Université de Nancy I U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

# APHODEX: UN SYSTEME EXPERT EN DECODAGE ACOUSTICO-PHONETIQUE DE LA PAROLE CONTINUE


## THESE

**présentée et soutenue publiquement le 14 mars 1986**

**A L'UNIVERSITE DE NANCY I**

**pour l'obtention du titre de**

**DOCTORAT de l'UNIVERSITE de NANCY I en INFORMATIQUE**

par

**Dominique FOHR**

Ingénieur E.N.S.E.M.

**Composition du Jury**

| | | |
|---|---|---|
| Président: | René | CARRE |
| Rapporteurs: | Pierre | LESCANNE |
| | Ramon | LOPEZ de MANTARAS |
| Examinateurs: | Jean-Paul | HATON |
| | Jean-Marie | PIERREL |
| Invité: | François | LONCHAMP |

à Arlette,
Nelly,
Suzanne
et Arthur.

Jean-Paul Haton a accepté de m'accueillir dans son laboratoire et de diriger mon travail de recherche. Qu'il trouve ici l'expression de ma profonde gratitude pour m'avoir confié un sujet aussi passionnant que la reconnaissance de la parole.

Je tiens à remercier Monsieur Carré qui me fait l'honneur de bien vouloir présider ce jury.

Que Monsieur Lescanne et Monsieur Lopez de Mantaras soient remerciés d'avoir accepté de juger ce travail.

Jean-Marie Pierrel trouvera ici l'expression de ma très vive gratitude pour l'intérêt constant qu'il a manifesté au cours de mes travaux.

Je veux remercier Noelle Carbonell pour sa collaboration efficace au cours de nos réunions avec l'expert.

Enfin, je remercie tout particulièrement François Lonchamp sans lequel ce travail n'aurait pu être réalisé. Les nombreux échanges que nous avons eus m'ont apporté beaucoup dans la conception de ce projet.

Que tous les membres de l'équipe RFIA, et particulièrement Anne, soient remerciés pour l'aide qu'ils m'ont apportée et pour leur soutien amical.

# APHODEX

Acoustic

PHOnetic

Decoding

EXpert

# PLAN

## INTRODUCTION

## PARTIE A

## GENERALITES

**Introduction**

CHAPITRE 1
**Les systèmes à bases de connaissances**

I. Evolution de l'Intelligence Artificielle

II. Représentation des connaissances

III. Les systèmes experts: généralités
   1) Caractéristiques d'un système expert
   2) Architecture
      a) La base de connaissances
      b) La base de faits
      c) Le moteur d'inférences
   3) Développement

IV. Manipulation de connaissances incertaines

V. Les systèmes experts à règles de production

VI. Intérêts et inconvénients des systèmes à règles de production
- 1) Intérêts
- 2) Inconvénients
- 3) Conclusion

CHAPITRE 2
**La reconnaissance de la parole**

I. Introduction

II. Méthode globale
- 1) Mots isolés
- 2) Mots enchaînés

III. Méthode analytique
- 1) Rôle du décodage acoustico-phonétique dans le processus de reconnaissance
- 2) Difficultés
  - a) Choix des paramètres
  - b) Variabilité intra-locuteur
  - c) Variabilité inter-locuteur
- 3) Approche traditionnelle et limites
- 4) Approche intelligence artificielle

CHAPITRE 3
**Phonétique**

I. Introduction

II. Appareil phonatoire
- 1) L'appareil respiratoire
- 2) Le larynx
- 3) Les cavités supra-glottiques

III. Une représentation acoustique: les spectrogrammes

IV. Les phonèmes du français: production et analyse acoustique
  1) Voyelles
    a) Les voyelles orales
    b) Les voyelles nasales
  2) Les consonnes
    a) Les occlusives
    b) Les fricatives
    c) Les sonantes

---

# PARTIE B

# ACQUISITION DE L'EXPERTISE

---

**Introduction**

CHAPITRE 1
**Méthodologie d'acquisition**

I. Difficultés de l'acquisition
  a) Au niveau du raisonnement
  b) Au niveau de la compétence visuelle

II. Règles de phonétique générale

III. Règles contextuelles
- règles de classification
- règles d'identification
- règles d'affinement de la segmentation
- règles d'élimination
- règles de confirmation
- règles de compatibilité
- règles de prédiction

IV. Conclusion

CHAPITRE 2
**Stratégies**

I. Approche globale

II. Analyse locale
1) segmentation évidente
2) segmentation ambigue ou difficile

---

# PARTIE C

# FORMALISATION DE L'EXPERTISE

---

**Introduction**

CHAPITRE 1
**Outils et prétraitements**

I. Introduction

II. Les spectrogrammes
    1) Obtention des spectrogrammes numériques
    2) Visualisation

III. Le corpus

IV. Les prétraitements
    1) noyaux vocaliques
    2) plosives
    3) fricatives

V. Les algorithmes de calcul d'indices acoustiques
    1) La barre d'explosion
    2) Suivi de formants
    3) Calcul de la limite inférieure de friction

VI. Conclusion

CHAPITRE 2
**La base de connaissances**

I. Justification du choix d'un système à règles de production

II. Syntaxe des règles
        a) Numéro de la règle
        b) Contexte
        c) Résultat
        d) Conditions sur les faits
        e) Conclusion

III. Option choisie

IV. Différents types de règles et exemples
    1) Règles de phonétique générale
    2) Règles de déduction
    3) Règles de confirmation
    4) Règles de discrimination
    5) Règles de prédiction
    6) Règles d'exclusion
    7) Règles de phonologie
    8) Règles de segmentation

CHAPITRE 3
**Le moteur**

I. Le raisonnement de l'expert

II. Le premier test

III. Le moteur actuel
    1) Le cahier des charges
    2) Description statique
        a) Prétraitement et mesures
        b) Le treillis
    3) Fonctionnement du moteur
        a) Cas général
        b) Cas du premier segment inconnu
        c) Cas d'une double hypothèse de segmentation
        d) Cas d'une double mesure
        e) Cas d'une règle de négation
        f) Exemple de fonctionnement

IV. Conclusion

# PARTIE D

# REALISATION ET RESULTATS

**Introduction**

CHAPITRE 1
**Les résultats des prétraitements**

I. Les noyaux vocaliques

II. Les plosives

III. Les fricatives

CHAPITRE 2
**Les résultats sur les sonantes**

I. But

II. Analyse factorielle discriminante
    1) Ensemble d'individus
    2) Résultats

III. Approche globale: modèles de référence

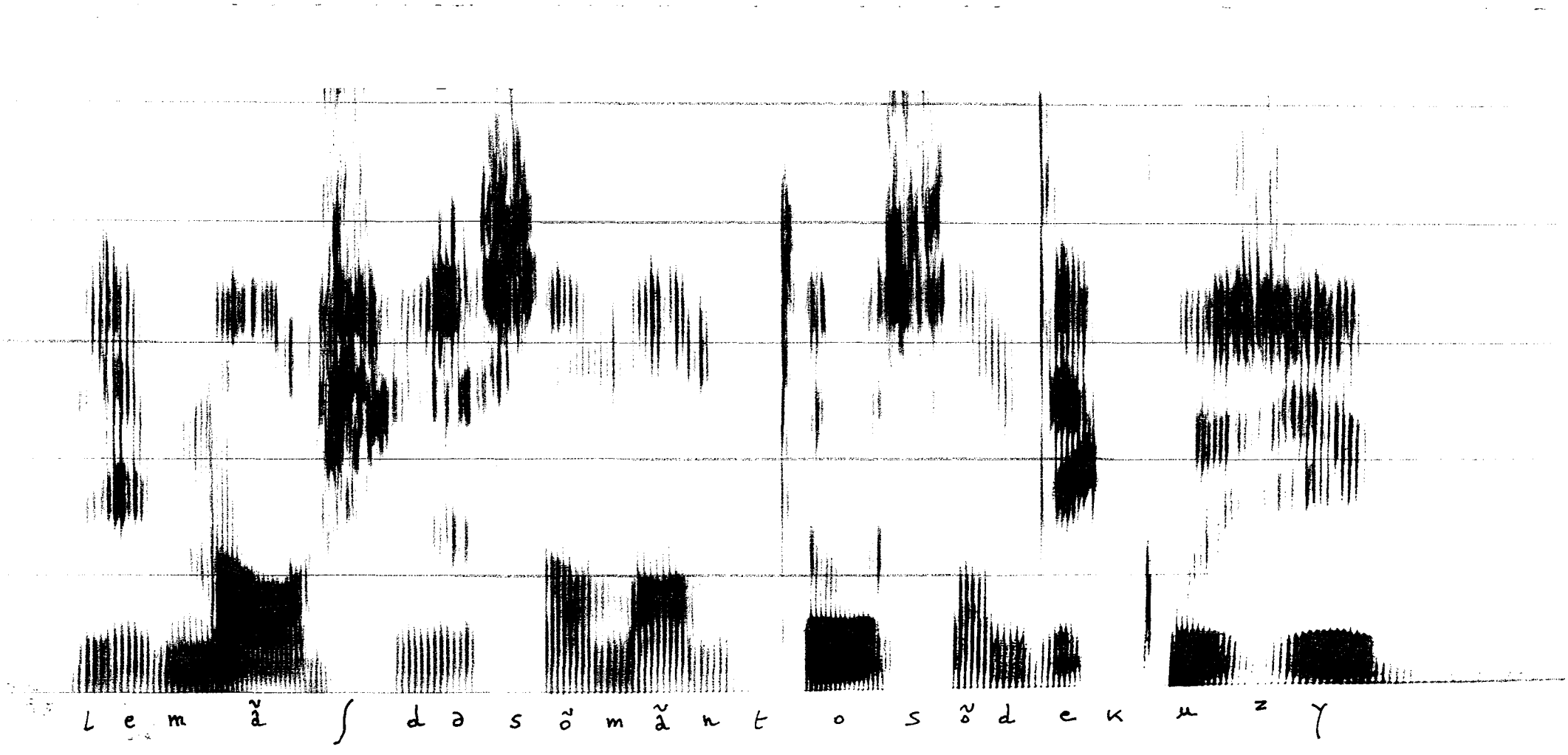
CHAPITRE 3
**Résultats du moteur d'inférences sur les plosives**

# CONCLUSION

# ANNEXES

# INTRODUCTION

**Spectrogramme de la phrase:** Les manches de son manteau sont décousues



# INTRODUCTION

Des moyens que l'homme utilise pour échanger de l'information, la parole est le plus spontané, le plus direct. Mais le décodage acoustico-phonétique représente une des difficultés majeures en reconnaissance de la parole continue en contexte multilocuteur. On peut noter que la compréhension de la parole est un processus largement inconscient; par conséquent l'introspection ne fournit guère d'informations sur les mécanismes de cette activité humaine. Nous ignorons même sur quels éléments acoustiques précis porte le processus de reconnaissance chez l'homme.

Pour améliorer les algorithmes existants de décodage acoustico-phonétique, nous avons entrepris l'analyse et la modélisation du savoir-faire d'un expert en lecture de spectrogrammes; la compétence mise en jeu à ce niveau étant plus accessible à l'analyse, car plus consciente. Nous travaillons en collaboration avec François Lonchamp de l'Institut de Phonétique de Nancy II. A partir de la représentation sous forme de spectrogrammes d'un énoncé de parole continue française, notre expert obtient un taux de reconnaissance (en environnement multilocuteur) voisin de 80%, ce qui est bien supérieur aux performances des systèmes automatiques actuels.
Compte-tenu du caractère particulier de l'expertise (examen visuel de formes complexes avec raisonnement contextuel), nous avons été conduits à expérimenter plusieurs méthodes pour acquérir cette expertise (partie B chap. 2).

Nous avons obtenu des règles à la fois générales et contextuelles, et, simultanément, nous avons essayé de recueillir les différentes stratégies utilisées par l'expert (chap. 3).

Nous avons formalisé cette expertise sous la forme d'un système expert. Dans un premier temps, nous segmentons le signal de parole en grandes classes phonétiques, à savoir: noyaux vocaliques, plosives et

fricatives. Ensuite, à chaque segment obtenu nous associons une base de faits qui résulte de l'application de procédures de prétraitement mettant en oeuvre les techniques classiques de traitement du signal ainsi que, dans certains cas particuliers, des algorithmes de reconnaissance de formes. En effet, la compétence de l'expert comporte deux aspects: une expertise visuelle qui lui permet de segmenter le spectrogramme et d'en extraire les indices pertinents, et une expertise au niveau du raisonnement grâce à laquelle il interprète ces indices et étiquette les segments.

Le but du système consiste à proposer pour chaque segment une liste d'hypothèses phonétiques incluant l'interprétation correcte du segment, quitte à augmenter pour y parvenir la taille du treillis phonétique correspondant à l'énoncé. Limiter l'indéterminisme au niveau du décodage afin de faciliter les traitements supérieurs (reconnaissance et compréhension) constitue notre second objectif.

La nature de nos objectifs ainsi que la complexité de l'expertise imposent des contraintes à notre système qui doit être capable de:

- remettre en cause la segmentation à tout moment,
- dérouler en parallèle une analyse sur plusieurs segmentations possibles,
- prendre en compte les phénomènes contextuels ; le contexte gauche d'un segment est connu, puisqu'on travaille de gauche à droite, mais il est constitué de plusieurs phonèmes (on utilise plusieurs étiquettes pour un même segment), d'où l'introduction d'un indéterminisme à ce niveau,
- tenir compte de l'incertitude en ce qui concerne l'interprétation des mesures (détection d'indices),
- déterminer des seuils relatifs qui correspondent le mieux possible à la démarche experte.

Les règles sont du type règles de production: les prémisses mettent en jeu des indices acoustiques discriminants, les conclusions sont constituées soit par des actions soit par des listes de phonèmes pondérés.
Les caractéristiques du moteur d'inférences sont les suivantes:

- il fonctionne en chaînage avant et en chaînage arrière,
- les règles du système sont compréhensibles et modifiables facilement par un expert,
- il est facilement incrémentable,
- il fournit une trace de son raisonnement.

Pour résoudre ces différents problèmes, nous avons développé notre propre moteur d'inférences, sa principale caractéristique étant de pouvoir suivre plusieurs lignes de raisonnement en parallèle.

Pour que le moteur puisse fonctionner, nous avons dû développer un certain nombre d'outils (Partie C, chap.2):

- **réalisation de spectrogrammes numériques**

Les prémisses des règles fournies par l'expert faisant intervenir les indices acoustiques présents sur des spectrogrammes produits par un Voiceprint A 700, il nous a paru nécessaire de développer une représentation machine du signal aussi proche que possible de celle fournie par les spectrogrammes utilisés par l'expert; ces spectrogrammes numériques sont affichables sur un système de traitement d'images et sur des consoles graphiques de type Tektronix;

- **programmation d'algorithmes de traitement du signal** (FFT, LPC, calcul de la fréquence fondamentale) et leur intégration, sous forme de prétraitement, dans le système expert;

- **acquisition et segmentation d'un corpus**

Pour mesurer les performances de notre système, nous avons acquis un corpus de 57 phrases phonétiquement équilibrées de Combescure [Combescure 81] prononcées à un rythme naturel d'élocution par cinq locuteurs masculins non professionnels. Les phrases étaient prononcées par un locuteur expérimenté et répétées par les locuteurs (mémoire immédiate). Nous avons adopté ce mode opératoire pour éviter un rythme d'élocution trop lent et une intonation artificielle de type lecture. Nous avons de ce fait homogénéisé la prosodie et le rythme des locuteurs (environ 14 phonèmes par seconde). Ces phrases ont été numérisées à 12 kHz sur 10 bits.

De plus, nous avons réalisé, outre les spectrogrammes analogiques sur Voiceprint, les spectrogrammes numériques correspondant à ces 57 phrases. L'expert a placé manuellement, à partir des spectrogrammes numériques et de la courbe de l'énergie globale, les marques de segmentation sur le corpus numérisé. Deux segmentations ont été effectuées: une segmentation phonémique et une intra-phonémique (par èvénements phonétiques). Les taux de reconnaissance que nous donnerons par la suite sont tous fondés sur ces segmentations manuelles.

Par ailleurs, l'expert ayant décodé les spectrogrammes analogiques, nous connaissons ses performances sur ce corpus (80,2% sur les consonnes) ce qui permet de comparer les erreurs du système avec celles commises par l'expert;

- **élaboration d'algorithmes de segmentation** de la parole continue en grandes classes phonétiques (noyaux vocaliques, fricatives, plosives).

Dans la partie D, nous analyserons et commenterons les résultats que nous avons obtenus.

# PARTIE A

# GENERALITES

# INTRODUCTION

Notre étude fait appel à des connaissances empruntées à différents domaines, à savoir: la reconnaissance de la parole, la phonétique, les systèmes à base de connaissances. Etant donné que nous devons manipuler des connaissances incertaines, nous avons été conduits à choisir une approche de type intelligence artificielle. Notre choix s'est porté sur le formalisme des systèmes experts dont nous exposerons les avantages et les inconvénients (chap 1).

Le décodage acoustico-phonétique s'inscrit dans le cadre plus général de la reconnaissance de la parole: nous essaierons de montrer comment il peut s'insérer dans un système de dialogue oral homme-machine (chap 2). Signalons que nous avons retenu l'approche analytique, une analyse globale n'étant pas envisageable en parole continue multilocuteur.

Il nous est apparu que, pour concevoir un système de décodage acoustico-phonétique, il était indispensable de comprendre comment étaient produits les sons et quels étaient les effets de la coarticulation. C'est pour cette raison que le chapitre 3 est consacré à la description de l'appareil phonatoire ainsi qu'aux différents phonèmes du français.

# CHAPITRE 1

# LES SYSTEMES A BASES DE CONNAISSANCES

## I. Evolution de l'intelligence artificielle

S'il n'y a pas de définition universellement admise de l'intelligence artificielle, on peut dire avec Alain Bonnet que c'est ''la discipline visant à comprendre la nature de l'intelligence en construisant des programmes d'ordinateur imitant l'intelligence humaine '' [Bonnet 84]. Plus précisément, cette science s'est donné pour objectif d'analyser les comportements humains dans les domaines de la compréhension, de la perception, de la résolution de problèmes, afin de pouvoir ensuite les reproduire à l'aide d'une machine.

Dans une tentative datant d'une trentaine d'années, les premiers chercheurs ont essayé de résoudre le problème de l'intelligence artificielle de manière générale, de façon globale. Ils pensaient alors que l'intelligence était avant tout une aptitude, très générale, à raisonner dans tous les domaines. Aujourd'hui, cette approche apparait présomptueuse, et on cherche, au contraire, à résoudre des problèmes dans des domaines spécifiques très restreints.

Au début, les pionniers ont surtout utilisé le raisonnement logique formel: soit un raisonnement déductif, soit un raisonnement par l'absurde. Mais le monde réel n'obéit pas à une logique aussi simple; la connaissance de la réalité est floue et incomplète, donc le raisonnement est approché ou incertain. Pour remédier à cette insuffisance, on a associé aux conclusions du raisonnement des coefficients de vraisemblance ou de plausibilité qui se combinent selon des lois probabilistes (Théorème de BAYES, par exemple) ou de façon plus heuristique.

De plus, on s'est rendu compte qu'il est nécessaire d'accumuler une masse énorme de connaissances pour qu'une machine soit capable de raisonner intelligemment sur un sujet donné: cette constatation est à l'origine des systèmes experts, cas particulier de systèmes à base de connaissances. Pour représenter et gérer ces nombreuses connaissances, les

chercheurs ont développé des formalismes appropriés, tels que: les règles de production, les frames, les scénarios.

## II. Représentation des connaissances

Les performances d'un système expert sont essentiellement liées à la qualité de sa base de connaissances. Il est nécessaire d'avoir une représentation qui soit à la fois:

- compréhensible par l'expert, pour qu'il puisse lui-même modifier la base de connaissances,
- facile à modifier et à augmenter, donc aisément constructible de manière incrémentale,
- efficace, c'est-à-dire susceptible d'optimiser les temps de réponse du système.

De nombreuses méthodes sont utilisées, parmi lesquelles on peut citer:

- la logique mathématique [PROLOG, Colmeraurer 83],
- les systèmes à règles de production [DENDRAL, Feigenbaum 71],
- les réseaux sémantiques [Pinson 81],
- les frames, et les représentations objets [Damestoy 85],
- le ''blackboard'' [HEARSAY II. Reddy 73].

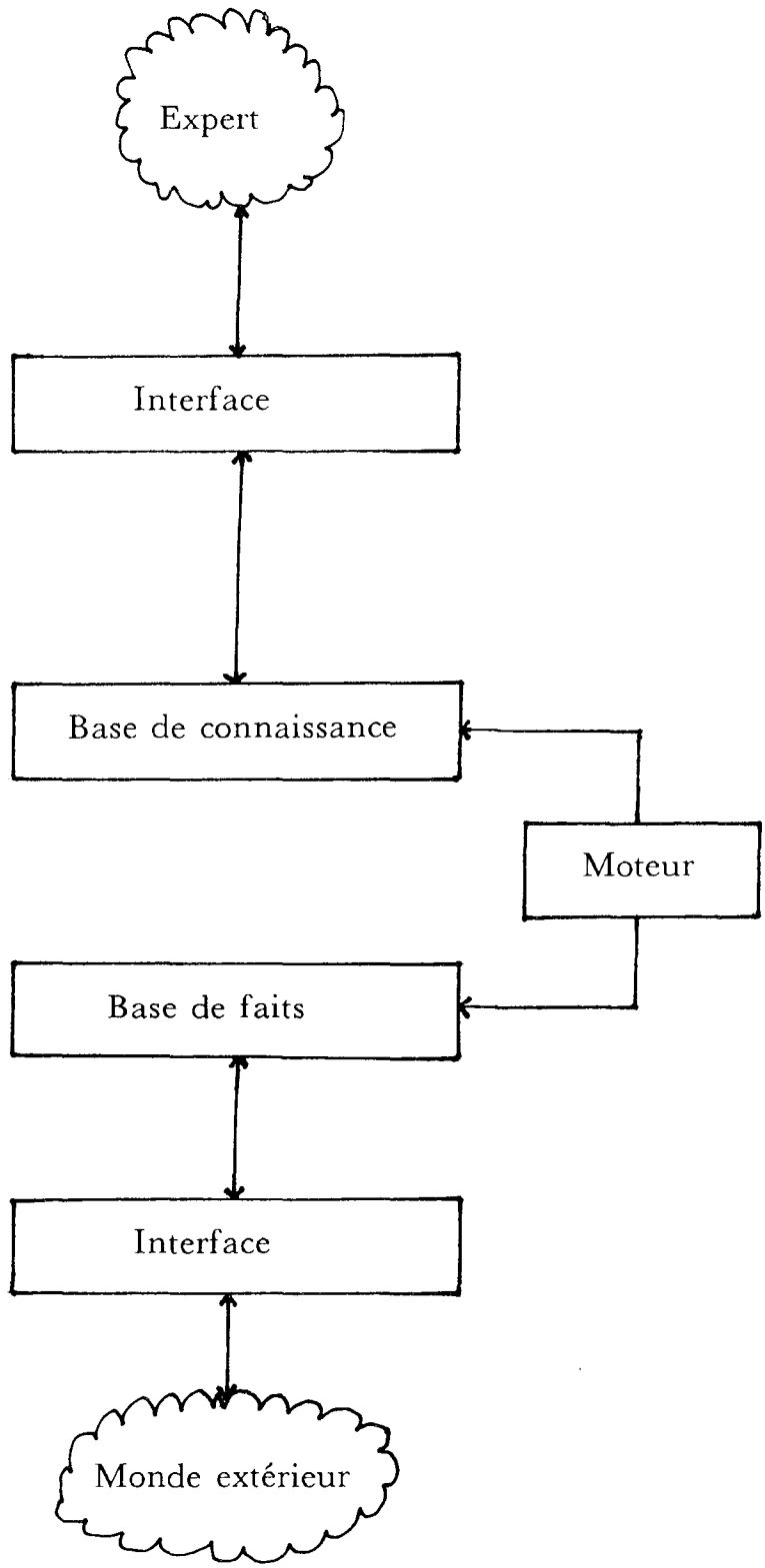
Le système que nous avons réalisé se fonde sur des règles de production, formalisme assez bien adapté à la nature de l'expertise que nous voulons modéliser, comme nous le verrons (cf. Parties B et C).

## III. Les systèmes experts: généralités

### 1) Caractéristiques d'un système expert

Ces systèmes se caractérisent par leur capacité à raisonner sur un ensemble de faits, à partir de connaissances plus ou moins spécifiques, en suivant une démarche comparable à celle adoptée par un spécialiste lorsqu'il résoud un problème relevant de sa discipline. Un tel système doit pouvoir expliciter la manière dont il a obtenu les conclusions auxquelles il a abouti.

**Figure A.1** : Architecture d'un système expert.



## 2) Architecture

Un système expert se compose principalement d'une base de connaissances, d'une base de faits, et d'un moteur d'inférences (cf figure A.1). Examinons maintenant chacune de ces trois parties.

### a) La base de connaissances

Elle contient les connaissances expertes ou non, qui peuvent être exprimées de façon factuelle ou de manière déductive.

Les connaissances factuelles (objets) peuvent être représentées par:

- des triplets (objet,attribut,valeur),
- des réseaux sémantiques,
- des frames ou prototypes.

Les connaissances déductives peuvent être traduites sous forme de:

- règles de production,
- procédures classiques,
- démons,
- clauses de Horn.

Ces deux représentations peuvent cohabiter simultanément dans la base de connaissances [Laurent 85]. Par exemple: la connaissance abstraite ''/p/ est une plosive '' peut être représentée sous forme d'un triplet (/p/,classe,plosive) ou sous forme déductive: ''si X = /p/ alors X est une plosive ''. L'ordonnancement de ces éléments peut être quelconque - ce qui permet un système facilement incrémentable - ou, au contraire, leur ordre peut être significatif et refléter une stratégie particulière.

### b) La base de faits

Elle contient les données propres aux différents problèmes à résoudre. En outre, elle peut renfermer les faits déduits par raisonnement, le but à atteindre, et les sous-buts à prouver.

### c) Le moteur d'inférences

A partir d'un état de la base de faits et à l'aide des connaissances et des heuristiques contenues dans la base de connaissances, le moteur d'inférences va générer un nouvel état de la base de faits.

transition

état de la BF ----------> nouvel état de BF

Ce cycle va se reproduire jusqu'à ce que la condition d'arrêt soit vérifiée. Cette condition d'arrêt peut correspondre soit à la résolution du problème, soit au fait qu'il n'y a plus d'opérateur applicable. Une partie de la stratégie se trouve donc contenue dans le moteur qui doit choisir les connaissances à activer et leur ordre d'application (métaconnaissances). Une transition est caractérisée par:

- l'état auquel elle s'applique,
- un couple objet-action de cet état,
- l'état qu'elle engendre.

Le moteur choisit d'abord un état parmi l'ensemble des états activables, c'est-à-dire parmi ceux pour lesquels toutes les transitions possibles n'ont pas encore été activées. Une fois l'état choisi, il sélectionne un couple objet-action applicable à cet état pour obtenir un nouvel état; c'est ce qu'on appelle la résolution de conflits [OPS - Forgy,79]. Dans le cas de moteurs ''avec variables '' (prédicats du 1er ordre), le nombre de ces couples peut être grand; dans le cas des moteurs sans variables (calcul des propositions), il est beaucoup plus faible. Les critères de choix peuvent porter sur:

- la transition,
- l'objet lui-même,
- l'action: par exemple, le nouvel état de la base de faits (plus ou moins proche du but final).

Remarque:
Certaines transitions suppriment des éléments dans la base de faits; l'élément peut être un objet, une propriété, une relation... Dans les cas, les plus nombreux, où on ne garde pas la trace de tous les états successifs de la base de faits, il faut prévoir un mécanisme qui permet, en cas de retour arrière, de restaurer les états antérieurs de la base de faits. Le retour arrière se produit, la plupart du temps, lorsque le système tombe sur une impasse, c'est-à-dire lorsqu'aucune transition ne peut plus être appliquée sur la base de faits et que le problème n'a pas été résolu.

### 3) Développement

Pour concevoir un système expert, il faut tout d'abord analyser le raisonnement de l'expert humain, puis formaliser la connaissance acquise. Il s'agit donc de développer une méthodologie d'acquisition des connaissances expertes. La qualité du système est directement liée à la

valeur de l'expert humain, spécialiste du domaine. Celui-ci doit pouvoir s'exprimer le plus librement possible, oralement, par écrit, ou graphiquement. Le cogniticien, spécialiste d'intelligence artificielle, devra noter toutes les remarques de l'expert, ses hésitations, et même ses erreurs. Ensuite, il lui faudra formaliser les connaissances et les stratégies acquises, toujours en coopération avec l'expert. Dans une dernière phase, le système devra être testé sur un ensemble d'exemples, l'expert étant toujours présent pour modifier et améliorer la base de connaissances.

## IV. Manipulation de connaissances incertaines

De nombreux systèmes experts prèsentent la particularitè de raisonner sur des connaissances incertaines. Dans MYCIN, chacune des prèmisses est affectèe d'un coefficient de vraisemblance:

de -1 (complètement faux),
à +1 (complètement vrai),
(0 reprèsentant l'ignorance).

L'opèration ''et'' est rèalisèe avec la fonction MIN,
l'opèration ''ou'' avec la fonction MAX.

