**a - syntaxique :** concernant la structuration d'une phrase en fonction des règles grammaticales du langage ;

**b - sémantique :** liées à la signification des mots et aux concepts qu'ils véhiculent ;

**c - pragmatique :** plus spécifiques que les précédentes, elles contiennent les informations relatives au contexte de l'univers de l'application, en liaison avec la notion du dialogue.

## 2.2. Diverses stratégies d'utilisation des sources de connaissances

Les sources de connaissances définies plus haut doivent être connectées entre elles pour participer à la compréhension en échangeant des informations. Il faut donc définir une stratégie de contôle pour parvenir au résultat recherché en présence de bruit et d'erreurs.

Deux grandes organisations sont alors possibles : une de type vertical, l'autre de type horizontal [ 6 ].

## 2.2.1. Stratégies verticales ou "en profondeur"

On distingue ici deux approches très différentes (Fig. A2.1) :

a - **stratégie ascendante** : qui consiste à partir du signal vocal ou de sa transcription phonétique pour construire la phrase en remontant les différents niveaux. Autrement dit, le **décodage** se fait progressivement des entités les plus élémentaires vers les plus larges (des sons aux phonèmes puis aux mots et enfin à la phrase).

L'avantage de cette méthode est de limiter les effets de bruits et les erreurs que l'on rencontre dans la transcription phonétique de la phrase. En revanche, elle nécessite la construction au préalable de cette chaîne des sons élémentaires qui atteint très vite des dimensions importantes dans le cas de traitement de langage à vocabulaire étendu et, par suite, une telle analyse strictement ascendante devient beaucoup trop coûteuse si ce n'est inopérante.

b - **stratégie descendante** : cette méthode, au contraire, part du plus haut niveau d'abstraction pour **prédire** les mots qui peuvent prendre place à un endroit donné de la phrase. Suit une recherche phonétique pour vérifier si ces mots sont bien à la place supposée. Cette prédiction permet d'éliminer un certain nombre de mots condidats. C'est donc intéressant dans l'exploitatioon de lexiques assez importants. Néanmoins, ce type de méthode présente deux inconvénients :

- Si le vocabulaire est important, le nombre d'hypothèses à considérer reste important ; on risque de se perdre dans des constructions érronées ;

- De plus, ce type d'analyse est beaucoup plus sensible au bruit qu'une méthode ascendante, et il faut mettre en oeuvre des procédures lourdes et coûteuses, pour se synchroniser dans la chaîne d'entrée.

c - **stratégie mixte :** le problème avec une stratégie purement ascendante ou descendante est que l'on arrive pas suffisamment à contrôler l'espace de recherche des solutions, vue les inconvénients et les limites de ces méthodes. C'est pour cette raison que des solutions, que nous pouvons qualifier de mixtes, ont été proposées dans plusieurs systèmes de compréhension de la parole.

Ces méthodes se reposent sur deux étapes :

**1ère étape :** des méthodes ascendantes seront essentiellement utilisées pour déterminer le point de départ et des points de reprise du traitement par une recherche des mots clés ;

**2ème étape :** des méthodes descendantes, par contre, permettront de lever en partie l'indéterminisme lié à la structure du langage par des tests s'effectuant directement sur le signal vocal.

Les méthodes mixtes, bien qu'elles donnent des résultats satisfaisants, sont très coûteuses et lourdes à mettre en oeuvre.

### 2.2.2. Stratégies horizontales

Les méthodes d'organisation verticales, déja exposées, sont toujours combinées avec une stratégie dite "horizontale" qui peut être essentiellement de deux types (Fig. A2.2) :

a - **stratégie "gauche-droite" :** qui consiste à traiter le signal suivant son ordre de production du début jusqu'à la fin. Une telle méthode permet de prendre en compte le caractère prédictif de la production de la parole; néanmoins, l'inconvénient de cette stratégie est que si le premier mot n'est pas identifié correctement, ou s'il n'est pas identifiable, la compréhension du reste de la phrase est retardée.

b - **stratégie insulaire ou "du milieu vers les côtés" :** qui consiste à trouver le mot qui peut être identifié immédiatement, puis étendre la recherche à chaque côté de tous ces mots jusqu'à la compréhension de toute la phrase. Les principales difficultés sont dans ce cas :

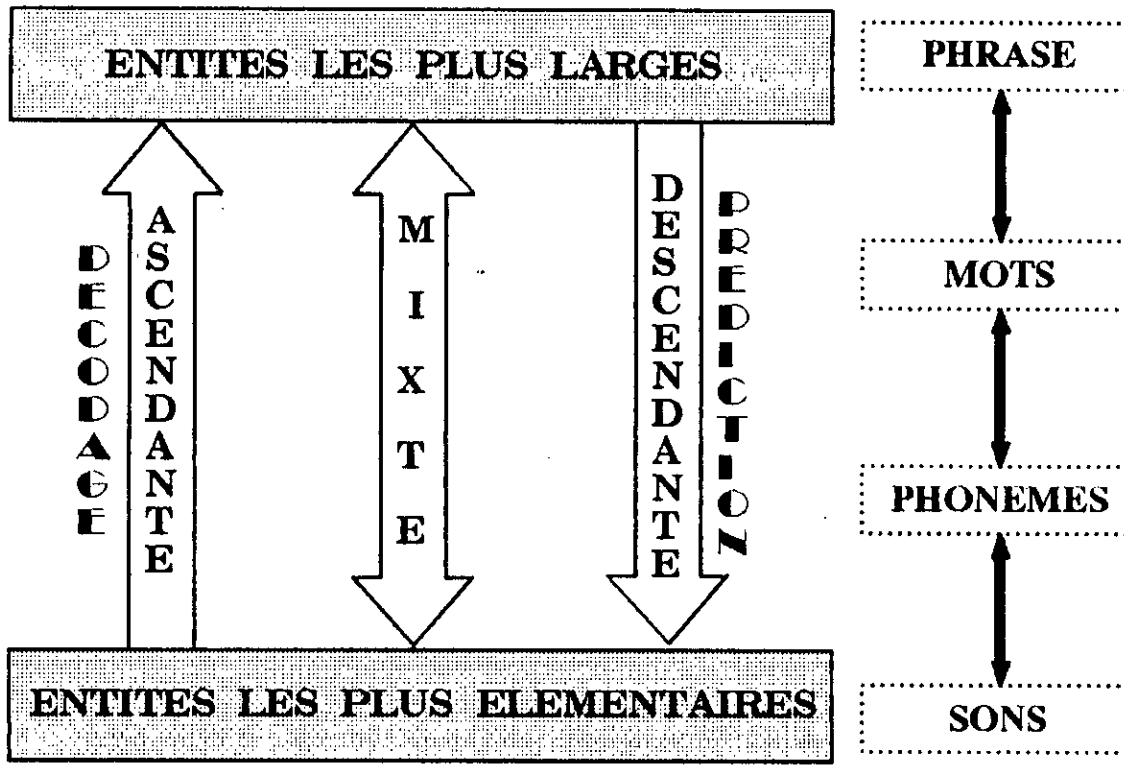



**Fig. A2.1 :** Schéma bloc de la stratégie verticale

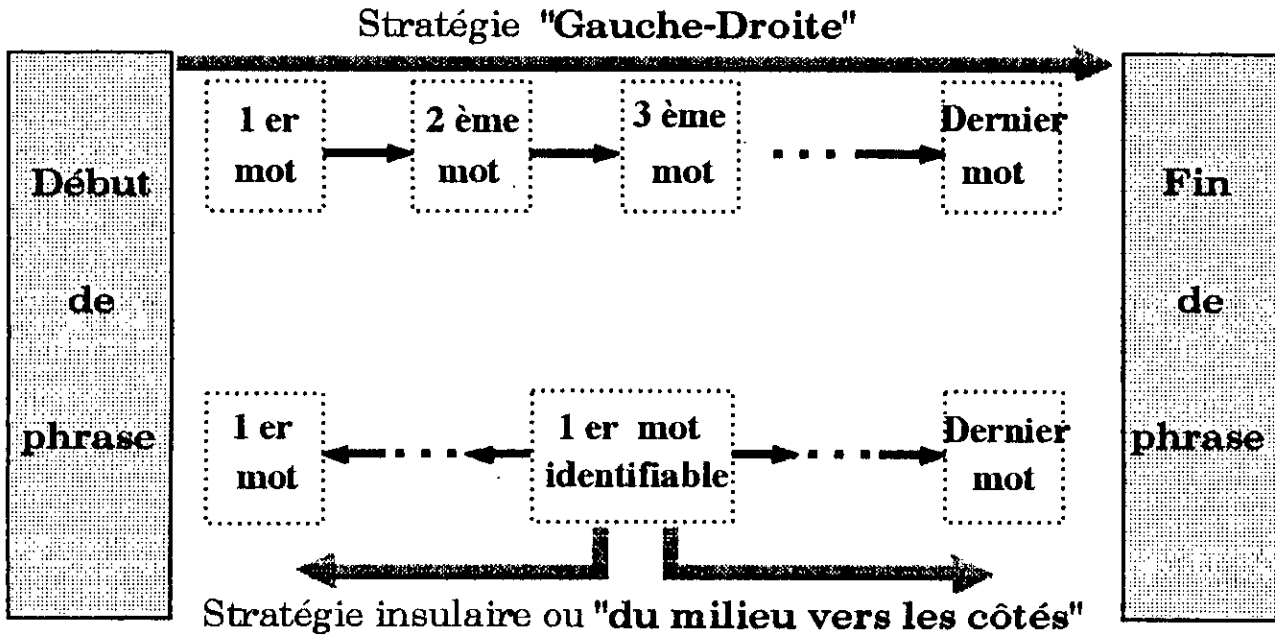



**Fig. A2.2 :** Schéma bloc de la stratégie horizontale

- Le nombre d'extensions hypothétiques des mots-clés ou ilôts peut être très grand,

- Le choix des mots-clés ou points d'ancrage est difficile et ne peut constituer des hypothèses réellement fiables.

c - **une stratégie hybride** entre la conduite insulaire et gauche-droite tente de comprendre n'importe lequel des trois ou quatre premiers mots. Alors l'extension de ce mot est alors dans une seule direction à la fois : premièrement à rebours vers le début de la phrase, et ensuite en continuant jusqu'à la fin. Ceci diminue de façon très importante le nombre d'hypothèses d'extension qui doivent être considérées en même temps.

## 3. DIFFERENTS MODELES EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Le choix d'un modèle adéquat pour représenter les diverses sources de connaissances disponibles est crutial pour ce qui est des performances d'un système de reconnaissance de la parole.

Les connaissances actuelles en linguistique, psychologique et perception ne permettent pas de s'appuyer sur un modèle, même élémentaire, du cerveau humain pour tenter d'implanter un système de reconnaissance de la parole.

En effet, la perception de la parole n'est pas seulement due à une extraction passive de traits phonétiques à partir du signal mais qu'elle met en jeu des phénomènes de mémorisation temporaire, de génération d'hypothèse, de synthèse interne avec comparaison au signal. Par exemple, il est prouvé que des mots isolés sont moins intelligibles que les mêmes mots prononcés dans une phrase, ce qui suggère que les informations ne sont pas seulement déduites du signal mais synthétisées intérieurement d'aprés le contexte syntaxique et sémantique.

Parmis les principaux modèles proposés, nous citerons les modèles stochastiques, connexionnistes et ceux à bases de connaissances et systèmes experts.

Le but de cette partie n'est pas de faire une étude exhaustive et complète de ces différents modèles, mais de donner une revue rapide des tendances générales de chacune d'elles ainsi que les avantages et les limites d'utilisation.

## a - Modèles stochastiques

Actuellement en reconnaissance automatique de la parole, une des directions de recherche les plus prometteuses consiste à modéliser l'ensemble des opérations de reconnaissance en termes de processus stochastiques. Les données statistiques sont alors apprises au préalable à partir de corpus importants de phrases du langage considéré.

Dans ces systèmes, le langage est modélisé par un ensemble de processus de MARKOV du premier ordre sous forme d'un réseau de transition stochastique intégrant les probabilités conditionnelles d'apparition de différentes unités linguistriques (phonèmes, diphones, mots). Ces probabilités sont évaluées à partir de très gros corpus de parole correspondant à plusieurs heures d'enregistrement. Cet apprentissage constitue une opération particulièrement critique pour les performances et très fastidieuse. L'organe central du système est un décodeur linguistique [ 7 ] dont le rôle est de trouver, à l'aide d'un algorithme de programmation dynamique, la phrase la plus probable en fonction de la chaîne de phonèmes en entrée. Le système **DRAGON** [ 8 ] réalisé à l'université de Carnegie-Mellon (U. S. A) présente un exemple des premiers systèmes fondés sur ces principes.

Le temps de calcul avec ce modèle est important, mais les performances figurent parmi les meilleures obtenues jusqu'à présent. Des travaux sont poursuivis sur la reconnaissance de phrases dictées avec de très grands vocabulaires (plusieurs milliers de mots) toujours à l'aide du même modèle, mais avec une prononciation mot à mot qui simplifie beaucoup le problème.

Une méthode un peu différente a été proposée récemment. Elle consiste à associer à chaque mot d'un vocabulaire un modèle de MARKOV caché (**H M M** : **H**idden **M**arkov **M**odel) [ 9 ] et à rechercher ensuite le modèle le plus ressemblant au mot à reconnaitre. Cette méthode, associée à un codage vectoriel spectral de la parole, a fourni de très bons résultats en reconnaissance de mots multi-locuteurs. Les performances sont presque aussi bonnes qu'avec le schéma classique comparaison dynamique et vocabulaire multi-références ; par contre, l'encombrement mémoire et le temps de calcul sont très largement réduits.

Il s'agit là d'une voie de recherche importante qui va être activement poursuivie à l'avenir.

## b - Modèles connexionnistes

Ce type de modèle est fondée sur une modélisation plus ou moins réaliste du cortex

humain. Il est constitué par l'interconnexion d'un très grand nombre de processeurs élémentaires inspirés du fonctionnement du neurone.

De façon générale, les modèles connexionnistes fournissent des résultats intérressants en reconnaissance automatique de la parole comme dans d'autres domaines de la perception [ 10 ]. Néanmoins, il faut espérer pour l'avenir une amélioration de ces modèles, en particulier par une meilleure prise en compte des données issues des neurosciences.

### c - Modèles à bases de connaissances et systèmes experts

Ces modèles mettent à profit l'expertise humaine dans un domaine précis, sous la forme des lois fondamentales et des relations régissant le phénomène considéré ainsi que des méthodes plus ou moins empiriques acquises par l'expérience. En ce sens, ces systèmes peuvent grandement aider à mieux comprendre et à formaliser les mécanismes humains de perception et d'interprétation dans des domaines comme celui de la compréhension de la parole.

Cette approche est donc très tentante dans ce domaine mais les difficultés sont très grandes, en particulier du fait de la rareté des vrais experts : comprendre une phrase est une tâche banale et naturelle mais qui fait cependant appel à des mécanismes les plus souvent inconscients et encore largement mal compris.

Le problème crutial dans ce type de modèle est donc celui de l'acquisition et la formalisation des connaissances, pour l'essentiel inconscientes chez l'auditeur humain. L'intelligence artificielle, par le biais des systèmes experts, propose une solution, à condition toutefois que l'expertise humaine existe. Dans de tels systèmes, la représentation des connaissances utilise souvent le formalisme ***des règles de production*** [ 11 ].

L'approche experte utilisant le formalisme des règles de production représente à notre sens la voie la plus prometteuse. C'est cette technique que nous adopterons pour notre modèle et fera l'objet d'une étude détaillée dans la partie B consacrée à l'étude et la réalisation du modèle proposé.

Néanmoins, remarquons que ces principaux modèles cités ne sont pas incompatibles : une solution optimale consisterait à faire coopérer deux d'entre eux, par exemple, un modèle stochastique avec un système à bases de connaissances pour bénéficier des propriétés complémentaires des deux modèles.

# 4. ETAT ACTUEL DES TRAVAUX ET PRINCIPAUX SYSTEMES

## 4.1. Objectifs et résultats du projet A R P A (Advance Reseach Project Agency)

Les programmes de reconnaissance et de la compréhension de la parole ont été développés principalement aux Etats-Unis dans le cadre du projet **A R P A** (Advance Research Project Agency, organisme faisant partie du Département de le Défence), qui a alloué 15 millions de dollars à la recherche dans cet axe.

Les objectifs de ce projet étaient les suivants [ 6 ] :

- reconnaître la parole continue ;
- plusieurs locuteurs, mais coopératifs ;
- environnement calme, avec un bon microphone et une lègère adaptation au locuteur ;
- vocabulaire de 100 mots strucuré éventuellement par une syntaxe artificielle ;
- moins de 10 % d' erreurs de compréhension ;
- réponse à quelques fois le temps réel.

Plusieurs systèmes ont été proposés dans le cadre de ce projet, quatre seulement ont été retenus et téstés sur un nombre de phrases ; il s' agit de :

- **B. B. N** (Bolt, Baranek and Newman) [ 12 ].

- **S. D. C** (System Devepment Corporation) [ 13 ].

- **HARPY** [ 14 ] et **HEARSAY II** [ 15 ] de **C. M. U** (Carnegie Mellon University).

Nous n' allons pas faire une description de ces systèmes, nous nous contenterons seulement de donner les résultats des tests qu' ils ont subis ainsi que les pricipaux enseignements tirés de ce projet.
Les résultats des tests étaient les suivants :

- la reconnaissance de la parole continue était reconnue "possible" ;

- les meilleurs systèmes nécessitaient une adaptation au locuteur par une phrase d' apprentissage de 20 à 60 phrases et étaient testés avec seulement 5 locuteurs. L' aspect multilocuteur était donc loin d' être maitrisé ;

- Les vocabulaires testés faisaient effectivement plus de 1000 mots ;

- les réponses des systèmes les plus rapides intervenaient en quelques minutes, celles de B.B.N parfois en quelques heures ; on était donc loin du temps réel ;

- seuls, les systèmes HARPY et HEARSAY II réalisaient moins de 10 % d'erreurs sur la signification du message.

- Bien qu'aucun des systèmes proposés n'ait réussi à répondre à tous les objectifs, les progrés réalisés au cours de cette période grâce au projet ARPA ont été remarquables. Le principal apport a été l'intégration et le contrôle de plusieurs sources de connaissance différentes (acoustique, phonétique, syntaxique, sémantique et pragmatique) participant à divers niveaux à la compréhension finale d'une phrase. Cela nécessitait pour permettre une coopération fructueuse entre ces diverses sources de connaissance d'importantes études sur l'architecture des systèmes.

## 4.2. Principaux enseignements tirés des systèmes existants

Bien que des progrés aient réalisés ces dernières années, il reste encore beaucoup de travail à réaliser dans le domaine de la compréhension de la parole. Il représente un potentiel important et nécessite encore de nombreuses années de recherche.

Néanmoins, un certain nombre d'enseignements généraux doivent être dégagés car ce sont eux qui vont nous guider à construire notre système modèle qui fera l'objet d'une étude complète dans la prochaine partie.

Les paragraphes qui suivent dressent la liste d'un certain nombre d'idées tirées des différents systèmes existants et spécialement ceux du projet ARPA. Ces idées portent sur :

### a - Organisation du système

Beaucoup a été appris du projet **ARPA** sur l'organisation et la représentation des hypothèses à différents niveaux de connaissance ; l'un des cadres de travail les plus flexibles à avoir émergé est l'organisation en tableau du système HEARSAY qui permet l'execution en parallèle de modules asynchrones, sur un ou plusieurs processeurs, communiquant par une base de données commune pour des applications plus générales. A l'autre extrémité, la structure de connaissance précompilée de HARPY la rend difficile à modifier mais a pour résultat une très haute performance pour des applications limitées.

## 4.3. Conclusion

Ces différentes remarques montrent, qu'on est encore loin de pouvoir résoudre le problème de la reconnaissance de la parole dans toute sa totalité, faute de connaissances suffisantes sur le processus de compréhension, mis en oeuvre dans la communication orale.

Tous les systèmes actuels sont encore dans les laboratoires et rien ne permet d'envisager dans un avenir proche la commercialisation d'un appareil complet de reconnaissance de la parole continue, c'est-à-dire reconnaissant des phrases prononcées naturellement et utilisant un vocabulaire relativement peu contraint tant au niveau de la syntaxe que de la sémantique. Néanmoins, les recherches en ce domaine, sont donc encore largement ouvertes, mais vraissemblablement pour plusieurs années, voire plusieurs décennies.

# 5. SCHEMATISATION DU PROCESSUS DE LA COMPREHENSION DE LA PAROLE

Le processus de la compréhension de la parole peut être schématisé très simplement selon la figure A2.3. On distingue les modules des niveaux inférieurs et les modules des niveaux supérieurs [ 16 ].

La distinction entre haut et bas niveaux peut être vue en terme des connaissances disponibles dans chaque cas. En effet, les processus de haut niveau raisonnent à partir des connaissances spécifiques au domaine (syntaxique, sémantique, pragmatique). En revanche, les traitements bas niveau opèrent sur les données physiques du signal (acoustique, phonétique) en mettant en jeu des modèles généraux (techniques de traitement du signal, segmentation, etc.).