Si une règle R1 conclut le fait F1 avec un coefficient vraisemblance V1, et une règle R2 conclut le même fait F1 avec un coefficient de vraisemblance V2, on calcule le coefficient de vraisemblance du fait F1 par la formule:

CV(F1,R1,R2) "= CV(F1,R1) + CV(F1,R2)"
" - CV(F1,R1) * CV(F1,R2)"
= V1 + V2 - V1 * V2

Ainsi: Si V1 = 0,75
la vraisemblance de F1 devient ègale à 0,9.
et V2 = 0,6

Ainsi le coefficient de vraisemblance de F1 devient pus forte que si l'on n'applique qu'une seule des deux règles. Il y a renforcement de l'hypothèse. De plus, chaque règle peut avoir un coefficient d'attènuation (C) compris entre 0 et 1, qui indique le degrè de confiance accordè à la règle.

CV(F1,R1,R2) = "CV(F1,R1)*C(R1) + CV(R2,F2)*C(R2)"
"- CV(R1,F1)*CV(R2,F2)*C(R1)*C(R2)"

Quelques rares systèmes, tel Prospector, utilisent la formule de BAYES et se fondent donc purement sur le calcul des probabilitès.

## V. Les systèmes à règles de production

Dans de tels systèmes, les connaissances sont formalisées sous forme d'entités élémentaires:

Prémisses -----> Conclusions

Il existe deux stratégies utilisables:

- le chaînage avant,
- le chaînage arrière.

### - le chaînage avant

Dans cette démarche, le système part des données initiales et, en appliquant les règles, il déduit de nouveaux faits. Quand, au cours d'un cycle complet d'examen des règles, plus aucune d'entre elles ne peut être appliquée, le système s'arrête. Pour qu'une règle soit appliquée, il faut que toutes ses prémisses soient vérifiées; à ce moment là, les conclusions de la règle sont placées dans la base de faits (si la conclusion de la règle est une action, elle est effectuée). La base de faits ayant été modifiée, on réexamine l'ensemble des règles pouvant être activées.

### - Le chaînage arrière

Dans ce mode de fonctionnement, le système essaie de prouver un but. Il y a création d'un arbre ''et-ou'', c'est-à-dire qu'il faut que toutes les prémisses d'une règle déduisant le but soient vérifiées. Les prémisses deviennent à leur tour des sous-buts, et le cycle recommence.

Remarque:

Certains systèmes utilisent à volonté le chaînage avant ou le chaînage arrière, il en est ainsi de Tango [Cordier 84], de TG1 et de beaucoup d'autres.

## VI. Intérêts et inconvénients des systèmes à règles de production

### 1) Intérêts

On peut noter que, lors de la phase d'acquisition de l'expertise, l'expert s'exprime souvent sous forme de règles du type :

''si prémisses, alors conclusions''.

En outre, cette formulation de la compétence de l'expert permet à ce dernier d'effectuer lui-même corrections et modifications. Elle autorise également le développement incrémental du système. Par ailleurs, elle permet une représentation de connaissances complexes sous forme de parcelles élémentaires beaucoup plus faciles à manipuler: les règles de production. Cette propriété peut être mise à profit en reconnaissance de la parole pour tenter d'apporter une solution à des sous-problèmes bien précis.

Enfin, ce formalisme offre la possibilité d'intégrer au système les stratégies de l'expert, soit sous forme de métarègles, soit par l'intermédiaire d'algorithmes de résolution de conflits. On peut alors tester diverses stratégies avec la même base de règles. La plupart des moteurs d'inférences permettent la manipulation de connaissances appartenant à des domaines tout à fait différents; il suffit de changer la base de connaissances pour pouvoir utiliser un même moteur d'inférences dans le cadre d'applications différentes.

**2) Inconvénients**

Tout d'abord, pour construire un système expert, il faut trouver un expert qui connaisse parfaitement son domaine pour l'avoir pratiqué pendant plusieurs années. Un théoricien ne suffirait sans doute pas. Le domaine à modéliser doit être circonscrit et le vocabulaire permettant de décrire les situations envisageables, limité. On imagine mal un système expert en oeuvres d'art. En outre, l'expert utilise quelquefois, pour expliquer sa démarche, des représentations de type graphique (dessins, diagrammes) qui sont parfois difficiles à exprimer sous forme de règles.

Par ailleurs, les connaissances étant entrées en machine, dans le désordre, les systèmes experts à règles de production ne permettent une structuration de la base de connaissances qu'a posteriori. Il peut donc s'ensuivre une perte de temps, lors de l'exploitation.

On notera enfin que ce formalisme offre un mode de représentation des informations peu économique.

**3) Conclusion**

Malgré ces inconvénients, les systèmes experts sont promis à un grand développement, car leurs avantages prédominent largement. On peut citer, parmi les réussites, les exemples de Prospector (découverte d'un gisement de Molybdène), de Mycin (actuellement utilisé au Pacific

Medical Center de San Francisco) ou d'Eurisko (dans une compétition nationale, Eurisko a battu tous les joueurs humains à un wargame). Dans le cadre de notre travail, le formalisme des systèmes experts à règles de production semble être particulièrement adapté pour résoudre les problèmes posés par la reconnaissance de la parole.

# CHAPITRE 2

# LA RECONNAISSANCE DE LA PAROLE

## I. Introduction

Depuis longtemps, l'homme rêve de créer des machines capables de parler et d'obéir à la voix humaine. On sait déjà réaliser, en laboratoire, de la synthèse de très bonne qualité [Allen 83], même si de nombreux problèmes restent à résoudre. Dans le domaine de la vie courante, nombreuses sont les applications de la synthèse: qui n'a pas entendu la voix d'une calculatrice, d'un ordinateur, voire même d'une voiture?

Quant au deuxième problème (la reconnaissance), il est loin d'être résolu dans son intégralité. Si le marché de la synthèse de la parole est assez développé, celui de la reconnaissance de la parole en est encore à ses premiers balbutiements. Dans le cadre de la reconnaissance de mots isolés, on commercialise déjà des systèmes; cependant, remarquons qu'ils sont soumis à de nombreuses contraintes: vocabulaire restreint, système le plus souvent monolocuteur, ambiance non bruitée, coût relativement élevé. Pour les systèmes de compréhension acceptant un langage artificiel (reconnaissance mettant en jeu des connaissances syntaxiques), le stade de la précommercialisation est atteint (Myrtille I. [Pierrel 81]). Pour des langages ''pseudo-naturels'' la faisabilité a été démontrée (Myrtille II.). Quant à la reconnaissance d'un véritable dialogue (multilocuteur, parole continue, langage naturel), le stade de la recherche n'a pas été dépassé. Toutefois, les progrès sont rapides et les perspectives prometteuses. Il est intéressant de noter qu'International Resource Development estime qu'en 1992, le marché de la reconnaissance de la parole atteindra 2,6 milliards de dollars (soit environ 20 milliards de francs) !

Les recherches se multiplient, étant donnés l'enjeu capital que représente le problème de la communication homme-machine et les avantages qui découlent de l'utilisation de la parole par rapport à l'utilisation d'un écran-clavier:

- l'utilisateur a les mains et la vue libérées, et peut accéder à distance à un service,

- la vitesse est plus grande car un utilisateur occasionnel parle plus vite qu'il ne tape à la machine,

- cela ne nécessite pas de connaissances particulières (il suffit de savoir parler).

Il est clair que les méthodes à appliquer pour la reconnaissance de la parole sont étroitement liées aux différents types de problèmes à résoudre: taille du vocabulaire, mono ou multilocuteur, mots isolés, mots enchaînés ou parole continue. Pour la reconnaissance ''monolocuteur, mots enchaînés'' ou ''multilocuteur, petit vocabulaire, mots isolés'', une approche globale semble parfaitement convenir, tandis que pour de ''grands vocabulaires, multilocuteur, parole continue '', la méthode analytique s'impose.

## II. Méthode globale

Dans cette approche, il s'agit de comparer globalement le mot, ''mot'' englobe ici aussi bien les mots au sens courant du terme que des expressions figées ou même des propositions simples et stéréotypées, aux différents mots de référence stockés dans un dictionnaire. Ce stockage peut s'effectuer sous différentes formes: sorties d'un banc de filtres, cepstres, FFT... parfois même à l'aide de traits [Haton 84] [Lockwood 84]. Le fonctionnement d'un tel système comprend deux phases: une phase d'apprentissage durant laquelle le locuteur prononce isolément, en général plusieurs fois, l'ensemble des mots du vocabulaire, suivie d'une phase de reconnaissance.

### 1) Mots isolés

Durant la phase de reconnaissance, le locuteur prononce un des mots du vocabulaire et le système compare ce mot avec l'ensemble des références stockées dans le dictionnaire. Le mot inconnu est étiqueté

comme étant le mot le plus proche selon un critère de distance. Pour ce faire, une méthode de programmation dynamique est fréquemment utilisée, son principal intérêt étant de permettre d'effectuer une normalisation temporelle (cf figure A.3).
Cette technique présente deux inconvénients majeurs:

- tout d'abord l'obligation de séparer les mots par un silence suffisant (au moins 300 millisecondes ), ce qui oblige le locuteur à s'exprimer d'une façon très contraignante car non naturelle,

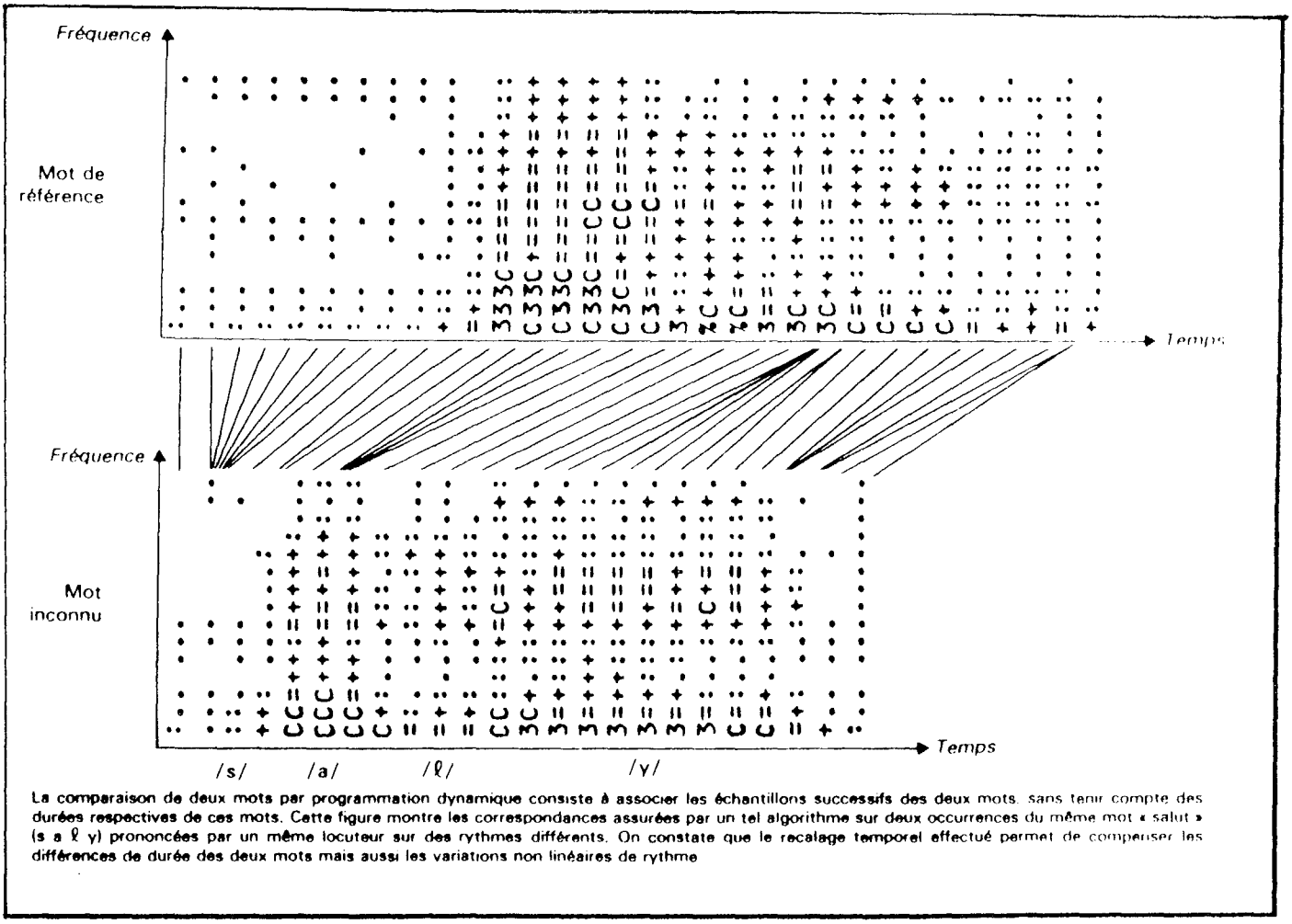
- ensuite, afin que le système offre des performances satisfaisantes (aussi bien en qualité qu'en temps de réponse ), il est nécessaire de limiter la taille du vocabulaire à quelques centaines de mots en monolocuteur, et à quelques dizaines en multilocuteur. Cette contrainte limite donc fortement les applications possibles.

### 2) Mots enchainés

Pour s'affranchir de la contrainte de séparer les mots par des silences suffisants, des chercheurs ont généralisé les algorithmes de programmation dynamique. Pour ce faire, on construit toutes les concaténations possibles de mots du vocabulaire pour les comparer à la forme inconnue prononcée. On peut citer l'algorithme en deux passes de Sakoe [Sakoe 78], Myers [Myers 81], ou plus récemment ceux de Bridle [Bridle 82] et Nakagawa [Nakagawa 83]. Mais un des inconvénients majeurs réside dans la limitation du vocabulaire réduit à quelques dizaines de mots. Pour traiter des applications nécessitant des vocabulaires de grande taille, et une élocution naturelle, il est clair que la méthode analytique s'impose.

## III. Méthode analytique

Dans cette approche, on cherche à décomposer une phrase en sons élémentaires tels que diphonèmes, syllabes, demi-syllabes, phonèmes... Cette méthode présente l'avantage de permettre la reconnaissance de parole continue pour de grands vocabulaires (plus de 1000 racines) car on ne mémorise qu'un nombre restreint d'éléments, indépendant de la taille du vocabulaire. Il se pose donc un double problème: segmenter et identifier une forme inconnue en ses éléments constituants.




La comparaison de deux mots par programmation dynamique consiste à associer les échantillons successifs des deux mots, sans tenir compte des durées respectives de ces mots. Cette figure montre les correspondances assurées par un tel algorithme sur deux occurrences du même mot « salut » (s a ℓ y) prononcées par un même locuteur sur des rythmes différents. On constate que le recalage temporel effectué permet de compenser les différences de durée des deux mots mais aussi les variations non linéaires de rythme

**Figure A.3** : Comparaison par programmation dynamique.

## 1) Rôle du décodage acoustico-phonétique dans le processus de reconnaissance

Le décodage acoustico-phonétique constitue la première étape dans un système de dialogue homme-machine. En effet, le role de cette étape consiste, à partir du signal sortant d'un micro, à fournir une description phonétique d'un énoncé. Dans ce but, ce module utilise des connaissances phonétiques, prosodiques, phonologiques et des algorithmes de traitement du signal. Rappelons les différents niveaux qui interagissent dans un système de compréhension de parole (cf. figure A.4).

### - le niveau acoustique

Son rôle consiste à détecter les début et fin des messages (séparation parole-non parole) et à extraire et calculer les paramètres utilisés par les niveaux suivants à l'aide de techniques de traitement de signal et d'heuristiques.

### - niveau prosodique

Il a pour tâche d'étudier les variations des paramètres prosodiques: mélodie, intensité, rythme. A l'aide de la courbe de ces variations, il est possible de définir des marqueurs prosodiques qui peuvent correspondre à des frontières de syntagmes. On peut aussi, de cette façon, déterminer si une phrase est affirmative ou interrogative.

### - le niveau phonétique et phonologique

Les niveaux phonétique et phonologique regroupent les informations qui seront exposées au chapitre 3 de cette même partie. Les descriptions phonétiques fournies par ces niveaux (en général sous la forme d'un treillis) sont ensuite traitées par les niveaux supérieurs qui utilisent des connaissances sémantiques, syntaxiques, grammaticales, lexicales et pragmatiques pour parvenir à une compréhension plus ou moins complète de l'énoncé prononcé.

### - le niveau lexical

Le lexique contient tous les mots du vocabulaire sous forme phonétique (les différentes prononciations possibles) à laquelle peut s'ajouter la forme orthographique. De plus, il peut contenir la classe grammaticale et les traits syntaxiques des mots et même des informations d'ordre sémantique.
On peut:

**Figure A.4** : Interactions entre les différents niveaux pour le dialogue ''homme-machine ''.

Processeurs — Destinataires ↓

Sources d'informations ⟶

| | PHONETIQUE | PROSODIE | LEXIQUE | SYNTAXICO-SEMANTIQUE | DIALOGUE |
|---|---|---|---|---|---|
| PHONETIQUE | hypothèses :<br>- phonétiques<br>- de segmentation<br>faits :<br>- variations de l'énergie (bandes)<br>- informations spectrales | | hypothèses :<br>- patrons phonétiques (mots)<br>- formes de référence lexicales | | |
| PROSODIE | hypothèses :<br>- signal segmenté<br>- durée vocalique moyenne | hypothèses :<br>prosodiques<br>faits :<br>- variations de l'énergie globale<br>- variations de F0 | | | |
| LEXIQUE | hypothèses :<br>phonétiques | hypothèses :<br>frontières de mots | hypothèses :<br>lexicales | hypothèses :<br>sous-lexique déterminé à l'aide de critères grammaticaux ou sémantiques | hypothèses :<br>sous-lexique (mots-clefs liés à l'application ou à l'étape du dialogue) |
| SYNT/SEM | | hypothèses :<br>- frontières syntagmatiques<br>- assertion/interrogation (au niveau de l'énoncé) | hypothèses :<br>lexicales | hypothèses :<br>syntaxico-sémantiques | hypothèses :<br>structure de l'énoncé |
| DIALOGUE | fait :<br>longueur de l'énoncé | | hypothèses :<br>mots-clefs liés à l'application | hypothèses :<br>sémantiques | hypothèses :<br>sur le dialogue en cours (historique) |

- soit inclure dans le lexique les descriptions des principales variantes d'un même mot, c'est à dire précompiler les variations phonétiques ou phonologiques dues aux différentes prononciations, et aux altérations morphologiques (pluriel, conjugaisons),
- soit générer à partir de règles ces différentes réalisations d'une même unité lexicale.

Le processeur lexical peut utiliser une stratégie du type ''meilleur d'abord'' ou par îlots de confiance [Mari 81]. Pour déterminer la suite de mots qui correspond le mieux à un treillis phonétique donné, des méthodes de programmation dynamique ont été testées [Charpillet 85].

- **le niveau syntaxique**

Il contient les informations relatives à l'ordre des mots, à la construction des phrases et à la grammaire du langage utilisé. Il associe une structure syntaxique au treillis lexical correspondant à un énoncé. Au cours de la reconnaissance d'une phrase, il peut également fournir au niveau inférieur des hypothèses sur la nature des mots non encore reconnus (contexte droit).

- **le niveau sémantique**

Ce module, à partir de la structure syntaxique d'un énoncé plus ou moins bien reconnu et des mots qui le composent, construit une ou plusieurs interprétations sémantiques de l'ensemble d'un énoncé. Au cours du processus de la reconnaissance, il produit également des hypothèses syntaxiques et lexicales sur la portion de l'énoncé non encore analysé; ces informations peuvent être utiles aux niveaux inférieurs.

- **le niveau dialogue**

Ce niveau, à partir de une ou plusieurs interprétations sémantiques du dernier énoncé produit par le locuteur, de l'historique du dialogue et des connaissances liées à l'application, génère la réponse du système. Plus généralement, il gère un dialogue finalisé avec l'utilisateur, donc organise les informations fournies par le locuteur dans un historique du dialogue qu'il construit progressivement, raisonne sur ces informations afin de satisfaire une requête ou une commande de l'utilisateur.
A chaque étape du dialogue, ce module peut également fournir, après avoir généré la réponse du système, des hypothèses sur la structure syntaxique et le contenu sémantique du prochain énoncé du locuteur; ces hypothèses peuvent guider utilement les modules de niveaux inférieurs dans l'analyse de cet énoncé.

Etant donné que le décodage acoustico-phonétique est la première ètape du processus de compréhension, ses erreurs éventuelles vont fortement pénaliser les niveaux supérieurs. De plus, si le décodage acoustico-phonétique n'obtient pas des performances suffisantes, quelles que soient les techniques utilisées par les niveaux supérieurs, la machine ne pourra pas comprendre la phrase prononcée. On peut raisonnablement penser qu'un taux d'erreur supérieur à 25% pour le niveau phonétique (dans un contexte multilocuteur, parole continue, vocabulaire de 1000 racines et grammaire peu restrictive) serait catastrophique pour l'ensemble du processus de reconnaissance. Il faut donc tendre vers un taux de reconnaissance phonétique d'au moins 80%.

## 2) Difficultés

La reconnaissance de la parole continue présente de multiples problèmes tels que la masse d'informations à traiter (que l'on tente de réduire par l'utilisation de techniques de traitement du signal vocal: FFT, LPC, etc...), la variabilité inter et intra-locuteur, le choix des paramètres à extraire du signal acoustique. Le problème majeur est donc un problème de réduction d'informations. En effet, on passe d'un débit de 100.000 bits par seconde en entrée (signal acoustique) à un débit de quelques dizaines de bits par seconde en sortie (contenu sémantique).

### a) Choix des paramètres acoustiques

Une des principales difficultés réside dans le fait qu'on ne sait pas sur quels paramètres l'homme s'appuie pour reconnaître la parole. En effet, les travaux sur la perception effectués sont encore fragmentaires et incertains. Des études ont montré que la position des trois premiers formants d'une voyelle constitue un critère déterminant dans la perception des voyelles. On peut donc affirmer que les valeurs des formants seront des caractéristiques importantes pour la reconnaissance des voyelles. Pour les fricatives, par exemple, c'est la limite inférieure du bruit de friction qui se trouve être un élément important. Mais cette limite est fonction du phonème qui suit (ou précède) la fricative. Il faut donc, non seulement déterminer les paramètres acoustiques pertinents, mais aussi leurs variations contextuelles.

### b) Variabilité intra-locuteur

La parole est un phénomène extrêmement variable. C'est ainsi qu'une phrase prononcée plusieurs fois par un même locuteur peut

présenter des variations notables (cf. figure A.5). Un autre problème est dû au fait qu'un même phonème peut prendre des formes très différentes en fonction du contexte qui l'environne (cf. figure A.6). De plus, un phonème caractérisé par plusieurs indices principaux peut, dans un même contexte, présenter seulement quelques uns de ces indices. Par exemple, le phonème /t/ peut présenter une barre d'explosion et des transitions typiques avec la voyelle suivante ou seulement l'un de ces indices caractéristiques (cf. figure A.7).

### c) Variabilité inter-locuteur

Il existe non seulement une variabilité intra-locuteur mais également une variabilité inter-locuteur encore plus importante. Par exemple, chaque voix est caractérisée par son timbre, sa hauteur,... Ces caractères découlent de différences morphologiques propres à chaque individu. De plus, suivant la région habitée, les différences s'accentuent, le débit, l'accent, la prononciation, l'intonation varient fortement. Par exemple, une des caractéristiques du sud de la France réside dans la prononciation des '' e '' muets. C'est grâce à ces caractéristiques qu'il est possible d'identifier une personne à sa voix. Bien sûr, dans notre système, il faudra s'affranchir de ces différences et trouver des invariants.

Tous les problèmes que nous venons d'évoquer nécessitent une meilleure connaissance du français parlé et des mécanismes de la communication orale. Cela implique une étude systématique et pluridisciplinaire de grandes bases de données de sons, telles que BDSONS constituée par le GRECO Communication Parlée.

## 3) Approche traditionnelle et limites

Dans l'approche analytique, les méthodes de reconnaissance employées jusqu'alors consistaient principalement à comparer le signal de parole à des formes de références. Le locuteur enregistrait des phonèmes pendant une phase d'apprentissage, ensuite, lors de la phase de reconnaissance, on comparait les phonèmes prononcés avec ces références. En reconnaissance monolocuteur, ces méthodes présentent l'inconvénient de nécessiter de nombreuses formes de références en raison de la grande variabilité intra-locuteur, et entraînent un apprentissage relativement long [Lazrek 83]. Un exemple permet d'illustrer ce point: le phonème /l/ ou /R/ doit être prononcé dans plusieurs contextes en raison de sa très grande variabilité. De plus, une telle approche ne peut différencier la fricative initiale des logatomes ''chi'' et ''sy'' (cf. figure A.8). Seule une approche contextuelle permettrait de résoudre le problème. Dans un cadre multilocuteur, ces méthodes sont difficilement applicables, étant donné le

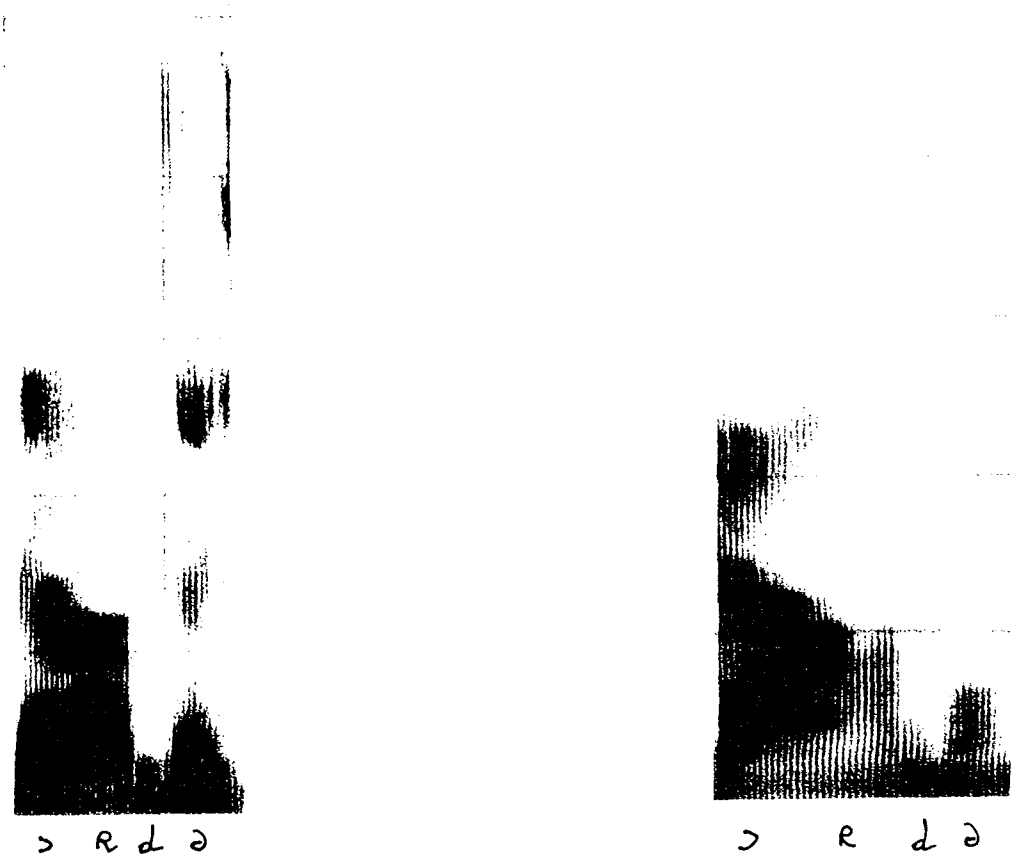



**Figure A.5** : Variation intra-locuteur pour le phonème /d/.

Noter la présence ou l'absence de la barre d'explosion

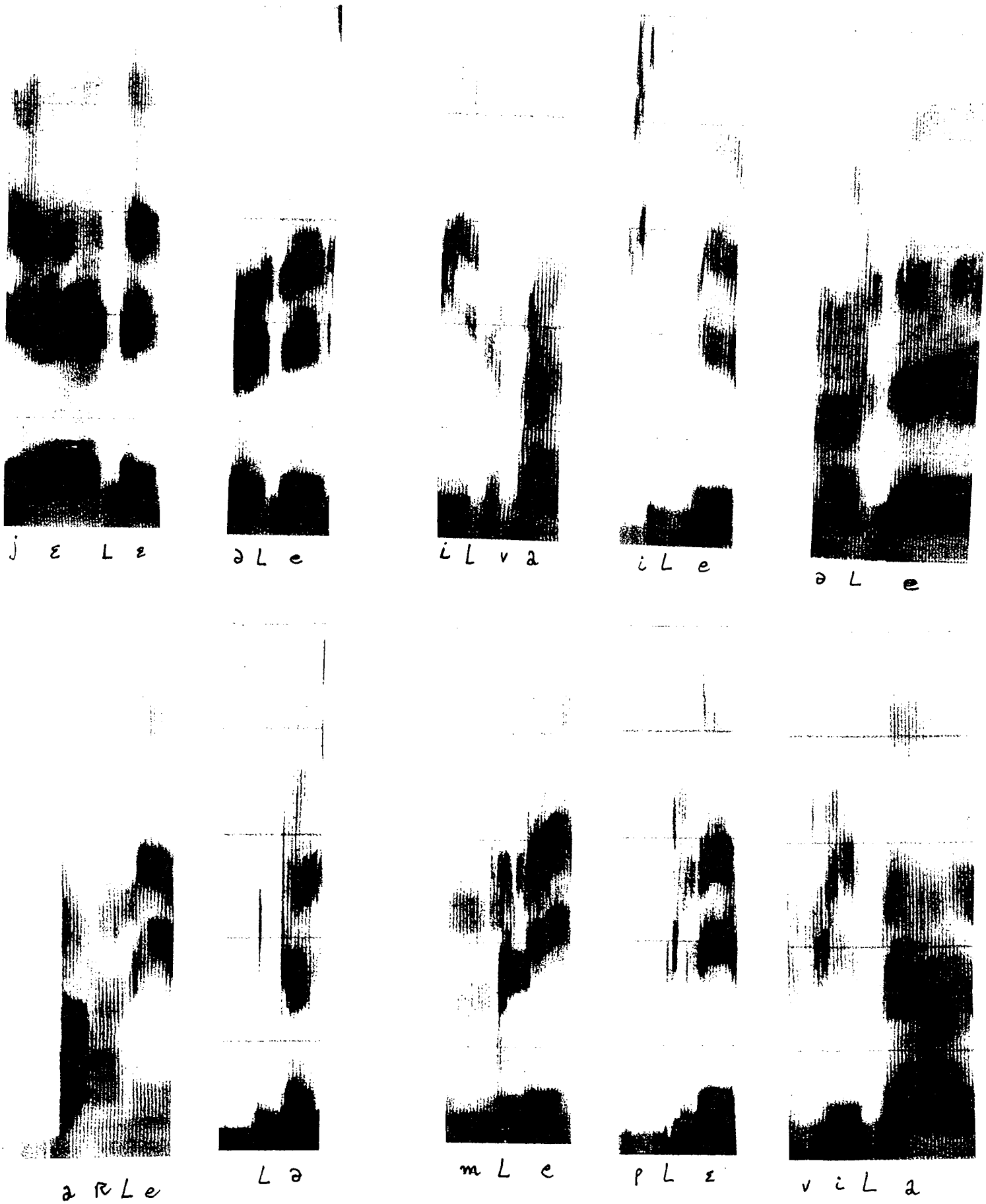



**Figure A.6** : Variation du phonème /l/ dans différents contextes.

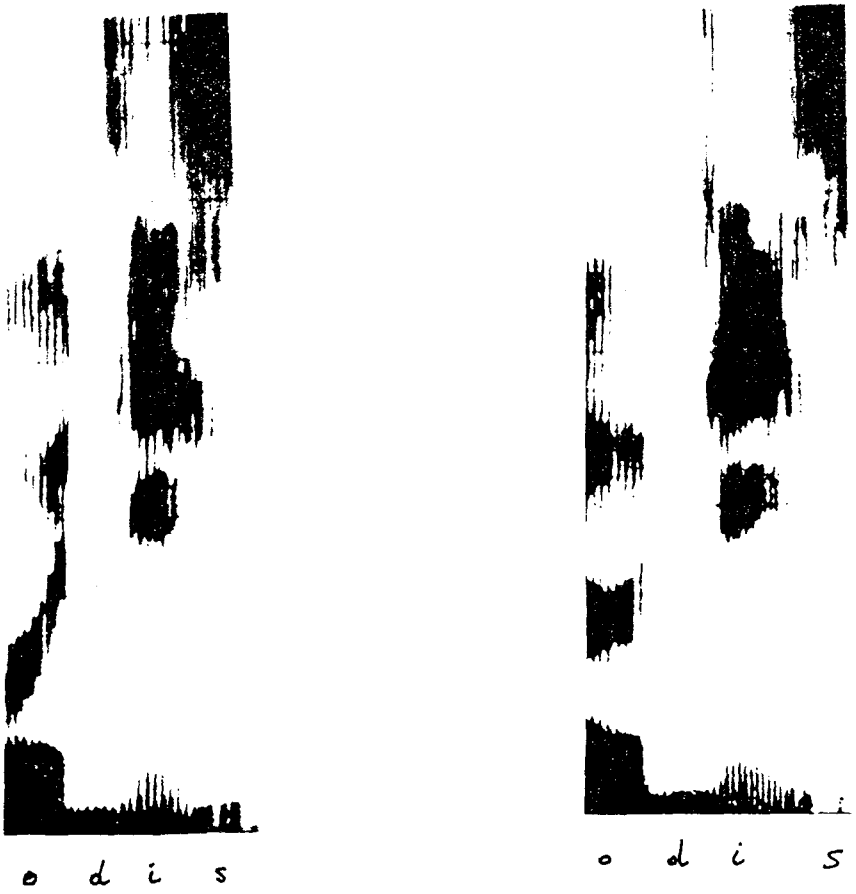



**Figure A.7** : Deux réalisations du phonème /d/ dans le logatome /odis/ :

Même phrase, même locuteur.

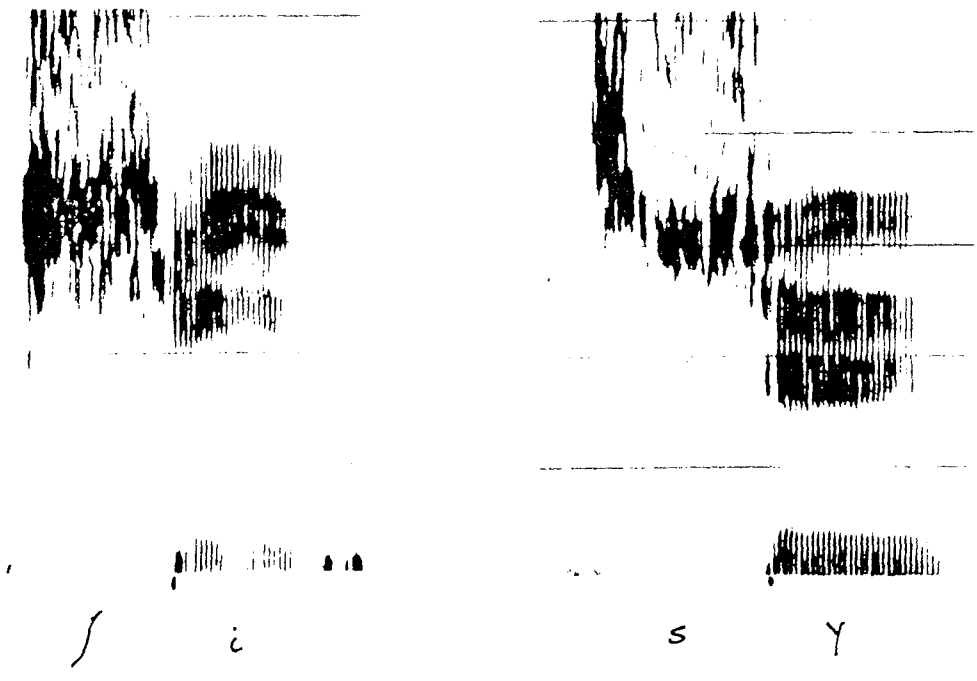



**Figure A.8** : Spectrogrammes des logatomes / i/ et /su/ :
On peut remarquer que la limite inférieure du bruit
est à la même fréquence dans les deux cas.

nombre prohibitif de références nécessaires. Pour ces multiples raisons, de nombreux chercheurs se sont orientés vers une approche intelligente mettant en oeuvre des raisonnements complexes s'appuyant sur des bases de connaissances diversifiées.

### 4) Approche intelligence artificielle

Une approche de type intelligence artificielle permet de pallier les difficultés que nous venons d'énumérer, en particulier d'éviter une explosion combinatoire des hypothèses à gérer par les différents niveaux du processus de reconnaissance. Certains chercheurs ont compilé l'ensemble des connaissances (depuis l'acoustique jusque la sémantique) dans un réseau qui contient les transcriptions phonétiques de toutes les phrases du langage et leurs variantes phonologiques. La reconnaissance de la parole se résume donc à la recherche d'un chemin dans un graphe: HARPY [Lowerre, 76]. Bien sûr, si cette solution est rapide, cette approche ne permet pas de traiter un langage quasi naturel, ni de changer facilement le langage de l'application. Une approche radicalement différente consiste à définir une structure dans laquelle chaque source de connaissances constitue un module indépendant qui peut communiquer avec les autres modules. Le système HEARSAY II [Lesser, 75] a été réalisé sur ce principe (''Blackboard'' cf.figure A.9). On peut également utiliser des langages orientés objets (LOO) pour représenter les connaissances du niveau acoustico-phonétique [Damestoy,85] [De Mori, 83]. Mais les systèmes experts semblent aussi constituer une approche très prometteuse.

#### Approche système expert

Comme nous l'avons déjà vu, on ne sait pas précisément quels paramètres l'homme utilise pour reconnaître la parole. Nous avons tous acquis cette faculté de façon quasi inconsciente dans notre jeune âge. Or, il existe des experts phonéticiens qui peuvent lire certaines représentations visuelles de la parole (telles les spectrogrammes) et qui ont acquis cette expertise à l'âge adulte. Ils peuvent en partie expliquer leur raisonnement et leurs stratégies. Pour formaliser les connaissances ainsi obtenues, une approche de type système expert semble naturelle. Le décodage acoustico-phonétique est caractérisé par la présence de sources de connaissances multiples, éventuellement erronées et par une base de faits souvent entachés d'erreurs. Il en résulte un fort indéterminisme nécessitant la mise en oeuvre de structures de contrôle et de stratégies élaborées et robustes. Le formalisme des systèmes experts semble bien approprié pour aider à la formalisation des connaissances et des stratégies nécessaires. En séparant clairement le contrôle et les connaissances, il permet une représentation

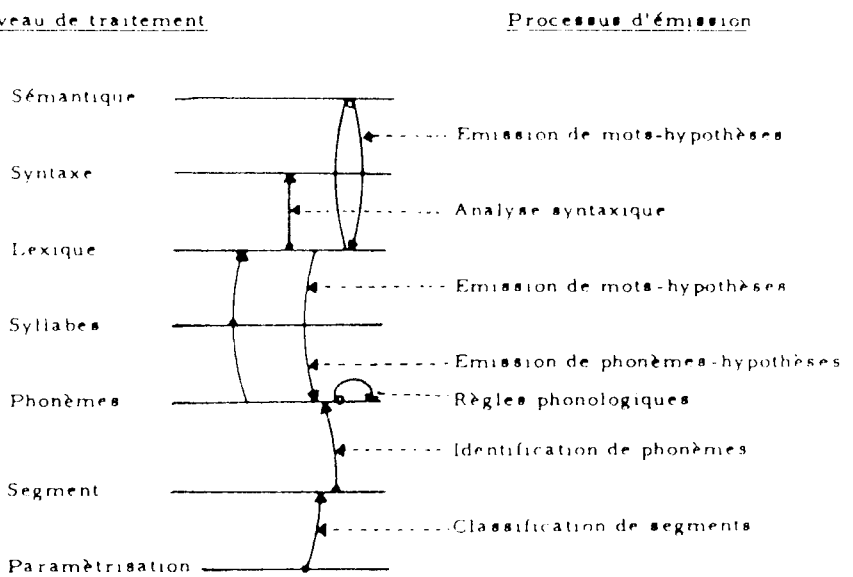



**Figure A.9** : Structure interne du blackboard dans HEARSAY II.

déclarative et modulaire qui est bien adaptée à l'acquisition incrémentale de connaissances dans des domaines mal formalisés au départ. Grâce à sa modularité, un tel système est relativement facile à modifier et à étendre, notamment en rajoutant des règles. On peut expérimenter diverses stratégies de contrôle sans avoir à changer tout le reste du système. Le formalisme déclaratif des règles de production permet à des experts humains même non informaticiens d'exprimer leurs connaissances petit à petit, de manière claire et naturelle. Parmi les travaux sur ce sujet, citons Zue [Zue, 83], Memmi [Memmi, 84], Johannsen [Johannsen, 83], Jonhson [Jonhson, 83], Gillet [Gillet, 83], De Mori [De Mori, 83],... Plus précisément, l'équipe du CNET à Lannion intègre dans le système SERAC (Système Expert de Reconnaissance ACoustico-phonétique) des techniques d'intelligence artificielle au niveau acoustico-phonétique. Le but poursuivi est la possibilité, non seulement d'exprimer plus facilement les connaissances, mais aussi de faire évoluer aisément ces connaissances sans un travail fastidieux de réécriture ou de modification de programmes. Le LIMSI construit un système expert en prolog SONEX pour lire des sonogrammes de parole en modélisant les connaissances d'une phonéticienne M.Eskenazi. La structure de contrôle est un moteur d'inférences général à chaînage avant avec variables et négation. Au MIT, Victor Zue travaille sur l'amélioration du décodage acoustico-phonétique en intégrant des connaissances expertes.[Zue 82]

**Notre Approche**

Au CRIN, nous avons constitué une équipe pluridisciplinaire (informaticien, phonéticien, linguiste) pour améliorer le décodage acoustico-phonétique. Dans cette équipe, c'est François Lonchamp, de l'Institut de Phonétique de Nancy, qui est l'expert phonéticien. Il est capable, à partir d'une représentation visuelle du signal acoustique de parole (spectrogramme) de décoder des phrases prononcées de manière naturelle avec un taux de reconnaissance nettement supérieur à celui des algorithmes actuellement disponibles. Notre travail a consisté tout d'abord à acquérir et formaliser l'expertise de François Lonchamp et ensuite à l'intégrer dans le système APHODEX (Acoustic Phonetic Decoding Expert). Le but essentiel de ce système expert est d'aider à acquérir et à formaliser un ensemble de connaissances phonétiques sur le français de façon à exploiter au mieux ces connaissances pour améliorer les systèmes de décodage acoustico-phonétique.

# CHAPITRE 3

# PHONETIQUE

## I. Introduction

Nous allons tout d'abord présenter dans ce chapitre l'appareil phonatoire. En effet, des connaissances articulatoires permettent de comprendre et d'expliquer les coarticulations et même de prédire les transitions entre phonèmes. Ainsi, par exemple, connaître le lieu d'articulation d'un /t/ et d'un /a/ permet de comprendre les transitions d'un /ta/. Un phonéticien peut donc prédire le spectre et les transitions de deux phonèmes en contact, même s'il n'a jamais observé auparavant cette séquence. Après une courte description des différents articulateurs intervenant dans la production de la parole, nous décrirons les différents phonèmes du français. Pour illustrer cet exposé, nous utiliserons des représentations de type spectrographique dont nous donnerons un bref aperçu.

## II. L'appareil phonatoire

On peut distinguer trois parties dans l'appareil phonatoire (cf figure A.10):

- l'appareil respiratoire qui se comporte comme un générateur d'air,
- le larynx permettant la formation de l'onde glottale pour les sons périodiques,
- les cavités supra-glottiques qui jouent le rôle de résonateurs et où se produisent également la plupart des bruits apériodiques utilisés dans la parole.

### 1) L'appareil respiratoire

Il comprend les poumons et la trachée. Les poumons ont pour fonction de créer une pression sub-glottique dont la valeur peut varier de 0.6 à 1.1 cm de mercure; la trachée relie les poumons au larynx [Boe 78].

### 2) Le larynx

Il est formé de cartilages: le cricoide, le thyroide, deux arythénoides et l'épiglotte, réunis entre eux par des liens fibro-élastiques. Les diverses pièces de cette charpente sont reliées entre elles par de petits muscles. Les plus importants sont les deux thyroarythénoidiens inférieurs qui, bordés de deux ligaments, forment ce qu'on appelle les cordes vocales. La glotte est l'espace compris entre les cordes vocales.

#### Vibration des cordes vocales

Pour expliquer la vibration des cordes vocales, nous allons exposer la théorie myo-élastique [Van Den Berg 70]. Les cordes vocales vibrent sous l'effet du passage de l'air à travers la glotte (cf figure A.11). Au départ de la phonation, les cordes vocales sont légèrement écartées, le flux d'air venant des poumons provoque un effet d'aspiration (effet Bernouilli); les cordes vocales se rapprochent puis se touchent, le passage de l'air est interrompu, la pression sub-glottique étant supérieure à la pression supra-glottique, les cordes vocales s'écartent à nouveau. Le cycle peut alors recommencer, ce qui explique la périodicité de l'onde glottique (cf figure A.12).

### 3) Les cavités supra-glottiques

Elles comprennent le pharynx et la cavité buccale; pour certains sons il est fait usage des fosses nasales et de la cavité labiale, prolongeant la cavité buccale. La cavité buccale renferme la langue, organe musculaire qui, en se déplaçant, fait varier la forme de cette cavité et permet donc d'en modifier les fréquences de résonance. La partie supérieure de la bouche est formée par le palais qui permet de mettre en communication la cavité buccale et les fosses nasales.

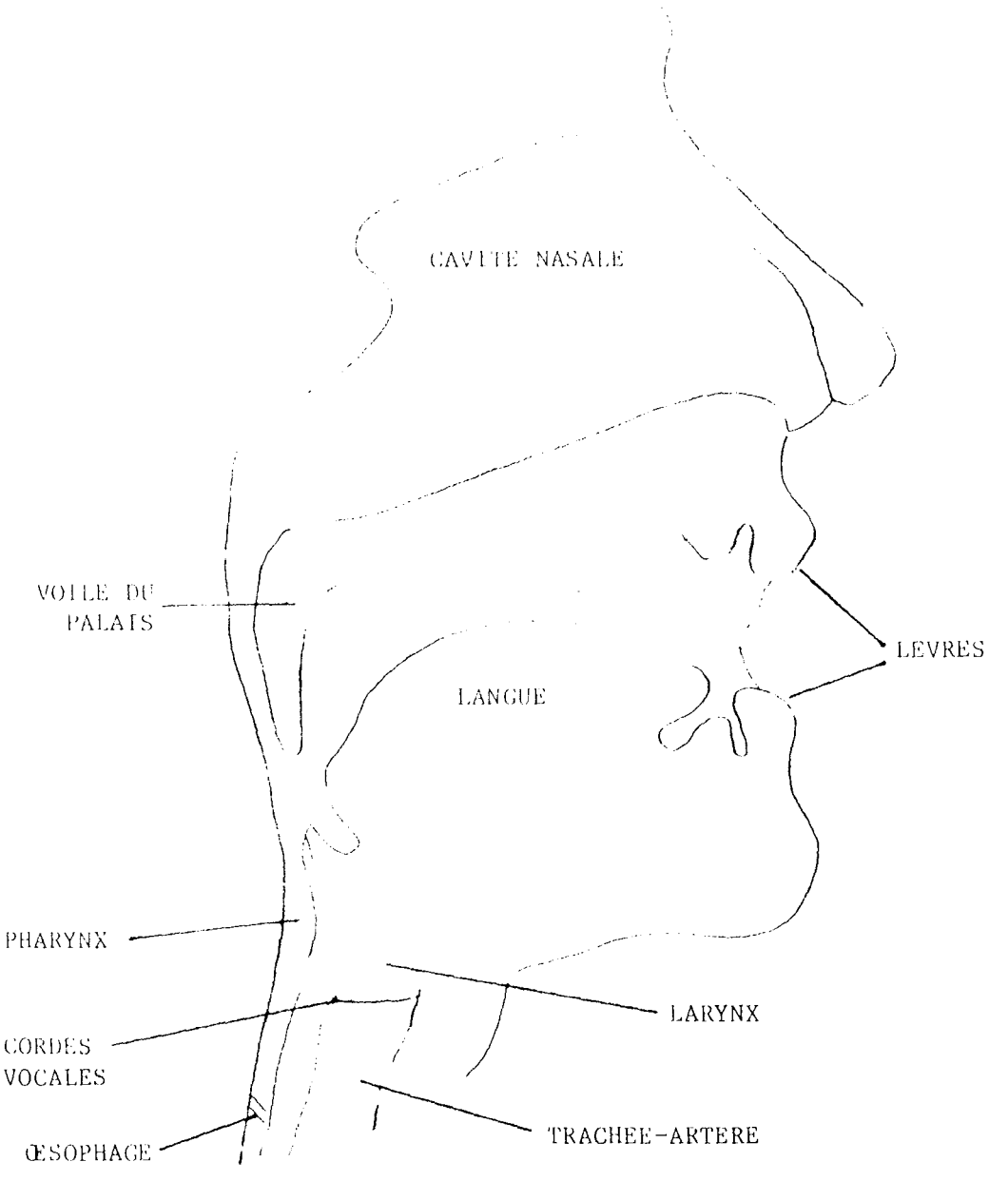



**Figure A.10** : L'appareil vocal (d'après Sundberg 80).

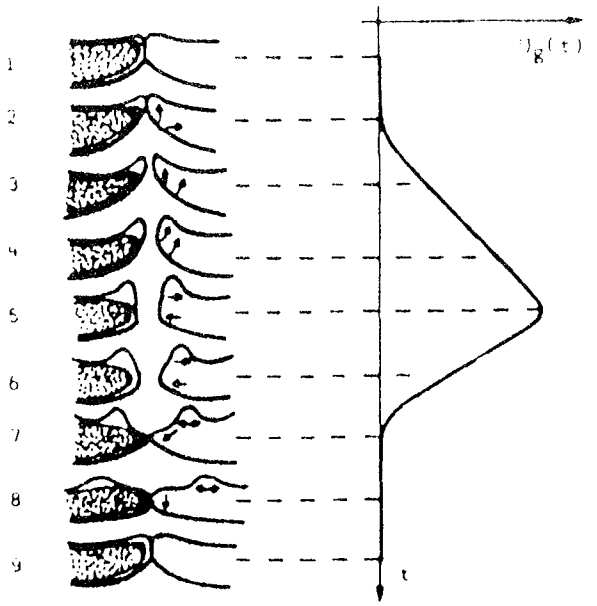


**Figure A.11** : Mouvement de cordes vocales.

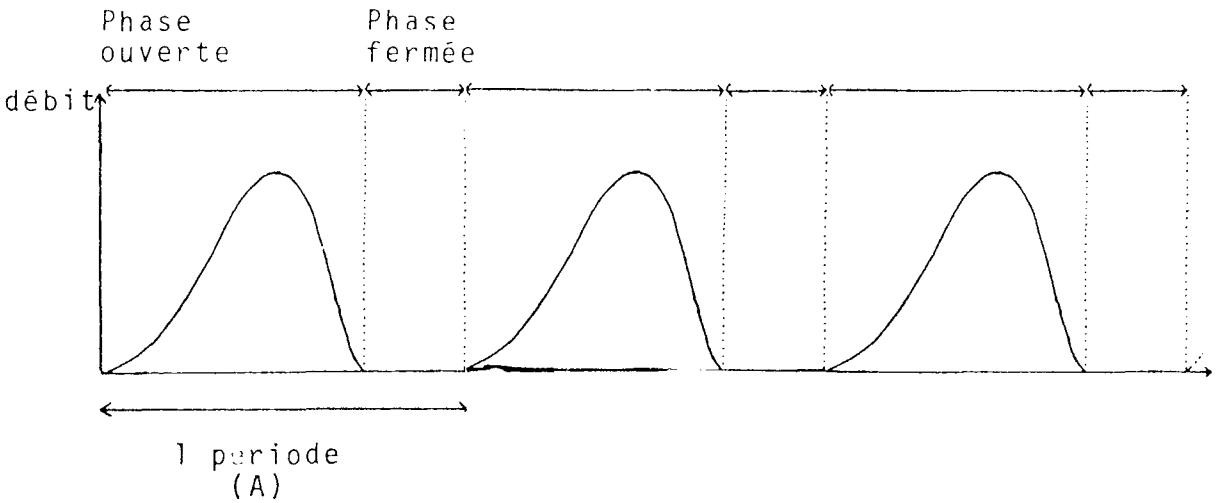
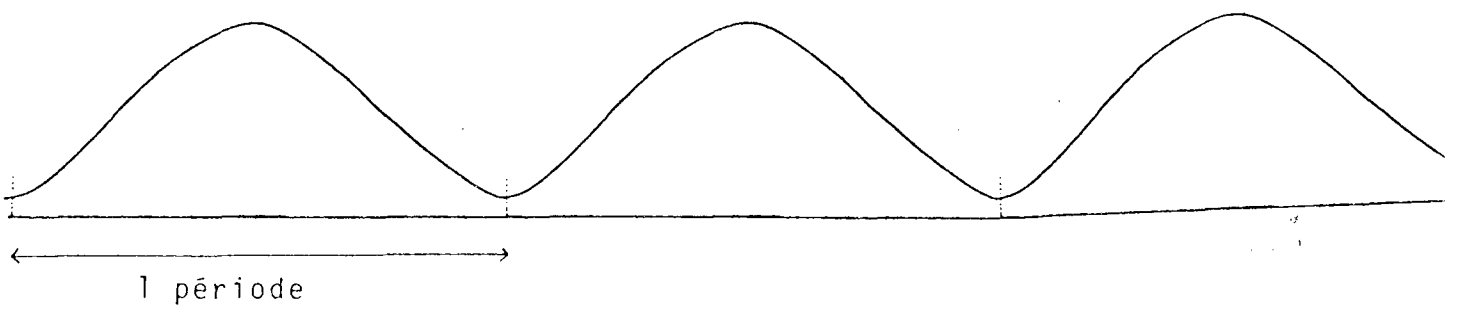



**Figure A.12** : L'onde de débit pour une voix intense (A) possède une phase où le débit est nul alors que pour une voix faible (B) le débit reste toujours positif.



## III. Une représentation acoustique: les spectrogrammes

Pour visualiser le signal de parole, on peut songer tout d'abord à représenter le signal numérisé provenant d'un micro dans un plan temps-intensité. Mais cette représentation présente l'inconvénient d'être dépendante de la phase des composantes du signal. En effet, sur la figure A.13, on peut voir deux représentations temporelles différentes d'un son résultant de la composition des trois mêmes fréquences mais déphasées. Or, l'oreille humaine étant peu sensible à la phase (les hauts-parleurs et les micros ne la conservent pas), il n'y a donc aucun intérêt à avoir deux représentations différentes pour des sons indiscernables à l'oreille.

De plus, on sait par des études de perception, que l'oreille est extrêmement sensible à des variations fréquentielles. Il est donc plus judicieux de disposer d'une représentation fréquence-temps-intensité: les spectrogrammes. A l'aide des spectrographes, on peut analyser en moins d'une minute quelques secondes de parole (2 à 3 s) pour des fréquences comprises entre 50 et 16000 Hz. On se limite en général à 5000 Hz pour les hautes fréquences, car on privilégie ainsi la définition des basses fréquences (zone de très grande sensibilité pour l'oreille et dans laquelle se situe la quasi totalité des indices perceptifs); on perd toutefois, au moins partiellement, les bruits de friction des /s/ féminins et le sommet des barres d'explosion des dentales /t/ ou /d/.

Le spectre chutant environ de -6 dB par octave, il est nécessaire de ''remonter'' (augmenter le gain) le spectre en fonction de la fréquence. On peut voir (ou plutôt ne rien voir!), sur la figure A.14, un spectrogramme réalisé sans préaccentuation sur lequel les fréquences supérieures à 1000 Hz sont très affaiblies et n'apparaissent presque plus. Notons enfin que le modèle analogique utilisé pour illustrer cette thèse possède une dynamique de 32 dB, ce qui est faible.

Pour mettre en évidence la structure formantique des sons, on utilise un filtre de 300 Hz de largeur de bande (filtrage large) et pour révéler la structure harmonique des sons voisés, on emploie un filtre de 45 Hz de largeur de bande (cf figure A.15). Le temps de réponse d'un filtre est, en première approximation, inversement proportionnel à sa largeur de bande exprimée en Hertz. Pour un filtre de 300 Hz de large, le temps de réponse est largement inférieur à la durée de la période fondamentale pour une voix masculine, la représentaion spectrographique sera périodiquement modulée en niveaux de gris correspondant aux instants de forte et faible ènergie dans une période. Si le fondamental s'élève (dans le cas de voix enfantines ou féminines) ou si on utilise un filtre de 45 Hz, les raies alternées noires et blanches disparaissent des sons voisés.