La tâche des niveaux inférieurs dans un tel système est d'utiliser ses propres connaissances qui conduisent à des hypothèses sur les mots prononcés. Ces hypothèses sont confirmées par les modules des niveaux supérieurs qui lèvent les ambiguités subsistantes, en fonction de la grammaire autorisée, de la signification des mots condidats et du contexte de l'application considérée.

L'interêt de l'application de ces niveaux supérieurs est d'une part d'éviter que les niveaux inférieurs génèrent un grand nombre d'hypothèses lexicales linguistiquement incorrectes et qui devraient ensuite être rejetées par l'analyseur linguistique représenté par les niveaux supérieurs (tâche à la fois coûteuse et difficile), et d'autre part, permet de

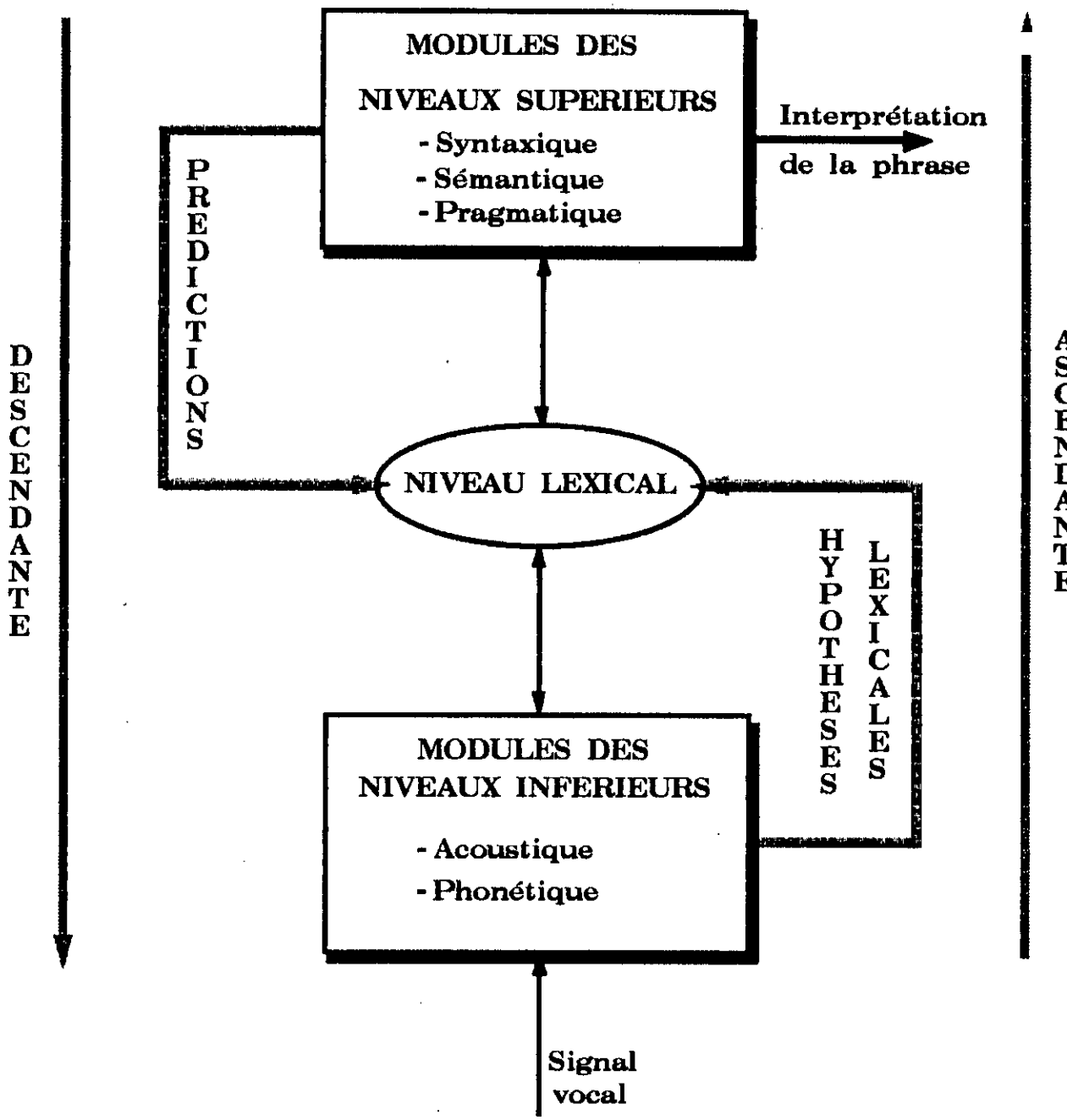



Fig. A2.3 : Vue simplifiée du processus de la compréhension de la parole d'après [16]

prédire la suite de la phrase, et donc d'accélérer la reconnaissance de celle-ci. On arrive ainsi à une boucle "reconnaissance phonétique - compréhension - prévision". Si la partie dominante de la boucle est la reconnaissance phonétique, on a une ***stratégie montante*** (ascendante) de reconnaissance. Si la partie dominante est la partie "compréhension-prévision", on a en revanche une ***stratégie descendante***. Ce processus paraît présent dans le mode de reconnaissance d'un auditeur humain, puisqu'il a été montré que ceux-ci peuvent détecter la présence d'un ou plusieurs mots dans une phrase avant même qu'ils n'aient été complètement prononcés.

L'un des principaux soucis des chercheurs est qu'il est nécéssaire d'utiliser au mieux ces différentes sources de connaissances afin d'aboutir à un modèle qui peut représenter le mieux le processus de la compréhension de la parole.

## 6. CONCLUSION

L'ampleur du problème et la nécessité d'apporter des solutions spécifiques font de la reconnaissance de la parole un champ d'application important et souvent original de l'intelligence artificielle. L'intéraction entre les aspects perceptifs et cognitifs constitue une des particularités du domaine.

Après un premier chapitre consacré à une brève revue de l'ensemble de la reconnaissance de la parole, nous avons tenté dans ce second chapitre de dresser un bilan de la reconnaissance-compréhension de la parole continue à la lumière des grandes techniques employées. Nous avons, dans la mesure du possible, illustré les principes généraux et présenté quelques exemples de systèmes réalisés dans divers laboratoires en dressant une liste des principaux enseignements dégagés de ces systèmes.

Le bilan total est mitigé. En effet, des systèmes, parfois très ambitieux, ont été réalisés et ont montré que le problème était soluble. Néanmoins, de gros problèmes demeurent, que l'intelligence artificielle devrait aider à resoudre. Il s'agit de questions concernant surtout les aspects particuliers du décodage acoustico-phonétique (une des faiblesses majeures), de l'utilisation de la prosodie, etc, mais aussi de problèmes plus globaux tels que la conception de modèles et d'architectures adaptés au dialogue oral homme-machine.

# *PARTIE B*

## ETUDE ET REALISATION DU SYSTEME DE RECONNAISSANCE ACOUSTICO-PHONETIQUE PROPOSE

B1 : ETUDE DU TRAITEMENT ACOUSTICO-PHONETIQUE

B2 : MISE EN OEUVRE DU DECODEUR ACOUSTICO-PHONETIQUE

B3 : EXPERIENCES, TESTS ET RESULTATS

B4 : PRESENTATION ET UTILISATION DU SYSTEME **S R A P H** REALISE

# *CHAPITRE B1*

# ETUDE DU TRAITEMENT ACOUSTICO-PHONETIQUE

Après avoir étudié au cours de la partie A les divesrs sources d' informations à prendre en compte dans un système complet de reconnaissance et de compréhension de la parole continue, nous allons consacrer cette seconde et dernière partie à l' étude, la mise en oeuvre ainsi que la présentation du système de reconnaissance Acoustico-Phonétique que nous avons réalisé sans oublier les tests et les résultats obtenus.

## 1. PRESENTATION GENERALE

Comme nous l' avons déja souligné précédemment, deux méthodes conceptuellement différentes permettent d' aborder la reconnaissance de la parole. Toutes deux sont fondées sur la segmentation du signal en unités :

- mots pour les méthodes globales,

- syllabes, diphones, phonèmes pour les méthodes analytiques.

Pour la reconnaissance de vocabulaires limités, l' extraction directe de mots dans le signal a été utilisée avec succès. Cependant, dès lors la reconnaissance de grands vocabulaires ou de parole continue est envisagée, il devient nécessaire d'extraire du signal des unités plus fines : c' est le rôle du décodage acoustico-phonétique. Ce niveau de traitement constitue une étape majeure et reste encore à l' heure actuelle un problème-clef, du fait de la redondance en informations linguistiques et extra-linguistiques du signal vocal qu' il faut réduire dans des proportions importantes.

Le Décodage Acoustico-Phonétique (D A P) représente le premier maillon de reconnaissanc. Tous les traitements linguistiques ultérieurs sont largement tributaires de la qualité de cette transcription phonétique, même s' ils peuvent, dans une certaine mesure, l' améliorer. Toute erreur à cette étape augmente donc de façon significative l' indéterminisme des modules supérieurs.

Le but recherché par cette étape de traitement est d' obtenir une chaîne ou un treillis d' unités minimales qui le plus souvent sont de type pseudo-phonèmes. Nous parlerons en effet plutôt de pseudo-phonèmes que de phonèmes, car les unités ainsi obtenues ne correspondent pas forcément à des phonèmes au sens précis utilisé par les phonéticiens.

La difficulté de ce problème tient pour une grande part à la variabilité inter- et intra-locuteur et à la nature continue de la parole qui, par le fait de la coarticulation, montre une forte intéraction des sons les uns sur les autres.

Historiquement, la première approche de ce problème a été la mise en oeuvre de techniques de reconnaissance de formes par extraction et classification d' un certain nombre de paramètres acoustiques. Le décodeur complète l' analyse par l' utilisation de techniques de comparaison à des prototypes. Ceux-ci se calculent lors d' une phase d' apprentissage pendant laquelle un locuteur prononcera plusieurs fois, dans des contextes différents, chacun des phonèmes de la langue choisie. Un prototype sera par exemple constitué par un spectre moyen du phonème qu' il décrit.

Un tel traitement n' est cependant pas totalement satisfaisant. En effet, ces décodeurs se caractérisent par un nombre élevé de paramètres liés au locuteur. De ce fait, l' aspect multilocuteur est difficilement intégrable, même s' il existe des procédures d' adaptation automatique [ 17 ].

Par ailleurs, le phonème est une unité très variable et fortement influencée par le contexte phonétique dans lequel il apparait. En effet, pour des problèmes liés à l'articulation, l' image acoustique d' un phonème sera modifiée par les deux phonèmes antérieur et postérieur du contexte. Dans certains cas, deux phonèmes différents peuvent se réaliser quasi identiquement, et seul le contexte permet la discrimination. Dans une approche de type reconnaissance des formes, la prise en compte de phénomènes contextuels nécessite de recourir à un nombre prohibitif de références. Le français, par exemple, comportant environ 30 phonèmes, il faudrait étudier de l' ordre de 30 X 30 X 30 contextes différents pour chaque locuteur.

Des méthodes de décodage en unités plus grandes que le phonème permettent de prendre en compte partiellement les problèmes de coarticulation. Parmis ces méthodes, citons l'analyse par diphonèmes [ 18 ] ou par demi-syllabes [ 19, 20 ].

Dans tous ces systèmes, les taux de réussite sont encore insuffisants (moins de 70 % en monolocuteur). Si cette approche est satifaisante dans les systèmes de reconnaissance travaillant sur les langages artificiels très contraints, le passage à la reconnaissance de grands vocabulaires et l'utilisation de langages quasi-naturels nécessitent des taux de reconnaissance bien supérieurs.

Une nouvelle voie de recherche s'est récemment développée. Elle consiste à incorporer dans les décodeurs des connaissances phonétiques de la parole. Cette orientation vers les techniques d'intelligence artificielle est encouragée par les performances obtenues par des experts humains en phonétique dans la lecture des spectrogrammes. De l'observation d'un ensemble de lecteurs de spectrogrammes, il ressort que [ 21 ] :

- le signal acoustique est **riche en informations phonétiques,**

- de telles informations peuvent être **extraites à partir d'une représentation spectrographique** du signal,

- le processus d'extraction de l'information **fait référence à des règles explicites** qui peuvent être enseignées.

De ce fait, l'amélioration des algorithmes de décodage peut être entreprise par l'analyse et la modélisation du savoir faire d'un expert en lecture de spectrogrammes. En effet, les performances atteintes par un expert humain dépassent les 80 % en contexte multilocuteur. La figure B1.1 donne un exemple de spectrogramme de parole sur les transcriptions phonétiques fournies par un expert phonéticien extrait du projet **SYSTEXP** développé par [ 22 ].

## 2. DIFFERENTES ETAPES DU DECODAGE ACOUSTICO-PHONETIQUE

Traditionnellement, dans un système de reconnaissance et de compréhension de la parole, le module de décodage acoustico-phonétique s'insère entre le module de traitement du signal et les modules linguistiques, dans une perspective montante ou descendante et remplit les fonctions suivantes :

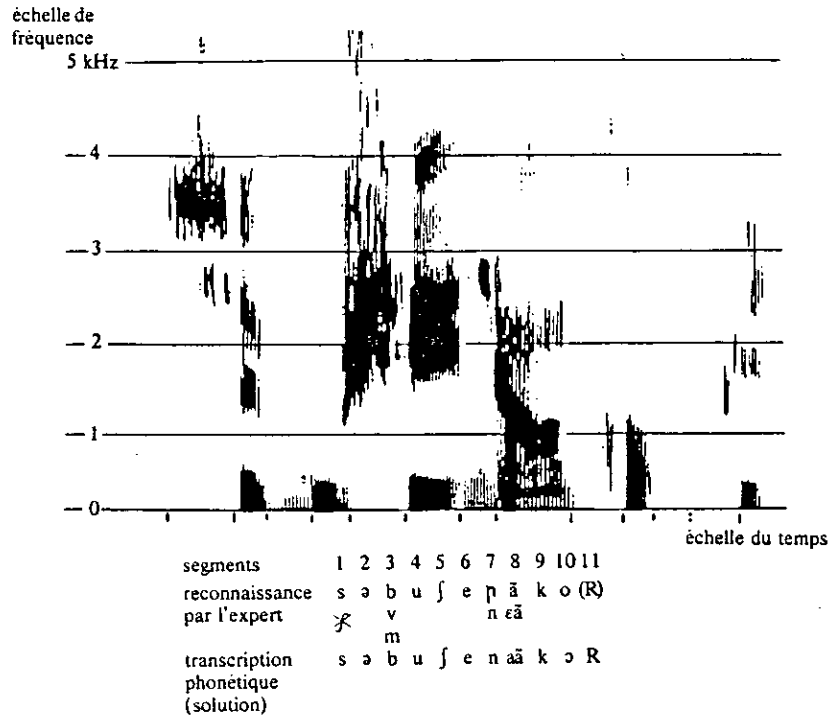



**Fig. B1.1** : Exemple de spectrogramme de la parole avec les transcriptions phonétiques exactes, fournie par l' expert et le système expert SYSTEXP [ 22 ]

a - extraction des paramètres pertinents (indices, correlats, etc.) ;

b - segmentation de la parole en unités définies ;

c - identification de ces unités ou du moins leur donner des attributs significatifs (étiquettes de classe ou de macro-classe, traits, etc.).

Ces différentes fonctions interagissent entre elles et avec les modules linguistiques environnants.

## 2.1. Extraction des paramètres du signal de la parole

Cette étape consiste à extraire du signal vocal un nombre restreint de paramètres pertinents, tout en éliminant l'information redondantes. Il n'existe pas de solution mathématique générale à ce problème et il est de toute façon impossible de séparer exactement information utile et information redondante dans la parole néanmoins un certain nombre de méthodes ont été proposées [ 23, 24 ], les algorithmes pouvant être réalisés de façon logicielle ou matérielle (processeurs câblés).

Nous nous contenterons ici de donner qu'un aperçu général des principales méthodes d'analyse utilisées.

### 2.1.1. Méthodes temporelles

Elles consistent à travailler directement sur le signal fonction du temps. On distingue :

**a - passage par zéro :** C'est la méthode la plus simple qui consiste à compter le nombre de passage par zéro du signal, ce qui permet d'en déduire approximativement les deux premiers formants de la parole qui sont des paramètres intéressants en reconnaissance. En effet, de telles mesures sont utiles lorsqu'il s'agit de classer les sons vocaliques.

Economiquement, cette méthode, est avantageuse dans la mesure où sa mise en oeuvre est très simple, cependant, son principal inconvénient est qu'elle est peu précise à cause de l'instabilité de ces formants.

**b - prédiction linéaire :** Cette méthode est particulièrement bien adaptée à l'étude de la parole. Elle consiste à prédire approximativement la valeur d'un échantillon de parole à partir d'un développement limité des échantillons précedents sur une base polynomiale orthogonale. La méthode permet d'obtenir des coefficients de prédiction et la valeur des formants. C'est le plus précis et le plus puissant des procédés actuellement utilisés mais les algorithmes mis en oeuvre s'avèrent le plus souvent trop sophistiqués pour être implémentés en temps réel, sauf dans le cas des processeurs câblés.

## 2.1.2. Méthodes spectrales

Ces méthodes s'intéressent à la représentation dans le domaine fréquentiel du signal vocal, ce qui correspond dans une certaine mesure à ce que fait l'oreille humaine lorsqu'elle effectue une analyse spectrale grossière de la parole. On ditingue :

**a - l'analyse par banc de filtres :** plus souvent utilisée pour des raisons de rapidité. Le spectre de la parole est divisé en 20 à 30 bandes de fréquence et un filtre correspondant à chacune de ces bandes. Le signal sonore passe à travers ces filtres qui mesurent l'énergie contenue dans chaque bande de fréquence. On peut comparer directement ces niveaux d'énergie avec ceux d'une empreinte. Le banc de filtres peut être aussi bien numérique (logiciel) que matériel (le vocodeur à canaux par exemple).

**b - Transformée de FOURIER Rapide (T. F. R) :** l'analyse s'effectue dans ce cas par calcul à partir d'un signal échantillonné. Cette méthode donne le donne le spectre amplitude d'un signal à partir des variations temporelles, ce qui permet d'évaluer les formants et la fréquence fondamantale. Cet algorithme constitue l'une des nombreuses façons de traiter l'information donnée par une batterie de filtres sous forme purement numérique (logiciel).

Les avantages majeurs de cette méthode résident dans la facilité de mise en oeuvre avec un ordinateur.

A ces méthodes on peut ajouter aussi l'analyse spectrographique de la parole grâce aux données issues d'un spectrographe. Les informations ainsi extraites peuvent être très diverses (distribution de l'énergie, bandes formantielles, transitions formantiques, etc.) et les résultats expérimentaux montrent que certains ont un bon pouvoir discriminant entre différents types de segments minimaux. Cette technique fera l'objet d'une étude particulière dans le chapitre B4 et sera adoptée, par la suite, dans les différents tests que nous avons effectué.

Enfin, notons qu'un autre type de méthodes de traitement du signal de parole, moins utilisé mais fournissant de bons résultats correspond à l'utisation d'un modèle d'oreille [ 25 ].

C'est sur la base de résultats fournis par ces différentes méthodes qu'on va tenter d'une part, la segmentation de parole en unités minimales et d'autre part, l'identification de ces unités.

## 2.2. Segmentation

### 2.2.1. but

Le but de la segmentation est de décomposer automatiquement le signal en une suite de segments tels que chacun d'eux corresponde à la réalisation d'un phonème, d'une syllabe, d'un mot, d'un syntagme, etc.

A ce niveau, deux problèmes se présentent : d'une part, choisir l'unité de décision, d'autre part, opérer une segmentation correcte, ou du moins cohérente.

Nous ne nous intéresserons ici qu'à la segmentation en phonèmes, mais malheureusement, il n'existe pas de liens directs simples entre un phonème et sa représentation acoustique. Les deux propriétés essentielles qui caractérisent le signal vocal sont sa continuité (les frontières entre les différents types de segments qui composent le signal vocal ne correspondent pas toujours au découpage phonémique) et sa variabilité (un phonème n'est pas associé de manière biunivoque à un type de son auquel correspondraient des propriétés acoustiques bien définies. Il est au contraire sujet à des différentes réalisations dues à des influences contextuelles très fortes et diverses, telles que la coarticulation, l'accent, l'émotion, le débit d'élocution, le locuteur, etc.).

Ces obstacles ne sont cependant pas insurmontables et des solutions tout au moins partielles existent.

Parmis les méthodes de segmentation les plus utilisées, la méthode de préclassification en grandes classes phonétiques [ 26 ] s'avère très intéressante dont nous proposerons ici d'en exposer les principes.

### 2.2.2. Principe de la segmentation par préclassification

Il s'agit dans cette méthode de réaliser une préclassification en grandes classes phonétiques par une stratégie de tests sur des paramètres liés au spectre, les zones soutenues (stables) du spectre : voyelles, fricatives, occlusions de plosives et les zones transitoires : explosion de plosives et transitions. L' algorithme, ainsi constitué, effectue une première partition de l'espace des formes a pour but de :

- réduire l'explosion combinatoire lors de la reconnaissance ce qui se traduit par un gain de temps apprécible.

- permettre un cadrage en vue d'une identification automatique.

Trois grandes classes ont été retenues :

* les noyaux vocaliques (voyelles : / i /, / a /, / u /),
* les fricatives {/ f /, / s /, / ʃ /, / z /, / Z / } + burst,
* les occlusives {/ p /, / t /, / k /, / b /, / d /, / g / }.

Le phonème / v / est une fricative, mais il n'a pas été retenu car il présente, pour de nombreux locuteurs, des caractéristiques proches de celles d'un / b /.

La procédure dans chacune des classes est la suivante :

**a - Noyaux vocaliques**

Pour déterminer les noyaux vocaliques, deux courbes d'énergie sont utilisées :

- valeur de l'énergie dans une bande de fréquences comprises entre 250 et 2500 Hz (zone où sont localisés les deux premiers formants),

- énergie totale.

La première courbe est obtenue en sommant parmi les zones correspondantes aux fréquences comprises entre 250 et 2500 Hz celles qui atteignent le seuil de visibilité sur le spectrogramme.

La seconde est obtenue en calculant l'énergie du signal temporel.

Sur ces deux courbes les maximas recherchés doivent vérifier :

- une intensité au moins égale à la moitié de l'énergie du pic précédent,
- une vallée droite et gauche suffisante,
- la présence du voisement.

Cette procédure permet en outre de calculer une durée vocalique moyenne qui donne une indication sur la vitesse d'élocution.

**b - Fricatives**

Deux courbes sont à calculer dans ce cas :

- la courbe des passages par zéro sur le signal filtré par un filtre passe haut dont la fréquence de coupure est de 800 Hz,
- la courbe des centres de gravités calculés sur les parties du spectre visibles sur les spectrogrammes numériques.

Une fricative est détectée si un maximum local est mis en évidence sur ces deux courbes.

**c - Occlusives**

Une courbe d'énergie est calculée sur le signal temporel préaccentué et filtré par un filtre passe haut dont la fréquence de coupure à pour valeur 600 Hz. Les occlusives correspondent à un minimum local sur cette courbe.

En fait, les performances d'un algorithme de segmentation ne peuvent être évaluées qu'en fonction de celle des étapes ultérieures du traitement de reconnaissance. Les niveaux successifs corrigent éventuellement des erreurs de segmentation ou demandent une nouvelle segmentation par retour en arrière (backtracking).

## 2.3. Identification

L' identification est l' opération qui consiste à attribuer aux unités acoustiques (segments) issues de l' étape de segmentation, des unités de natures phonétiques (traits, macro-classes, phonèmes, etc.).

Cette étape est caractérisée par la présence de multiples sources de connaissances, éventuellement érronées et souvent entachées d' erreurs. Il en résulte à ce niveau un fort indéterminisme nécessitant la mise en oeuvre de structures de contrôle et de stratégies élaborées et robustes.

Pour répondre à ces exigences, un certain nombre de solutions sont proposées . Parmis les méthodes les plus utilisées à l' heure actuelle, on trouve les méthodes de classifications automatiques (algorithmiques) qui dérivent des procédures générales de reconnaissance de formes (codage vectoriel [ 27 ], modèles de Markov [ 28 ], réseaux de neurones [ 29 ], D T W (Dynamic Time Warping Algorithms) [ 30 ], etc.) et celles qui proviennent des systèmes experts. Nous ne donnerons qu' un aperçu général sur le premier type utilisant l' algorithme de comparaison dynamique, par contre, nous nous focaliserons dans notre étude sur l' approche de type système expert qui sera décrite plus en détail et mise en oeuvre dans notre système.

### 2.3.1. Algorithme de comparaison dynamique

Il s' agit de comparer un segment inconnu aux segments de référence stockés en mémoire (une trentaine pour les phonèmes). Le calcul d' un taux de similitude globale entre deux segments pose le problème de leur normalisation en durée car les variations que l' on constate d' un segment à l' autre peuvent être très importantes (du simple au double par exemple). Un algorithme de comparaison dynamique permet d' effectuer une normalisation non linéaire, et optimale des segments [ 31 ].

La méthode permet de mettre dynamiquement en correspondance les échantillons de parole en fonction de leur similitude, sans tenir compte de leurs positions respectives dans le temps. On obtient ainsi une normalisation non linéaire en cours de comparaison, adaptée au traitement des variations de rythme, fréquentes dans la parole.

Cet algorithme s' est révélé très performant, d' abord sur les mots isolés, puis sur les segments. Plusieurs optimisations y ont été apportées, et un processeur câblé a été réalisé à cet effet [ 32 ].

### 2.3.2. Approche système expert

Les méthodes classiques d' identification (comparaison à des formes de références ou traitements statistiques) ne peuvent pas fournir des performances satisfaisantes dans le cas du codage multi-locuteur de la parole continue. Le nombre de références serait prohibitif et, de plus, la solution au problème ne peut pas être codée sous forme d' un algorithme de taille raisonnable. Des approches de type système expert [ 33 ], similaire dans la démarche, mais différentes en ce qui concerne la méthodologie adoptée, sont avérées très intéressantes pour introduire des connaissances phonétiques explicites a priori. Cette solution se fonde sur l' utilisation explicite des connaissances acquises par l' étude de l' activité de lecture de spectrogrammes vocaux par un expert phonéticien. Cependant cette dernière soulève les questions de stratégies et de représentation de connaissances.

#### 2.3.2.1. Représentation des connaissances

Le processus d' identification met en jeu plusieurs sources de connaissances : acoustiques, phonétiques et articulatoire. Cependant toutes ne sont pas toujours facile à formaliser. Néanmoins, on doit sélectionner, à partir de ce flôt de connaissances, celles qui paraissent les plus fiables et choisir un modèle de représentation.

Le développement récent de l' intelligence artificielle a favorisé l' éclosion d' outils facilitant les problèmes de représentation, de formalisation et de manipulation de connaissances, d' où l' apparition, ces dernières années des approches du type système expert utilisant le formalisme des ***règles de production*** pour la représentation de ces connaissances, à cause de leur facilité d' implantation et du caractère naturel de cette représentation.

Rappelons brièvement qu' une règle de production s' écrit sous la forme :

si **Condition** alors **Conclusion**

Où "**Condition**" représente, en général, une conjonction de prédicats qui doivent être vérifiés pour que la règle soit applicable. Son application résulte alors en "**Conclusion**", émission ou modification d' hypothèses permettant de se rapprocher du but.

Cette utilisation d' une règle est appelée *chaînage avant*. En fait, le formalisme des règles se prête aussi bien à un *chaînage arrière* consistant à considérer la partie droite d' une règle (Conclusion) comme un but à atteindre et la vérification des prémisses de la règle (partie gauche) comme autant de sous-problèmes à résoudre pour atteindre ce but.

Les modèles fondés sur un ensemble de règles se sont déja avérés performants, particulièrement si l'on sait intégrer dans les règles les connaissances accumulées par l'étude de nombreux cas particuliers. Une façon d'intégrer ces connaissances consiste précisément à formaliser l'expertise du phonéticien capable de lire des spectrogrammes de parole grâce à l'expérience qu'il a accumulée [ 34 ]. L'acquisition d'une telle expertise et, plus généralement les connaissances sur la parole devrait donc permettre d'améliorer de façon significative les systèmes de segmentation et d'identification phonétique.

### 2.3.2.2. Nature des connaissances

Les connaissances expertes sont essentiellement (mais pas seulement) formalisées sous forme de règles [ 35 ]. Les prémisses portent sur le contexte gauche, sur le contexte droit et sur les déductions déja faites sur le segment sur lequel on travaille; elles comprennent également une expression logique qui porte sur les indices acoustiques mesurés sur le signal. Les mesures peuvent concerner le segment sur lequel on fait l'analyse, le segment précédent ou le suivant.

Il existe deux types de règles, d'une part des règles qui donnent en conclusion une liste des phonèmes pondérés, d'autre part des règles qui modifient la segmentation. Plusieurs centaines de règles sont nécessaires pour obtenir des résultats tangibles. Ces règles se composent de plusieurs parties, pouvant être facultatives :

* un numéro de règle,
* une partie contexte gauche (liste de phonèmes),
* une partie contexte droit (liste de phonèmes),
* une partie des déductions déjà faites (liste de phonèmes),
* une partie prémisses ; conditions sur des mesures effecuées sur le segment,
* une partie conclusion ; soit une action à déclencher pour modifier la segmentation ou le treillis, soit une liste de phonèmes pondérés.

### 2.3.2.3. Stratégie employée

La stratégie de déduction employée dans cette étape doit se rapprocher le plus possible de celle de l'expert. Elle doit permettre :

- la remise en cause de la segmentation à tout moment,

- le déroulement en parallèle de l'analyse sur plusieurs segmentations (plusieurs lignes simultanées de raisonnement). Si l'expert se trouve en présence d'un segment trop long pour être un segment unique, il va essayer deux segmentations : premièrement un seul segment allongé par le contexte droit, deuxièmement deux segments. Il ne se décidera que lorsque le segment suivant aura été identifié,

- l'introduction d'un indéterminisme dans le raisonnement,

- le mode de chaînage doit être mixte :

  * chaînage avant (analyse guidée par les indices présents dans le signal),

  * chaînage arrière pour vérifier la validité des hypothèses.

En effet, les règles sont contextuelles. Elles portent à la fois sur le contexte gauche et le contexte droit du segment analysé. L'analyse étant de gauche vers la droite, les éléments du contexte droit sont inconnus et donnent lieu à des hypothèses qu'il conviendra par la suite de vérifier. Par ailleurs, un contexte gauche peut être constitué de plusieurs hypothèses qu'il conviendra de confirmer ou d'infirmer.

Pour répondre à ces différentes exigences, on utilise généralement [ 22 ], pour la mise en oeuvre de ce module, une structure d'arborescence dans laquelle chaque noeud contient un numéro de segment, la liste des règles activées et une base de faits. Le père d'un noeud contient les mêmes informations pour le segment précédent. Chacun des fils d'un noeud contient les différents contextes droits possibles pour ce segment. Nous obtenons à la fin de l'analyse un treillis de phonèmes et non pas simplement une suite de listes d'hypothèses phonétiques.

**La syntaxe d'une règle est la suivante :**

- Numéro de règle (NR)
- **SI** CONTEXTE_GAUCHE (liste de phonèmes) (LCG)
- **SI** CONTEXTE_DROIT (liste de phonèmes) (LCD)
- **SI** DEDUCTIONS_DEJA_FAITES (liste de phonèmes) (LDF)
- **SI** Conditions sur des mesures effectuées sur le signal
- **ALORS** (liste de phonèmes résultat) (LPR)

Nous indiquerons entre parenthèses les oms que nous donnons dans la suite aux différentes variables qui interviennent.

**Stratégie de choix des règles à activer**

Chaque noeud N de l'arbre de raisonnement contient :

- le numéro du segment sur lequel on travaille ( N° ).
- la liste des phonèmes déja trouvés ( LT ).
- la liste des phonèmes correspondant au contexte droit supposé ( LCDS ).
- la liste des règles déjà appliquées ( LR ).

Les règles dont l'intersection entre la LPR (résultat de la règle) et la LCDS du père (le contexte droit du père c'est le segment actuel) est non vide sont selectionnées en premier (fonctionnement en chaînage arrière). On élimine de cet ensemble toutes les règles dont :

- l'intersection entre la LCDS (contexte droit du noeud) et la LCD (contexte droit de la règle) est vide,

- l'intersection entre la LT du père du noeud (contexte gauche) et la LCG (contexte gauche de la règle) est vide,

- l'intersection entre la LDF (les déductions déja effectuées) et la LT (phonèmes résultats du noeud) est vide.

Si une règle est activée et que son contexte droit est plus restrictif que l'actuel contexte droit supposé, On crée un noeud frère : toutes les caractéristiques des deux frères sont identiques, excepté pour la liste LDS (contexte droit supposé).

- pour l'un des noeuds LCDS = intersection entre LCDS et LCD.

- pour l'autre LCDS = différence entre LCDS et LCD.

Si aucune règle ne s'applique pour un noeud, on détruit ce noeud et tous les fils. C'est le cas quand on a hypothétisé un contexte droit qui n'est pas confirmé lorsqu'on passe au segment suivant. Seuls les chemins de l'arbre qui atteignent le dernier segment seront conservés pour les niveaux supérieurs.

**Illustrons cela par un exemple [ 22 ] :**

Supposons que nous sommes à un noeud dont la liste CONTEXTE_DROIT (LCDS) est ( / p /, / t /, / k /, / b /, / d /, / g / ) (classe des plosives). Nous appliquons une règle dont la liste CONTEXTE_DROIT est ( / d /, / t /, / n / ) (classe des dentales). On va donc modifier la LCDS du noeud en ( / p /, / b /, / k /, / g / ) (différence) et créer un frère dont la LCDS sera ( / d /, / t / ) (intersection).

Pour résumer le fonctionnement du système nous donnons ci-dessous un exemple portant sur l'identification d'une suite de quatre segments extraits d'une phrase réelle.

La base de faits concernant ces segments est la suivante :

| | | | |
|---|---|---|---|
| segment 1 | formant 1 | = 800 Hz | formant 2 = 1000 Hz |
| segment 2 | fréquence_burst | = 3000 Hz | |
| segment 3 | formant 1 | = 200 Hz | formant 2 = 1900 Hz |
| segment 4 | bruit | = 3000 Hz | limite_bruit_descendant = vrai |

Le sous ensemble de règles activées est :

***Règle 1***

SI CONTEXTE_DROIT / y, u, w /
SI silence
SI 3000 Hz < fréquence_burst < 4000 Hz
ALORS / t /

***Règle 2***

SI CONTEXTE_DROIT / i, e /
SI silence
SI fréquence_burst > 4000 Hz
ALORS / t /

***Règle 3***

SI CONTEXTE_DROIT / i, e /
SI silence
SI 2500 Hz < fréquence_burst < 3500 Hz
ALORS / k /

***Règle 4***

SI CONTEXTE_DROIT / y, u, w /
SI silence
SI fréquence_burst < 2500 Hz
ALORS / k /

***Règle 5***

SI CONTEXTE_DROIT / y, u, w /
SI friction
SI 3000 Hz < bruit < 4000 Hz
ALORS / s /

***Règle 6***

SI CONTEXTE_DROIT / y, u, w /
SI friction
SI 1500 Hz < bruit < 2500 Hz
ALORS / ʃ /

***Règle 7***

SI CONTEXTE_DROIT / i, e /
SI friction
SI 2000 Hz < bruit < 3000 Hz
ALORS / ʃ /

***Règle 8***

SI 200 Hz < formant1 < 300 Hz
SI 1800 Hz < formant2 < 2200 Hz
ALORS / a /

***Règle 9***

SI formant1 - formant2 < 250 Hz
ALORES / a /

***Règle 10***

SI 200 Hz < formant1 < 300 Hz
SI 1900 Hz < formant2 < 2200 HZ
ALORS / i /

On obtient alors un treillis de phonèmes représenté par l'arborescence de la figure B1.2 :

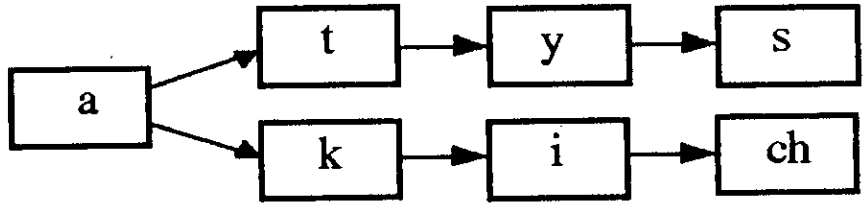



**Fig. B1.2 :** Treillis phonétique fourni au niveau du D A P [ 22 ]

Si on ajoute une règle qui tient compte du fait qu'une limite de bruit descendante ne peut se produire que si le contexte gauche est labial ( / u /, / y /, / w /), on obtient un nouvel arbre donné en figure B1.3 :

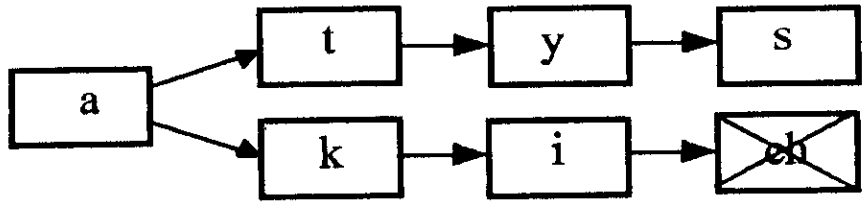



**Fig. B1.3 :** Arbre de décodage final [ 22 ]

### 2.3.2.4. Détection des traits et macro-classes

Cette opérution cherche à localiser sur une fenêtre de signal des indices ou des paramètres de base dont l'utilisation dans des règles d'interprétation permettra d'identifier le trait correspondant. L'ensemble des phonèmes d'une langue peut être décrit par un arbre établissant une hiérarchie sur les traits [ 36 ]. Certains d'entre eux plus stables et moins influencés par le contexte figurent au niveau supérieur de l'arbre. Le passage d'un niveau de l'arbre au niveau inférieur constitue l'essentiel du décodage acoustico-phonétique. Ce processus utilise soit séquentiellement soit conjointement trois types de connaissances : **acoustiques** (corrélats, indices), **articulatoires** (le plus souvent implicites), et **phonétiques** (description des phonèmes en termes de traits, contexte, etc.). Ces connaissances peuvent être mises sous forme de règles et l'on peut employer à partir de là une stratégie de système expert.

Généralement, on procède en deux temps pour l'identification :

- on cherche avec une stratégie ascendante, les **macro-classes phonétiques** (voyelles, fricatives, occlusives, autres consonnes) pour obtenir une classification large mais robuste ;

- on cherche avec une stratégie descendante de type **hypothèse / test**, à affiner la décision à l'intérieur d'une macro-classe. On se sert alors de règles et de méta-règles spécifiques à chaque macro-classe, de règles contextuelles et éventuellement d'informations phonologiques.

### a - Etiquettage des macro-classes : Voyelles-Consonnes

l'attribution de l'étiquette voyelle ou consonne à un spectre met en jeu les indices suivants :

E [ $\alpha - \beta$ ] : énergie dans la bande de fréquence [ $\alpha - \beta$ ].
Max [ $\alpha - \beta$ ] : amplitude du maximum d'énergie dans la bande [ $\alpha - \beta$ ].
Min [ $\alpha - \beta$ ] : amplitude du minimum d'énergie dans la bande [ $\alpha - \beta$ ].
C. D. G : centre de gravité spectral du spectre.
FOR max : amplitude du plus grand formant sur toute la partie de phrase déja analysée.
F [ $\alpha - \beta$ ] : fréquence du maximum d'énergie dans la bande [ $\alpha - \beta$ ].

Une vingtaine de règles comparant ces indices entre eux ou à des seuils normalisés permet la distinction voyelle-consonne.

Ainsi, par exemple, la règle suivante appliquée dans le système Keal [ 37 ] est l'une des règles qui permet d'affecter l'étiquette *"voyelle"* à un spectre.

Si pour un spectre, les conditions suivantes :

( E [ 250 - 450 Hz ] >= 8 dB ) et ( E [ 250 - 450 Hz ] >= E [ 450 - 650 Hz ] )
et ( E [ 250 - 450 Hz ] >= FOR max - 24 dB ) et ( C.D.G. >= 7 ) et ( C.D.G. >= 8.9 )
et ( 2 X MAX [ 250 - 850 Hz ] - E [ 1300 - 1900 Hz ] > 28 dB )
et ( E [ 1900 - 2800 ] - 2 X Max [ 250 - 850 ] <= 20 dB )

sont vérifiées. Alors ce spectre correspond à la Macro-classe = **VOYELLE.**

## b - Détection du trait de voisement

La distinction consonne voisée-consonne sourde est habituellement basée sur les indices suivants [ 36 ] :

- présence du fondamental ($F_0 > 0$),
- rapport entre les énergies basse fréquence et haute fréquence,
- rapport entre l'énergie basse fréquence et l'énergie totale.

Dans le cas des occlusives un autre indice important est le **V. O. T** (Voice Onset Time) c'est-à-dire la durée d'établissement du voisement au moment de l'explosion.

## c - Détection des consonnes fricatives

Les principaux indices facilitant la reconnaissance de cette classe de consonnes sont :

- Le nombre de passage par zéro du signal ou de la dérivée du signal,
- Les rapports entre l'énergie haute et basse fréquence,
- La valeur du centre de gravité spectral.

Les consonnes / f /, / s /, / ʃ /, / z /, / Z / sont bien détectées par ces indices.

### d - Détection des consonnes occlusives

Les occlusives sont habituellement caractérisées par plusieurs phases plus ou moins marquées dans la parole continue : l'implosion, l'occlusion ou burst et le bruit de friction et la transition vers la voyelle suivante. Les principaux indices descriptifs de ces consonnes sont liés à l'occlusion et à l'explosion.