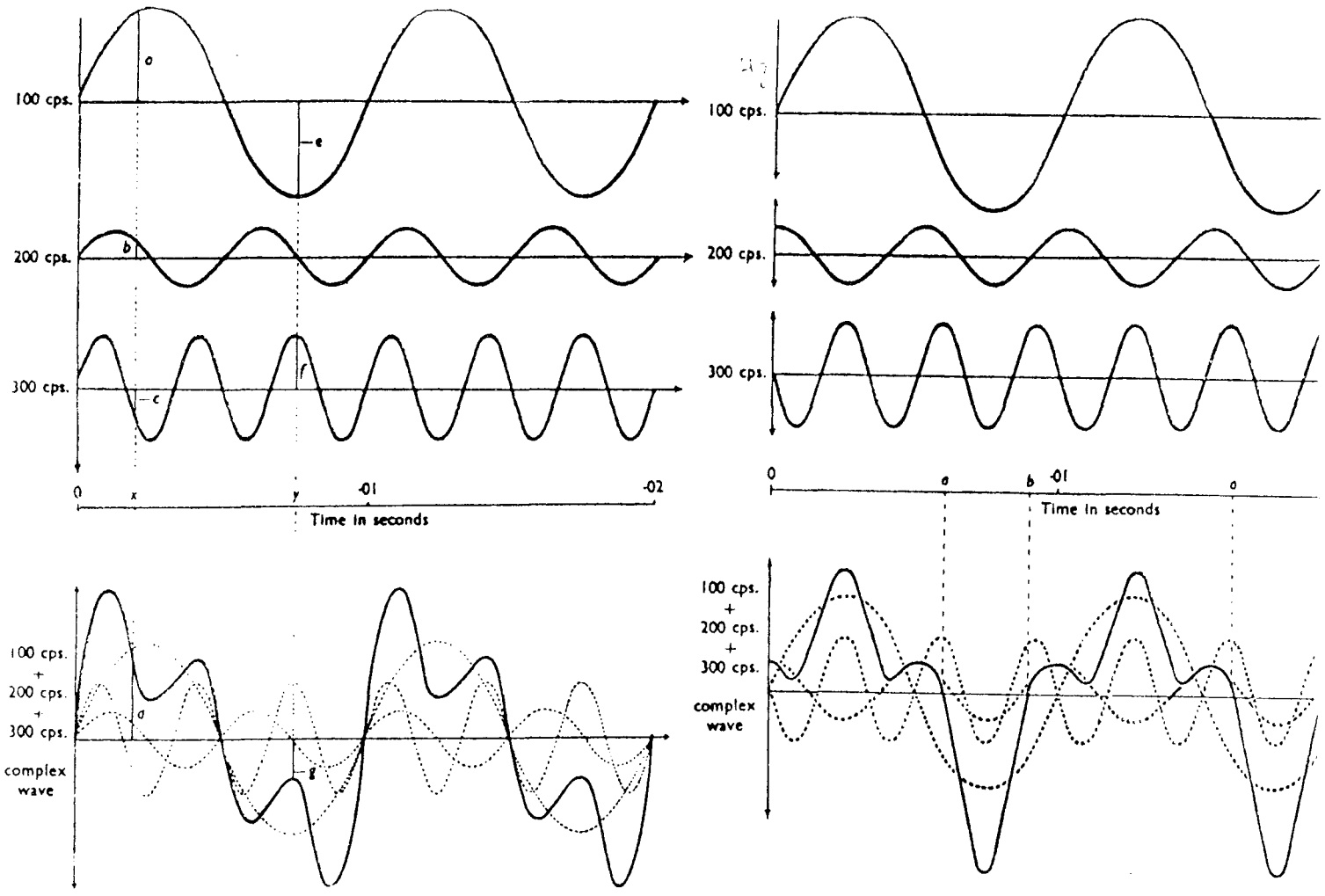
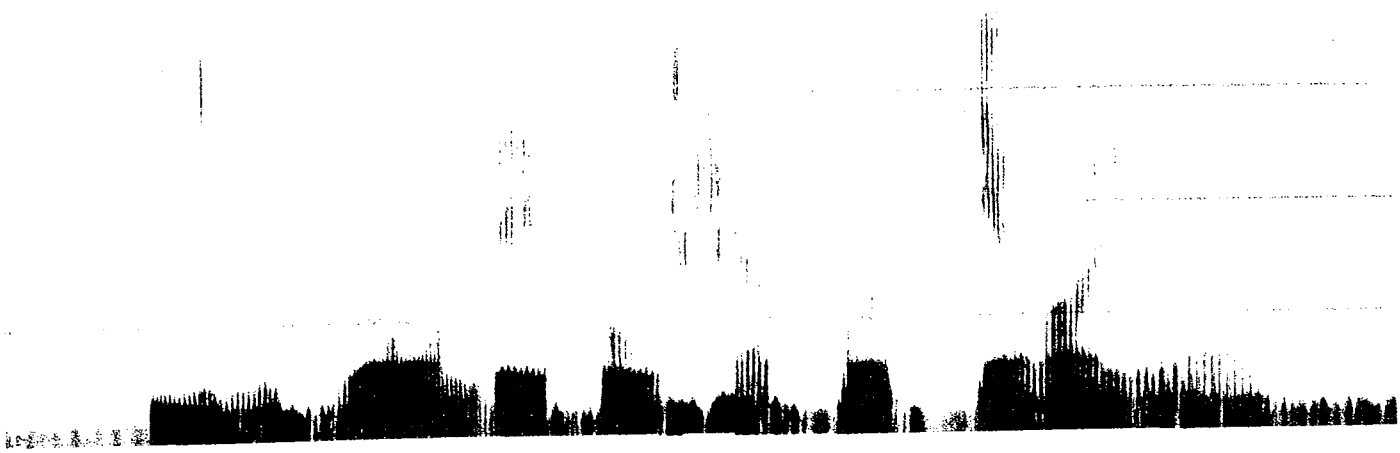



**Figure A.13** : Signaux obtenus en composant les trois mêmes fréquences 100, 200 et 300 Hz mais déphasées

N.B. ces deux signaux sont indiscernables à l'oreille.


**Figure A.14** : Spectrogrammes de la phrase : ils ont de beaux chapeaux tyroliens

avec préaccentuation de 6 dB.



sans prèaccentuation

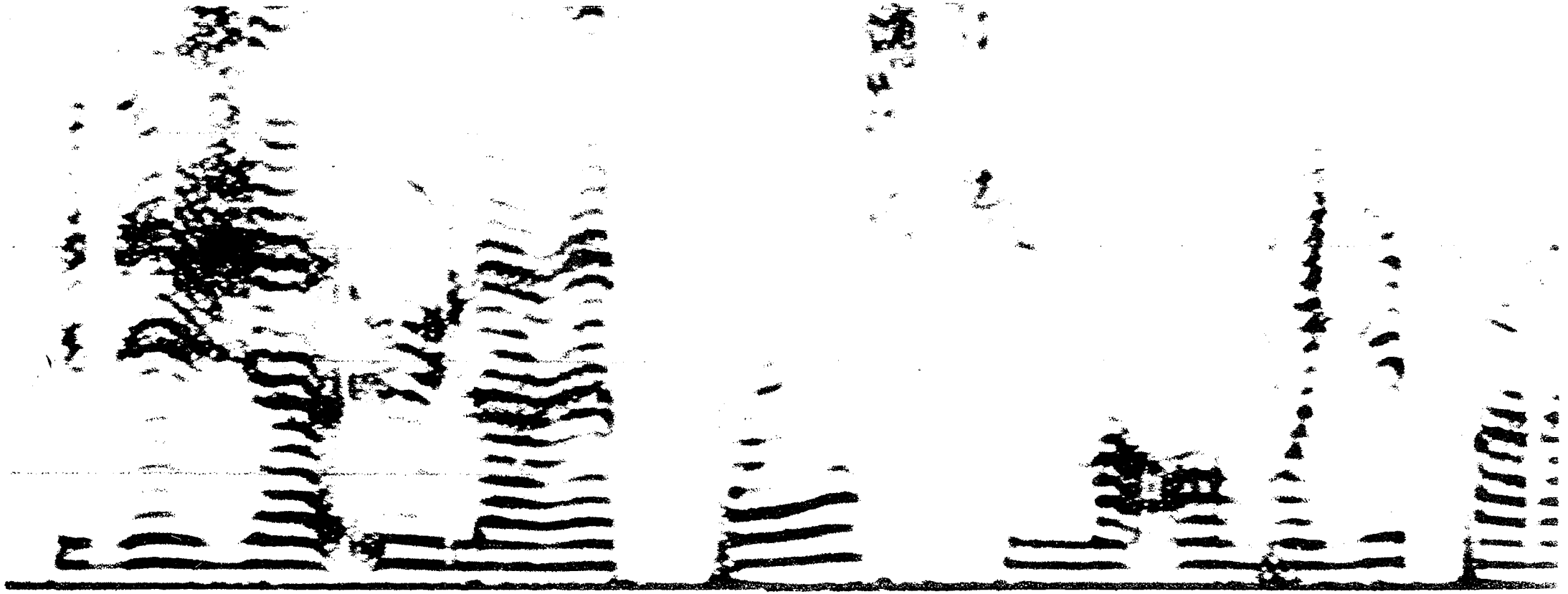


**Figure A.15 :** Spectrogramme réalisé avec un filtre de 45 Hz de largeur de bande.

## IV. Les phonèmes du français: production et analyse acoustique

### 1) Les voyelles

On distingue en français les voyelles orales /i/ /ɛ/ /ɔ/ /e/ /a/ /o/ /y/ /oe/ /u/ /o/ et les voyelles nasales /$\tilde{a}$/ /$\tilde{o}$/ /$\tilde{\varepsilon}$/ /$\tilde{oe}$/ .

#### a) Les voyelles orales

La présence de formants constitue la principale caractéristique des voyelles. Un formant est un maximum spectral apparaissant comme une bande noire, large d'environ 300 Hz sur le spectrogramme. La fréquence de ce maximum correspond à une fréquence de résonance du conduit vocal, c'est-à-dire approximativement à la fréquence d'un pôle dans la fonction de transfert. On a souvent l'habitude de représenter les voyelles dans le plan premier formant, deuxième formant. On distingue:

- les voyelles antérieures /i/ /e/ /y/ /o/ /ɛ/ /oe/ (langue massée en avant),
- la voyelle centrale /a/ (langue basse et étalée loin du palais),
- les voyelles postérieures /u/ /o/ /ɔ/ (langue massée postérieurement vers le voile ou vers le pharynx).

De plus, les voyelles /u/ /ɔ/ /oe/ sont labialisées (arrondissement des lèvres) ce qui produit une baisse des deuxième et troisième formants. Les formants étant donc la caractéristique perceptive essentielle des voyelles, il est intéressant de les mesurer. Des mesures ont été effectuées pour les logatomes /p + voyelle/ [Tigziri 85] (figure A.16). La figure A.17 présente les polygônes convexes englobant tous les points mesurés pour les sujets masculins et féminins. La figure A.18 montre pour chaque voyelle le vecteur reliant les valeurs des médianes formantiques des deux sexes. On peut expliquer ces différences formantiques par des raisons d'ordre physiologique. En particulier, le conduit vocal et le pharynx sont en moyenne plus longs chez l'homme que chez la femme.

#### - Coarticulation

Les fréquences formantiques varient fortement avec le contexte. Par exemple, dans le mot ''doute '', le deuxième formant du /u/ va être beaucoup plus élevé que dans le cas d'un /u/ isolé (cf figure A.19). En effet, la langue part d'une position antérieure (proche des dents) puis doit reculer rapidement jusqu'au voile, et enfin, elle doit être ramenée en

| | | Sujets masculins | | | | | Sujets féminins | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| voyelles | | F1 | F2 | F3 | F4 | | F1 | F2 | F3 | F4 |
| i | Med. | 308 | 2064 | 2976 | 3407 | Hz | 306 | 2456 | 3389 | 3966 |
| | σ | 34 | 134 | 147 | 208 | Hz | 42 | 111 | 68 | 169 |
| | Ki | 0.99 | 1.19 | 1.14 | 1.16 | % | | | | |
| e | Med | 365 | 1961 | 2644 | 3362 | | 417 | 2351 | 3128 | 4161 |
| | σ | 31 | 119 | 107 | 155 | | 31 | 52 | 115 | 121 |
| | Ki | 1.14 | 1.20 | 1.18 | 1.24 | | | | | |
| ɛ | Med | 530 | 1718 | 2558 | 3300 | | 660 | 2080 | 2954 | 4231 |
| | σ | 49 | 132 | 103 | 221 | | 46 | 108 | 156 | 210 |
| | Ki | 1.25 | 1.21 | 1.15 | 1.28 | | | | | |
| a | Med | 684 | 1256 | 2503 | 3262 | | 788 | 1503 | 2737 | 4143 |
| | σ | 47 | 32 | 131 | 155 | | 51 | 86 | 174 | 192 |
| | Ki | 1.15 | 1.20 | 1.09 | 1.27 | | | | | |
| ɔ | Med. | 531 | 998 | 2399 | 3278 | | 634 | 1180 | 2690 | 3960 |
| | σ | 39 | 60 | 116 | 155 | | 48 | 59 | 198 | 201 |
| | Ki | 1.19 | 1.18 | 1.12 | 1.21 | | | | | |
| o | Med. | 383 | 793 | 2283 | 3256 | | 461 | 855 | 2756 | 3865 |
| | σ | 22 | 63 | 126 | 161 | | 38 | 73 | 240 | 183 |
| | Ki | 1.20 | 1.08 | 1.21 | 1.17 | | | | | |
| u | Med. | 315 | 764 | 2027 | 3118 | | 311 | 804 | 2485 | 3550 |
| | σ | 43 | 59 | 136 | 172 | | 43 | 53 | 284 | 197 |
| | Ki | 0.99 | 1.05 | 1.23 | 1.14 | | | | | |
| y | Med. | 300 | 1750 | 2120 | 3145 | | 305 | 2046 | 2535 | 3570 |
| | σ | 37 | 121 | 182 | 141 | | 68 | 124 | 139 | 216 |
| | Ki | 1.02 | 1.17 | 1.20 | 1.14 | | | | | |
| ø | Med. | 381 | 1417 | 2235 | 3215 | | 469 | 1605 | 2581 | 4005 |
| | σ | 44 | 106 | 113 | 201 | | 36 | 90 | 148 | 168 |
| | Ki | 1.23 | 1.13 | 1.15 | 1.25 | | | | | |
| œ | Med. | 517 | 1391 | 2379 | 3353 | | 647 | 1690 | 2753 | 4038 |
| | σ | 42 | 94 | 91 | 149 | | 58 | 47 | 155 | 202 |
| | Ki | 1.25 | 1.21 | 1.16 | 1.20 | | | | | |

**Figure A.16** : Valeurs en Hz des quatre premiers formants des voyelles corpus de 10 hommes et 9 femmes

Med : médiane

σ : écart type

Ki : rapport formants féminins sur formants masculins (en %).

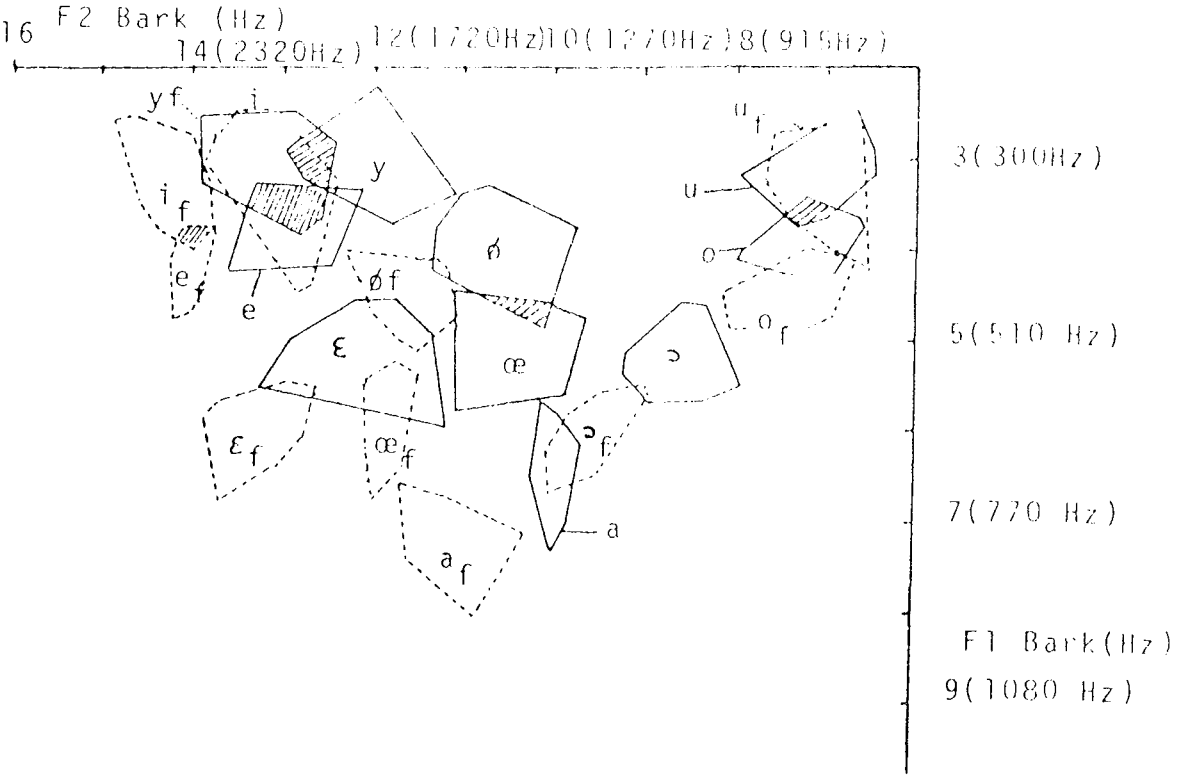



**Figure A.17** : Zone de dispersion des voyelles orales dans le plan F1/F2 (échelle Bark).

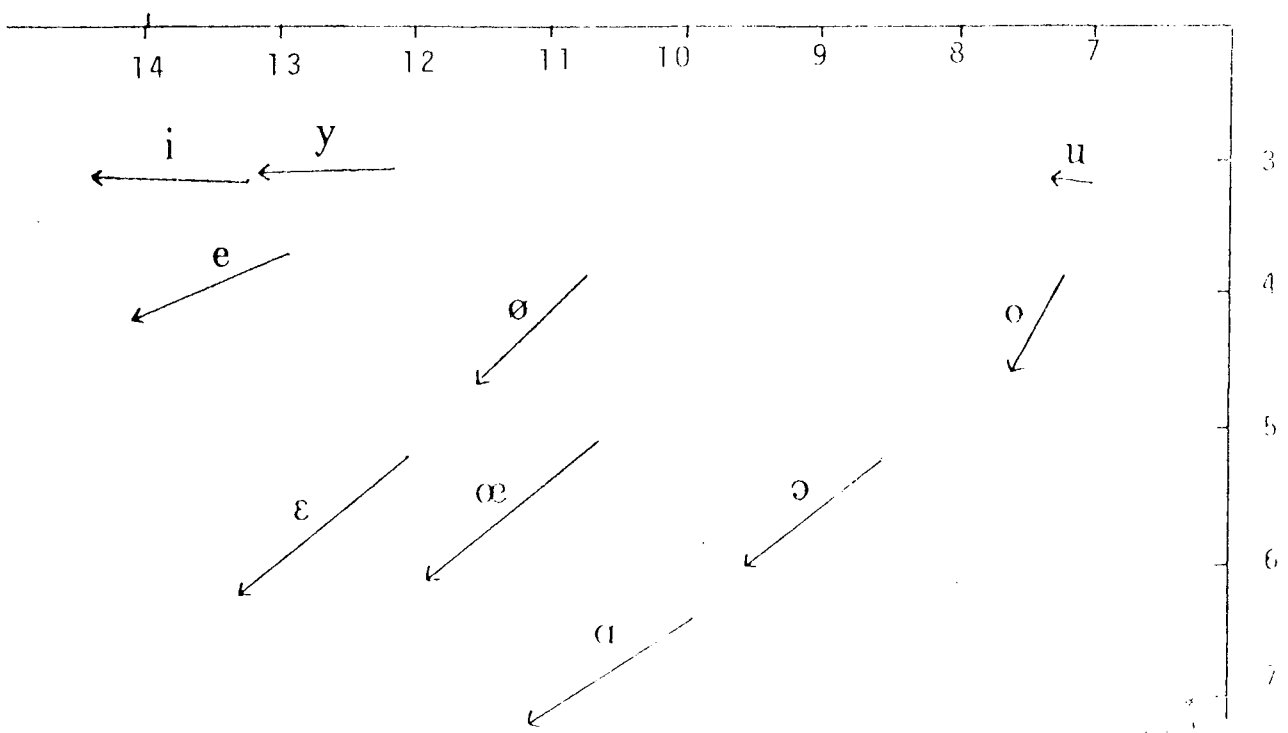



**Figure A.18** : Chaque vecteur relie les valeurs des médianes formantiques masculines et féminines (échelle Bark).

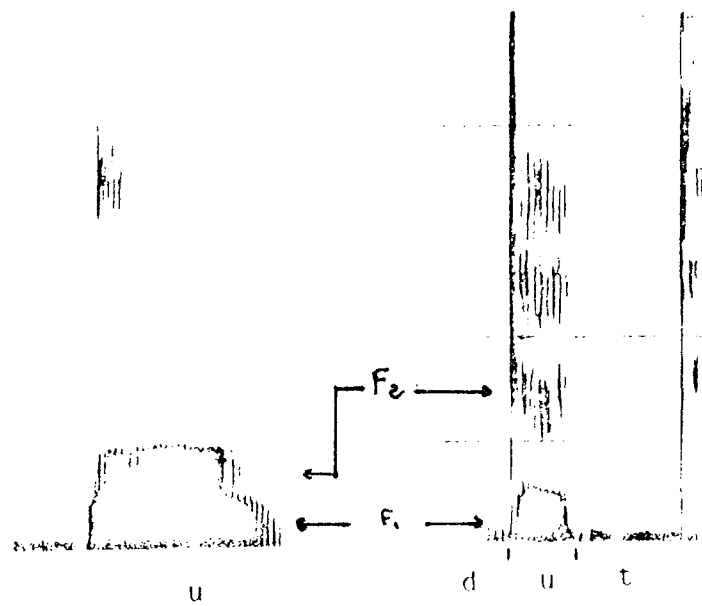



**Figure A.19 :** Spectrogramme de la voyelle /u/ hors-contexte et en contexte dental (/dut/) près de 35% de variation.

position antérieure. En débit rapide, des contraintes mécaniques (inertie) empêchent la langue d'atteindre une position vélaire. Le /u/ va donc posséder un deuxième formant entre 1200 et 1500 Hz, ce qui représente une variation, due à la coarticulation, de près de 35 %. Le contexte joue donc un rôle extrêmement important et on trouvera sur le tableau de la figure A.20 les variations fréquentielles des différents formants de sept voyelles dans cinq contextes différents.

### b) Les voyelles nasales

Le français comporte quatre voyelles nasales /ɑ̃/ /ɔ̃/ /ɛ̃/ /œ̃/. Elles sont obtenues par l'abaissement du voile du palais (mise en communication du conduit nasal et du conduit oral). Du fait de cette cavité supplémentaire (cavité nasale), il y a adjonction de paires formant-antiformant (pôles et zéros); de plus, on peut noter une modification de la fréquence des formants associés au seul conduit oral. De basse en haute fréquence, l'ordre des pôles est le suivant:

F1n, A1n, F'1, F'2, F2n, A2n, F'3.

- F1n: premier formant nasal situé entre 250 à 400 Hz,
- le premier formant oral F'1 est plus élevé que le F1 obtenu sans nasalisation, son intensité est en général plus faible et sa largeur de bande plus grande,
- F'2 est très voisin de F2,
- F'3 est plus élevé que F3.

Un anti-formant apparaît comme une forte vallée profonde en raison de la proximité d'un pôle. La figure A.21 contient, outre les spectrogrammes des quatre voyelles nasales du français, un tableau qui donne des indications quantitatives sur les voyelles nasales du français

## 2) Les consonnes

Parmi les consonnes, on distingue:

- les fricatives /f/ /s/ /ʃ/ /v/ /z/ /ʒ/ ,
- les occlusives /p/ /t/ /k/ /b/ /d/ /g/,
- les sonantes /l/ /R/ /m/ /n/ /ɥ/ /ɲ/ /j/ /w/ .

### a) Les occlusives

Une occlusive est caractérisée par un silence qui correspond à l'occlusion complète du conduit vocal, suivi d'une barre d'explosion

| Consonnes: | | labiales | dentales | vélaires | uvulaires | palatale |
|---|---|---|---|---|---|---|
| | | b | d | g | R | ʃ |
| Voyelles i | F1 | 330 Hz | 0 Hz | -25 Hz | 75 Hz | 25 Hz |
| | F2 | 2225 | 25 | 100 | -200 | -25 |
| | F3 | 3100 | -75 | -25 | -200 | -350 |
| ɛ | F1 | 520 | 0 | 0 | 100 | 0 |
| | F2 | 1830 | 25 | 50 | -150 | -125 |
| | F3 | 2585 | 100 | 25 | 0 | -50 |
| a | F1 | 615 | -50 | -50 | 25 | 0 |
| | F2 | 1350 | 200 | 250 | -125 | 100 |
| | F3 | 2500 | 125 | -250 | 75 | -125 |
| ɔ | F1 | 560 | -25 | -25 | 25 | 0 |
| | F2 | 1115 | 250 | 150 | -100 | 50 |
| | F3 | 2275 | 200 | 75 | 250 | -150 |
| u | F1 | 375 | -25 | -25 | -50 | -50 |
| | F2 | 875 | 300 | 200 | -50 | 175 |
| | F3 | 2560 | -50 | -50 | 150 | -200 |
| y | F1 | 285 | 50 | 0 | 75 | 25 |
| | F2 | 1750 | 0 | -50 | -175 | -50 |
| | F3 | 2375 | 200 | -200 | 100 | -125 |
| œ | F1 | 475 | -25 | -25 | 75 | 25 |
| | F2 | 1470 | 0 | 100 | -250 | 25 |
| | F3 | 2465 | 125 | -125 | 50 | -125 |

**Figure A.20** : Ecarts fréquentiels mesurés entre les fréquences formantiques de 7 voyelles dans les entourages de type dental vélaire uvulaire et palatal en fonction d'un entourage labial (contexte C1 V C1).

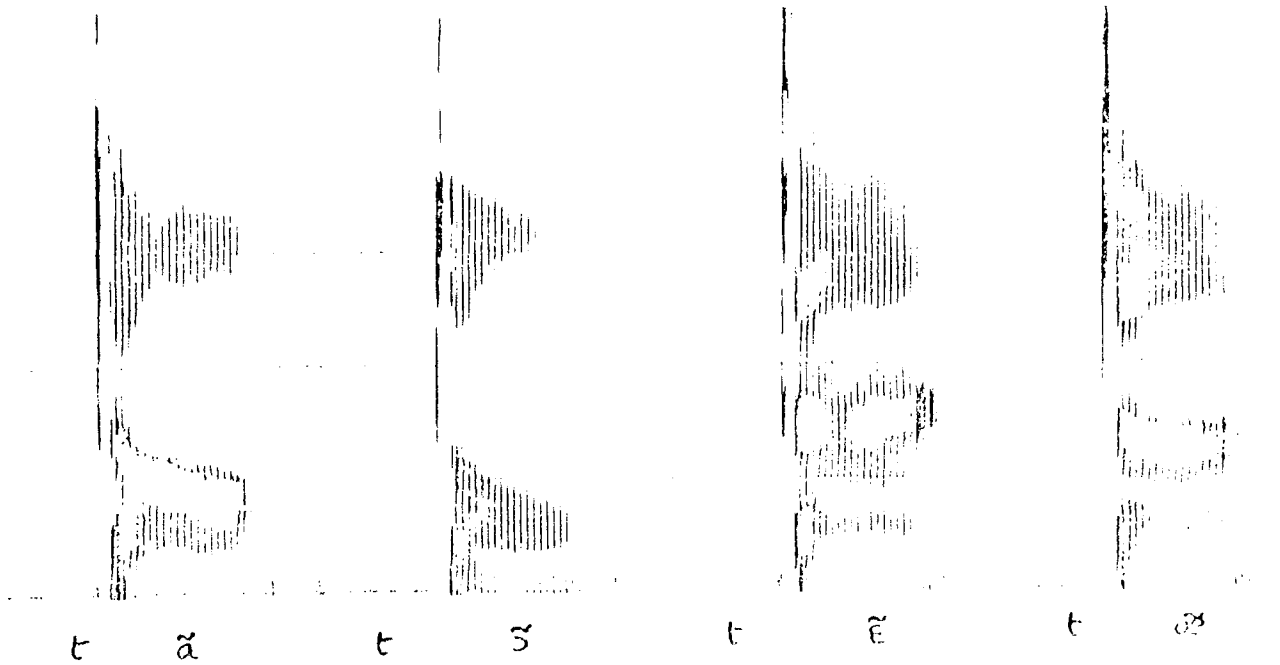


**Figure A.21** : spectrogrammes et fréquences formantiques des 4 voyelles nasales.

| | F1n | F'1 | F'2 | F2n | F'3 | A1n | A2n | |
|---|---|---|---|---|---|---|---|---|
| ɑ̃ | 300 | 700 | 950 | 1850 | 2300 | 600 | 2000 | Hz |
| ɔ̃ | 300 | 500 | 950 | 2000 | 2800 | 900 | 2350 | |
| ɛ̃ | 300 | 700 | 1500 | 2200 | 2700 | 600 | 2400 | |
| œ̃ | 300 | 750 | 1250 | 2000 | 2600 | 600 | 2100 | |

(burst). Parfois, lorsque l'occlusive est suivie d'un son /i/ /e/ /y/ /j/, on note un bruit de friction assez long. Pour une occlusive sonore, le silence n'est pas total car les cordes vocales continuent à vibrer pendant l'occlusion. La barre d'explosion est de courte durée et sa position et sa forme dépendent de l'occlusive elle même et du phonème qui suit. Pour reconnaître les occlusives, les transitions formantiques sont un indice important. On trouvera sur la figure A.22 la description de la barre d'explosion et des transitions en contexte vocalique.

**b) Les fricatives**

Elles sont caractérisées par un rétrécissement du passage de l'air, qui produit un bruit de friction. Pour les dentales /s/ /z/, la constriction, très étroite, est située derrière les incisives supérieures. Pour /ʃ/ et /ʒ/, la langue est légèrement relevée et les lèvres sont projetées en avant (labialisation). Quant aux fricatives labio-dentales /f/ /v/, elles sont produites par la mise en contact des incisives supérieures et de la lèvre inférieure. Le spectre d'une fricative /s/ ou /z/ se caractérise par la présence d'un bruit de friction continu entre 5 et 8 kHz (cf figure A.23). Cette limite inférieure est dépendante de la voyelle adjacente à la fricative, elle peut même descendre à 3 kHz pour des labiales (/u/ par exemple). Pour un /ʃ/ ou un /ʒ/, le bruit de friction est visible de 1.5 à 2.5 kHz (limite contextuelle) jusqu'à 7 kHz, le maximum d'énergie étant généralement situé vers 3 ou 4 kHz (cf figure A.24). Les fricatives /f/ et /v/ sont des sons beaucoup moins intenses; on peut noter une différence d'énergie de -20 à -35 dB par rapport à la voyelle adjacente.

**c) Les sonantes**

Les consonnes sonantes /l/ /R/ /m/ /n/ /ɥ/ /ɲ/ /j/ /w/ présentent une structure formantique lorsqu'elles ne sont pas adjacentes à une consonne sourde. Dans ce cas, elles se transforment en fricatives sourdes, avec un fort affaiblissement sous 1 KHz.
On distingue:

- les nasales /m/ /n/ /ɲ/,
- les liquides /l/ /R/,
- les semi-consonnes /j/ /ɥ/ /w/ .

| | Barre d'explosion | Transitions |
|---|---|---|
| pi | - < F2 (2KHz)<br>- < F3 (2.5 KHz)<br>bruit et friction faible | F3: plat ou +100 Hz |
| pe | - presque identique à pi | |
| pɛ | - de 1.5 à 4 KHz, renforcée sous F2, F3 et F4, ou entre F2-F3 et F3-F4<br>trace ≤ 5 KHz<br>trace vers F1 | F1: plat, +100 ou -100<br>F2: +200<br>F3: plat ou +100 |
| pa | - plus intense entre F2 et F3 faible vers F1, F3 et 4.5 KHz | F2: plat ou +100<br>F3: +100 |
| pɔ | - renforcée entre F3 et F4 (2.5 - 3.5 KHz)<br>- faible entre F2-F3 (2KHz)<br>- très faible vers F1 | |
| po | - renforcée entre F2 et F3 (1.7- 2.3 KHz), entre F3 et F4 (3.2 KHz)<br>- très faible vers F1, et > 5KHz | |
| pu | - entre F2 et F3, F3 et F4 (2 et 3 KHz), ou continue de 1.5 à 3.5 KHz | |
| py | - de F2 à F4 (1.5 à 4.5 KHz) renforcée aux (ou sous les) formants (1.8, 2.3, 3.5 KHz) | F2: +150 |
| pø | - de 1.5 à 4.5 KHz, renforcée vers les formants ou entre F2-F3 et F3-F4 (1.9, 3 KHz) | |
| pœ | - léger de 0.5 à 4.5 KHz renforcée entre F2-F3 et F3-F4 (1.9, 3 KHz)<br>- très faible ≈ F4 | |

**Figure A.22** : Caractéristiques acoustiques de /p/.


**Figure A.23** : Spectrogramme de /s/.


**Figure A.24** : Spectrogramme de /ʃ/.


**Figure A.25** : Spectrogramme de /m

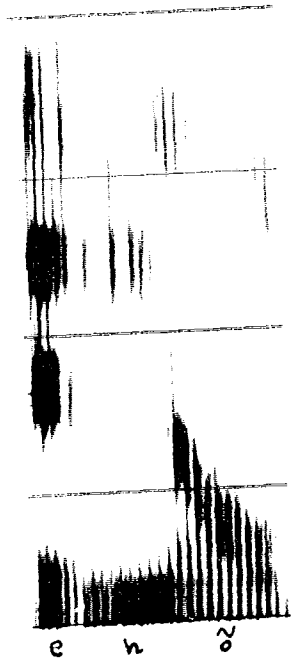



**Figure A.26** : Spectrogramme de /n/.

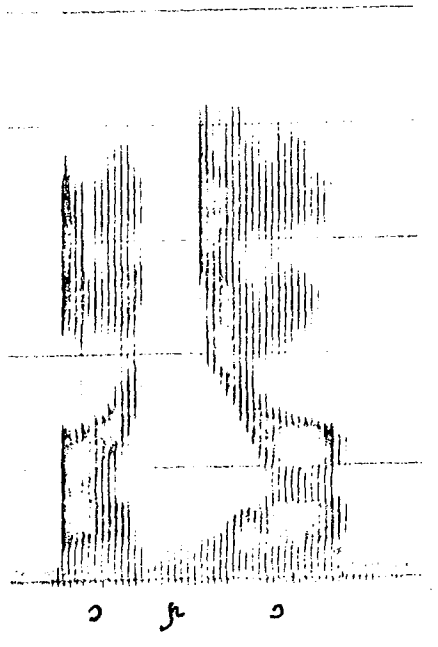



**Figure A.27** : Spectrogramme de /ʁ/.

- **les consonnes nasales**

Comme le voile du palais est abaissé pendant leur phonation, il y a couplage entre la cavité buccale et le conduit nasal, donc leur fonction de transfert possède des zéros. De plus, elles présentent une occlusion de la partie buccale:

- aux lèvres pour /m/,
- dans la zone dentale pour/n/,
- dans la zone palatale pour /ɲ/.

La consonne /m/ possède 4 formants entre 0 et 2.5 kHz: 0.3, 1, 1.3, 2 kHz et un zéro apparaît vers 2.4 kHz (cf figure A.25). Il existe souvent un formant aux alentours de 2400 Hz. La consonne /n/ présente également quatre formants entre 0 et 2.5 kHz: 0.3, 1, 1.5, 2.1 (cf figure A.26). Un zéro se trouve aux alentours de 2.8 kHz.
Sur un spectrogramme, on observe deux formants bien visibles à 300 et 2000 Hz pour un /ɲ/ (cf figure A.27). Mais la consonne /ɲ/ est souvent réalisée comme un /n/ suivi d'un /j/. De toute façon, l'élément terminal ressemble fortement à un /j/.

- **la consonne /l/**

Cette consonne est dite latérale car, si la langue est en contact avec la zone alvéo-dentale, l'air peut sortir des deux côtés du lieu d'articulation. On peut voir, sur les spectrogrammes, un premier formant vers 300 Hz et un deuxième très dépendant du contexte (1300-1900 Hz) et parfois un troisième (1600-2700 Hz), voire même un quatrième (2500-3200 Hz). Le formant 2 est élevé au contact d'une voyelle antérieure (1750-1900 Hz), il est plus bas avec une voyelle postérieure (1300-1600 Hz) comme on peut le constater sur la figure A.28.

- **la consonne /R/**

Elle présente une grande variabilité acoustique due à la coarticulation (cf figure A.29). Le formant 1 peut varier de 450 à 600 Hz, le formant 2 de 800 à 1600 Hz, le formant 3 de 2000 à 2800 Hz. Lorsque le /R/ est au contact d'une consonne sourde, il se dévoise et devient fricatif. La fonction de transfert possède alors des zéros (1.8 et 4 kHz).

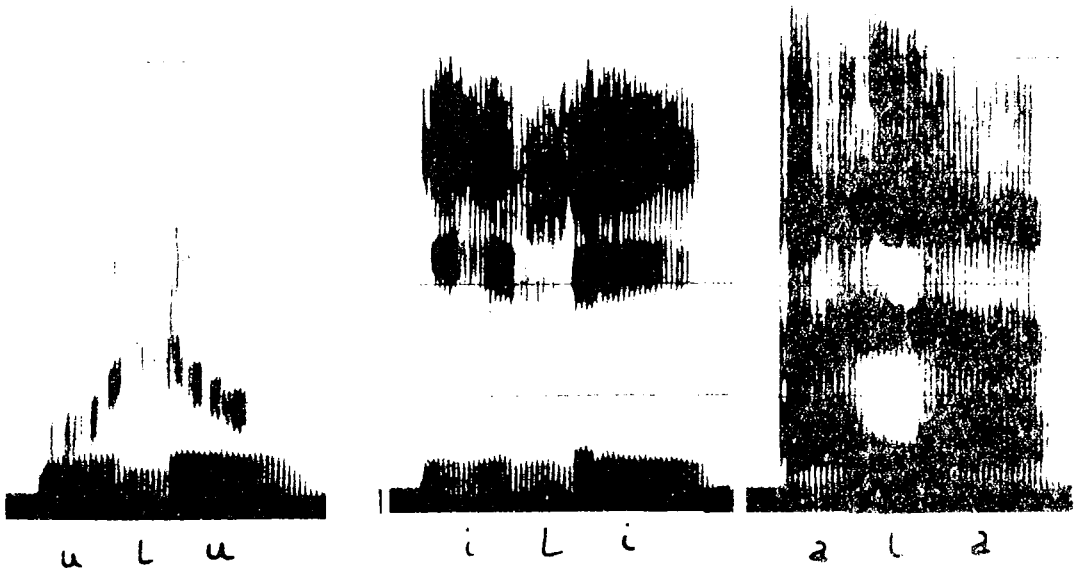



**Figure A.28** : Spectrogrammes de /l/ dans différents contextes.

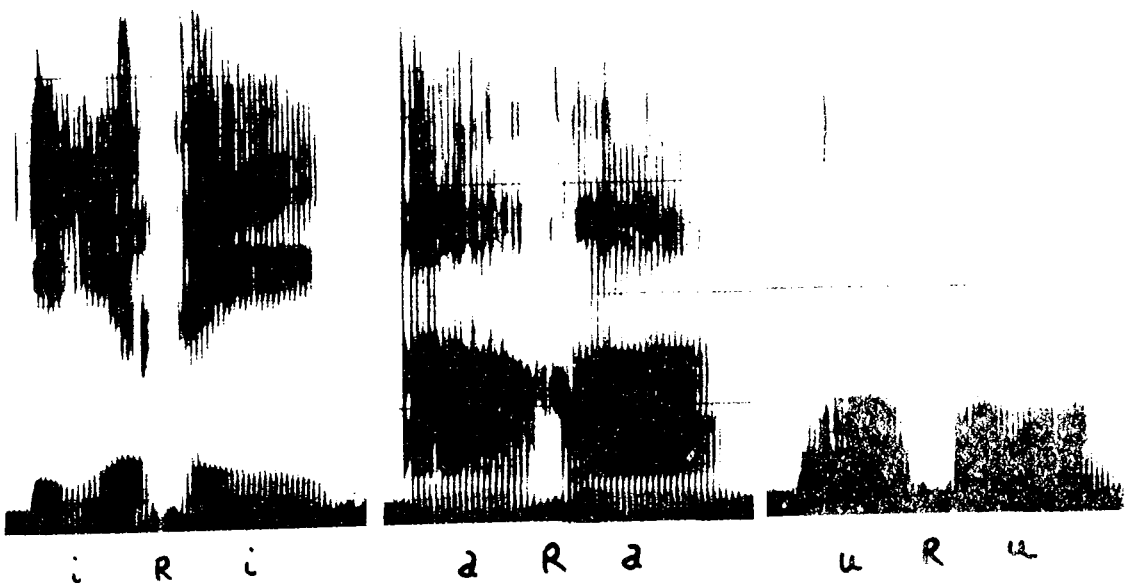



**Figure A.29** : Spectrogrammes de /R/ dans différents contextes.

- **les semi-consonnes**

**/j/**

Sa structure acoustique ressemble à celle d'un /i/; les valeurs des quatre premiers formants sont 0.3, 2.1, 3.0, 3.5 KHz.
Sur un spectrogramme (cf figure A.30), /j/ apparaît comme un /i/ dont le deuxième formant serait très faible.

**/w/**

Cette consonne est proche d'un /u/. Elle est généralement suivie d'un /a/, d'un /$\tilde{\varepsilon}$/ , d'un /ɔ/ ou d'un /i/. Le premier formant est présent vers 300 Hz, le deuxième vers 700; ce sont en général les seuls visibles (cf figure A.31).

**/ɥ/**

Cette consonne est proche d'un /y/. En français, cette consonne est toujours suivie d'un /i/. Le premier formant apparaît vers 300 Hz, le deuxième entre 1500 et 1700 Hz, les troisième et quatrième, souvent faibles, respectivement vers 2100 et 3200 Hz (cf figure A.32).


**Figure A.30** : Spectrogramme de /j/.

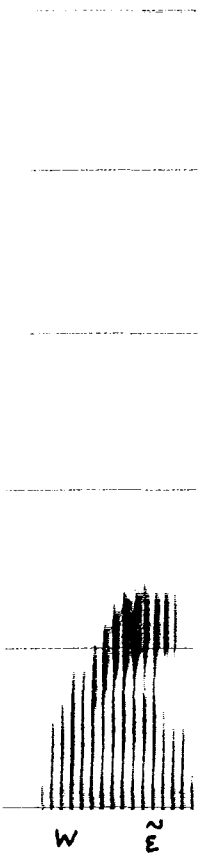



**Figure A.31** : Spectrogramme de /w/.

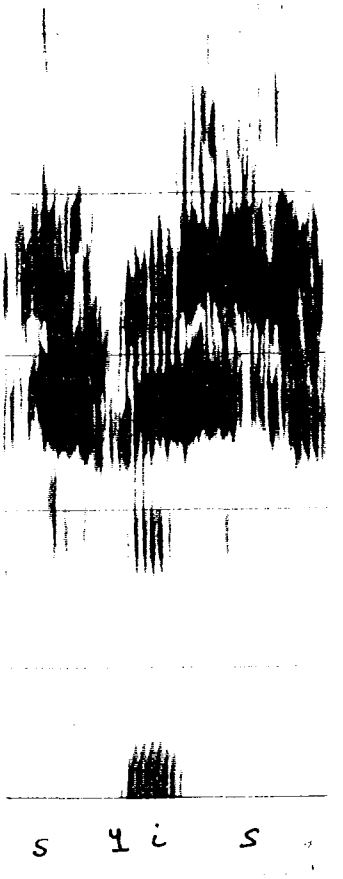



**Figure A.32** : Spectrogramme de /ɥ/.

# PARTIE B

# ACQUISITION DE L'EXPERTISE

# INTRODUCTION

Notre première tâche a consisté à collecter l'expertise de notre expert. La difficulté provient du fait que l'expertise comprend deux aspects : mise en oeuvre de raisonnements plus ou moins complexes d'une part et analyse visuelle d'autre part.

Nous avons recueilli tout d'abord des règles de phonétique générale et ensuite des règles contextuelles. Mais une importante partie de l'expertise réside dans la stratégie d'application de ces règles, stratégies parmi lesquelles on peut distinguer une approche globale et une analyse locale.

# CHAPITRE 1

# METHODOLOGIE D'ACQUISITION

# DE L'EXPERTISE

Nous allons étudier brièvement les différentes méthodes utilisées successivement ou concurremment pour acquérir l'expertise de François Lonchamp.

Dans un premier temps, l'expert, tel un professeur, décrivait les règles qui lui permettent d'interpréter un spectrogramme en illustrant son exposé d'exemples ponctuels. Cette méthode n'était pas sans présenter de nombreux inconvénients: les choses évidentes pour l'expert étaient passées sous silence, par exemple nous n'avons recueilli que des règles d'identification et pas de règles de segmentation. En outre, nous avons surtout obtenu des règles classiques de phonétique générale et très peu de règles contextuelles. Par ailleurs, l'expert définissait des formes de références plutôt que des règles permettant de connaître les différentes réalisations et coarticulations d'un phonème. Evidemment, aucune indication de stratégie n'était fournie. En appliquant les règles ainsi obtenues nous étions incapables de décoder un spectrogramme, ce qui montre qu'une infime partie de l'expertise avait été recueillie.

Par la suite, nous avons étudié l'activité de l'expert au cours du décodage d'un spectrogramme. Nous lui avons demandé de décoder des spectrogrammes en notre présence afin de préciser sa démarche par un jeu de questions-réponses (cf figures B.1 B.2 B.3). L'expérience a montré qu'une dizaine d'énoncés bien choisis (phrases phonétiquement équilibrées de Combescure) permettent d'obtenir des réalisations de tous les phonèmes du français dans un nombre important de classes de contextes. Nous avons donc enrichi très rapidement le volume des règles recueillies au cours de la première étape.

**Figure B.1.: Décodage d'un spectrogramme par l'expert**
Transcription d'une séance de décodage

Ca commence par une sonante.
Alors, cette sonante n'a pas de troisième formant visible donc on va admettre que f1 f2 sont confondus, comme il y a un formant intense vers 250, je choisis l.
On s'aperçoit d'ailleurs qu'il y a ensuite des transitions extrêmement rapides vers les formants, le formant trois de la voyelle qui suit.

La voyelle qui suit est du type, oh la la, alors voyelle labialisée peut être puisque le formant trois est bas, formant 1 relativement bas et formant 2 bon, /ø/ tout bêtement.
donc c'est /ø/ ou /oe/

Bon, ensuite, il y a une occlusive.
Y-a-il une sonante entre la voyelle et l'occlusive, non: alors cette occlusive est un /k/ puisqu'elle a une barre d'explosion tout à fait concentrée vers 900 Hz,
Les formants à gauche, on ne pourrait pas en dire grand chose, ils se dirigent vers le deuxième formant de la voyelle suivante, encore une transition parfaitement normale, donc c'est un /k/, avec beaucoup de bruit de friction... cinq centièmes de seconde de bruit de friction.
La voyelle qui suit est un /u/ sans doute, un /u/ ou un /o/ aussi,
Tiens, est-ce que ça ne serait pas un /on/.
C'est pas impossible que ce soit un /on/, mais...bon! voyelle postérieure, très labialisée, non nasale.

Ensuite une sonante, cette sonante pourrait être un /R/ dans la mesure où il y a pas mal d'énergie heu entre 0 et 1000 HZ, on voit mieux sans doute sur l'original que sur la photocopie et ensuite, il y aurait une transition super rapide vers une voyelle comme /y/ ou /i/, à mais d'accord, c'est un /j/ en fait, c'est pour ça que derrière où la pente est extrêmement rapide, on pensait à une voyelle comme /i/, /y/ mais alors on n'expliquerait pas pourquoi il n'y a pas de deuxième formant près de 2000 HZ, pourquoi il n'y a qu'un peu d'énergie vers 2500 et même pas beaucoup vers 250. Donc, une explication possible, c'est que c'est un /j/.

**Suite de la figure B.1**

Ensuite il y a un certain nombre de sons vocaliques, un très très gros paquet là, alors, combien est-ce qu'il peut y avoir de voyelles ou de sonantes, alors bon, je ne peux pas faire grand chose sauf reconnaitre que j'ai un formant 1 et un formant 2 très proches vers le milieu du paquet. Vers le milieu du paquet, bon alors je vais dire que c'est /a/, ensuite je semble me diriger de quelque chose comme /i/ vers /a/ en passant par tous les intermédiaires, alors je pense à quelque chose comme /e a/, /oe a/, parce que je peux difficilement avoir de nouveau un /j/ et un /i/, /j e a/, mais j'ai tous ces intermédiaires acoustiques donc il faut que je choisisse quelque chose, si je n'avais pas le /j/ devant, je dirais bien /i e a/, comme j'ai un /j/, je vais garder quelque chose comme /e a/.

Mais j'ai peut être encore un troisième, parce que c'est long tout ça, bon, après le /a/ je mets un /r/ parce qu'il me semble qu'il y a une espèce de transition bizarre, alors j'enlève le /l/ parce qu'a mon avis, le formant 1 serait beaucoup plus bas, je garde un /r/ donc /ari/ ou /are/, alors ça a l'intérêt de m'expliquer pourquoi il y a des transitions bizarres vers f3 f4, une sorte de non continuité des transitions.

Après le /r/, je place une voyelle, je dis que cette voyelle est du type /i/, donc ça ferait /e a r i/, espérant n'avoir rien oublié.

Ensuite, j'ai une occlusive et il y a un /an/.
Pourquoi /b/ ? Un /b/, parce qu'il n'y a pratiquement pas de barre d'explosion. Alors çà ne peut pas être un /v/, je ne crois pas parce qu'on redémarre un peu trop vite...etc...

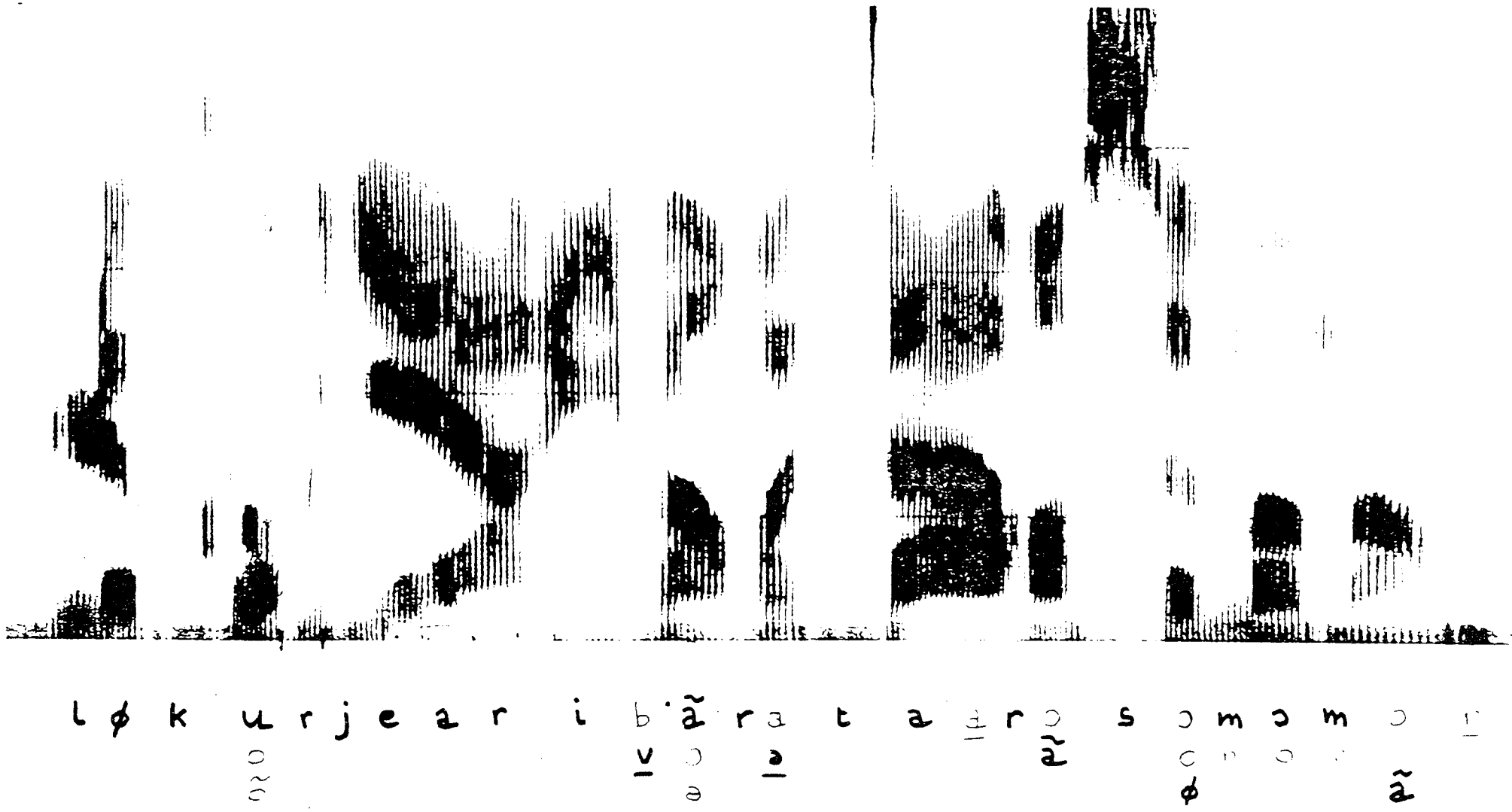



**Figure B.3** : Spectrogramme de la phrase correspondant à la transcription de la figure B.1

''Le courier arrive en retard en ce moment''


**Figure B.2** : Exemple de phrase décodée par l'expert
"Ce chasseur projette de partir d'ici"
En gras: les phonèmes exacts

Ensuite, nous avons mené une autre expérience pour élargir l'ensemble des règles obtenues au cours des étapes précédentes, de façon à saisir les différentes stratégies utilisées par l'expert et connaître avec précision ses performances. Nous lui avons fait décoder cinquante spectrogrammes (cf Partie C.1.III.). Chacun des cinq locuteurs, choisis parmi les membres du laboratoire, a enregistré dix phrases phonétiquement équilibrées prononcées de manière naturelle (rythme d'élocution de 14 phonèmes environ par seconde). Afin de ne pas perdre d'informations pertinentes de l'expert, nous avons décidé d'enregistrer ses commentaires pendant la phase du décodage des spectrogrammes et lors de l'analyse de ses erreurs, qu'il effectue après avoir pris connaissance de la transcription phonétique exacte de chacune des phrases.

## I. Difficultés de l'acquisition

Dans le recueil de l'expertise, nous avons rencontré de grandes difficultés dûes aux lacunes importantes du discours de l'expert et à la difficulté que présentait pour lui la verbalisation, l'explicitation de sa démarche.

### a) Au niveau du raisonnement

- Parfois l'expert se borne à énoncer ses conclusions sans les justifier. Par exemple: je vois un /l/. Ceci se produit lorsqu'il est sûr de son jugement, lorsqu'il n'hésite pas. Dans ce cas, on ne peut reconstituer la démarche: y a-t-il eu analyse, puis déduction, ou simplement comparaison de la forme observée avec une forme de référence? (cf les nasales /m/ /n/).
  A noter que s'il utilise des formes de références, ces formes sont contextuelles: elles englobent les transitions et tiennent compte du contexte; à la limite, elles englobent peut-être également le contexte.
- Souvent, il n'explicite qu'une partie du raisonnement, l'étape qui précède immédiatement la conclusion: c'est-à-dire l'argument décisif, l'indice déterminant.
  Exemple: ''c'est un /l/ parce que le premier formant est bas'', l'expert n'a pas identifié le /l/ simplement à la vue du premier formant, mais il a dû utiliser d'autres indices pour reconnaître la sonante (nombre de formants, intensité ...) et c'est la position du premier formant qui a permis de lever la dernière ambiguité: ce n'est pas /R/ mais /l/ parce que le premier formant est trop bas.

- L'expert utilise souvent le mot ''alors'' mais il y a ambiguité sur la signification de ce terme: parfois il l'emploie dans un sens voisin de l'implication mathématique, parfois la contraposée (non B --> non A) est fausse. Dans certains cas, ''alors'' signifie simplement ''penser à'', ''c'est peut-être '' ou ''ce ne peut être que ''. Il faut donc faire très attention lors de l'écriture des règles, à faire préciser, par l'expert, ce qu'il entend par ''alors''.
- Souvent la règle est énoncée de manière incomplète ou ambigue. Par exemple, omission, pour une règle contextuelle, du contexte dans lequel elle s'applique:
  '' /l/ car concentration d'énergie autour de 1200 Hz et le premier formant est bas''. Cette règle n'est valable qu'en contexte labial; généralement le deuxième formant de /l/ se situe vers 1500 Hz.

**b) Sur l'analyse perceptive**

L'expert est encore moins explicite sur ce qu'il ''voit'' que sur la manière dont il utilise l'information visuelle. Lorsqu'il déclare: ''c'est trop long pour qu'il s'agisse d'un seul segment'', quelle est sa durée de référence? S'agit-il d'une valeur absolue ou relative? On peut se poser la même question en ce qui concerne son utilisation des niveaux de gris sur le spectrogramme, par exemple, lorsqu'il affirme ''formant 1 trop faible pour que ce soit /a/''.

C'est en essayant de répondre à ces questions que nous nous sommes aperçus qu'avant d'aborder le spectre de la gauche vers la droite, l'expert reste quelques secondes, silencieux à le regarder. On peut penser que pendant ce temps, il porte un regard global qui lui permet de:

- déterminer la durée moyenne d'un segment à partir, essentiellement, des noyaux vocaliques,
- déterminer l'amplitude des variations significatives d'énergie, c'est-à-dire étalonner les niveaux de gris du spectrogramme.

## II. Règles de phonétique générale

Ces règles ont été obtenues principalement dans la première phase d'acquisition de l'expertise. L'expert décrivait les phonèmes hors contexte. Pour les voyelles, il indiquait la hauteur des trois premiers formants. Pour les plosives, il précisait les indices principaux qui caractérisent ces phonèmes: silence, présence d'un burst, position des formants. Mais il

n'indiquait pas lesquels, parmi ces indices, étaient toujours présents ou facultatifs, ni lesquels étaient influencés ou non par les phonèmes adjacents. De plus, ces règles étaient souvent incomplètes, car l'expert ne précisait pas des prémisses évidentes pour lui.
Exemple: ''Si bruit de friction intense entre 4000 et 5000 Hz, Alors /s/''
Cette règle ne précise pas qu'il ne doit pas y avoir d'énergie visible inférieure à 4000 Hz.
Pour les phonèmes fortement variables ( /R/, /l/ ), l'expert ne les décrivait que dans un contexte intervocalique. Des réalisations telles que /pl/, /tr/, étaient passées sous silence. Avec de telles règles, nous nous sommes vite rendu compte que nous étions incapables de segmenter et de décoder des spectrogrammes. Elles permettent simplement de reconnaitre les principales classes de sons: plosives, noyaux vocaliques, fricatives.

## III. Règles contextuelles

Ce sont des règles qui ont été obtenues au cours de la deuxième phase. En effet, comme nous l'avons déjà signalé, l'expert utilise pratiquement toujours le contexte droit et le contexte gauche pour prendre une décision. Ces règles servent souvent à déterminer les limites fréquencielles de certains paramètres en fonction des différents contextes. Par exemple: la limite inférieure du bruit de friction d'une fricative, dans un contexte labial ou non labial.

### Règles de classification

Elles permettent de donner la classe phonétique à laquelle appartient le segment: noyau vocalique, plosive, fricative, sonante.
Exemple: ''Si quatre formants bien marqués, alors noyau vocalique ''.
Elles sont en général utilisées en premier lieu, ensuite on applique les autres règles pour affiner et identifier les phonèmes.

### Règles d'identification

Ces règles sont appliquées quand l'expert essaie de distinguer deux phonèmes extrêmement proches. En général, il y a une règle par contexte possible.
Par exemple: pour distinguer /m/ et /n/ dans un contexte vocalique, on pourra se référer au tableau de la figure B.4.

**Règles d'affinement de la segmentation**

Le but de ces règles est de diviser un segment en deux sous-segments. Les prémisses de ces règles peuvent comporter des conditions sur la longueur du segment ou sur des instabilités spectrales.
Exemple: ''Si durée du segment supérieure à une fois et demie la durée vocalique moyenne, Alors envisager deux segments''.