- **l'occlusion** est marquée soit par une zone de silence (occlusion complète du conduit vocal dans le cas des occlusives sourdes), soit par une zone contenant de l'énergie dans les basses fréquences, caractéristique de la barre de voisement (Buzz-Bar) des occlusives voisées.

- **l'explosion** ou **BURST** est marquée par un brusque saut d'énergie et une forte instabilité spectrale. Cet ensemble d'indices et éventuellement la présence d'un bruit de friction relativement court par rapport à la friction des fricatives conduisent à une bonne détection des occlusives.

## 3. CONCLUSION

Nous avons présenté dans ce chapitre les principes généraux et les approches possibles pour la mise en oeuvre de la composante acoustio-phonétique dans les systèmes de reconnaissance de la parole. Il reste encore beaucoup de problèmes fondamentaux à résoudre pour accroître les possibilités et les performances de tels systèmes ; il nous semble en particulier qu'un effort soutenu doit se poursuivre dans deux axes complémentaires.

- recensement et formalisation des connaissances adéquates dans ce domaine,

- conception de modèles et d'architectures adaptés dans lesquelles les sources de connaissances seront intégrés et mises en oeuvres de façon efficaces.

Malheuresement la constitution d'un catalogue aussi complet que possible des sources de connaissances est un travail de longue haleine nécessitant la collaboration d'experts de diverses disciplines telles que la linguistique, la psycologie, l'intelligence artificielle etc.

Néanmoins, ce qui apparaît le plus clair c'est qu'aucune démarche sérieuse n'étant refusable en l'état actuel des expérimentations.

# CHAPITRE B2

# MISE EN OEUVRE DU DECODEUR ACOUSTICO-PHONETIQUE

Nous avons étudié au chapitre précédent différentes méthodes et outils dans le domaine du Décodage Acoustico-Phonétique. Il s'agit maintenant de mettre en oeuvre cette composante pour réaliser notre système, conformément à nos objectifs.

## 1. ROLE DU DECODAGE ACOUSTICO-PHONETIQUE

L'objectif au niveau acoustico-phonétique, est de fournir, pour chaque segment, une liste éventuellement longue d'hypothèses (généralement 3 ou 4 phonèmes) incluant le phonème prononcé par le locuteur, plutôt que l'interprétation la plus plausible au risque d'exclure l'interprétation correcte. En effet, il est plus facile, pour les niveaux supérieurs, d'éliminer des hypothèses phonétiques à l'aide de critères ressortant de leur compétance que d'en créer des nouvelles qui leur paraîtront pertinantes, mais qu'ils seront dans l'incapacité de valider sur le plan phonétique. En choisissant de représenter chaque énoncé analysé non pas par une simple chaîne mais par un treillis phonétique, on s'autorise non seulement à donner, pour un segment, plusieurs interprétations, mais aussi à proposer, pour une suite de segments, différentes interprétations, ce qui permet de prendre en compte les altérations contextuelles très marquées que l'on observe en parole continue au niveau de la réalisation des différents phonèmes.

Le décodage acoustico-phonétique a été longtemps assuré par des méthodes classiques de comparaison de formes paramétriques ou de modèle stochastiques [ 38 ]. Compte tenu des phénomènes phonologiques et du caractère contextuel de l'identification des unités phonétiques, il est intéressant de mêler méthodes de comparaison et experte fondée sur des connaissances pour résoudre ce problème. C'est dans cette voie que notre modèle a été réalisé.

## 2. DIFFERENTES FONCTIONS DU DECODAGE ACOUSTICO-PHONETIQUE

Le niveau acoustico-phonétique assure les fonctions suivantes :

- l'**acquisition des paramètres** du signal de la parole,

- la **segmentation du signal en unités discrètes** correspondant aux réalisations des phonèmes successifs que le locuteur a voulu prononcer,

- l'étiquetage phonétique des segments détectés, qui constitue une phase d'interprétation et d'**identification des évènements acoustiques.**

La segmentation et l'identification s'effectuent en deux temps :

a - d'abord, une **étude globale de l'énoncé**, sorte de prétraitement destiné à :

- effectuer une segmentation grossière de l'énoncé en noyaux vocaliques et groupes consonantiques ;

- regrouper en segments homogènes, où chacun d'eux est constitué par un ou plusieurs échantillons qui se suivent sans discontinuité. Nous considérons qu'ils traduisent les mouvements articulatoires successifs du conduit vocal. Ce traitement conduit à une classification des segments qui ne prend pas en compte l'incidence du contexte sur la réalisation des phonèmes et fournit des informations qui seront utilisées au cours de la seconde étape.

b - ensuite, une **analyse fine, segment par segment,** au cours de laquelle chaque segment est interprété de manière précise en fonction du contexte et la représentation de l'énoncé sous forme d'un treillis d'hypothèses phonétiques construites ; à noter qu'au cours de cette étape il peut s'avérer nécessaire d'enrichir les informations dont on dispose sur le segment par d'autres paramètres complémentaires (durée d'un segment par exemple) ou de remettre en cause la segmentation fournie par la phase initiale.

La première étape est réalisée par une méthode algorithmique. La seconde prend la forme d'un système expert opérant de manière itérative sur chaque segment de l'énoncé.

# 3. EXPLICATION ET MISE EN OEUVRE DES DIFFERENTES ETAPES

## 3.1. Acquisition des paramètres

On divise le signal de parole en échantillons de 10 ms et pour chaque échantillon, on extrait des paramètres acoustiques du type fréquentiel et énergétique.

Nous n' envisagerons que les paramètres utiles à la segmentation, les paramètres propres à l' analyse ne seront pas abordés ici. Notons que les paramètres choisis sont fonction de la méthode de segmentation utilisée.

Nous avons retenus les paramètres suivants :

- Energie dans certaines bandes de fréquences.

- Energie totale du signal.

Pour cela, nous utiliserons 12 filtres dont les caractéristiques sont les suivantes :

**$F_1$ : 100 Hz à 200 Hz**

**$F_2$ : 200 Hz à 400 Hz**

**$F_3$ : 400 Hz à 600 Hz**

**$F_4$ : 600 Hz à 800 Hz**

**$F_5$ : 800 Hz à 1000 Hz**

**$F_6$ : 1000 Hz à 2000 Hz**

**$F_7$ : 2000 Hz à 3000 Hz**

**$F_8$ : 3000 Hz à 4000 Hz**

**$F_9$ : 4000 Hz à 5000 Hz**

**$F_{10}$ : 5000 Hz à 6000 Hz**

**$F_{11}$ : 6000 Hz à 7000 Hz**

**$F_{12}$ : 7000 Hz à 8000 Hz**

Pour chaque sortie de filtre, on prélève l'énergie moyenne du signal sur 10 ms. Ces énergies seront désignées par : **P1, P2,. ., P12,** qui correspondent respectivement aux filtres : $F_1, F_2, \ldots, F_{12}$. On calcule, par la suite, l'énergie totale "**EN**" de chaque échantillon consideré. Ces différentes énergies vont nous servir pour faire demarrer la segmentation.

## 3.2. Segmentation

Pour effectuer cette segmentation, nous avons choisi une méthode qui procède par étapes successives jusqu'à l'obtention d'une segmentation en phonèmes tenant compte des réalisations particulières de certains phonèmes et en leur intégrant les segments de transition qui peuvent s'insérer à leurs frontières. Ces étapes se résument par :

- **Segmentation du signal en échantillons** de 10 ms et attribution d'une classe à chaque échantillon.

- **Constitution de segments homogènes** en regroupant tous les échantillons (voisins directs) ayant été affectés de la même classe et affectation d'une nouvelle classe à chaque segment ainsi obtenu.

- **Découpage en phonèmes** en tenant compte de la réalisation particulière de certains phonèmes et par absorption de segments de transition dépendant de l'entourage phonique.

### a - Première étape de segmentation

A partir des paramètres fournis par les 12 filtres de l'étape d'acquisition, on détermine d'abord les indices acoustiques **V, F, B, Z, M, D,** dans lequel :

**V** : est un indice de voisement.
**F** : est un indice de friction.
**B** : est un indice de Buzz.
**Z** : est un indice précisant si l'énergie atteint un seuil $\theta$.
**M** : est un indice de variation positive de l'énergie.
**D** : est un indice de variation négative de l'énergie.

- l'indice de voisement **V** dépend du filtre $F_1$.

- l'indice de Buzz **B** dépend des filtres $F_1$, $F_2$, $F_3$.

- l'indice de friction **F** est obtenu à l'aide des paramètres fournis par l'ensemble des filtres : $F_3$ à $F_{12}$.

- les indices **M** et **D** ne dépendent que de l'énergie totale du signal.

- le seuil $\theta$ dont dépend l'indice **Z** à été choisi à 7 (d'après les expériences).

A chaque échantillon de 10 ms, on associe le doublet :

$$I(j) = [\ C_i(j),\ P_i(j)\ ] \qquad \text{dans lequel :}$$

- $C_i$ est la classe affectée à l'échantillon **j**.

$C_i$ (**FR, FV, SI, BZ, VF, VS, VM, VD**)

- $P_i$ est l'ensemble des paramètres énergétiques spécifiques de la classe $C_i$.

Les caractéristiques associées à chaque classe sont les suivantes :

- Pour un échantillon classé **FRICATIF** :
  $C_1$ = **FR** si les indices présents sont : **V~, F**

- Pour un échantillon classé **FRICATIF VOISE** :
  $C_2$ = **FV** si les indices présents sont : **V, F**

- Pour un échantillon classé **SILENCE** :
  $C_3$ = **SI** si les indices présents sont : **V~, F~**

- Pour un échantillon classé **BUZZ** :
  $C_4$ = **BZ** si les indices présents sont : **V, F~, B**

- Pour un échantillon classé **VOISE FAIBLE**, c'est-à-dire dont l'énergie ne dépasse pas le seuil $\theta$ :
  $C_5$ = **VF** si les indices présents sont : **V, F~, B~, Z~**

- Pour un échantillon classé **VOISE STATIONNAIRE** en énergie :

$C_6$ = **VS** si les indices présents sont : **V, F~, B~, Z, M~, D~**

- Pour un échantillon classé **VOISE MONTANT** (à variation énergétique positive) :

$C_7$ = **VM** si les indices présents sont : **V, F~, B~, Z, M, D~**

- Pour un échantillon classé **VOISE DESCENDANT** ( à variation énergétique négative):

$C_8$ = **VD** si les indices présents sont : **V, F~, B~, Z, M~, D**

Le tableau B2.1 résume les différentes classes utilisées ainsi que leurs indices correspondants.

| **Indices présents** | **Classe correspondante** |
|---|---|
| V~ , F | FR : Fricatif |
| V , F | FV : Fricatif Voisé |
| V~ , F~ | SI : Silence |
| V , F~ , B | BZ : Buzz |
| V , F~ , B~ , Z~ | VF : Voisé Faible |
| V , F~ , B~ , Z , M~ , D~ | VS : Voisé Stationnaire |
| V , F~ , B~ , Z , M , D~ | VM : Voisé Montant |
| V , F~ , B~ , Z , M~ , D | VD : Voisé Descendant |

**Tableau B2.1 :** Les différentes classes utilisées dans la première étape de segmentation

L'organigramme de la première étape de segmentation est représenté par la figure B2.1.

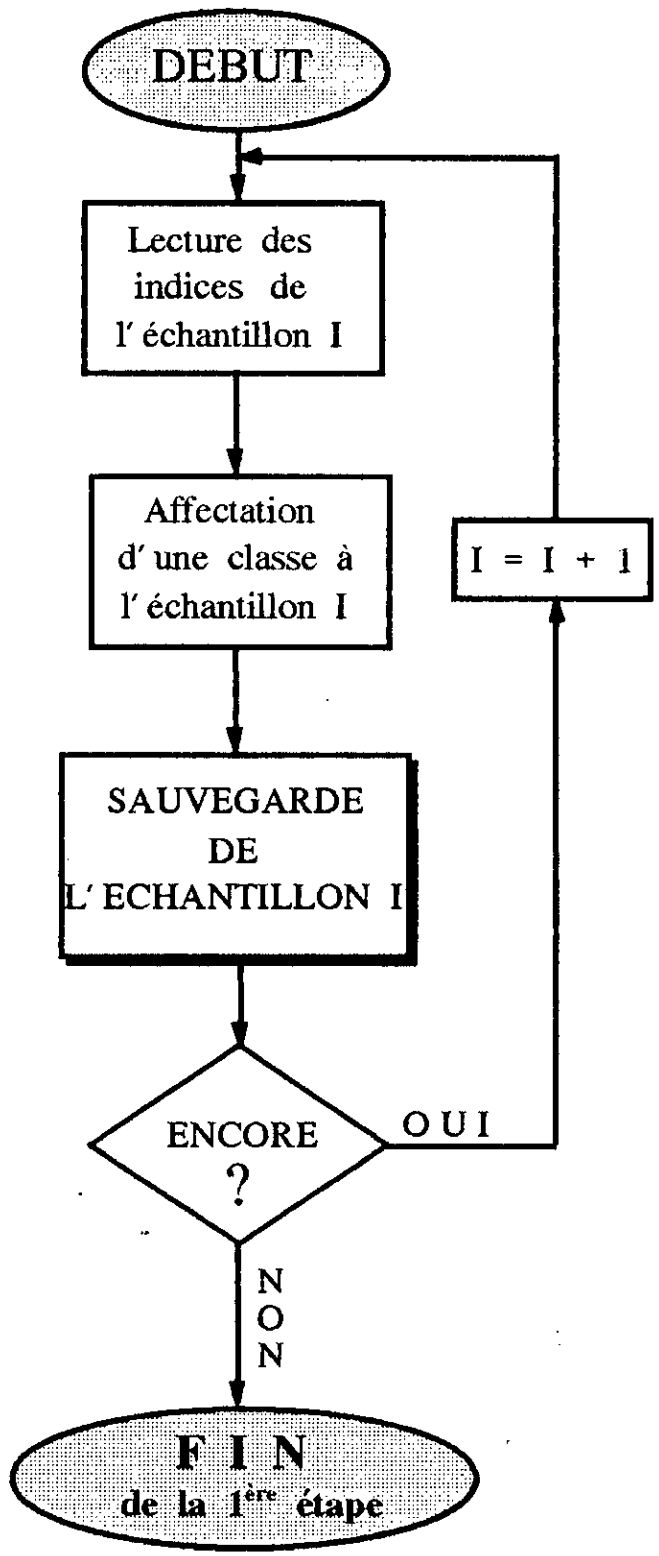



**Fig. B2.1** : Organigramme de la 1ère étape de segmentation

## b - Deuxième étape de segmentation

A partir des échantillons **I ( j )** de la première étape, on constitue des segments homogènes en concaténant des échantillons affectés de la même classe. Cette concaténation n' est possible que si ces échantillons sont adjacents c' est-à-dire voisins directs. Chaque segment homogène sera affecté d' une nouvelle classe. Cette étape dépend des classes des échantillons précédents ainsi que leur nombre concaténés (Fig. B2.2).

L' affectation d' une nouvelle classe à un segment se fait selon des règles de productions qui sont résumées dans le tableau B2.2. On notera que les principales règles de production nécessaires pour la segmentation sont utilisées, en particulier les règles concernant les fricatives et les plosives.

| **SI** l' ancienne classe de l' échantillon **est** | **ET** le nombre "N" d' échantillons concaténés **est** | **ALORS** la nouvelle classe affectée au segment homogène **sera** |
|---|---|---|
| FR | N ≥ 6 | FR |
| FR | N < 6 | (FR) ~ |
| FV | N ≥ 4 | FV |
| FV | N < 4 | (FV) ~ |
| BZ | N ≥ 3 | BZ |
| BZ | N < 3 | (BZ) ~ |
| VF | N ≥ 3 | VF |
| VF | N < 3 | (VF) ~ |
| VS | N ≥ 4 | VS |
| VS | N < 4 | (VS) ~ |
| SI | N ≥ 15 | SI |
| SI | N < 3 | SI |
| SI | N ≥ 3 et N < 15 | PL |
| VM | pour tout N | VM |
| VD | pour tout N | VD |

**Tableau B2.2 :** Les principales règles de production utilisées par la deuxième étape de segmentation

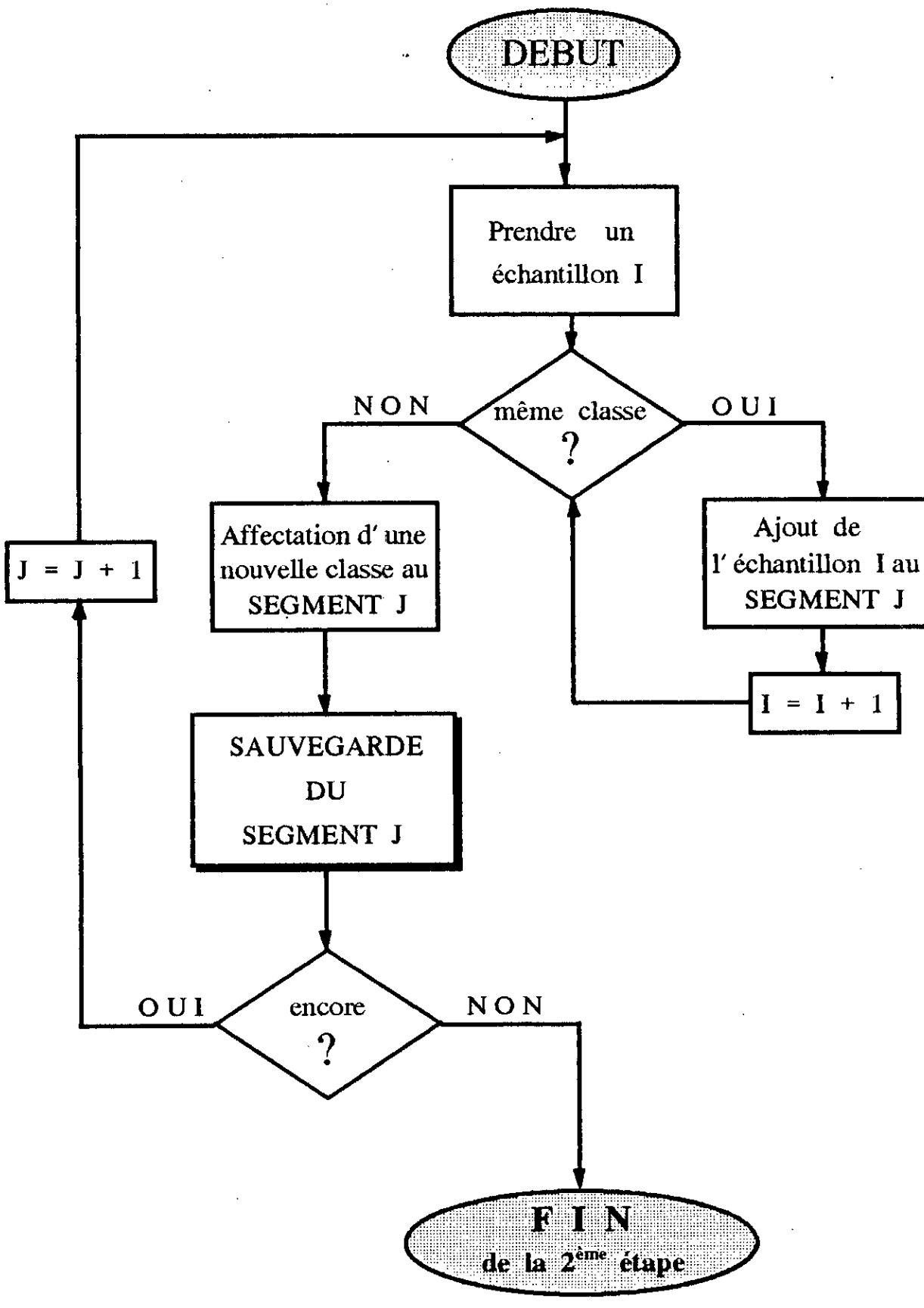



**Fig. B2.2 :** Organigramme de la 2ème étape de segmentation

Ces règles de production font apparaître des classes que l'on ne différencie que par la durée du segment. La classe des occlusives sourdes **PL** est ainsi créée.

## C - Troisième étape de segmentation

Le découpage phonémique ne se borne pas à la constitution de segments homogènes car la réalisation des phonèmes est beaucoup plus complexe. C'est pour cette raison, une troisième et dernière étape de segmentation est nécessaire et qui consiste à regrouper certains segments en segments phonémiques suivant des règles de production. Cette étape a pour but de créer de nouvelles classes en tenant compte de la structure morphologique de certains phonèmes, leur rattacher des segments de transition dépendant de l'entourage phonique, éliminer des parasites dûs, soit à une mauvaise extraction de paramètres, soit à des perturbations issues de l'enregistrement.

Le module qui constitue cette étape reçoit en entrée deux segments (engendrés par l'étape précédente), qu'il tente de concatener en une seul, en appliquant une règle de la forme :

**SI** l'étiquette du segment 1 est "A"
**SI** l'étiquette du segment 2 est "B"
**SI** relation entre "A" et "B" donne "C"
**ALORS** le segment résultat aura comme étiquette "C"

Si la concaténation est possible, on tente de construire le segment résultat en réévaluant les paramètres retenus sur le signal et on tentera d'en concaténer encore d'autres segments avec celui déjà trouvé. Si par contre cette concaténation est impossible, on passe à un autre groupe de segments et l'identification de celui déjà trouvé sera immédiate (Fig. B2.3).

Dans le cas ou plusieurs règles s'appliquent à deux segments, le résultat de la concaténation serait un segment avec une liste d'étiquettes possibles (hypothèses), on laissera alors le soin au module d'identification pour décider de l'étiquette exacte en se basant sur le contexte dans lequel il est présenté.

L'étiquette attribuée à un segment lors de cette étape va donc être une étiquette de macro-classe et le segment obtenu sera un segment phonémique qui fera l'objet d'une identification à la prochaine étape. Notons que la base de règles peut être définie indépendamment de ce module, une interface d'acquisition est réalisée à cet effet.

| ▶Z◀ | (SI)~ | (BZ)~ | (FR)~ | (FV)~ | (VF)~ | (VS)~ | VF | VS | VM | VD | SI | BZ | PL | FR | FV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (SI)~ | | | | | | | VF | | | | | | | FR | FV |
| (BZ)~ | | | | | | | VF | VS | VSM | VSD | SI | | PL | | FV |
| (FR)~ | (SI)~ | | | FV | FV | FV | | | | | | | | | FV |
| (FV)~ | | | FV | | | | | | | | | | | FR | |
| (VF)~ | | | | | | VS | | VS | VM | | SI | BZ | PL | | FV |
| (VS)~ | | | | FV | VS | | VF | | VSM | VSD | | | | | FV |
| VF | VF | VF | | | VF | | | | | | | | | | |
| VS | | | | FV | VS | VS | VS | VS | VSM | VSD | | | | | FV |
| VM | | | | | | VMS | | VMS | | VSD | | | | | |
| VD | | | | | | VDS | | VDS | | | | | | | |
| SI | SI | SI | | | SI | | | | | | SI | | | | |
| BZ | PL | | | | BT | BT | | | | | | BZ | PL | | |
| PL | PL | PL | PF | | PT | PT | | | | | | | PL | | |
| PF | | | PF | PFT | PFT | PFT | | | | | | | PL | | |
| FR | FR | FR | FR | | | | | | | | | | | FR | |
| FV | FV | FV | FV | FV | FV | FV | | | | | | | | | FV |

**Tableau B2.3 :** Table de l'automate utilisée par par la 3ème étape de segmentation

Pour réaliser ces deux phases, nous avons décidé de distinguer deux types de règles propres à chaque phase : des règles d'informations assez grossières, et celles qui comportent des informations plus fines.

* Dans la première phase, on utilisera des informations sur l'identification de la macro-classe phonétique correspondante à une liste de phonèmes caractérisant cette classe.

Ces informations seront formalisées sous forme de règles du type :

| | | | |
|---|---|---|---|
| **Si** | Etiquette du segment | **Alors** | Liste de phonèmes |

* La seconde phase utilisera par contre, des informations spécifiques aux différents contextes gauche et droit ainsi que les déductions déja faites. Cette phase agit comme filtre pour la première phase en attribuant un poids à chaque phonème résultat pour ne retenir à la fin que les trois meilleurs condidats.

Ces règles sont du type :

| | | |
|---|---|---|
| **Si** | liste de phonèmes contexte gauche | **Et** |
| **Si** | liste de phonèmes contexte droit | **Et** |
| **Si** | liste de phonèmes déja supposée | |
| **Alors** | liste de phonèmes pondérés | |

Le moteur d'inférences accède aux informations contenues dans la base de faits relative au segment en cours d'analyse, chaque fois qu'il tente d'appliquer une règle ; il doit en effet consulter cet ensemble pour déterminer si les prémisses sont satisfaisantes. Si la règle est conditionnelle, il peut être amené à examiner le treillis phonétique du segment précédent et la base de faits du segment suivant.

# CHAPITRE B3

# EXPERIENCES, TESTS ET RESULTATS

## 1. INTRODUCTION

Malgré l'importance et l'adaptation d'un certain nombre de *méthodes d'analyse* du signal de la parole, la technique d'analyse spectrographique a été retenue pour nous servir comme support de travail pour l'établissement de la base de connaissances, que nous n'avions pas pu obtenir avec une autre méthode.

Ainsi, avec l'introduction du spectrographe, un progrés considérable a pu être fait dans la compréhension des corrélats acoustiques de la perception phonétique. Les expériences de lecture automatique de spectrogrammes de parole [ 22] ont montré que les humains étaient capables de segmenter correctement et ensuite d'interpréter la suite des événements acoustiques. Il semble que ces experts utilisent à la fois des règles, des procédures et tout un ensemble de connaissances acoustiques, phonétiques, perceptives, articulatoires et prosodiques, pas toujours faciles à formaliser pour élaborer progressivement les différentes hypothèses phonétiques.

L'analyse de l'activité de l'expert en lecture des spectrogrammes nous a permis d'enrichir notre savoir-faire en matière de décodage acoustico-phonétique :

- en augmentant le nombre et la qualité des informations acoustiques pertinentes relatives à la segmentation et à l'identification phonétique de la parole continue,

- en mettant en évidence des stratégies de décodage plus efficaces que celles utilisées jusqu'à présent dans les systèmes automatiques.

## 2. METHODOLOGIE D'ACQUISITION DE L'EXPERTISE EN LECTURE DE SPECTROGRAMMES DE PAROLE

Ces dernières années de nombreux travaux s'intéressent au problème de l'expertise des spectrogrammes de parole. Leur objectif principal est de formaliser et d'évaluer les connaissances de l'expert humain, en vue de leur utilisation ultérieure dans des systèmes de reconnaissance automatique de la parole, plus particulièrement au niveau du décodage acoustico-phonétique.

Dans ce type d'analyse, les données sont constituées de spectrogrammes de parole que l'expert peut traduire en suite de phonèmes. Cette image est la représentation visuelle de l'évolution spectrale des sons prononcés par un locuteur.

Durant la tâche de décodage d'un spectrogramme, l'expert ne relève dans l'image que les motifs qui lui semblent pertinents. Cependant, les éléments de description qu'il fournit comportent déjà une part d'interprétation.

Pour décoder un spectrogramme, l'expert adopte une démarche qui peut être décomposée en deux étapes successives [ 40 ] :

### 2.1. étude globale

L'expert effectue une observation globale avant de décoder le spectrogramme. Cette étape est destinée à :

- déterminer la durée vocalique moyenne (vitesse d'élocution),
- étalonner les niveaux de gris (spectrogramme "très contrasté" ou "rès clair").

### 2.2. étude locale (segment par segment)

Dans cette étape, deux cas peuvent se présenter :

a - dans le cas où la **segmentation ne pose pas de problème**, la démarche de l'expert se résume par :

- analyse visuelle pour détecter un ou plusieurs indices clairs (non ambigus),
- émission d'une ou plusieurs hypothèses,

- validation et sélection ou classement des hypothèses émises.

La validation consiste à vérifier que les caractéristiques acoustiques de chaque "phonème hypothès" sont compatibles avec les propriétéss du segment observé dans le contexte identifié et, éventuellement, avec les transitions avec les phonèmes voisins. Il y aura compatibilité s' il n' y a pas d' indice contradictoire avec l' hypothèse émise et si la plupart des indices attendus sont présents.

b - dans le cas où la **segmentation est difficile** (paquets vocaliques notamment) il utilise une autre stratégie. Il démarre souvent du centre du paquet (ou d' un endroit qui semble plus clair) et il réduit progressivement les incertitudes. Une segmentation n' est acceptée que si l' identification qui en découle est satisfaisante. Il y a des allers et retours entre segmentation et identification, puis choix de la meilleure hypothèse.

## 2.3. Importance du contexte

En parole continue, lorsque le rythme d' énonciation est relativement rapide, les phénomènes de co-articulation influent sensiblement sur la réalisation des unités phonétiques. Or, s' il existe des descriptions relativement précises et complètes des réalisations possibles des différents phonèmes en fonction du contexte, on ne dispose que d' études très partielles sur la nature des faits acoustiques qui permettent d' interpréter un segment comme une réalisation particulière d' un phonème donné.

Le rôle de l' expert a été ici de fournir un ensemble de règles contextuelles permettant d' identifier, dans des contextes très variés, chacun des différents phonèmes à partir d' indices acoustiques relativement indépendants du locuteur. Ces règles peuvent porter sur un contexte très large, jusqu' à trois segments plus loin dans certains cas.

# 3. DESCRIPTION SPECTROGRAPHIQUE DES VOYELLES

Les voyelles sont les phonèmes les plus simples dans la mesure où ce sont des états stationnaires. Leur structure acoustique se caractérise principalement par la présence de maxima spectraux, c' est-à-dire de zones de fréquences où les harmoniques sont particulièrement intenses, que l' on appelle formants et qui apparaissent spectrographiquement sous la forme de bandes noires plus ou moins parallèles à l' axe des temps (Fig. B3.1).

Des expériences de synthèse ont montré que seuls les trois premiers formants **F1, F2,** et **F3** caractérisaient le timbre vocalique, et de ce fait les plus importants pour leur identification. On peut donc se contenter de ne considérer qu'eux pour établir une première classification acoustique des voyelles. Cependant, il est à remarquer, que ces fréquences de formants peuvent varier assez considérablement d'un locuteur à l'autre, et ne sont vraiment quasi-constantes que s'il s'agit de la voix d'un même locuteur. Les formants des voix feminines sont de fréquences plus élevées que les formants des voix masculines, d'environ 15 %, l'invariant restant généralement est le rapport entre ces fréquences. Toujours est-il que ces formants sont dans l'état actuel, le seul critère valable sur lequel nous puissions nous appuyer pour la distinction entre les différentes voyelles et pourront nous servir de base pour la lecture de spectrogrammes de parole.

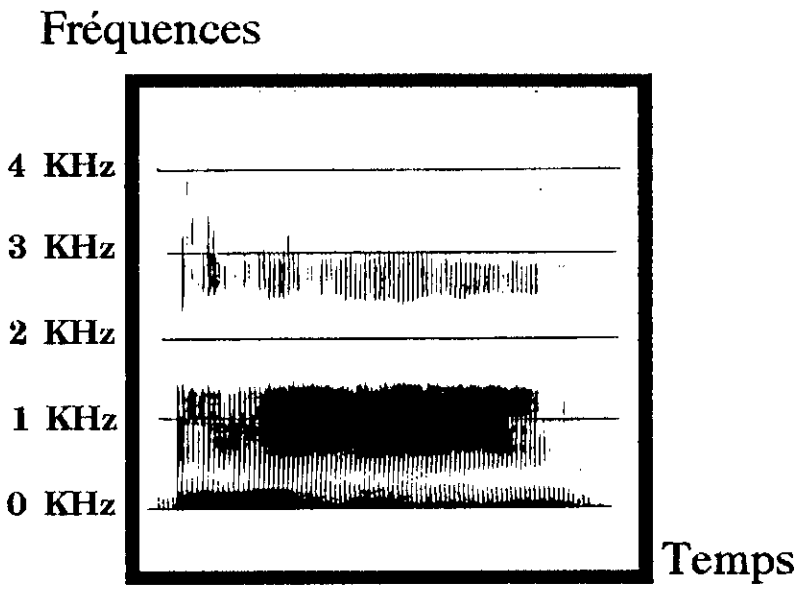



**Fig. B3.1** : Spectrogramme d'une voyelle /a/

## 4. DESCRIPTION SPECTROGRAPHIQUE DES CONSONNES

Sur le plan acoustique, là où les voyelles étaient des sons purs, les consonnes sont soit des bruits purs où un très grand nombre de fréquences apparaissent à la fois, soit des bruits combinés à un ton larygien (un formant du niveau des basses fréquences). Dans le premier cas, on parle de consonnes sourdes, dans le deuxième, de consonnes sonores.

Contrairement aux voyelles, les consonnes sont difficiles à identifier isolément. C'est que les bruits qui en sont la base sont produits de plusieurs manières différentes et qu'ils sont fortement influencés par les voyelles qui les entourent. Le spectrographe n'est cependant pas inutile dans ce cas. On peut en analyser la morphologie, la simplifier et dégager les caractères pertinents. On voit ainsi émerger des corrélats acoustiques bien caractéristiques des consonnes [ 41 ]. On peut par exemple (Fig. B3.2) :

- Etudier la **distribution de l'énergie** (un /s/ se distingue par une concentration de l'énergie dans les fréquences élevées, se manifestant par des stries verticales dans la partie supérieure du spectrogramme).

- Analyser les **bandes formantielles** quand elles sont présentes (plus le son est vocalique, plus ces bandes sont claires).

- Etudier les **transitions formantiques**, c'est-à-dire les déflexions fréquentielles rapides des formants que l'on observe au passage d'une consonne à une voyelle et réciproquement.

Nous avons retenu, pour nos expériences, les classes des consonnes occlusives et fricatives que nous décrirons tel que nous le révèlent les spectrogrammes.

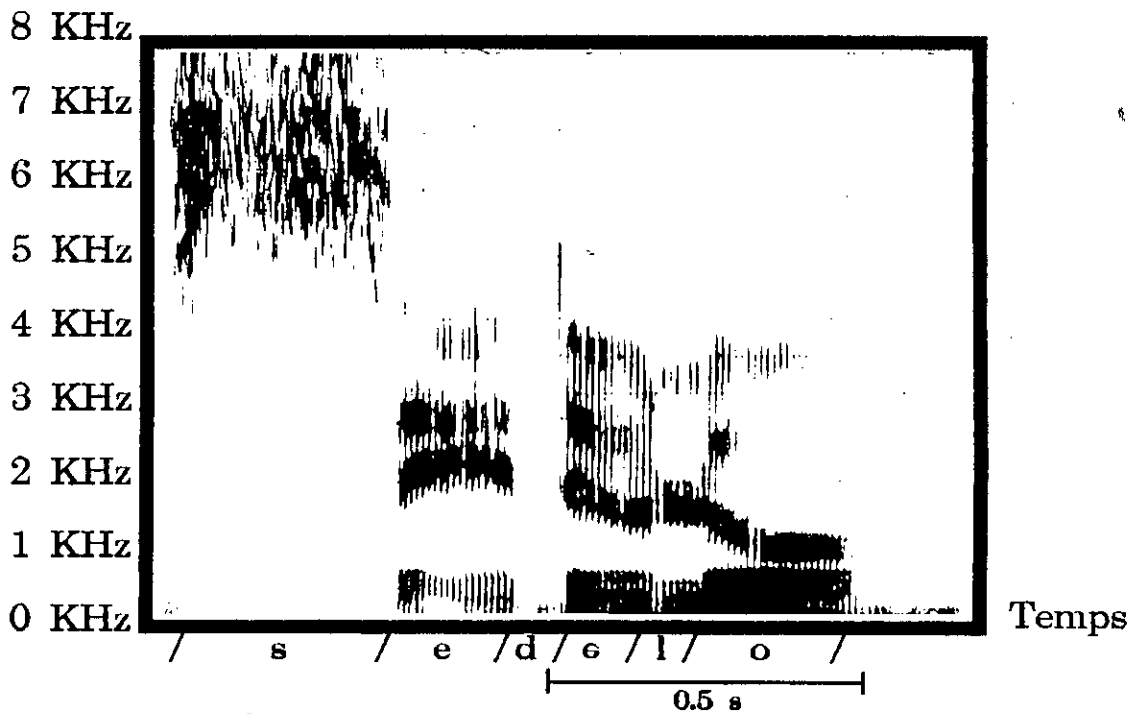



**Fig. B3.2** : Spectrogramme de la phrase "c'est de l'eau" sur la bande 0 - 8 KHz [ 41 ]

## 4.1. Consonnes occlusives

a - Les occlusives labiales (/ p /, / b /) : Ces occlusives se caractérisent par une barre. Ce sont les paires de consonnes : (/ p /, / b /) ; (/ t /, / d /) et (/ k /, / g /). Chacune de ces paires comporte une variante sourde (/ p /, / t /, / k /) et une variante sonore (/ b /, / d /, / g /). Leur brève [illegible] généralement répartie dans une large bande de fréquences (diffuse) et les basses fréquences prédominent (descendante). Les transitions

Du point de vue articulatoire, les occlusives se distinguent principalement par un silence provenant de la fermeture totale du conduit vocal (occlusion) en un lieu bien défini, pendant leur tenue (c'est-à-dire le temps pendant lequel les organes se mettent en place pour les prononcer). Acoustiquement, une occlusive se compose d'une suite d'événements acoustiques [42] (Fig. B3.3). [illegible] de l'observation de plusieurs spectrogrammes [illegible] sujet. Les [illegible] sont donc pas [illegible], mais illustrent simplement les [illegible] [illegible], un sujet unique l'est vraisemblable que ces variations [illegible]

- Un silence, qui correspond à la phase de tenue articulatoire de l'occlusion complète du conduit vocal. Celle-ci est bilabiale pour (/ p /, / b /), dentale pour (/ t /, / d /) et vélaire pour (/ k /, / g /). Cependant, dans le cas des occlusives sonores, le silence n'est pas total dans la mesure où les vibrations des cordes vocales pendant la tenue articulatoire se traduisent par une concentration d'énergie faible en très basse fréquence (100 - 300 Hz) appelée "barre de voisement". En général, une consonne sonore est plus brève qu'une consonne sourde et dure généralement plus de 50 ms. La nature de l'entourage influe également sur la durée de l'occlusion.

- Une barre d'explosion (BURST) : L'air comprimé retenu lors de l'occlusion est relâché provoquant ainsi une perturbation acoustique de courte durée (5 à 35 ms). Cette perturbation qui peut être intense, se manifestera sur les spectrogrammes par une mince barre verticale, qu'on appelle, de ce fait, barre d'explosion ou BURST (impulsion de bruit).

- Des transitions de formants particulières à chaque mode articulatoire.