**Règles d'élimination**

Ces règles ne concluent pas sur la présence d'un phonème, mais au contraire l'infirment. Ces règles sont particulièrement utiles lorsque l'expert est arrivé à établir une liste de phonèmes possibles et qu'il désire réduire cette liste.
Par exemple: ''Si énergie entre 1000 et 2000 Hz, Alors pas /m/ ni /n/''.

**Règles de confirmation**

Lorsque l'expert a constitué une liste de phonèmes compatibles, il cherche à valider, à confirmer ces hypothèses. Il utilise alors des règles portant sur des indices qui ne sont pas utilisés dans la première phase de raisonnement, en particulier des indices dont l'absence n'est pas significative mais dont la présence est porteuse d'informations.
Par exemple: ''Si transition rapide descendante sur le formant 3 de la voyelle précédente, Alors /l/''.

**Règles de compatibilité**

Parfois, quand l'expert a décelé un indice qui lui parait ''clair'', il essaie d'énoncer une liste de phonèmes compatibles avec cet indice.
Exemple: ''Si formant 1 bas vers 300 hz, Alors /l/ /m/ /n/ /gn/ /j/ /w/''.

**Règles de prédiction**

Il arrive parfois que l'expert, en remarquant un indice particulier, puisse prédire des informations sur le segment suivant (la classe phonétique par exemple).
Par exemple: ''Si la limite inférieure du bruit d'une fricative est

descendante (gauche -droite), Alors le contexte droit est labial''.

## IV. Conclusion

L'acquisition de l'expertise est une étape fondamentale dans la réalisation d'un système expert, la base de connaissances en étant la pièce maîtresse. Cette étape a demandé beaucoup de temps et elle n'a pu être réalisée que grâce à la bonne collaboration qui s'est établie entre notre èquipe et l'expert.
Parallèlement, pendant que nous amassions cette expertise, nous avons acquis nous mêmes une partie de cette expertise. Mais, si nous avons atteint rapidement un taux de reconnaissance moyen en lecture de spectrogrammes, nous nous sommes rendu compte que nous sommes loin de rivaliser avec l'expert. On n'acquiert pas facilement une expertise obtenue par des années d'expérience et c'est cela qui sépare le ''bon amateur'' du grand expert.

# CHAPITRE 2

# LES STRATEGIES

## I. Approche globale

Après un regard global (pour déterminer la durée vocalique moyenne), l'expert procède au décodage de la phrase, de la gauche vers la droite, mais c'est à notre demande, car il lui serait possible de commencer à partir de n'importe quelle frontière claire.

## II. Analyse locale

Surtout dans les cas difficiles (en particulier, zones vocaliques comprenant plusieurs sons), l'expert effectue en parallèle segmentation et identification. Pour analyser les stratégies utilisées par l'expert, nous allons distinguer deux éventualités:

- le cas où la segmentation est évidente,
- le cas où elle ne l'est pas.

### 1) Segmentation évidente

Cette situation se présente lorsque l'expert remarque de belles frontières bien nettes de part et d'autre d'un segment et lorsque la longueur de ce segment n'excède pas une fois et demie la durée vocalique moyenne. A ce moment, l'expert émet l'hypothèse que ce segment représente un seul son, et il axe essentiellement son raisonnement sur l'identification de ce segment.

- **Cas où l'identification est immédiate**

Dans de telles situations, on ignore tout de la stratégie utilisée par l'expert; en effet, il se borne à dire: ''Je vois un ... , cela ne peut être que ... ''. Peut-être y-a-t-il eu comparaison avec des formes de références de nature complexe, car elles doivent englober les transitions avec les phonèmes voisins?

- **Cas où l'identification résulte d'une analyse visuelle et d'un raisonnement explicite**

Cela se produit chaque fois que l'information spectrale est ambigue, donc que l'expert hésite. Le champ d'observation est généralement composé du segment lui-même et de ses deux voisins. Normalement, à ce stade, l'analyse du voisin gauche est terminée (analyse gauche-droite), mais la nature phonétique du voisin droit est inconnue. Pour pouvoir prendre en compte dans son raisonnement sur le segment courant l'influence du contexte droit, l'expert va être obligé d'analyser partiellement au moins ce segment. Cette opération de classification, en particulier si un des voisins n'est pas une voyelle, peut conduire à élargir le contexte.
Par exemple, pour déterminer la nature d'une occlusive sourde /p/ /t/ /k/ dont le contexte droit est une fricative, pour être en mesure d'interpréter la barre d'explosion caractéristique de l'occlusive, il lui faut déterminer la nature de cette fricative, donc examiner son contexte droit. Pour identifier certains phonèmes, il est nécessaire d'identifier complètement le contexte droit (il ne suffit pas de déterminer sa classe). Il y a donc interaction entre l'identification des deux segments, qui peut être menée de façon parallèle ou par un jeu d'aller et retours.
Par exemple: Pour analyser la barre d'explosion d'une occlusive, il est nécessaire de connaître la voyelle qui suit cette occlusive. S'il y a hésitation sur la voyelle entre /i/ et /y/ et que le burst soit vers 2500 Hz, on peut hésiter entre /t/ et /d/.

Sur le segment lui-même, l'expert procède à l'analyse visuelle pour détecter ainsi un ou plusieurs indices. A partir d'un ou plusieurs de ces indices, il sélectionne une liste de phonèmes candidats. Ensuite, il essaie de classer ces phonèmes, par ordre de plausibilité décroissante, ou de réduire cette liste; pour cela, il vérifie que les caractéristiques acoustiques de chaque phonème hypothétisé sont compatibles avec les propriétés du segment observé dans le contexte identifié. Au cours de cette phase, l'expert procède à une analyse visuelle fine aussi exhaustive que possible, il

y aura compatibilité s'il n'existe pas d'indices contradictoires avec l'hypothèse émise et s'il ne manque pas d'indices que l'on sait être toujours présents.

Mais la notion de compatibilité est plus complexe. L'expert dira que la forme observée est compatible avec un phonème hypothétisé s'il peut fournir une explication articulatoire qui permette de rendre compte de la distance entre la forme observée et la réalisation attendue.

Parmi ses connaissances, l'expert utilise un modèle articulatoire complexe dont il faudrait essayer de déterminer avec lui le rôle exact dans l'identification. Il faudrait préciser en particulier dans quels cas il fait intervenir cette source d'informations: l'utilise-t-il uniquement pour l'identification des réalisations atypiques résultant d'altérations dues à de fortes coarticulations?

Nous nous sommes aperçus que l'expert ne se fonde jamais sur un seul indice acoustique, aussi clair soit-il, pour justifier une interprétation phonétique.

A partir d'un indice bien marqué, il émet une hypothèse sur la nature du segment, puis tente d'infirmer ou de confirmer cette interprétation à partir de l'examen des autres indices (ceci tient à la nature pluri-indicielle du codage phonétique). Dans certains cas, l'expert itère la démarche émission d'hypothèses - validation: ''Qu'est-ce que ça pourrait être d'autre?'' Cette itération est un moyen de pallier les insuffisances du raisonnement humain (peur de l'oubli), mais elle traduit un objectif fondamental. Il cherche à obtenir un treillis de phonèmes qui contienne tous les phonèmes de l'énoncé, même si, pour y parvenir, il doit augmenter sensiblement le nombre d'hypothèses émises pour chaque segment.

Il est en effet plus facile, pour les niveaux supérieurs, de sélectionner parmi des hypothèses nombreuses la solution exacte, que de reconstruire l'énoncé prononcé à partir d'un treillis entaché de nombreuses omissions.

**Remarque 1. Adaptation de l'expert**

Il arrive parfois que l'expert rencontre, au cours du décodage d'une phrase, une forme semblable à celle d'un segment précédemment analysé. S'il parvient à identifier la deuxième forme avec plus de certitude que la première, il peut remettre en cause les conclusions établies sur la première forme en la comparant en détail avec celle qu'il a réussi à identifier avec certitude.

Arrivé au milieu du spectrogramme, il lui arrive aussi parfois d'émettre une remarque du type ''Normalement c'est un /y/ mais comme j'ai remarqué que les voyelles de ce locuteur présentent depuis le début du spectrogramme des formants 3 très bas, ce segment pourrait aussi être un /ɛ/ / ''.

**Remarque 2. Raisonnement par défaut**

Lorsque l'expert ne voit aucun indice discriminant, il raisonne par défaut. Par exemple: S'il est en présence d'un segment ne présentant pas d'énergie visible (excepté la barre de voisement), il propose la liste /b/ /v/ /m/ /n/ /l/ .

**2) Segmentation ambigue ou difficile**

Lorsque l'expert hésite sur le nombre de segments présents dans une partie du spectrogramme, il applique des stratégies différentes de celles employées dans les cas précédents.

Dans les paquets vocaliques, par exemple, c'est-à-dire dans les zones vocaliques dont la durée correspond à trois ou quatre segments et qui ne présentent aucune frontière nette interne, l'expert conduit son analyse à partir des extrémités.

Il essaie d'identifier complètement les voisins de gauche, de droite, du paquet, et d'émettre des hypothèses en partant de ces zones sûres vers le centre. Une hypothèse se compose d'une segmentation de la zone et de l'identification des segments obtenus.

Une segmentation n'est acceptée que si l'identification est cohérente, en particulier si elle respecte les règles de coarticulation. Parfois, s'il existe un extremum à l'intérieur de la zone (un maximum ou un minimum des variations formantiques), on part de cet extremum vers les extrémités de la zone. La segmentation s'effectue à partir de la détection des variations importantes du signal acoustique et/ou de la duré vocalique moyenne.

Pour conclure, on peut remarquer que, sauf dans les cas clairs, il ne peut pas y avoir séparation entre segmentation et identification, lesquelles sont étroitement liées pour l'expert. Une autre caractéristique très importante est le fait que l'analyse se déroule le plus souvent en parallèle.

# PARTIE C

# FORMALISATION DE L'EXPERTISE

# INTRODUCTION

Après avoir acquis l'expertise, il nous faut maintenant la formaliser pour l'intégrer dans notre système. Pour ce faire, nous avons créé des outils: un spectrogramme numérique, des algorithmes de prétraitement, des algorithmes sur les plosives... Mais la plus grande partie de l'expertise a été traduite sous forme de règles de production, dont nous donnerons la syntaxe ainsi que des exemples. Nous détaillerons ensuite le fonctionnement du moteur d'inférences, chargé de la construction du treillis phonétique.

# CHAPITRE 1

# OUTILS ET PRETRAITEMENTS

## I. Introduction

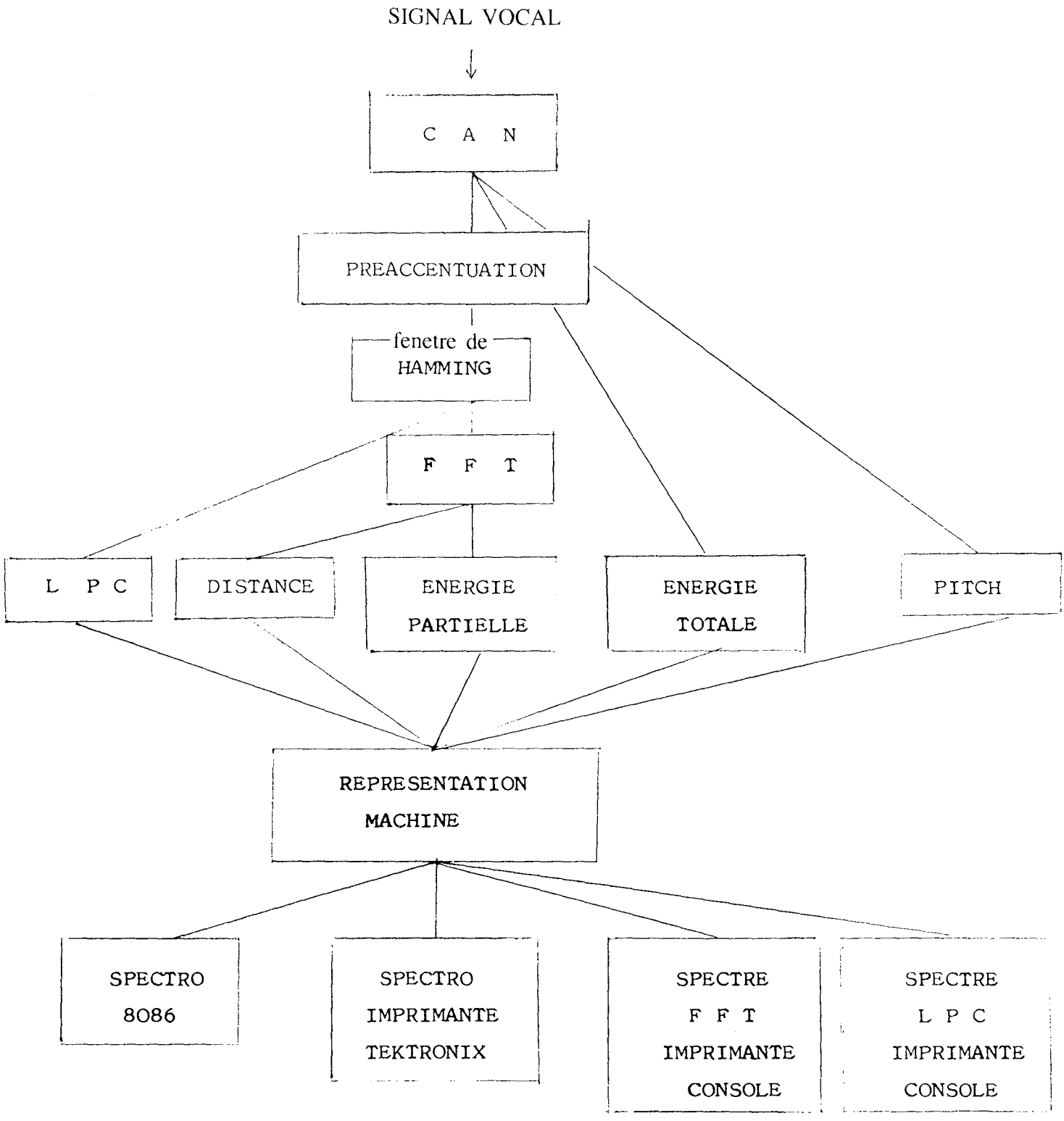
Pour formaliser les connaissances de l'expert, nous avons dû développer un certain nombre d'outils, et tout d'abord des spectrogrammes numériques, afin que notre représentation machine soit proche des spectrogrammes utilisés par l'expert. Pour valider notre système, nous avons acquis et segmenté un corpus de parole continue en contexte multilocuteur de 57 phrases. De plus, certaines connaissances de l'expert n'ont pas été formalisées sous forme de règles (cf chapitre 3) mais sous forme de procédures de prétraitement. Enfin, au vu du résultat de l'expert sur les liquides (taux d'erreur le plus important), nous avons essayé de voir si une méthode plus classique (méthode stastistique et méthode par formes de références) pourrait donner de meilleurs résultats.

## II. Les spectrogrammes

Les règles de François Lonchamp s'appuyant sur la présence ou l'absence de détails propres aux spectrogrammes Voiceprint A 700, nous devions avoir une représentation machine aussi proche que possible des spectrogrammes utilisés par l'expert. Il nous a donc fallu développer des spectrogrammes numériques.

### 1) Obtention des spectrogrammes numériques

Nous partons d'un signal de parole continue échantillonné à 12 kHz sur 10 bits (cf figure C.1). Ensuite, sur des fenêtres de 256 points, nous effectuons une différentiation du signal (préaccentuation de 6 dB par octave) et utilisons une fenêtre de Hamming:




**Obtention des spectrogrammes numériques**

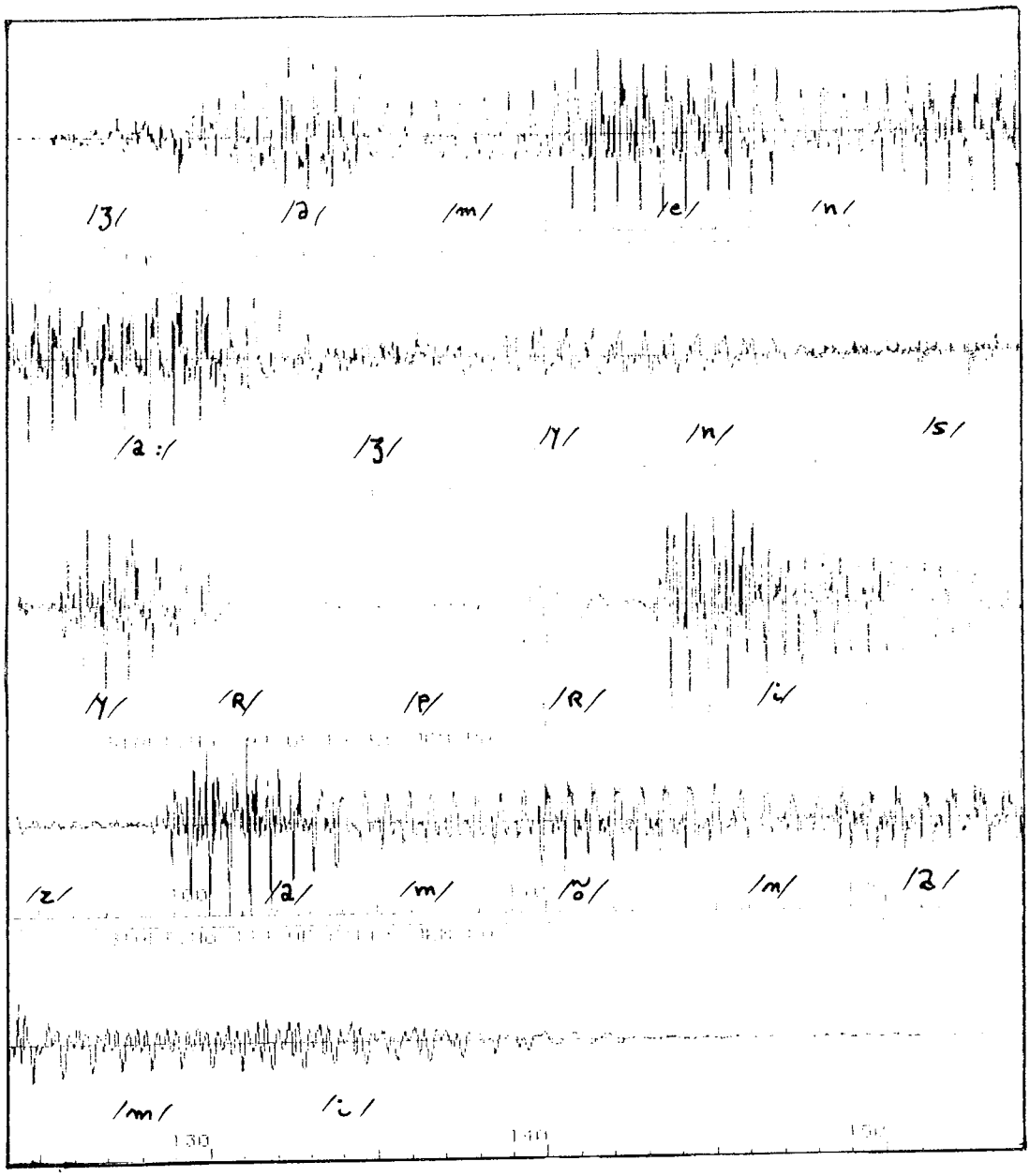
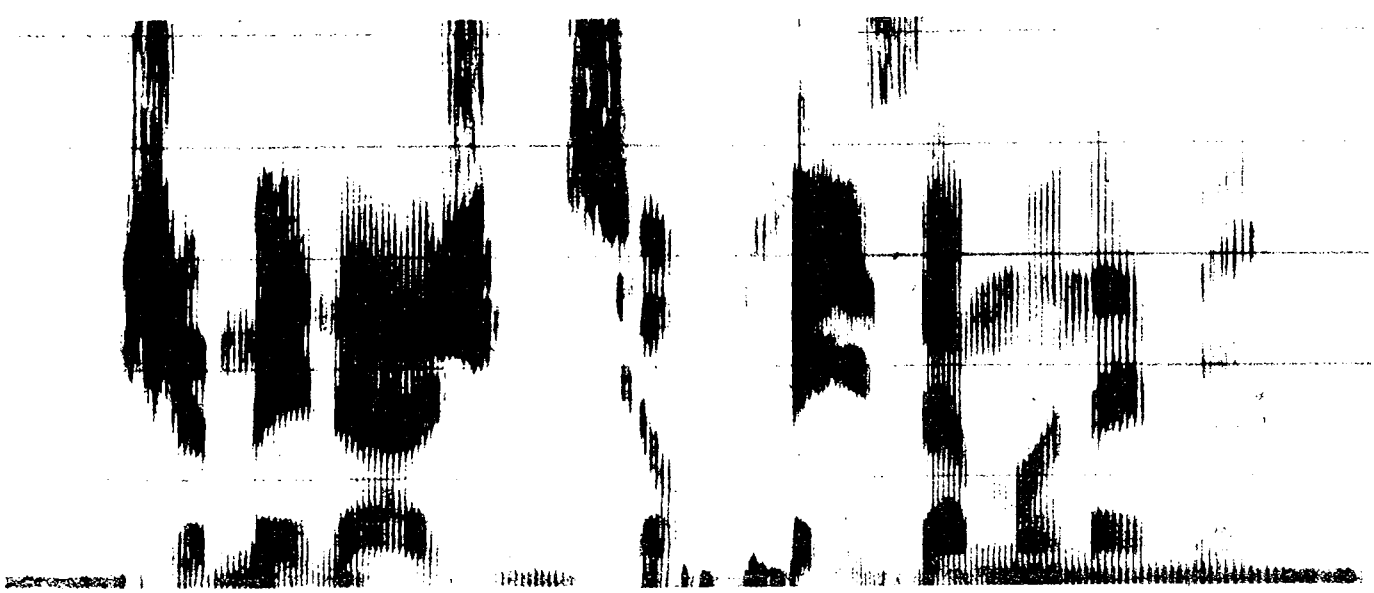



**Figure C.1** : Spectrogramme et signal acoustique de la phrase Je ménage une surprise à mon ami.



$$y(i) = [\, x(i) - x(i-1) \,] * [\, 0.54 - 0.46 * \cos(\, 6.28 * (\, i-1 \,) / 255 \,) \,]$$

Après une FFT (Fast Fourrier Transform), nous prenons le logarithme, puis effectuons un lissage, c'est-à-dire une moyenne sur une fenêtre glissante de 400 Hz de large. La fenêtre d'analyse est déplacée à chaque fois de 64 points (environ 5 ms) pour réaliser une analyse fine du signal.

Pour permettre une normalisation des phrases, nous ajoutons une constante C1 aux valeurs obtenues. Pour obtenir cette constante, nous calculons la courbe de l'énergie comprise dans la bande 250-2200 Hz. Puis, nous prenons les maxima M1 M2 M3 des trois spectres (représentation fréquence-intensité) qui corrrespondent aux trois plus grands pics sur cette courbe. C1 est donné par la formule:

$$C1 = 60 - ( M1 + M2 + M3 ) / 3$$

Ainsi, toutes les phrases ont leur maximum aux alentours de 60 dB. Ceci constitue donc notre représentation machine.

## 2) Visualisation

Pour contrôler la qualité des spectrogrammes ainsi obtenus, la meilleure façon était encore de les afficher sur les différentes consoles graphiques disponibles au Laboratoire. Nous avons donc dû affecter un niveau de gris à chaque valeur fournie par la F.F.T. Pour ce faire, nous avons copié le mode de fonctionnement du spectrogramme qui affecte aux fortes intensités le noir et aux faibles intensités le blanc. Nous avons utilisè la même dynamique: 32 dB. Pour qu'un point du spectre soit affiché, il faut qu'il satisfasse deux critères; son énergie doit:

- être supérieure à SEUILINF; limite de visibilité absolue: 27 dB (les spectres ayant été normalisés grâce à C1),
- être supérieure à Max_du_spectre - DYNAMIQUE (avec DYNAMIQUE = 32 dB).

### - Tektronix 4014

Sur cette console noir et blanc, les niveaux de gris sont simulés en noircissant une matrice de points. La résolution disponible autorisait 8 niveaux de gris.

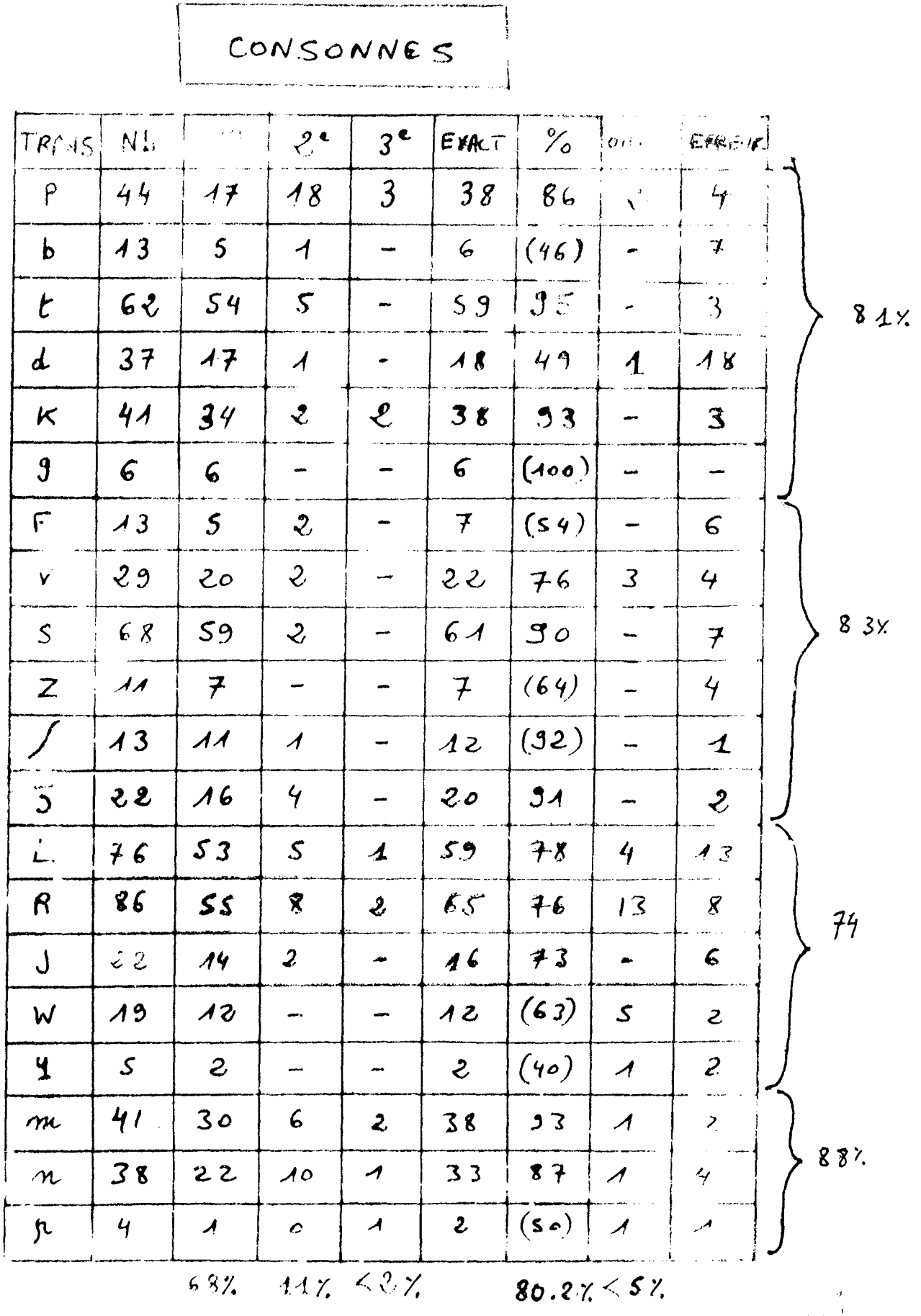



**Figure C.2** : Résultats de l'expert sur les consonnes
(les pourcentages sur moins de 20 items sont indiqués entre parenthèses)

Les résultats de reconnaissance que nous donnerons dans la suite de cette thèse seront tous fondés sur ces segmentations manuelles.

Ce corpus a également servi à mesurer les performances de notre expert. En effet, il nous semblait intéressant de connaître avec précision, le pourcentage de phonèmes exacts décodés par l'expert. Pour cette expérience, nous avions imposé à l'expert :

- de ne pas utiliser de connaissances linguistiques ou lexicales,
- de décoder de gauche à droite (de l'avis même de l'expert, cette contrainte ne modifie pas ses performances: l'expert peut procéder de la droite vers la gauche ou par îlots de confiance sans que les résultats s'en ressentent),
- de ne pas utiliser de réglet pour mesurer sur le spectrogramme,
- de commenter le décodage complet des deux premières et des deux dernières phrases de chaque locuteur ainsi que les erreurs commises; ces commentaires ont été enregistrés.

Les résultats sont donnés sur la figure C.2.

On peut noter que le pourcentage global de reconnaissance sur les consonnes (80,2%) est nettement supérieur à celui atteint par les meilleurs systèmes automatiques pour ce qui est de la parole continue. Remarquons que ces résultats sont obtenus à partir des 3 premières étiquettes phonétiques proposées par l'expert pour un segment, mais que la troisième augmente les performances de moins de 2%. On peut alors se poser la question de savoir s'il est nécessaire, dans un système automatique d'autoriser une troisième étiquette pour un si faible gain.

C'est la classe des nasales qui est la mieux reconnue (88%), par contre, les phonèmes /l/ /r/ /d/ /j/ et /v/ sont ceux qui posent le plus de problèmes. Le tableau C.4 permet de se rendre compte du nombre d'étiquettes erronées pour chaque phonème. On peut noter que le phonème /g/ n'est jamais confondu avec un autre phonème. En revanche, le phonème /p/ qui apparait 44 fois dans le corpus, a été étiqueté par l'expert avec 54 autres labels (on autorise plusieurs labels par phonème) dont 43 sont des plosives (une seule erreur sur la classe phonétique).