- Un bruit de friction : La durée de ce bruit dépend de la vitesse à laquelle les articulateurs s'écartent. Cette durée est brève pour les labiales, où les articulateurs sont très mobiles. Elle est longue pour les vélaires en raison de l'inertie de l'articulateur mobile qui est le dos de la langue. La durée pour les dentales est intermédiaire, car la pointe de la langue est un articulateur assez véloce. Le spectre de la partie fricative du bruit de friction est assez proche du spectre de la barre d'explosion.

| Occlusive | Barre d'explosion | Transitions de formants | Bruit de friction |
|---|---|---|---|
| [illegible] | [illegible] | -F2: [illegible] ou [illegible]<br>-F3: [illegible] ou [illegible] | faible |
| [illegible] | - faible vers F1<br>- [illegible] entre F1 et F4, au dessus de<br>de 1.5 à 3.5 KH | -F1: [illegible] ou [illegible]<br>-F3: [illegible] | [illegible] |
| [illegible] | - au dessous de F3<br>- vers 2.5 KHZ<br>- au dessus de F3 | -F1: [illegible]<br>-F2: [illegible] ou [illegible]<br>-F3: [illegible] | nul |
| [illegible] | [illegible] | [illegible] | faible ou [illegible]<br>-F2: [illegible]<br>-F3: [illegible] |

Tableau B3.1 : [illegible] des occlusives [illegible]

**b - Les consonnes dentales (/ t /, / d /)** : La barre d'explosion, intense, est décrite comme "diffuse-montante", l'énergie est répartie dans une large bande de fréquence et les fréquences élevées dominent, dépassant parfois 8000 Hz ; quant aux transitions formantiques, elles seront peu marquées si la voyelle suivante est antérieure fermée (/ i /, / e /, / y /), et descendante pour les autres voyelles dont les formants F2 et F3 sont plus bas. F1 sera toujours plat ou montant. L'énergie importante de la barre d'explosion au delà de 5000 Hz se voit très aisément sur les spectrogrammes. On y notera également la durée importante du bruit de friction au contact des voyelles antérieures fermées (/ i /, / e /). Le tableau B3.2 suivant fournit une description plus détaillée.

| Occlusives dentales | Contexte | Barre d'explosion | Transitions formantiques | Bruit de friction |
|---|---|---|---|---|
| / t / | / i / | - continue entre 4.5 et 8 KHz (max vers 5 KHz)<br>- vers 3 KHz | -F1: ___ ou ⌜<br>-F2: ⌜<br>-F3: ___ ou ⌜ | intense |
| | / a / | - continue entre 1.5 et 6 KHz<br>- max vers 2, 3, 5 et entre 5 et 6 KHz | -F1: ⌟<br>-F2: ⌞<br>-F3: ⌞ | intense |
| | / u / | - continue entre 1.5 et 3 KHz<br>- max vers 1.7 et 3 KHz | -F1: ___ ou ⌜<br>-F2: ⌝<br>-F3: ⌝ | intense |
| / d / | / i / | - vers F2 (# 2 KHz) ou un peu plus bas<br>- vers F3 (# 3 KHz)<br>- plus faible vers 4 KHz | -F1: ⌜<br>-F2: ⌜<br>-F3: ⌞ | intense |
| | / a / | - vers F2 (# 1.8 KHz), vers F3 (# 2.7 KHz) ou entre F2 et F3<br>- faible vers 5 - 6 Khz | -F1: ⌟<br>-F2: ⌞<br>-F3: ⌞ | faible |
| | / u / | - entre 2.5 et 4 KHz<br>- rarement entre F1 et F2<br>- vers F2 ou entre F2 et F3 | -F1: ___ ou ⌜<br>-F2: ⌝<br>-F3: ⌝ | friction vers les maximas de la barre |

**Tableau B3.2** : Caractéristiques essentielles des occlusives dentales [ 41 ]

**c - Les occlusives vélaires (/ k /, / g /)** : Les occlusives vélaires sont des "compactes". L'énergie de la barre d'explosion, intense et de longue durée, est concentrée dans une étroite bande de fréquence. Les transitions, surtout pour F1, sont plus longue que celles des dentales et labiales. Les transitions pour / a / sont visuellement caractéristiques. La barre d'explosion est parfois double. Les caractéristiques de / k / et / g / sont détaillées dans le tableau B3.3.

| Occlusives vélaires | Contexte | Barre d' explosion | Transitions formantiques | Bruit de friction |
|---|---|---|---|---|
| / k / | / i / | - entre 2 et 4 KHz (renforcée entre 3 et 3.5 KHz)<br>- vers 5 et 6 KHz (plus faible) | -F1: ⌜<br>-F2: ⌜<br>-F3: — ou ⌜ | intense |
| | / a / | - intense entre 1.8 et 3 KHz (surtout 1.8 et 2.5 KHz)<br>- entre 4 et 5 KHz (double explosion fréquente) | -F1: ⌜<br>-F2: ⌝<br>-F3: ⌜ | intense |
| | / u / | - très intense, longue vers 0.7 KHz<br>- vers 4 KHz | -F1: ⌜<br>-F2: ⌞<br>-F3: — ou ⌜ | intense |
| / g / | / i / | - continue entre 2 et 4 KHz (intense entre 2.5 et 3.5 KHz)<br>- vers 5 et 6 KHz (faible) | -F1: ⌟<br>-F2: ⌜<br>-F3: — ou ⌜ | intense (longue durée) |
| | / a / | - vers 0.5 et 1 KHz<br>- continue entre 1.5 et 3 KHz (surtout entre 2 et 2.5 KHz)<br>- double explosion fréquente | -F1: ⌜<br>-F2: ⌝<br>-F3: ⌜ ou — | intense |
| | / u / | - très intense entre 0.6 et 0.8 KHz<br>- légère vers 4 KHz<br>- double explosion fréquente | -F1: ⌜<br>-F2: — ou ⌞<br>-F3: — ou ⌜ | intense |

**Tableau B3.3 :** Caractéristiques essentielles des occlusives vélaires [41]

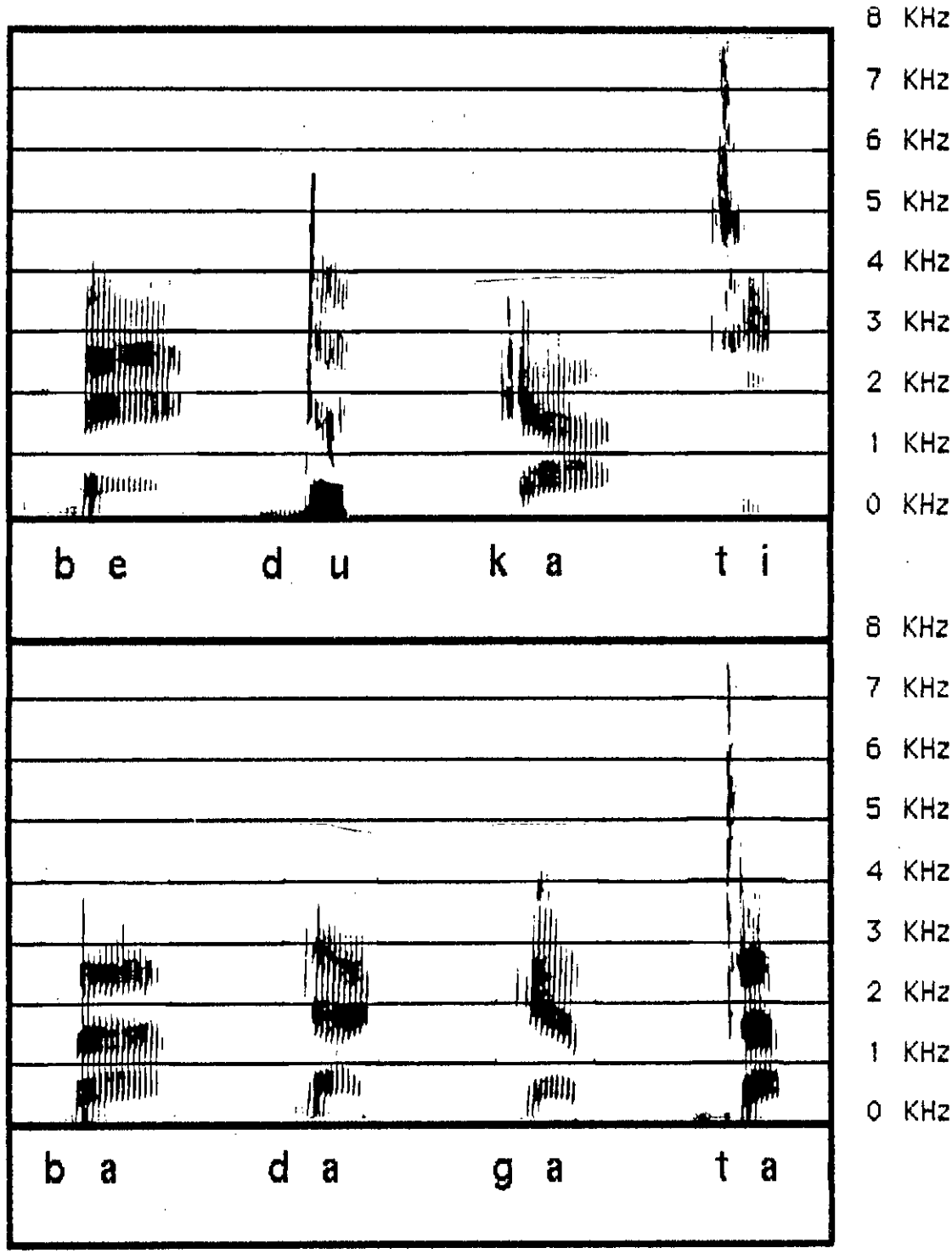



**Fig. B3.3 :** Exemples de spectrogrammes de diphones du type : "Consonne occlusive - Voyelle"

## 4.2. Les consonnes fricatives (ou constrictives)

Les fricatives ou constrictives sont des bruits, c'est-à-dire des événements apériodiques. Ce bruit résulte d'une turbulence aérodynamique qui prend naissance en un ou plusieurs points du conduit vocal en raison de la présence d'un fort resserrement (ou constriction) ou d'un obstacle placé dans le flot d'air expiratoire. Les fricatives peuvent être sourdes (/ s /, / ʃ /, / f /) ou sonores (/ Z /, / z /, / v /). Dans ce dernier cas, il y aura donc, comme pour les occlusives sonores, présence d'une barre de voisement.

Spectrographiquement, le bruit de turbulence apparaît comme un ensemble de petites stries verticales plus ou moins longues, d'intensité variable, disposées aléatoirement (Fig. B3.4).

**a - Les fricatives (/ s /, / Z /)** : En spectrographie, le bruit de turbulence est visible entre entre 4000 et 8000 Hz. Il apparaît quelquefois deux concentrations diffuses, l'une vers 5000 Hz et l'autre vers 8000 Hz, souvent appellées formants de bruit. La mesure la plus utile est celle de la limite inférieure du bruit qui reste fonction de l'entourage vocalique. Les transitions formantiques pour (/ s /, / Z /) sont voisines de celle de (/ t /, / d /) car ces deux couples sont des consonnes dentales.

**b - Les fricatives (/ ʃ /, / z /)** : Le bruit de turbulence est visible entre 2000 et 7000 Hz (ou plus) selon le degré de labialité des sons voisins. Deux concentrations peuvent être visibles, la première est comprise entre 2000 et 3000 Hz et la seconde est supérieure à 4000 Hz. L'énergie décroit souvent régulièrement vers les plus hautes fréquences avec un maximum vers 3000 - 4000 Hz. Généralement, il se produit de véritables petits silences (10 - 20 ms) encadrant la friction.

**c - Les fricatives (/ f /, / v /)** : Les fricatives labio-dentales (/ f /, / v /), pour lesquelles les incisives supérieures viennent au contact de la lèvre inférieure, se distinguent des autres fricatives principalement par leur faible intensité. Les transitions formantiques seront voisines de celles des labiales. Le bruit est rarement visible sous 1500 Hz, avec avec des variations rapides d'intensité.

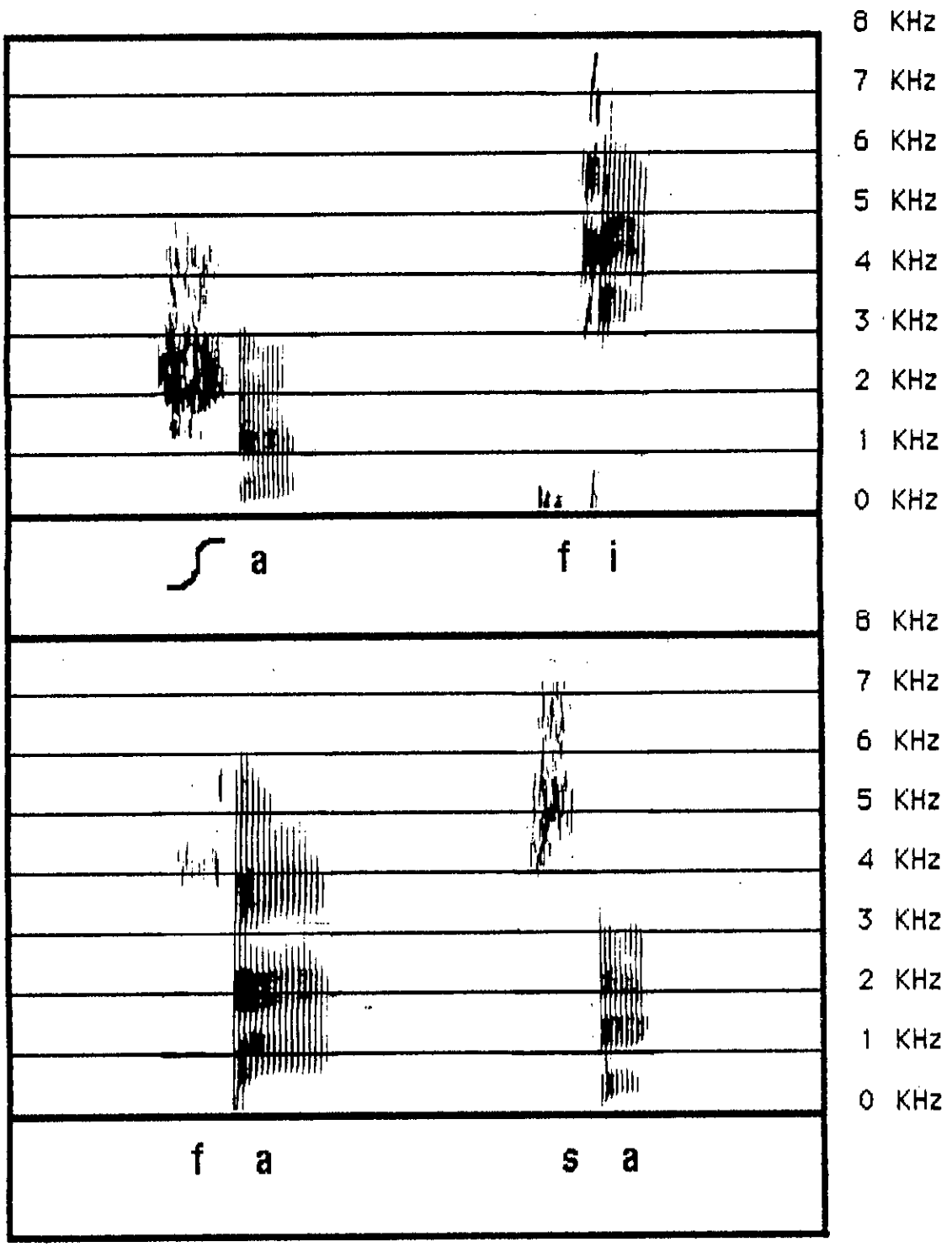



**Fig. B3.4 :** Exemples de spectrogrammes de diphones du type : "Consonne fricative - Voyelle"

# 5. RESULTATS

Plusieurs tests ont été effectués sur un corpus constitué d'un ensemble de diphones du type consonne-voyelle représenté par un certain nombre de spectrogrammes. Les consonnes choisies appartiennent aux classes des occlusives et fricatives et sont prises dans les contextes / **i** /, / **a** / et / **u** /.

Nous avons volontairement séparé la présentation des résultats obtenus en différents types à savoir : la courbe des énergies, la segmentation et l'identification afin d'obtenir une meilleure appréciation de ces résultats à chaque étape de traitement.

Dans le cadre de cet exposé nous ne pouvons donner que quelques résultats obtenus. Nous traiterons dans ce qui suit, à titre d'exemple, un échantillon représentatif des tests effectués et les résultats obtenus pour le diphone / ∫ a /.

## 5.1. La courbe des énergies

Les paramètres énergétiques **P1, P2,, ... , P12** et "**EN**" de l'échantillon / ∫ a / qui sont stockés dans le fichier ∫a.ech représenté par le tableau B3.4 :

```
reg2(q(1,p(1,1,1,2.5,1,1,7.75,13.75,8,5,10,1,4.4),"SI"))
reg2(q(2,p(1,1,1,2.5,1,1,7.75,13.75,8,5,10,1,4.4),"SI"))
reg2(q(3,p(1,1,1,7.5,1,3,12.25,16.5,18,9,12,7,7.4),"SI"))
reg2(q(4,p(1,1,1,7.5,1,3,12.25,16.5,18,9,12,7,7.4),"SI"))
reg2(q(5,p(1,1,1,6,4,8,21.25,22.5,21,15,21,15,11.4),"FR"))
reg2(q(6,p(1,1,1,4,1,1,20,24,20,20,25,19,11.4),"FR"))
reg2(q(7,p(1,1,1,4.5,5,1,17.75,23.25,26,26,29,14,12.5),"FR"))
reg2(q(8,p(1,1,1,4,7,1,16.5,28,28,23,26,20,13.1),"FR"))
reg2(q(9,p(1,1,1,1.5,6,5,18,24.5,28,27,26,15,12.8),"FR"))
reg2(q(10,p(1,1,1,4,6,7,17.5,27.75,21,15,25,13,11.6),"FR"))
reg2(q(11,p(1,1,1,3.5,9,6,19.5,25.75,22,15,17,10,10.9),"FR"))
reg2(q(12,p(1,1,1,9.5,11,9,18.5,21,15,9,14,13,10.2),"FR"))
reg2(q(13,p(1,1,1,8,8,5,16.25,21.75,20,11,12,7,9.3),"FR"))
reg2(q(14,p(1,1,1,10,11,7,11.5,16,9,4,5,1,6.46),"FR"))
reg2(q(15,p(33,34,32,24,11,11.5,5.5,7,1,1,1,1,13.5),"VM"))
reg2(q(16,p(42,43,41,34,30,26,6,5.25,6.33,1,1,1,19.7),"VM"))
reg2(q(17,p(42,44,43,39,39,33,7.5,10.75,3.33,1,1,1,22.1),"VM"))
reg2(q(18,p(42,42,42,40,36,34,9.25,9.25,12,1,1,1,22.5),"VM"))
reg2(q(19,p(45,45,44,42,40.5,30,12.75,11,12,1,1,1,23.6),"VS"))
reg2(q(20,p(45,45,44,42,40.5,30,12.75,11,12,1,1,1,23.8),"VS"))
reg2(q(21,p(45,45,44,42,40,30,12.75,11,12,1,1,1,23.7),"VS"))
```

**Tableau B3.4** : Les paramètres énergétiques des échantillons du diphone / ∫ a /

La courbe des énergies correspondante du même diphone est représentée par les figures B3.5 a, b et c. Sur cette courbe qui est constituée de 21 échantillons notés de q1 à q21, on remarque surtout une concentration d'énergie relativement grande des paramètres **P7, P8, P9, P10,** et **P11** des échantillons q3 à q14 correspondants normalement au phonème / ʃ / dans la bande de fréquence comprise entre 2 et 7 KHz. Les échantillons q15 à q21 correspondants au phonème / a / quant à eux sont caractérisés par une énergie plus intense des paramètres **P1** jusqu'à **P4** et un peu moins de **P5** et **P6.** Le tableau B3.5 donne des explications plus détaillées ainsi qu'une interprétation de la courbe des énergies du diphone / ʃ a /.

| PHONEME | Description à partir de la réponse du système | | | Description réelle | | Commentaires |
|---|---|---|---|---|---|---|
| | Echantillons | Paramètres | Energie | Bande de fréquence | Energie | |
| / ʃ / | q1 à q14 | P1, P2, P3 | nul | | | Les résultats sont très proches de la réalité |
| | | P4, P5, P6 | très faible | | | |
| | | P7, ... , P11 | intense | 2000 - 3000 Hz<br>4500 - 7000 Hz | intense | |
| | | P12 | faible | | | |
| / a / | q15 à q21 | P1, ... , P6 | intense | 500 - 700 Hz<br>1100 - 1300 Hz<br>2000 - 2500 Hz | intense | |
| | | P7, P8, P9 | faible | | | |
| | | P10,... , P12 | nul | | | |