Pour les voyelles, le résultat (cf figure C.3) est moins bon (73% seulement), mais on peut expliquer ce score par le fait que nous lui avions interdit de mesurer les fréquences formantiques avec un réglet. Si l'expert avait pu disposer des fréquences exactes des formants (avec une analyse L.P.C. par exemple), ces performances auraient été bien meilleures.

- **Tektronix 4105**

Cette console couleur permet l'utilisation de 8 couleurs simultanément.

- **Traitement d'images**

On utilise ici la résolution maximale de cette machine, c'est-à-dire 512*512 pixels avec 256 niveaux de gris.

Nous avons affiné la représentation spectrographique jusqu'à ce que l'expert affirme qu'il pourrait réaliser un décodage sur nos spectrogrammes numériques avec des performances comparables à celles obtenues avec son Voiceprint. Bien sûr, ce décodage ne pourrait s'exécuter que sur le système de traitement d'images, seul appareil ayant une définition qui lui permette de rivaliser avec les spectrogrammes analogiques. L'utilisation de la couleur n'apporte, semble-t-il, aucune aide à l'expert, trop habitué à ses spectrogrammes en noir et blanc. Nous ne voulions pas modifier l'expertise en changeant la représentation utilisée par l'expert: l'aquisition de l'expertise doit se faire sans gêner l'expert.

## III. Le corpus

Pour tester notre système, nous disposons d'un corpus de 57 phrases équilibrées de Combescure [ Combescure 81 ], prononcées à un rythme naturel d'élocution par cinq locuteurs masculins non professionnels. Les phrases étaient prononcées par un locuteur expérimenté et répétées de mémoire (mémoire immédiate). Nous avons adopté ce mode opératoire pour éviter un rythme trop lent et une intonation de lecture. Nous avons de ce fait homogénéisé la prosodie et le rythme du corpus (environ 14 phonèmes par seconde). Ces phrases ont été numérisées à 12 kHz sur 10 bits.

De plus, nous avons réalisé les spectrogrammes numériques (cf A.3.III.) de ces 57 phrases et l'expert a manuellement placé des marques de segmentation sur ce corpus. Deux segmentations ont été effectuées: une segmentation phonémique et une intra-phonémique. Par exemple, dans cette deuxième segmentation, une distinction est faite entre:

- la partie nasale d'une plosive sonore et la partie de silence dans les cas de nasalisation,
- le silence et la barre d'explosion dans le cas d'une plosive.

VOYELLES

| TRANS | Nb | CORRECT | % | OMIS | |
|---|---|---|---|---|---|
| i | 55 | 47 | 85 | 5 | |
| e | 70 | 66 | 94 | - | |
| ɛ | 42 | 32 | 76 | 4 | |
| a | 99 | 75 | 50 | 1 | |
| ɔ | 20 | 10 | 50 | 1 | 75 % |
| o | 21 | 16 | 76 | - | |
| u | 32 | 21 | 66 | 1 | |
| y | 22 | 14 | 64 | - | |
| ø | 4 | 3 | (75) | - | |
| œ | 7 | 1 | (14) | - | |
| ə | 67 | 43 | 64 | - | |
| ɑ̃ | 35 | 23 | 66 | 1 | |
| ɔ̃ | 29 | 23 | 79 | - | 61 % |
| ɛ̃ | 15 | 5 | (33) | - | |
| œ̃ | 4 | 0 | (0) | - | |

73% < 2.5%

**Figure C.3** : Résultats de l'expert sur les voyelles
(les pourcentages sur moins de 20 items sont indiqués entre parenthèses)

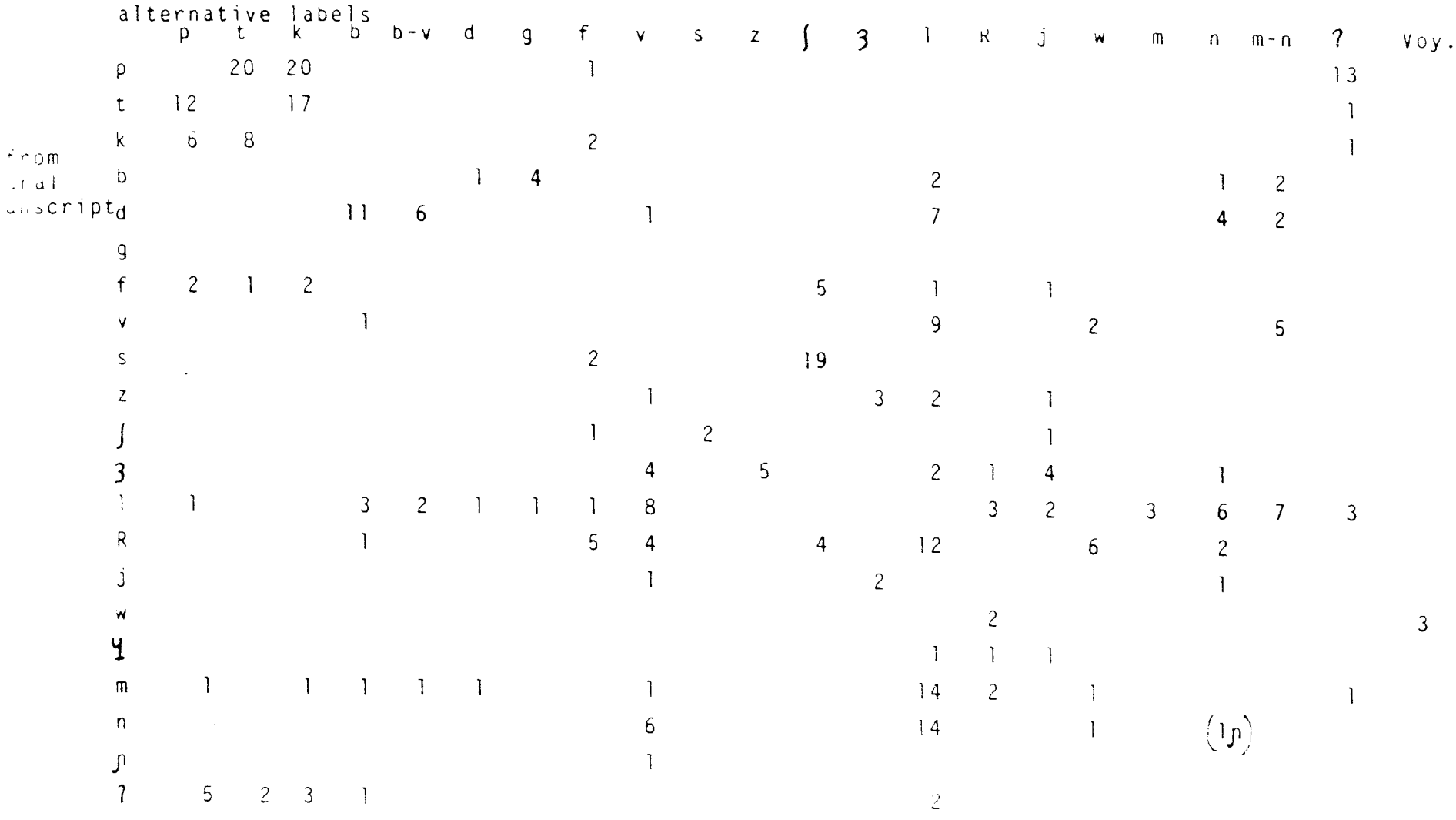



**Figure C.4** : Distribution des étiquettes phonétiques incorrectes.
ex: parmi les étiquettes associées au 44/p/ du corpus
on dénombre 20 /t/ 20 /k/ 1 /f/ 13 /ʔ/

## IV. Les prétraitements

Nous expliquerons l'utilisation des résultats fournis par les procédures de prétraitement dans le chapitre relatif au moteur d'inférences. Nous nous bornerons ici à exposer succintement les raisons pour lesquelles nous avons développé des algorithmes de prétraitement. Pour des raisons d'efficacité, il ne serait pas raisonnable d'utiliser une approche système expert pour analyser un énoncé sans aucune marque de segmentation. En effet, le nombre de segmentations erronées que le système envisagerait serait trop coûteux.

Nous avons donc décidé d'essayer de segmenter, à l'aide d'algorithmes, le signal en grandes classes phonétiques: noyaux vocaliques, fricatives, plosives. Ensuite, sur cette segmentation grossière, nous utiliserons le système expert pour identifier et segmenter finement la phrase.

En effet, cette segmentation grossière doit être facilement réalisable car l'expert peut, en quelques secondes, segmenter un spectrogramme en grandes classes phonétiques. Cette opération est non-contextuelle car grossière et repose sur des critères simples, facilement accessibles aux non-phonéticiens. Ainsi, au contact de l'expert, nous sommes parvenus facilement à effectuer cette opération.

### 1) Noyaux vocaliques

Le but de cet algorithme est de trouver tous les noyaux vocaliques contenus dans une phrase et de déterminer la durée moyenne des sons prononcés. Nous appelons ce paramètre relatif à un énoncé, la durée vocalique moyenne. L'expert nous a tout d'abord décrit les critères qu'il utilisait pour reconnaître les noyaux vocaliques sur un spectrogramme: forte intensité, présence de formants (pics d'énergie sur le spectre) dans une certaine bande de fréquence. C'est ainsi que nous avons décidé d'utiliser un critère reposant sur l'énergie.

Nous calculons tout d'abord l'énergie dans une bande de fréquence. Cette bande a été choisie de manière à défavoriser les sons ayant principalement de l'énergie en très basse fréquence (par exemple les nasales) et ceux qui ont de l'énergie en haute fréquence (par exemple les fricatives). Il fallait inclure dans cette bande la zone du premier et du deuxième formants des voyelles. Nous avons essayé différentes fréquences 250-1500 Hz, 250-2350 Hz, 250-2500 Hz. La bande que nous avons retenue, 250-2200 Hz, est celle qui fournit les meilleurs résultats. Ensuite nous cherchons les pics de cette courbe qui vérifient les critères ci-dessous:

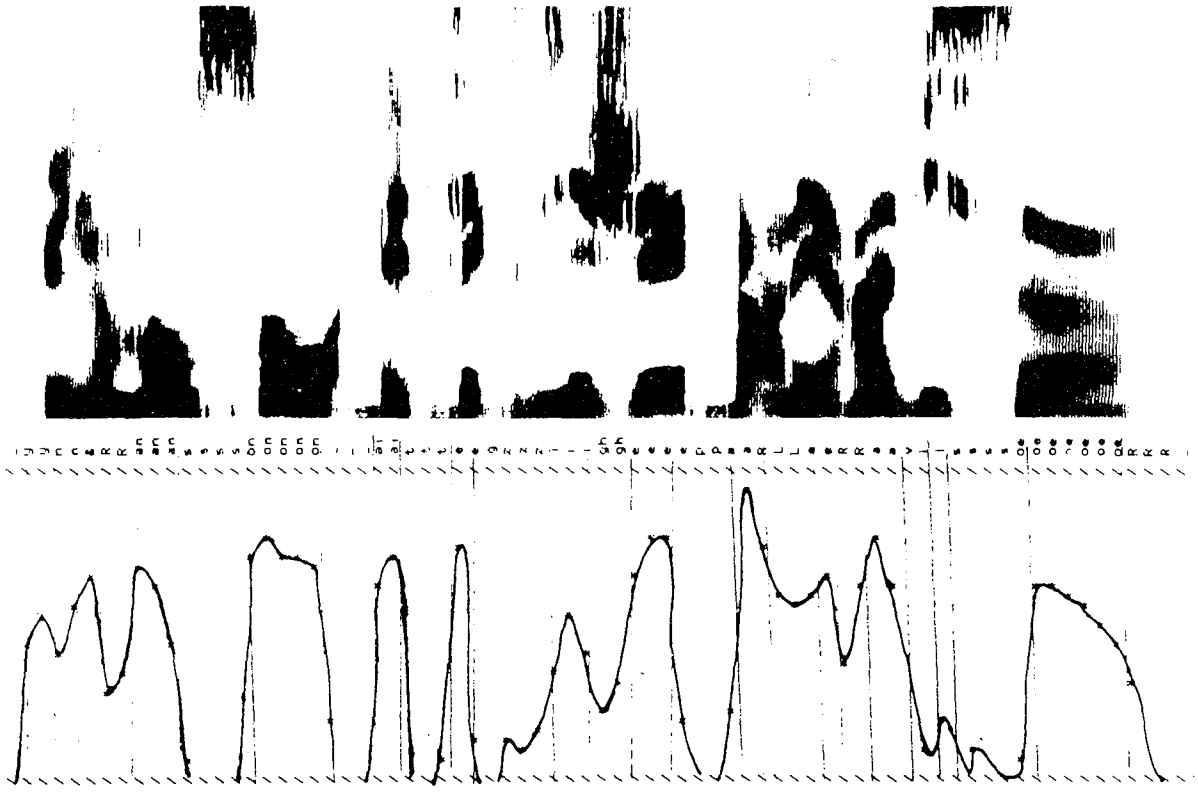


**Figure C.5** : Résultat de segmentation sur la phrase
Une rançon est exigée par les ravisseurs
(courbe d'énergie 250-2200 Hz).

- le pic doit, pour être retenu comme noyau vocalique, atteindre au moins 55 % du pic précédent; en effet, deux noyaux successifs ne peuvent avoir des maxima énergétiques trop différents, car même si les variations d'énergie sont importantes au cours de la phrase (cf en particulier les baisses d'énergie importantes en fin de groupe syntaxique ou en fin de phrase), elles ne présentent pas de discontinuité brutale.

- la vallée de part et d'autre du pic est fonction de la hauteur du pic; plus un pic est important, plus la vallée doit être importante (en fin de phrase l'intensité baissant, le locuteur dispose de moins de dynamique),

- au moins 50 % des échantillons du noyau vocalique doivent être voisés; nous n'avons pas imposé que tous les échantillons soient voisés car nous avons observé que certaines voyelles (/i/ en particulier) pouvaient être en partie dévoisées lorsqu'elles étaient en contact avec des consonnes sourdes.

Quand un pic (de:debut,pi:pic,fi:fin) vérifie tous ces critères, on recherche le début et la fin du noyau correspondant. On ne décrira que la recherche de la fin du noyau, car la recherche du début est symétrique.

(1) A partir du pic, on recherche le point D d'abscisse d et d'ordonnée E(d) qui vérifie:

$$\text{et} \begin{cases} E(d) <= E(pi) - SEUIL1 \\ d > pi. \end{cases}$$

On élimine ainsi le cas où le début de la descente ne serait pas situé juste après le point maximum (présence d'un plateau par exemple).

(2) On recherche ensuite le prélèvement R (r,E(r)) où l'énergie commence à remonter (voyelle ou sonante suivante):

$$\text{et} \begin{cases} r > d \\ E(r) < E(r+1) \end{cases}$$

(3) A partir de R, on recherche le prélèvement F qui vérifie:

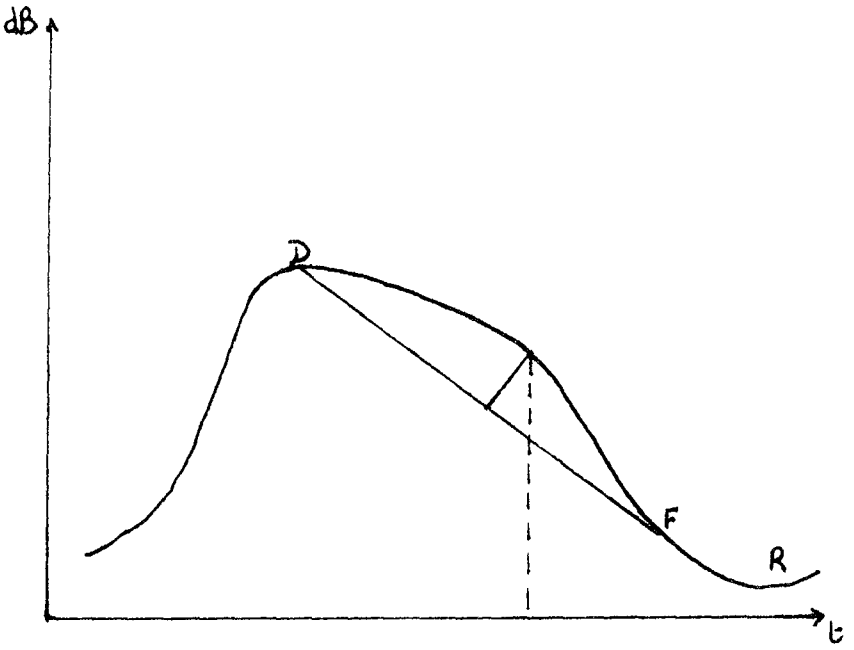
$$\text{et} \begin{cases} E(f) > E(r) + SEUIL2 \\ f < r \text{ (fin de la descente).} \end{cases}$$

(4) On trace le segment [ D,F ] et on cherche le prélèvement de la courbe de l'énergie situé au-dessus de cette droite et qui est à la plus grande distance de cette droite. Si le point trouvé est entre D + (F - D) / 4 et F - (F - D) / 4 (épaule), c'est le marqueur de fin de

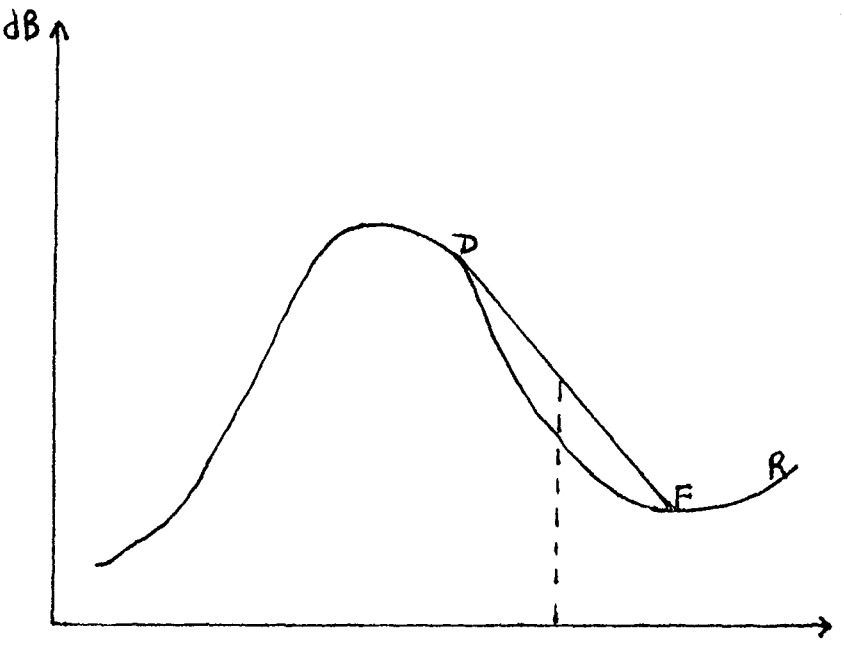
**Algorithme de recherche des noyaux vocaliques: calcul de D et F**

```
Tant que   E(i) < ( E(anc.pic) - min ) * .55 + min
       et E(i) < E(i+1)
       et E(i) < E(i-1)

faire i = i + 1
fin tant que
pic = i

vallee= seuil3 * ( E(pic) - min ) / ( max - min )
% max: maximum de la courbe E %
% min: minimum de la courbe E %
bon_pic = vrai
j = pic
tant que   bon_pic et E(j) > E(pic) - vallee

faire  j=j+1
       si E(j) > E(pic) alors bon_pic = faux
fin = j

j = pic
tant que   bon_pic et E(j) > E(pic) - vallee

faire  j=j-1
       si E(j) > E(pic) alors bon_pic = faux

debut = j

nb_voise = 0
pour k = debut a fin faire
       si voisement(k) alors nb_voise = nb_voise + 1
si nb_voise < ( fin - debut ) / 2 alors bon_pic = faux

si bon_pic alors
  i = debut
  tant que E(i) > E(pic)-SEUIL1 faire i = i + 1
  D = i
  j = fin
  tant que E(j) > E(j + 1) faire j = j + 1
  r = j
  tant que E(j) <= E(r) + SEUIL2 faire j = j - 1
  F = j
```

**Figure C.6 :** Calcul de la fin d'un noyau vocalique




Exemple avec épaule.




Exemple sans épaule

noyau, sinon c'est ( D + F ) / 2 . La figure C.5. présente deux exemples. Ce mode opératoire permet de segmenter correctement les noyaux situés en fin de groupes syntaxiques, car ces noyaux sont souvent allongés, et la courbe d'énergie présente souvent une décroissance lente puis une chute brutale (épaule).

On effectue 2 passes:

- une première passe en imposant une vallée importante de part et d'autre du pic (pics ayant une très forte probabilité d'être des noyaux mais taux d'omission de noyaux important).

- une deuxième passe en imposant une faible vallée de part et d'autre du pic (taux d'omission très faible mais quelques insertions)

Le taux de réussite est proche de 97%. La figure C.6. présente le résultat de la segmentation. Les résultats complets sont donnés au chapitre 2 de la partie D.

## 2) Les plosives

Le but de l'algorithme est de détecter les plosives (/p/, /t/, /k/, /b/, d/, /g/) et de délimiter dans un énoncé de parole continue les segments correspondants. Les plosives sont caractérisées par une occlusion momentanée du passage de l'air dans le conduit vocal. Les plosives sourdes /p/ /t/ /k/ apparaissent sur le spectrogramme comme des colonnes blanches suivies d'une barre d'explosion.
Dans le cas des occlusives sonores, il subsiste une vibration des cordes vocales qui se matérialise par une barre de voisement sur les spectrogrammes.

C'est pourquoi il vient à l'esprit, pour détecter les plosives, de rechercher les segments qui ne présentent pas d'énergie visible au-delà de 650 Hz. En conséquence, nous cherchons le maximum M1 du spectre 0-6000 Hz, le maximum M2 du spectre 650-6000 Hz, et celui de la voyelle adjacente M3.
Un spectre sera classé occlusif ou silence s'il vérifie l'un des critères suivants:

- M1 < SEUIL1 (rien de visible sur le spectrogramme)
- ( M1 - M2 ) < SEUIL2 (uniquement la barre de voisement visible)
- ( M3 - M2 ) < SEUIL3 (fort affaiblissement par rapport à la voyelle adjacente).

### 3) Les fricatives

Théoriquement, elles se reconnaissent aisément à leur répartition fréquentielle particulière: beaucoup d'énergie en haute fréquence et pratiquement pas d'énergie visible en basse et moyenne fréquence. Ceci est particulièrement vrai pour les sons /s/ /z/ /z/ /ʃ/ et précisément ce sont ces fricatives que nous essayerons ici de trouver et de segmenter. Après une première étude infructueuse qui utilisait des critères voisins de ceux utilisés par NOVOCA mais pour une bande de fréquence 3500-6000 Hz, nous avons essayé de calculer des centres de gravité. En effet, pour les fricatives, c'est la répartition d'énergie sur toute la gamme de fréquence qui est importante pour la reconnaissance (beaucoup d'énergie dans les hautes fréquences **et** pas d'énergie visible sur le spectrogramme en basses fréquences). Nous calculons donc le centre de gravité sur l'énergie visible (critère utilisé pour l'affichage des spectrogrammes numériques (cf II.)).

## V. Algorithme d'extraction d'indices

### 1) La barre d'explosion

Pour une plosive (/b/,/d/,/g/,/p/,/t/,/k/), un des indices pertinents pour la reconnaissance, est la barre d'explosion qui suit le silence. Dans un premier temps, nous allons chercher cette barre d'explosion, puis nous en calculerons les paramètres importants: intensité, maxima, continuité... Cette procédure ne peut être appelée que si on a détecté une zone silencieuse auparavant. Dans notre approche, nous avons essayé de ne lancer une procédure d'extraction d'indices que si on est sûr que cet indice est pertinent: on ne cherchera pas à calculer les formants d'une plosive ou la limite inférieure de friction d'une voyelle.
On va donc chercher une brusque bouffée d'énergie après le silence. Suivant le contexte et la plosive, le burst peut être présent de 800 à 6000 Hz. Nous allons donc calculer 5 courbes d'énergie correspondant aux 5 bandes de fréquences suivantes:

800-2000 1800-3000 2800-4000 3800-5000 4800-6000 Hz.

S'il existe au moins deux de ces courbes qui présentent la suite d'événements: ''montée_rapide creux montée_vers_le_son_suivant'', le creux étant situé au même prélèvement PR, alors on va lancer une analyse de burst sur le prélèvement PR.

Cette procédure retourne les paramètres suivants:

- un booléen indiquant si un burst a été trouvé

- le numéro du prélèvement correspondant au burst trouvé
- les différentes fréquences des maxima du spectre
- la valeur en dB du maximum
- la fréquence de ce maximum
- l'intensité moyenne du burst
- des paramètres précisant si le burst est continu ou concentré en fréquence
- le rapport énergétique des hautes fréquences sur les basses fréquences.

Ces paramètres sont intégrés à la base de mesures du segment considéré (cf partie C.3).
N.B. cette procédure peut trouver plusieurs burst. Il y a alors création de deux bases de mesures, chacune contenant les informations relatives à chacun des burst trouvés.

**2) Suivi de formants**

Parmi les indices souvent utilisés par l'expert, on peut citer le suivi de formants. Cet indice est utilisé pour la détection des /R/ (premier formant montant et deuxième formant descendant) et l'identification des plosives. Nous allons approximer la courbe formée par l'évolution temporelle de la fréquence d'un formant par une parabole

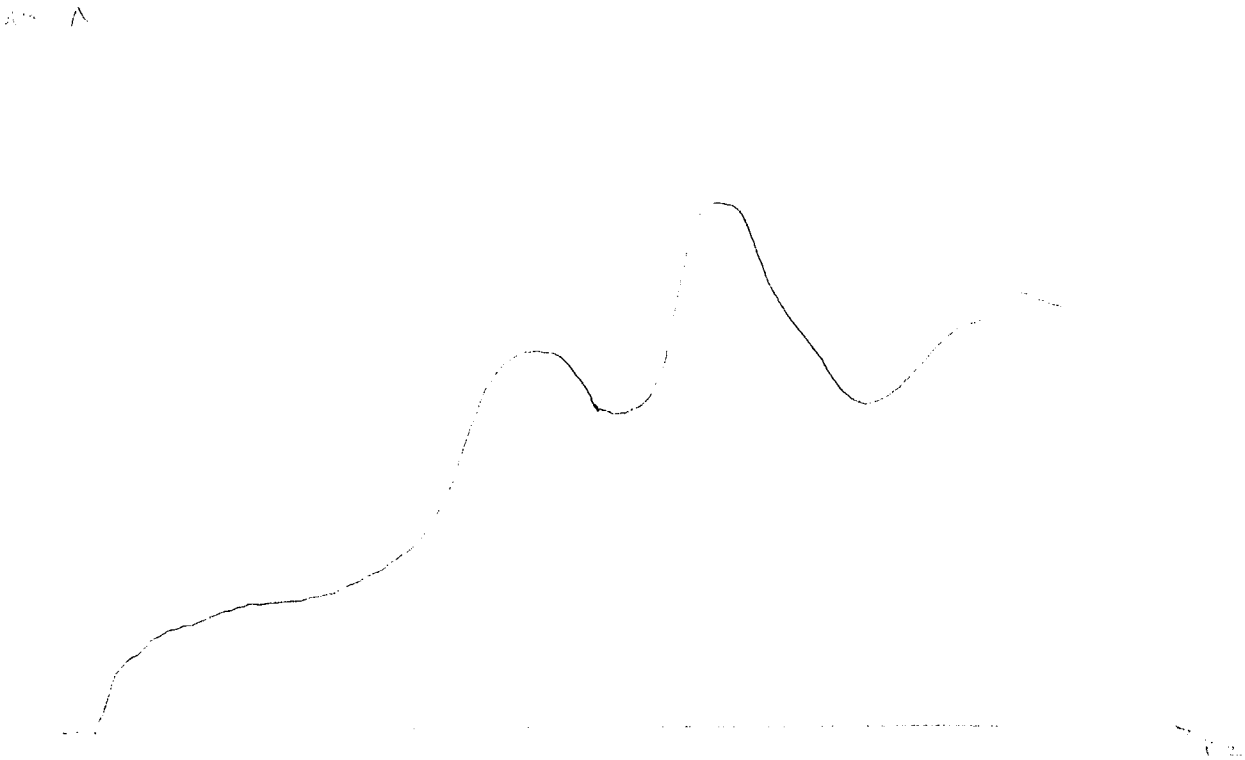
$$Y = A X^2 + B X + C \quad \text{(régression du 2ème ordre)}$$

Pour chacun des trois premiers formants, on calcule les trois paramètres A, B et C.