**Tableau B3.5 :** Résultats et interprétations de la courbe des énergies des échantillons du diphone / ʃ a /

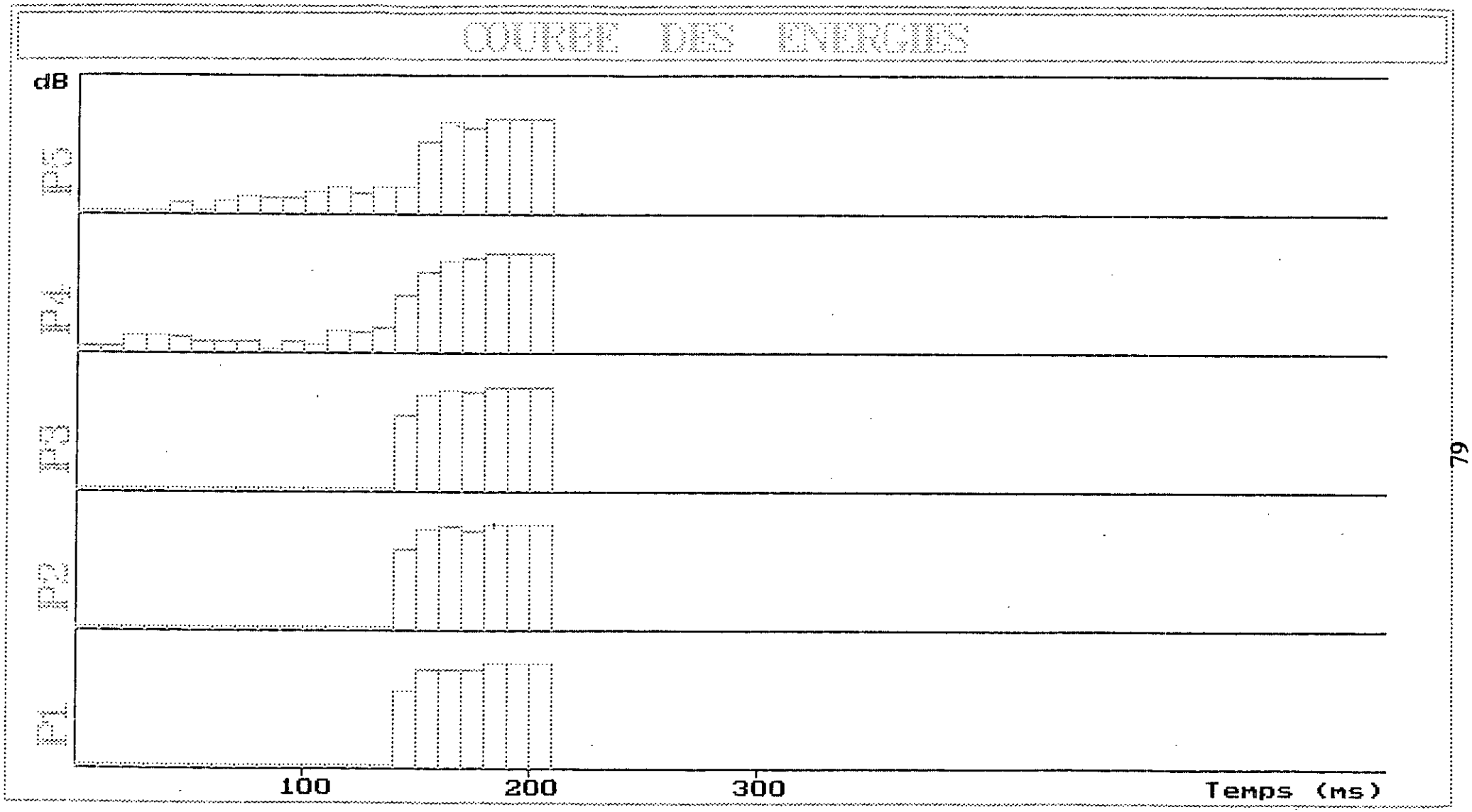





**Fig. B3.5 a) :** Courbe des énergies du diphone / ∫ a / obtenue par le système **SRAPH**

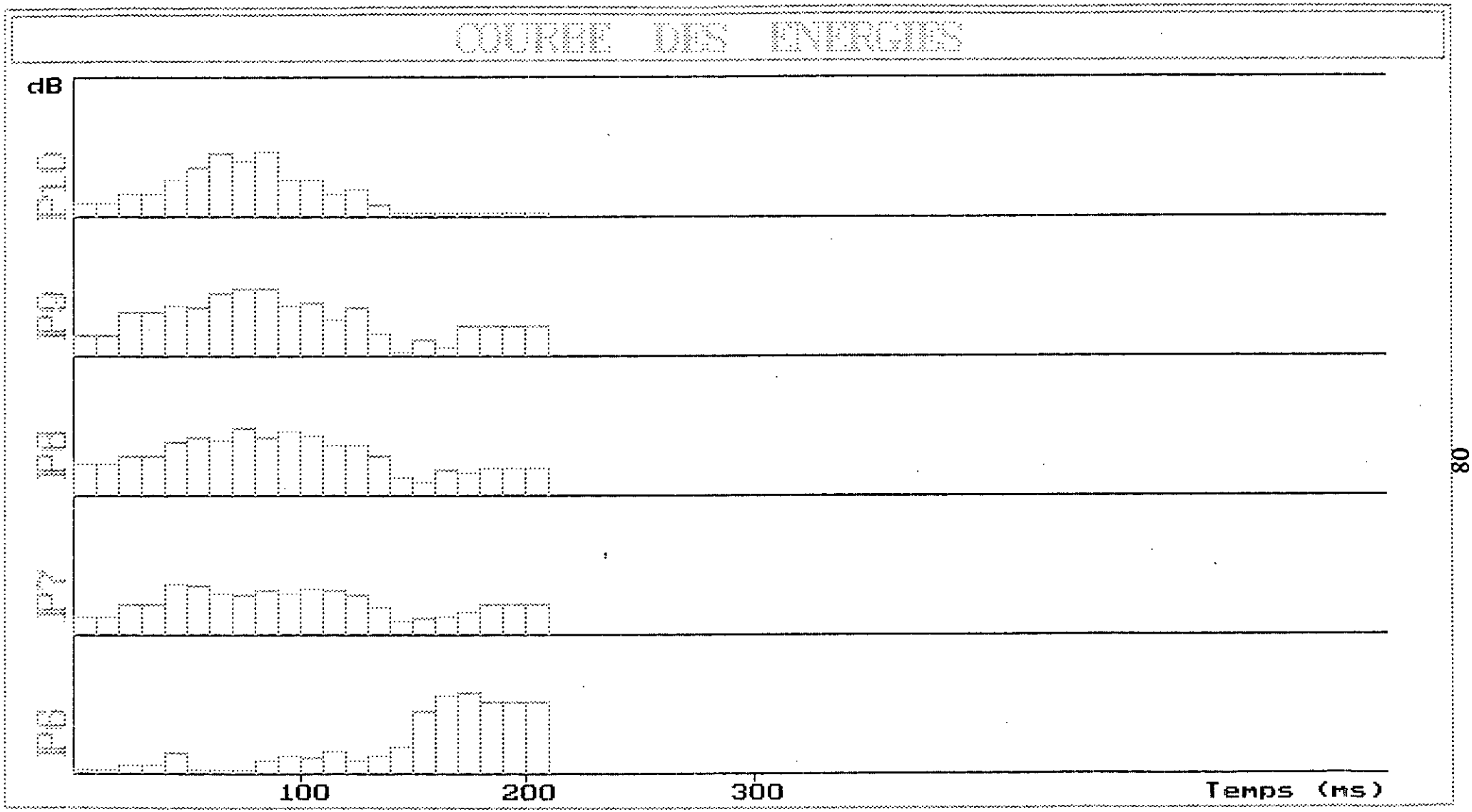



Fig. B3.5 b) : Courbe des énergies du diphone / ʃ a / obtenue par le système **SRAPH**

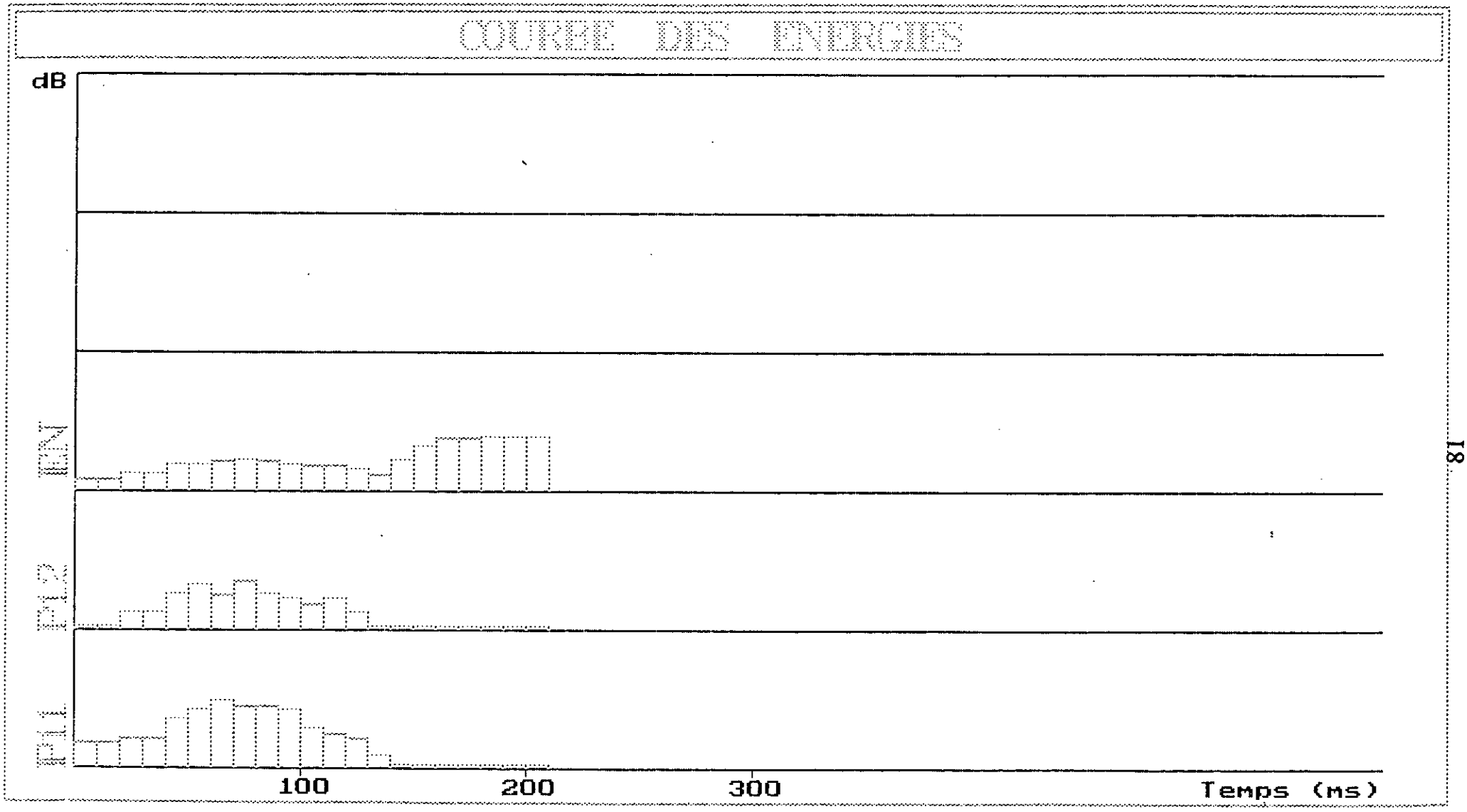



Fig. B3.5 c) : Courbe des énergies du diphone / ∫ a / obtenue par le système **SRAPH**

## 5.2. Résultats de la segmentation

Les tests que nous avons effectués lors de cette étape ont permis surtout de tester les différents algorithmes de segmentation mis en oeuvre. Or, les résultats préliminaires sur le corpus utilisé montrent que cette étape a été franchie avec succès. Néanmoins, un certain nombre de facteurs peuvent provoquer des écarts entre les résultats obtenus par le système et les données réelles : exactitude des valeurs des paramètres énergétiques, choix du banc de filtres (nombre de filtres et les fréquences de coupure), la valeur du seuil θ, etc.

Les tests de la segmentation du diphone / ∫ a / ont donné les résultats suivants (Tableau B3.6) :

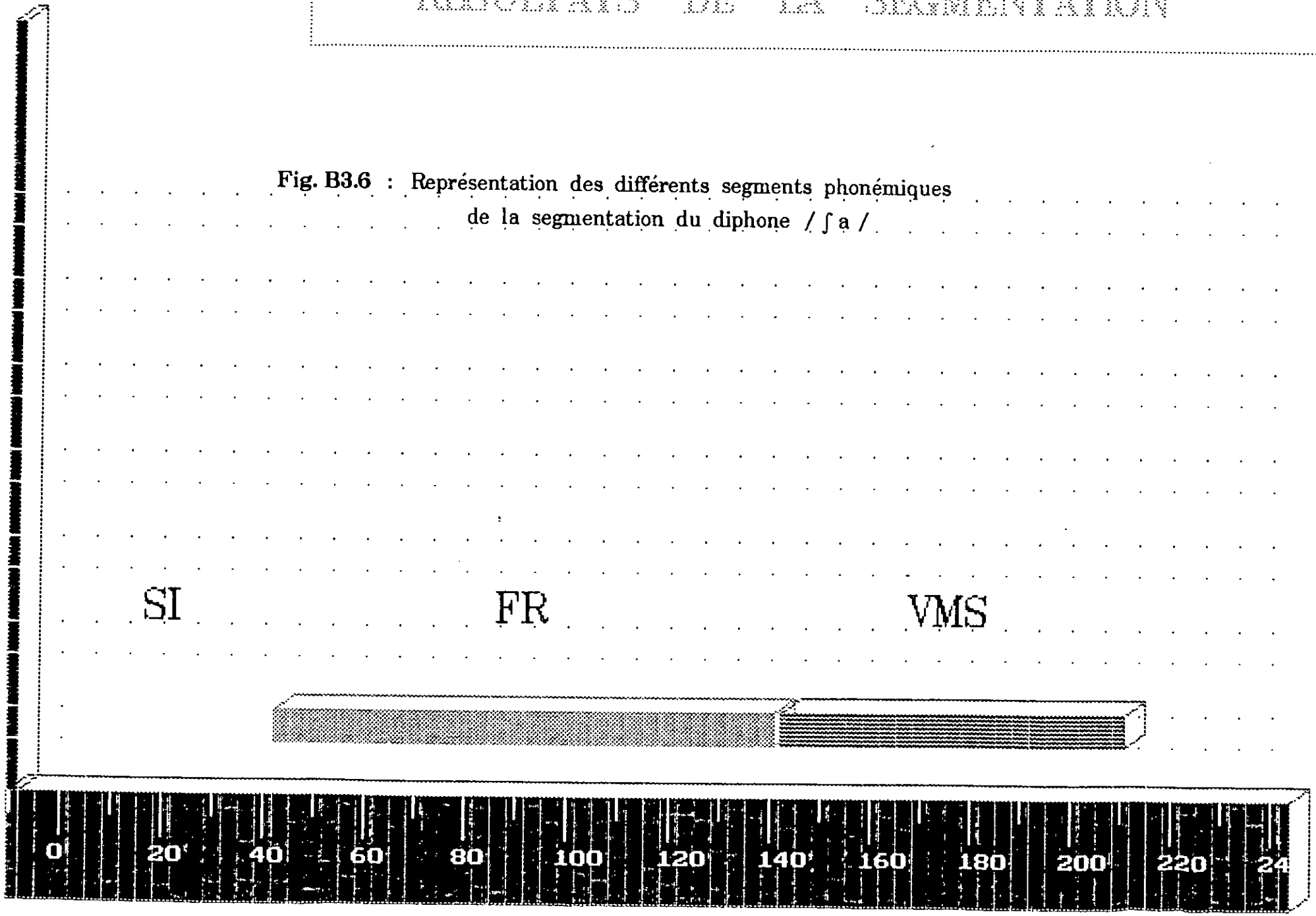
```
se( ss (1,210, [

   r(1,1,  40, "SI", p(   1,    1,    1,    5,
                          1,    2,   10, 15.1,
                         13,    7,   11,    4,  5.9 )),

   r(2,41, 100,"FR", p( 0.7,  0.7,  0.7,  3.9,
                        4.8,  3.5, 12.6, 16.7,
                         15, 11.7, 14.2,  9.1,  7.8 )),

   r(3,141,70, "VMS",p( 15.2, 15.5, 15.1, 13.6,
                        12.2, 10.1,  3.4,  3.4,
                         2.9,  0.3,  0.4,  0.4,  7.7 ))]))
```

**Tableau B3.6 :** Résultats de la segmentation du diphone / ∫ a /.

Ces résultats font apparaître ainsi trois types de segments phonémiques. Le premier est un SILENCE "**SI**" de durée 40 ms, le second est une FRICTION "**FR**" qui dure 100 ms et le troisième segment est un noyau vocalique du type VOISE MONTANT STATIONNAIRE "**VMS**" de durée 70 ms.

La figure B3.6 donne une meilleure représentation des résultats de la segmentation du diphone / ∫ a /. On remarque les différents segments phonémiques, leur durée ainsi que leur type représenté par une couleur caractéristique choisie.

Fig. B3.6 : Représentation des différents segments phonémiques de la segmentation du diphone / ʃ a /



## 5. 3. Résultats de l' dentification

Cette étape de traitement se caractérise par l' utilisation d' un certain nombre de règles. On donnera à titre d' exemple, une partie de quelque règles utilisées dans le cas des fricatives sourdes / f, s, ∫ /.

***REGLE 1***

**SI** Contexte droit est / /
**SI** Contexte gauche est / a /
**SI** déductions déjà faites sont / s, f, ∫ /
**SI** il existe deux zones de bruit
*) 3000 - 4500 Hz
*) 5000 - 7000 Hz
**SI** valeur d' énergie est comprise entre 8 et 20 dB
**ALORS** / s /

***REGLE 2***

**SI** Contexte droit est / /
**SI** Contexte gauche est / a, u /
**SI** déductions déjà faites sont / s, f, ∫ /
**SI** il existe une zone de bruit comprise entre 1000 et 7000 Hz
**SI** valeur d' énergie est comprise entre 8 et 20 dB
**ALORS** / f /

***REGLE 3***

**SI** Contexte droit est / /
**SI** Contexte gauche est / a /
**SI** déductions déjà faites sont / ∫ /
**SI** il existe deux zones de bruit
*) 2500 - 4000 Hz
*) 5500 - 7800 Hz
**SI** valeur d' énergie est comprise entre 8 et 20 dB
**ALORS** / ∫ /

Voici comme exemple, les résultats de l'identification du diphone / ʃ a / (Tableau B3.7) :

```
suit_id(id(1,[ id1(1, [ ]),
               id1(2, [ ph( 'c', 3.1, [1],    [], [2] ),
                        ph( 'f', 2.1, [3],    [], [2] ),
                        ph( 's',   1,  [],    [],  [] ) ]),
               id1(3, [ ph( 'a', 2.3, [2], [3,1],  [] ),
                        ph( 'o',   1,  [],    [],  [] ),
                        ph( 'u',   1,  [],    [],  [] )])]))
```

**Tableau B3.7 :** Résultats de l'identification du diphone / ʃ a /.

Ces résultats peuvent être interprétés de la manière suiuvante :

- le **premier segment** est un **SILENCE**

- le **deuxième segment** peut être un :

  - / ʃ / avec un score égal à 2.1
  - / **f** / avec un score égal à 1
  - / **s** / avec un score égal à 1

- le **troisième segment** peut être un :

  - / **a** / avec un score égal à 2.1
  - / **o** / avec un score égal à 1
  - / **u** / avec un score égal à 1

L'ensemble des résultats présentés ci-dessus du diphone / ʃ a / a été obtenu sans aucune modification de notre système de décodage acoustico-phonétique et confirme la validité des algorithmes et des règles utilisées. Toutefois, la grande précision observée de ces résultats est due en grande partie au peu de règles employées et que les indices utilisés sont sans chevauchement entre eux.

## 6. DISCUSSION DES RESULTATS

La tradition veut qu'on établisse des statistiques de résultats après un travail sur la parole. Dans notre cas, il nous paraît difficile au seul niveau acoustico-phonétique de fournir un taux de reconnaissance précis. Malgré son importance, Celui-ci ne peut constituer une référence pour l'évaluation des performances globales du système. Ce module, représenté par notre système, ne constitue qu'une étape vers la reconnaissance de la phrase qui a été prononcée par utilisation des niveaux supérieurs.

Cela est d'ailleurs en accord avec le processus humain de perception : l'homme comprend bien souvent une phrase prononcée par un locuteur sans avoir reconnu tous les phonèmes émis, en ayant recours à une analyse syntaxico-sémantique de cette phrase.

De plus, l'état de réalisation de ce système fait qu'il ne nous a pas été possible de faire des tests systématiques de reconnaissance, mais uniquement de tester un nombre de diphones pour lesquels nous avons obtenu des treillis de phonèmes acceptables, et incluant presque dans tous les cas, les réponses exactes parmi les trois premières reconnues par le système.

Nous avons tout de même essayé de relever un certain nombre d'observations qui nous ont permis de juger la qualité des résultats obtenus ainsi que le fonctionnement du système.

- Les essais que nous avons effectués ont permis surtout de **tester la validité du modèle** que nous avons proposé : algorithmes de la segmentation et d'identification phonétique en particulier. Celle-ci a été faite et Les résultats obtenus sont très encourageants.

- Le **temps de calcul** nous importait peu, et nous n'avons pas cherché à optimiser les algorithmes mis en oeuvre du point de vue temps d'exécution. Malgré cela, le temps de reconnaissance reste très acceptable : environ 4 secondes par traitement complet (segmentation et identification).

- Une des tâches primordiales du système est de **tester la logique des règles** décrivant les connaissances acoustiques et phonétiques mises en oeuvre. Or, d'après l'ensemble des résultats présentés ci-dessus, nous pouvons confirmer le bon fonctionnement des règles utilisées. Néanmoins, nous pensons que certaines d'entre elles peuvent être perfectibles, et qu'en conséquence les performances seraient améliorées significativement.

- les tests effectués sur notre système sont limités à une reconnaissance d'un certain nombre de diphones, ce qui n'a pas permis pour le moment de **juger le système dans un cadre réel de dialogue vocal.**

Les résultas obtenus permettent de valider notre approche du décodage acoustico-phonétique. Cette réalisation représente pour nous une étape et non le terme de nos recherches.

# *CHAPITRE B4*

# PRESENTATION ET UTILISATION DU SYSTEME S R A P H REALISE

## 1. PRESENTATION GENERALE DU SYSTEME

**SRAPH** est un Système de Reconnaissance Acoustico-Phonétique prévu pour s' insérer dans un système plus complet de reconnaissance de la parole continue. Son objectif principal est de fournir en sortie une chaîne de treillis phonétiques pour une éventuelle interprétation par les niveaux supérieurs.

Le système se présente sous la forme d' un ensemble de programmes écrit en langage **TURBO PROLOG** en raison de son efficacité et de son adaptation à ce type de traitement [ 43, 44, 45 ]. La mise en oeuvre de ces programmes a nécessité de gros efforts de programmation pour l' obtension de logiciels facilement utilisables et maintenables, tant pour ce qui concerne le système SRAPH proprement dit, que pour les différents fichiers et bases de données exploités.

Le système réalisé inclut, en plus de l' étape du décodage acoustico-phonétique, plusieurs autres fonctions de base puissantes et intéractives, déstinées toutes à faciliter la tâche de l' utilisateur spécialisé grâce au multifenêtrage et à un système complet de menus déroulants.

## 2. CONFIGURATION LOGICIELLE ET ORGANISATION DU SYSTEME

Pour disposer d' une architecture logicielle modulaire permettant une adaptation souple aux nouvelles applications et l' amélioration indépendante de chaque partie, nous avons organisé notre système en rubriques indépendantes où chacune d' elles est liée à une utilisation possible du système. La figure B4.1 schématise l' organisation générale ainsi obtenue du système SRAPH.

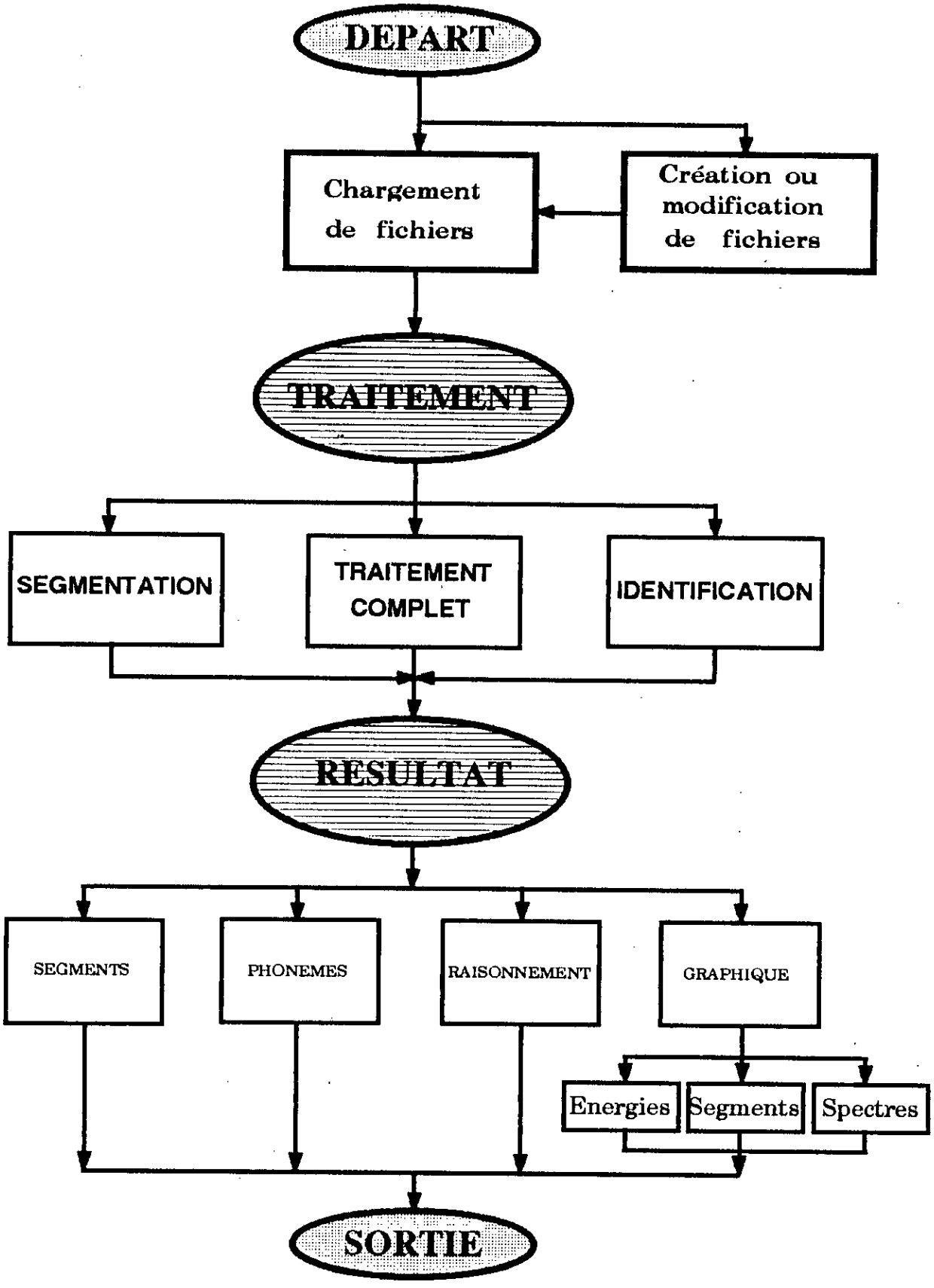



**Fig. B4.1 :** Organisation générale du système SRAPH réalisé

### 2.1. Rubrique AIDE

Comme presque tout système, un fichier d'aide contenant un texte visant à apporter des éclaircissements et des conseils d'utilisation, est mis à la disposition de l'utilisateur.

### 2.2. Rubrique INT-EXP

Cette rubrique permet de confier essentiellement aux spécialistes phonéticiens la manipulation et la formalisation des multiples connaissances dont ils disposent. Cette formalisation impose une énonciation déclarative aussi naturelle que possible des règles qui constitue le savoir souvent empirique, imprécis et en constante évolution sur ce difficile problème. Les stratégies d'utilisation de ces connaissances sont élaborées de manière indépendante. Elles peuvent donc s'adapter plus facilement aux exigences du contexte pour optimiser l'utilisation des bases de données disponibles. Les bases qui peuvent être manipulées sont de deux types :

#### a - Base acoustique :

Ces connaissances sont étroitement liées aux différentes étiquettes phonétiques et aux relations entre elles. Ces relations sont de la forme :

**Si** Etiq_1
**Et** Etiq_2
**Alors** Etiq_resultat

Chaque étiquette est représentée par : son numéro, la liste représentative des phonèmes probables ainsi qu'une couleur caractéristique choisie normalement pour la visualisation sur un écran couleur.

Les étiquettes phonétiques ainsi que les relations entres elles sont sauvegardées dans un fichier qu'il est possible de rappeler pour le consulter ou le modifier.

#### b - Base phonétique

Ce sont les connaissances les plus nombreuses du système ; elles permettent de caractériser les sons, par comparaison au moyen d'un certain nombre de paramètres stables et représentatifs des phonèmes associés et d'interpréter phonétiquement les données de manière inductive.

Les paramètres portent sur :

- le numéro de la règle,
- le méssage de la règle,
- la validité de la règle ( V / I ),
- le poid de la règle,
- le contexte gauche,
- le contexte droit,
- les déductions déjà faites,
- l' étiquette de la règle,
- les conditions du signal,
- les conclusions de la règle.

De la même manière que le premier type, plusieurs manipulations peuvent être faites dans ce cas à savoir :

- la consultation (de n' importe qu' elle règle donnée simplement par son numéro ou de toute la base),
- l' adjonction de règles,
- la modification de règles,
- la suppression de règles ou de bases,
- la création de bases.

### 2.3. Rubrique FICHIER

Cette rubrique sert de données pour les différents traitements. Elle est constituée de quatre types de fichiers qui peuvent être sélectionnés selon le type de traitement recherché. On distingue :

#### a - Fichiers ECHANTILLONS

appelé aussi fichier "parole". Chacun de ces fichiers est constitué de l' ensemble des échantillons de séquences de parole définies par leurs paramètres énergétiques décrit auparavant. Si cette option est sélectionnée, le système affiche tous les fichiers échantillons existants qui sont caractérisés par l' option *. ECH. L' utilisateur a la possibilité de choisir un fichier parmi eux, de le visualiser, pour faire demarrer la segmentation, mais il peut aussi faire des modifications sur ces fichiers ou en créer d' autres nouveaux.

### b - Fichiers SEGMENTS

Ce type de fichier représente des fichiers échantillons déja segmentés et portent les noms *. SEG. Ces fichiers peuvent être visualisés et identifiés à la demande de l' utilisateur. l' intérêt majeur de cette option réside dans la possibilité offerte aux chercheurs phonéticiens pour tester certaines segmentations déjà faites en modifiant certains paramètres.

### c - Fichiers BD_CONN

Cette option donne la main à l' utilisateur de sélectionner la base de connaissances nécessaire pour l' étape d' identification. Ces fichiers portent les noms * . CON. Elle permet aussi à l' expert phonéticien de bâtir et de tester sa propre base.

### d - Fichiers IDENTIF

Ce type de fichiers sont le résultat de l' étape d' identification et portent les noms *. IDT. Ces fichiers contiennent la liste des phonèmes identifiés avec un certains nombre d' informations utile à l' utilisateur que celui-ci peut visualiser.

## 2.4. Rubrique TRAITEMENTS

Cette rubrique constitue le noyau de notre système ; elle permet d' effectuer les principaux traitements prévus à savoir : la segmentation et l' identification. Cependant, pour une meilleure utilisation, nous avons choisi de donner à notre système la possibilité soit de séparer les deux traitements segmentation et identification ou de faire un traitement complet en une seule étape et ceci bien sûre à la demande de l' utilisateur. Chaque fois qu' un traitement est abordé, un message de début et fin d' exécution est affiché, en plus l' option **CONFIGURATION** nous permet de se renseigner sur l' état des opérations déjà effectuées sur les fichiers parole existants.

## 2.5. Rubrique RESULTATS

Dans cette rubrique, on retrouve les résultats de tous les traitements et sous différentes formes. Ainsi, comme on peut le deviner, d' abord la liste des segments et des phonèmes résultats qui peuvent être visualisés grâce aux options **SEGMENT** et **PHONEME**, par la suite, ces résultats peuvent être interprétés d' une manière graphique où l' on peut visualiser

les courbes des différentes énergies, des segments ou des spectres, ainsi qu' une fonction d' explication du raisonnement à l' aide de l' historique de la résolution retracée en langage pseudo-naturel. Cette facilité est d' un très grand secours dans la recherche des règles ayant amené des résultats jugés erronés ou douteux.

# CONCLUSIONS ET PERSPECTIVES

Cette étude s' insère dans le cadre général de la reconnaissance automatique de la parole continue. Elle avait pour but principal de réaliser un système de reconnaissance acoustico-phonétique dans l' optique de fournir des informations structurées à un système plus complet de reconnaissance intégrant d' autres composantes (syntaxe, sémantique, prosodie, pragmatique).

La première partie de cette thèse nous a permis de replacer nos recherches dans un contexte général. En présentant successivement l' importance de la communication orale et les recherches effectuées jusqu' alors en reconnaissance et en compréhension de la parole. Nous avons donné, en particulier, une vue d' ensemble des diverses informations à mettre en oeuvre dans un système de reconnaissance et de compréhension de la parole continue.

La seconde partie a été entièrement consacrée à l' étude et la réalisation du Système de Reconnaissance Acoustico-Phonétique **S R A P H**. Notre principal souci à travers cette réalisation a été de valider le modèle que nous avons proposé dans ce domaine tant pour la structure de la base de règles utilisée et les connaissances impliquées que pour l' architecture adoptée.

Le système réalisé a une double vocation. C' est d' abord un outil aidant à identifier des séquences de parole et donc à s' intéresser aux points critiques du décodage acoustico-phonétique, mais aussi un outil efficace d' acquisition et d' apprentissage des connaissances dans ce domaine.

Les résultats que nous avons proposés sont donnés à titre indicatif, le système réalisé n' a été testé qu' à partir d' un certain nombre de spectrogrammes. Il ne fait pas de doute que les vraies tests devront être faits en utilisant une carte d' acquisition du signal de parole et un corpus beaucoup plus large, laissant varier le locuteur, les contextes et les conditions d' élocution. C' est alors seulement, que nous pourrons mesurer, avec beaucoup plus de précision, les performances de notre système de reconnaissance acoustico-phonétique.

De plus, notre système, est un bon cadre pour la collaboration et la découverte de nouvelles connaissances ; il constitue une base de connaissances évolutive. Les objets de base des différents modules de connaissances présentés dans cette étude sont spécifiques au domaine du décodage acoustico-phonétique, mais la méthodologie employée, la structuration des règles et l' architecture de notre système sont indépendants du type de connaissances représentées, et donc du domaine.

L' expertise en décodage acoustico-phonétique ne nous semble pas encore aussi sûr et bien cerné que celle d' un certain nombre de domaines et qu' une réflexion sur la nature de cette expertise doit être essentiellement menée afin de pouvoir mieux exprimer et employer cette expertise. Néanmoins, ce qui apparaît le plus clair c' est qu' aucune démarche sérieuse n' étant réfusable en l' état actuel des expérimentations, l' accent doit être mis sur le développement d' outils et de base de données des sons pour développer des protocoles expérimentaux permettant de répondre aux besoins des machines aptes au décodage phonétique indépendant du locuteur et du contexte.

Dans ce sens et malgré les résultats obtenus, nous pensons que notre étude ne constitue qu' une contribution à la compréhension de la parole continue en général et à la recherche sur le décodage acoustico-phonétique en particulier. Ainsi nous avons jeté les bases d' un système ouvert qui puisse être amélioré, au fur et à mesure que sont trouvés les meilleurs compromis entre les exigences du traitement informatique et celles de la formalisation des données linguistiques.

Ce travail pourra se poursuivre dans différentes directions, notamment :

- élargir le système en lui rajoutant d' autres contraintes linguistiques de niveaux supérieurs (en particulier la composante lexico-syntaxique) ;

- améliorer la base de connaissances grâce à la contribution d' un vrai expert en phonétique afin d' obtenir un treillis de meilleure qualité ;

- relier le système à une carte spécialisée (acquisition et traitement du signal de la parole) afin d' améliorer les différents indices qui décrivent réellement le signal de la parole et pouvoir faire des tests réels sur un corpus beaucoup plus large améliorant ainsi les performances du système ;

- enfin faire appel aux techniques du traitement d' images dont l' objet consiste à incorporer les mécanismes propres à la perception et l' interprétation visuelle des spectrogrammes de parole.

[12] **W. A. WOODS** et al, *"Speech Understanding Systems - Final report, B B N"*, Rapport 3438, 1976.

[13] **B. RITEA**, *"Automatic Speech Understanding System"*, Proc. of the 11th IEEE Computer Society Conference, Washington D. C, pp. 319 - 322, 1975.

[14] **B. T. LOWERRE**, *"The HARPY Speech Recognition System"*, Carnegie-Mellon University (U S A), Ph. D., Computer Science, 1976

[15] **D. R. REDDY** et al., *"Speech Understanding Systems : Final report"*, Carnegie-Mellon University (U S A), Computer Science, 1977.

[16] **N. CARBONNEL**, et al., *"Les connaissances nécessaires dans un système de dialogue oral Homme-machine"*, 5ème congrés RF-IA, GRENOBLE, 1985.

[17] **A. LELIEVRE**, *"Codage et traitement de certains types de données phonétiques pour une reconnaissance automatique de la parole par classification"* Thèse de 3ème cycle, University de RENNES, 1981.

[18] **J. MARIANI**, *"Reconnaissance de la parole continue par diphonèmes ; Processus d'encodage et de décodage phonétique"*, Actes du Séminaire GALF, TOULOUSE, pp. 97 - 115, 1981.

[19] **A. E. ROSENBERG, L. R. RABINER, J. G. WILPON** et **D. KAHN**, *"Using Concatenated Demi-Syllabes in a Isolated Word Recognition Systeme"*, 11ème I. C. A, TOULOUSE, 1983.

[20] **L. R. RABINER , A. E. ROSENBERG , J. G. WILPON** et **T. M. ZAMPINI**, *"A Bootstrapping Technique to obtaining Demi - Syllabe Reference Patterns"*, 102 nd A S A Meeting, 1981.

[21] **V. W. ZUE** et **R. A. COLE**, *"Experiments in Spectrogram Reading"*, Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proc., pp. 116 -119, 1979.

[22] **N. CARBONELL, J. P. HATON, J. M. PIERREL** et **F. LONCHAMP**, *"Elaboration d'un sytème expert pour le décodage phonétique automatique"*, Speech Communication, Vol. 2, n° 2 - 3, pp. 231 - 233, 1983.

[23] **G. MERCIER**, *"Analyse acoustique et transcription phonétique du signal de parole"*. Ecole de l'I. R. I. A., Reconnaissance et compréhension du dialogue écrit et parlé, NANCY, 1977.

[24] **J. P. HATON et PERENNOU**, *"Reconnaissance et compréhension de la parole"*, Cours de l'école d'été Informatique de l'A. F. C. E. T., NAMU, Belgique, 1978.

[25] **J. CAELEN**, *"Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique"*, Thèse d'état, TOULOUSE, 1979.

[26] **D. AUTESSERRE** et **M. ROSSI**, *"Propositions pour une segmentation et un étiquetage hiérarchisé ; application"*, 14émes Journées d'Etudes sur la Parole, PARIS, pp. 147 - 151, 1985.

[27] **GALF**, *"La quantification vectorielle pour le traitement de la parole"*. Actes du séminaire GALF, L. Miclet (ed.) E. N. S. T - PARIS, 1985.

[28] **G. BAILLY** et **D. LIU**, *"Détection d'indices par quantification vectorielle et réseaux markoviens"*, 16èmes Journées d'Etudes sur la Parole, HAMMAMET, Tunisie, pp. 60 - 63, 1987.

[29] **A. WAIBEL** et al, *"Phoneme Recognition Using Time-delay Neural Netwoks"*, A T R Report, 1987.

[30] **C. S. MYERS** et **L. R. RABINER**, *"Connected Digit Recognition Using a Level Building DTW Algorithm"*, IEEE Trans. Acoust., Speech and Signal Proc., Vol 29, n° 3, pp. 351 - 363, 1981.

[31] **J. P. Haton** et **M. LAMOTTE**, *"Un algorithme de comparaison dynamique pour la reconnaissance automatique de la parole et son application pratique"*, Automatisme, XIX. n° 5, pp. 248 - 289, 1974.

[32] **G. PERENNOU**, *"Reconnaissance des mots isolés dans le cas d'un grand vocabulaire"*, 2ème Congrés A. F. C. E. T, pp. 334 - 342, 1979.

[33] **N. CARBONELL, J. P. DAMESTOY, D. FOHR, J. P. HATON, F. LONCHAMP** et **J. M. PIERREL**, *"Techniques d'intelligence artificielle en décodage acoustico-phonétique"*, 14èmes Journées d'Etudes sur la parole, PARIS, pp. 299 - 303, 1985.

[34] **D. MEMMI, M. ESKENAZI, J. MARIANI** et **A. NGUYEN-XUAN**, *"Un système expert pour la lecture de sonagrammes"*, Speech Communication, Vol. 2, n° 2 - 3, pp. 234 - 236, 1983.

[35] **J. P. HATON**, *"Les systèmes à bases de connaissances dans la communication Homme-machine"*, Revue COGNITIVA 85, pp. 211 - 224, 1985.

[36] **M. GRENIE**, *"Nature et hiérarchie d'indices acoustiques indépendants du locuteur : application à la reconnaissance automatiques des voyelles du francais"*, Thèse de 3éme cycle, AIX-EN-PROVENCE, 1987.

[37] G. MERCIER et al, "The KEAL Speech Understanding System, dans *"Spoken Language Generation and Understanding"*, J. C. Simon (Ed), D. Reidel, pp. 524 - 544, 1980.

[38] **J. P. HATON, J. M. PIERREL, G. RERENNOU, J. CAELEN, J. L. GAUVAIN**, *"Reconnaissance automatique de la parole"*, Edition DUNOD informatique, 1991.

[39] C. ABRY, D. AUTESSERRE, J. CAELEN, "Proposition pour la segmentation et l'étiquetage de la base de données des sons du Français", 14 ème Journées d'Etude sur la Parole, Paris, pp. 156 - 163, 1985.

[40] **P. E. STERN, M. ESKENAZI** et **D. MEMMI**, *"An expert system for speech spectrogram reading"*, IEEE, Proc. ICASSP, TOKYO, 1986.

[41] **CALLIOPE**, *"La parole et son traitement automatique"*, Edition MASSON, 1989.

[42] **P. DURAND**, *"Etude acoustique de consonnes occlusives du français commun"*, Doctorat de 3ème cycle, Université de provence, AIX-MARSEILLE, 1982.

[43] **C. TOWNSEND**, *"Introduction à **TURBO PROLOG**"*, Edition SYBEX, 1987.

[44] **J. LAPORTE** et **D. DELPORT**, *"**TURBO PROLOG** : Construisez des applications"*, Edition EYROLLES, 1987.

[45] **COPYRIGHT**, *"**TURBO PROLOG** version 2.0 : Reference guide and User's guide"*, Edition BORLAND INTERNATIONAL, 1988.

| VOYELLES | | | CONSONNES | | |
|---|---|---|---|---|---|
| **Sons ou phonèmes** | **Mots clefs** | **Classe ou nature** | **Sons ou phonèmes** | **Mots clefs** | **Classe ou nature** |
| a | passe | ORALE | p | pas | OCCLUSIVE SOURDE |
| i | il | | t | tas | |
| y | nu | | k | cas | |
| o | eau | | b | bon | OCCLUSIVE VOISEE |
| ə | le | | d | dans | |
| ɛ | lait | | g | gars | |
| e | et | | v | vie | FRICATIVE VOISEE |
| oe | heure | | Z | zéro | |
| ũ | ou | NASALE | z | je | |
| ã | an | | f | feu | FRICATIVE SOURDE |
| õ | on | | s | sons | |
| ɛ̃ | lin | | ʃ | chat | |
| œ̃ | un | | n | nous | NASALE |
| | | | m | ma | |
| | | | l | lent | LIQUIDE |
| | | | r | eue | |

ALPHABET PHONETIQUE INTERNATIONAL