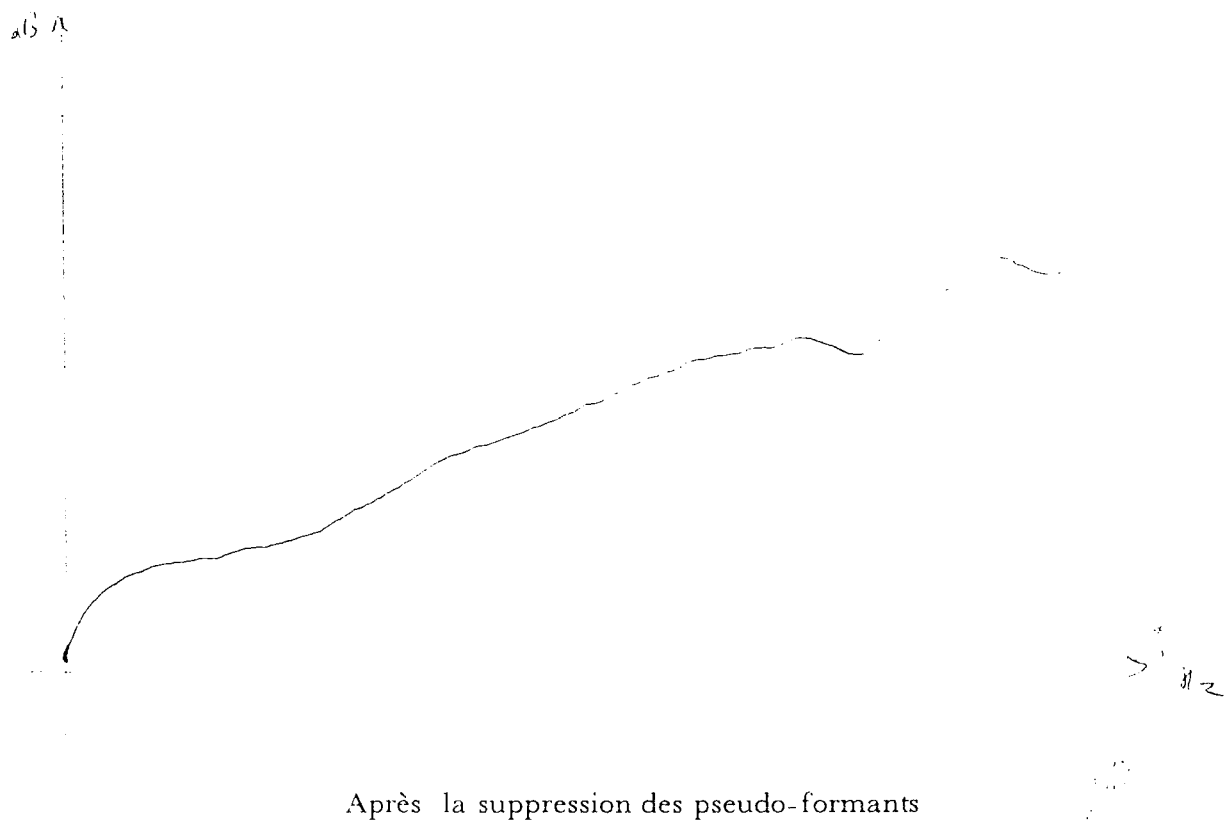
**3) Calcul de la limite inférieure de friction**

Pour une fricative, l'indice majeur pour l'identification, c'est la limite inférieure du bruit de friction. Avec des locuteurs professionnels, cette limite se calcule très facilement: il suffit de parcourir le spectre de la fricative depuis les basses fréquences, la limite de friction est la fréquence à laquelle on atteint le seuil de visibilité spectrographique. Mais, pour des locuteurs non professionnels (ce qui est le cas de notre corpus), on observe des pseudo-formants entre 1500 et 3500 Hz. Il faut donc éliminer ces pseudo-formants pour se retrouver dans un cas facile à résoudre.
Nous allons supprimer, dans le spectre de la fricative, des zones de 300 Hz de largeur de bande (à -3 dB). Le résultat de cette procédure est présenté en figure C.7.

Figure C.7



Avant la suppression des pseudo-formants



Après la suppression des pseudo-formants

## VI. Conclusion

Pour mettre en oeuvre les connaissances expertes recueillies lors de la phase d'acquisition, nous avons dû créer des outils:

- les spectrogrammes numériques (pour appliquer les règles sur la même représentation que celle utilisée par l'expert),

- des procédures d'extraction d'indices qui essaient de copier l'expertise visuelle,

- un corpus pour valider ces algorithmes.

De plus, nous avons développé des procédures de prétraitement pour segmenter (grossièrement) le signal de parole, avant d'appliquer les règles expertes. Mais ces procédures, qui ont été élaborées avec l'expert, sont pour nous une façon de formaliser l'expertise. En effet, c'est l'expert qui nous expliquait sa démarche, c'est-à-dire quels critères il utilise pour reconnaître les grandes classes phonétiques. Les erreurs des procédures ainsi élaborées étaient commentées par l'expert et comparées aux erreurs qu'il a lui-même commises sur ces phrases. Nous ne cherchons pas à corriger une erreur de nos algorithmes de prétraitement sur un segment dont l'expert n'a pas pu correctement identifier la classe.
Les procédures ont été affinées progressivement, au fur et à mesure des corrections proposées par l'expert.

A la vue des performances réalisées par l'expert sur les liquides (/l/,/R/), nous avons tenté de comparer les résultats obtenus avec des méthodes plus classiques: l'analyse discriminante. Les résultats complets sont donnés en partie D, et on peut dire que des méthodes purement statistiques, si elles peuvent parfois donner des résultats acceptables, sont ensuite très difficiles à améliorer (car elles sont aveugles) alors qu'une approche KBS (systèmes à bases de connaissances) au contraire permet en modifiant la base de connaissances d'accroître les performances; car on sait ce que l'on fait et surtout pourquoi. L'expert comprend ce qui ne va pas, pourquoi il y a une erreur et peut la corriger.

# CHAPITRE 2

# LA BASE DE CONNAISSANCES

## I. Justification du choix d'un système expert à règles de production

Il nous fallait un formalisme qui permette:

- d'entrer les connaissances acquises de façon incrémentale et dans le désordre, du fait du caractère long et délicat de l'acquisition de l'expertise,
- de retrouver le raisonnement par le système pour établir ses conclusions (pour une mise au point rapide et efficace),
- à l'expert de modifier lui-même la base de connaissances (grâce à un formalisme de représentation compréhensible par un expert du domaine).

Les systèmes à règles de production répondent en bonne partie à ces impératifs; en effet notre expert s'exprime naturellement, au cours de l'explicitation de sa démarche, sous forme de règles de production. Dans notre optique, le système expert ne sera pas la forme définitive d'un décodeur acoustico-phonétique; c'est pour nous une manière de formaliser la connaissance acquise de façon incrémentale au fur et à mesure de l'acquisition de connaissances nouvelles. En effet notre expert nous explique, sans ordre particulier, sa façon de procéder, étant donné qu'il commente les difficultés lors de leur apparition au cours de la phrase décodée. Ces aller-et-retours d'un problème à un autre sont inévitables en décodage phonétique de la parole, compte-tenu de l'ampleur du problème et du caractère contextuel des phénomènes qui entrent en jeu.

D'autres formalismes, comme celui des ''frames'' par exemple [Damestoy 86], que nous développons dans l'équipe ou plus généralement des représentations objets, permettent de structurer une base de connaissances comme celle que nous avons recueillie, mais cette structuration ne peut se faire, bien souvent, qu'a posteriori, lorsqu'on a

recueilli une grosse partie de l'expertise, tant en ce qui concerne les règles que les stratégies de l'expert. Par exemple, dans le cas d'un système à règles de production, on pourrait ordonnancer les règles ou les hiérarchiser; mais alors l'ajout d'une nouvelle règle doit respecter certaines contraintes, ce qui rend le système moins facilement incrémentable, surtout par un expert n'ayant pas participé à son élaboration.

Un système expert étant relativement coûteux en temps de calcul et en place mémoire, il faudra, lorsqu'on estimera la phase d'acquisition terminée, compiler la base de connaissances ainsi obtenue dans un formalisme permettant une réponse proche du temps réel, indispensable pour des essais sérieux en reconnaissance de la parole.

## II. Syntaxe des règles

La syntaxe générale d'une règle est la suivante:

R Numéro-de-la-règle

CONTEXTE_DROIT Liste-de-phonèmes
CONTEXTE_GAUCHE Liste-de-phonèmes
DEJA_SUGGERE Liste-de-phonèmes

SI
conditions-sur-les-faits

ALORS
conclusion

### a) Numéro de la règle

Il sert simplement à repérer les règles. On l'utilisera pour construire la trace du raisonnement du système.
Par la suite, nous autoriserons de labeller les règles par une chaîne de caractères mnémotechniques.

### b) Contexte

L'ensemble des contextes constitue une partie des conditions d'application d'une règle. En effet, les règles étant pour la plupart contextuelles, l'expert a précisé dans quel contexte particulier une règle sera applicable. Nous avions tout d'abord pensé à utiliser les grandes classes phonétiques classiques (plosives, labiales, sourdes...) utilisées par

une valeur seuil_2, la plausibilité tombe à 0, et pour une valeur de la variable comprise entre seuil_1 et seuil_2, la plausibilité décroît linéairement de 1 à 0. Nous avons choisi une décroissance linéaire pour plus de simplicité, l'expert nous donnant les valeurs de façon très approximative, cela convient parfaitement.

La fonction de cet opérateur est complétée par deux autres opérateurs notés ''<<'' et ''^'', qui réalisent les fonctions représentées sur la figure C.9. Nous assurons ainsi une pondération sur les indices, mais pas sur la partie contexte. En effet, même si le contexte gauche (déjà décodé) a une plausibilité faible, il se peut que ce chemin de raisonnement en cours soit quand même le meilleur. C'est la tâche du niveau supérieur d'explorer l'ensemble des différents chemins pour trouver la phrase prononcée (parcours de graphe). Ce n'est pas le rôle du seul niveau acoustico-phonétique de décider de détruire un chemin car sa plausibilité est plus faible qu'un autre. Notre politique est de retarder au maximum la prise de décision; plus on attend et plus on a d'informations pour prendre la bonne décision. Il s'agit d'une stratégie efficace en reconnaissance de la parole. Le système que nous avons réalisé permet de l'implanter plus aisément que dans une approche algorithmique de reconnaissance des formes classique.

### c) Conclusion

La conclusion de la règle peut être de deux types:

- soit une liste de phonèmes pondérés
- soit une action pour modifier l'arbre de déduction (le treillis).

En effet, les règles qui concluent sur les phonèmes sont des règles de déduction, et les pondérations peuvent être positives ou négatives, suivant que la règle affirme ou infirme des phonèmes:

1 pour affirmer
-1 pour infirmer
-9 pour éliminer définitivement.

Les règles d'action ont pour but de modifier la segmentation (scinder un segment en deux parties ou rassembler deux segments en un seul), ou d'activer des procédures particulières de traitement du signal qui peuvent être attachées aux règles.

## III. Option choisie

Nous avons été confrontés à deux formulations possibles pour les règles: soit utiliser des règles qui décrivent exhaustivement tous les

## LES PHONEMES DU FRANCAIS EN A. P. I.

| représentation machine | *VOYELLES* | | *CONSONNES* | | représentation machine |
|---|---|---|---|---|---|
| i | [ i ] | vie | [ p ] | soupe | p |
| e | [ e ] | blé | [ t ] | terre | t |
| ai | [ ɛ ] | merci | [ k ] | cou | k |
| a | [ a ] | patte | [ b ] | bon | b |
| | [ ɑ ] | pâte | [ d ] | dans | d |
| ) | [ ɔ ] | mort | [ g ] | gare | g |
| o | [ o ] | eau | [ f ] | feu | f |
| u | [ u ] | genou | [ s ] | sale | s |
| y | [ y ] | vêtu | [ ʃ ] | chat | ch |
| eu | [ ø ] | deux | [ v ] | vous | v |
| oe | [ œ ] | peur | [ z ] | zéro | z |
| & | [ ə ] | le | [ ʒ ] | je | gh |
| in | [ ɛ̃ ] | matin | [ l ] | lent | l |
| an | [ ɑ̃ ] | sans | [ r ] | roue | R |
| on | [ ɔ̃ ] | bon | [ m ] | main | m |
| un | [ œ̃ ] | lundi | [ n ] | nous | n |
| | | | [ ɲ ] | agneau | nj |

*SEMI-CONSONNES*

| | | |
|---|---|---|
| j | [ j ] | yeux |
| w | [ w ] | oui |
| ui | [ ɥ ] | huit |

Figure C.8

l'expert, mais nous nous sommes rendus compte qu'en fait, selon les cas, l'expert ajoutait ou retranchait des phonèmes ; pour une plus grande lisibilité, le contexte est ainsi toujours constitué d'une liste de phonèmes. Pour représenter ces phonèmes, nous n'avons pas pu utiliser l'alphabet phonétique international (A.P.I.),car nous ne disposons pas des caractères spéciaux nécessaires; nous donnons donc en figure C.8. la correspondance entre l'A.P.I. et notre représentation. Dans les exemples qui suivront, nous utiliserons toujours cette notation.

**c) Déjà_suggéré**

Certaines règles ne s'appliquent que si on a déjà pu prouver que le segment étudié appartient à un certain sous-ensemble de phonèmes. C'est le cas, par exemple, des règles qui discriminent /m/ et /n/ ou des règles de confirmation (cf IV.3.).

**d) Conditions sur les faits**

Quand il décode un segment, l'expert fonde son raisonnement à la fois sur le contexte, et sur un faisceau d'indices visuels. Ces derniers sont souvent des inégalités par rapport à des seuils fréquenciels. Dans les prémisses des règles, il faut donc combiner les différentes fonctions mathématiques usuelles:

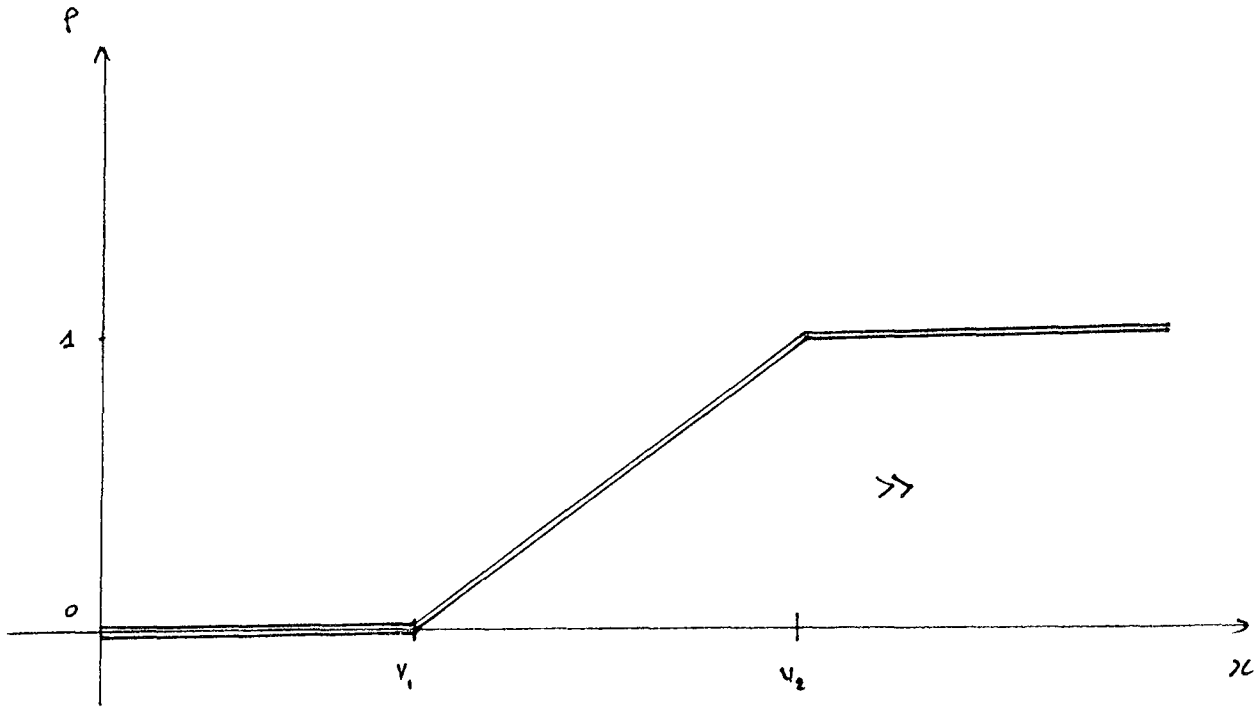
+, -, x, /,
partie entière, exponentielle, logarithme,
sinus, cosinus, tangente..

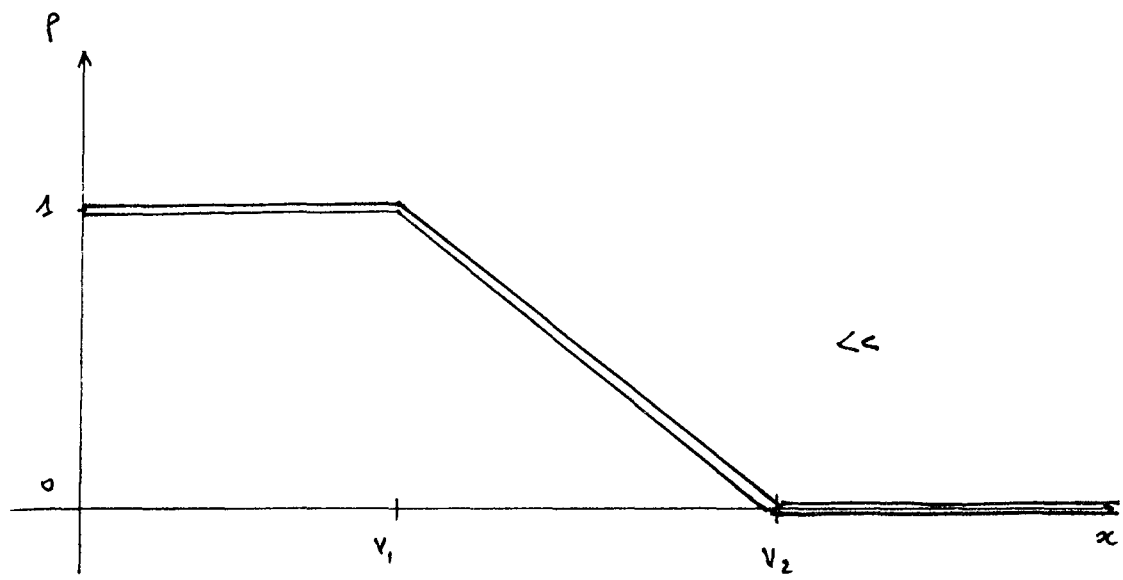
et les opérations booléennes:

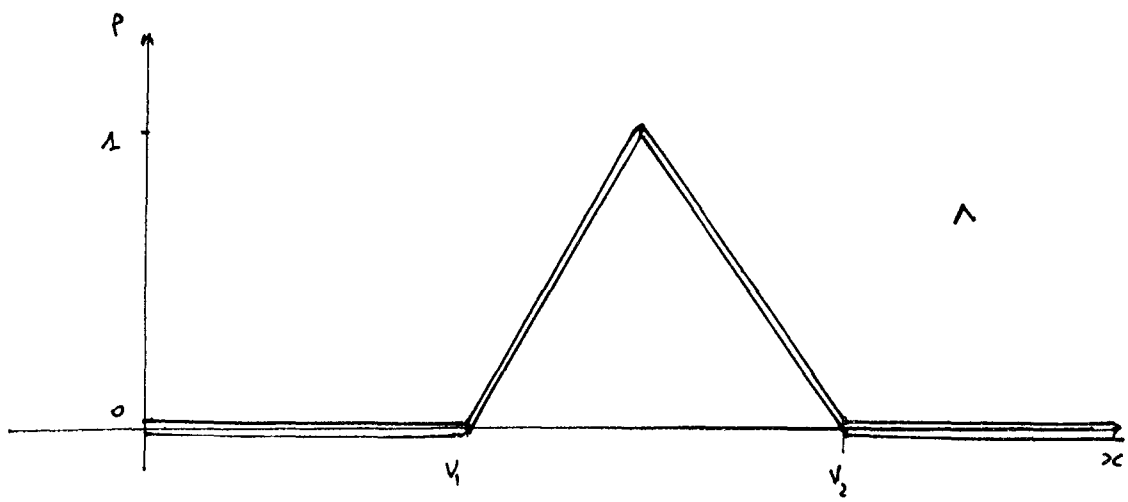
<=, <, >, >=, #, =, ''et'', ''ou''.

Peuvent apparaitre dans les prémisses, soit des constantes numériques, soit les valeurs des indices du segment en cours, du segment précédent ou du segment suivant (l'analyse contextuelle ne porte que sur le segment courant, le suivant ou le précédent). Pour savoir à quel segment appartient la variable, on la fait suivre d'un suffixe ''PRED'', ''ACT'', ou ''SUC'' respectivement pour le segment précédent,actuel ou suivant. Les seuls opérateurs booléens ne suffisent pas, car les limites données par le phonéticien ne sont pas absolues; quand il dit:''Si burst supérieur à 4000 Hz'', la règle peut s'appliquer même si le burst se situe à 3950 Hz, mais plus on s'éloigne des 4000 Hz, plus sa plausibilité diminue. Nous avons donc introduit la notion de seuil flou à l'aide d'un opérateur noté ''>>''. Par exemple: si une variable est supérieure à une valeur seuil_1, la plausibilité d'application de la règle est de 1. Si la variable est inférieure à

**Figure C.9 :** Opérateurs >> << ^



évènements acoustiques qui peuvent être présents dans un contexte donné,pour un phonème particulier, soit dédoubler ces règles pour obtenir pratiquement une règle par indice acoustique et par contexte. Dans la première approche, une règle ne pourra être déclenchée que si tous les évènements acoustiques apparaissent. Or, c'est rarement le cas, car certains indices peuvent ne pas apparaître clairement pour certains locuteurs;d'autres indices peuvent être absents à cause d'une vitesse d'élocution rapide ou d'une élocution peu soignée. Par exemple, nous avons remarqué souvent l'absence de barre d'explosion ou de transitions dans la réalisation d'un /p/. De façon générale, lorsqu'un des indices est absent, un autre est généralement bien marqué ( phénomène de compensation). Nous avons donc choisi de découper les règles, mais cette approche présente aussi des inconvénients:

- l'obtention d'un grand nombre de règles, ce qui augmente ainsi le temps de calcul et la mémoire utilisée,
- l'application de règles trop peu sélectives par exemple: la présence du premier formant vers 300 Hz va engendrer comme hypothèse les phonèmes /l/ /m/ /n/ /r/ /gn/ /w/ /j/
alors que, la présence du deuxième formant vers 1800 Hz èliminait /m/ /n/ /nj/.

Il faudra donc trouver une méthode pour éliminer les phonèmes en surplus (cf. chapitre C.3). Il est plus facile, pour les niveaux supérieurs, d'enlever des étiquettes phonétiques supplémentaires que de supposer des omissions.

## IV. Différents types de règles et exemples

### 1) Règles de phonétique générale

Ces règles traduisent des connaissances générales que possède le phonéticien.
Par exemple des règles sur le voisement:
(* /p/ /t/ /k/ sont des consonnes sourdes *)

```
R100
  SI
    pitch_ACT = 0

  ALORS [  a -9   e -9  i -9  o -9
           u -9   y -9  e -9 ai -9  oe -9
          eu -9  an -9 in -9 un -9
          on -9   b -9  d -9  g -9
           v -9   z -9 gh -9 ]
```

NB. -9 est une valeur arbitraire qui garantit que le phonème est définitivement éliminé.

Remarque: Pour exprimer que [p t k f s ch ] sont des phonèmes toujours sourds, nous allons établir une règle qui conclut que si un segment est sourd, les phonèmes toujours voisés sont à exclure.

(* règle complémentaire si le segment est voisé ce n'est pas /p/ /t/ /k/ /f/ /s/ /ch/ *)

```
R102
  CONTEXT_DROIT [ p t k f s ch ]
  SI
     vrai
  ALORS [  b -9  d -9  g -9
           v -9  z -9 gh -9 ]
```

**2) Règles de déduction**

Elles concluent sur un ou plusieurs phonèmes. Ce sont les plus courantes.

```
R25
  SI
      silence &
      burst faible
  ALORS  [ p 1 b 1 ]
```

NB. quand une règle comporte plusieurs prémisses elles sont séparées par un ''&''

```
R89
CONTEXTE_DROIT [ u ) y w & ]
  SI
     limite_friction_ACT >> (3500 5000) &
     forte friction
  ALORS [ s 1 z 1 ]
```

NB. on peut remarquer ici l'emploi de l'opérateur ''>>'' si limite_friction_ACT est inférieure à 3500 Hz, la règle ne s'applique pas,

si limite_friction_ACT est supérieur à 5000 Hz, la prémisse est vraie (plausibilité de 1),

si la limite_friction_ACT est comprise entre 3500 et 5000 Hz, la plausibilité de la prémisse est comprise entre 0 et 1.

**3) Règles de confirmation**

cet indice (transition-rapide-vers-F3-voyelle) n'est pas toujours présent et à lui tout seul, il n'est pas suffisant pour conclure sur une liste de phonèmes. Mais sa présence, avec d'autres indices pour /l/, est une confirmation.

**4) Règles de discrimination**

Elles sont utilisées pour distinguer des phonèmes très proches dans un contexte particulier

```
R80
DEJA_SUGGERE [ m n ]
CONTEXTE_DROIT [ i ]
  SI
    transition-F3-montante &
    transition-F2-montante &
    energie-visible ^ (2000 2400)
  ALORS [ m 1 n -1 ]
```

```
R81
DEJA_SUGGERE [ m n ]
CONTEXTE_DROIT [ i ]
  SI
    transition-F2-montante
    transition-F3-montante
    energie-visible ^ (2600 3000)
  ALORS [ n 1 m -1 ]
```

**5) Règles de prédiction**

Elles permettent de prédire une caractéristique du contexte droit en fonction de l'indice visible dans le segment étudié.

Ces règles permettent de prédire une liste de phonèmes possibles pour le segment suivant. Elles sont traduites sous forme négative par la procédure ''detruire ''; c'est-à-dire qu'un chemin est ouvert, dont le contexte droit est composé de l'ensemble des phonèmes supposés impossibles pour le segment suivant, et pondérés avec la valeur -9 pour indiquer l'impossibilité.

**6) Règles d'exclusion**

Ces règles indiquent une contradiction entre un ou plusieurs indices visuels et les caractéristiques acoustiques d'une famille de phonèmes.

```
R45
    SI
      non energie-visible ^ ( 800 1200 )
    ALORS [ w -9 ]

R46
CONTEXTE_DROIT [ u o oe ) ]
CONTEXTE_GAUCHE [ u o oe ) ]
    SI
       energie-visible ^ ( 1200 1500 )
    ALORS [ r -9 ]
```

**7) Règles de phonologie**

Il existe en français des contraintes sur la succession des sons. Ces contraintes sont peu nombreuses, car dans notre langue, un mot peut se terminer par pratiquement n'importe quel phonème, et le mot suivant peut commencer de même. Une exception importante concerne le phonème /w/, car il ne peut pas finir un mot:
/w/ ne peut être suivi que des sons /a/ /i/ /$\tilde{\varepsilon}$/ /$\varepsilon$/ . Donc, toutes les

règles concernant /w/ auront une partie contexte droit composée au maximum de quatre voyelles.

### 8) Règles de segmentation

Ces règles modifient la segmentation soit en fusionnant deux segments soit en scindant un segment en deux parties.
Exemple: règles d'allongement

(* si on a deux segments courts qui précèdent une des consonnes allongeantes du français (/R/ /z/ /ʒ/ /v/) alors essayer de les fusionner *)

### Conclusion

La base de connaissances a été intégrée sous la forme de règles de production. Leur syntaxe a été choisie pour qu'un expert du domaine puisse les comprendre et les modifier facilement: une partie contextuelle (non pondérée), une partie prémisse permettant l'utilisation de connaissances floues et une partie conclusion composée soit d'une action soit d'une liste de phonèmes pondérés.
Remarque: Pour accélérer les traitements ultérieurs sur la base de connaissances, les règles sont ''compilées'': les prémisses sont traduites en notation postfixée.

# CHAPITRE 3

# LE MOTEUR D'INFERENCES

## I. Le raisonnement de l'expert

On peut résumer ainsi la démarche de l'expert dans les cas où la segmentation ne pose pas de problèmes (cf partie B.1):

- analyse visuelle pour détecter un ou plusieurs indices clairs (non ambigus),
- émission d'une ou plusieurs hypothèses,
- validation, sélection ou classement des hypothèses émises,
- vérification de la compatibilité des caractéristiques acoustiques de chaque phonème hypothétisé avec les propriétés du segment observé dans le contexte identifié, et, éventuellement, avec les transitions de ce segment avec ses voisins.

Les deux principales caractéristiques de la démarche de l'expert sont:

- une analyse très fortement contextuelle et visuelle,
- la possibilité de mener plusieurs lignes de raisonnement en parallèle.

Nous devions donc concevoir un moteur capable de remplir ces deux fonctions.

## II. Le premier test

Dans un premier temps nous avons voulu tester rapidement les règles obtenues, en construisant un moteur classique écrit en Le_Lisp. Ce moteur fonctionnait uniquement en chaînage avant. N'étant pas relié au signal vocal, ce moteur posait, de manière conversationnelle, des questions sur la présence et la valeur des différents indices utilisés dans les règles. On répondait au clavier en analysant un spectrogramme.

Ce moteur, très simple, a permis de valider les règles. Mais il a surtout mis en évidence la nécessité, pour notre problème, de structures de contrôle plus complexes, permettant un raisonnement contextuel très élaboré, se rapprochant de celui de l'expert humain.

## III. Le moteur actuel

### 1) Le cahier des charges

Nous nous sommes fixés le cahier des charges suivant:

- la segmentation en unités phonétiques peut être remise en cause à tout moment,
- l'analyse doit se dérouler parallèlement sur plusieurs segmentations,
- le contexte gauche d'un segment est connu, puisqu'on travaille de gauche à droite, mais il peut être constitué de plusieurs phonèmes (on utilise plusieurs étiquettes pour un même segment) d'où l'introduction d'un indéterminisme à ce niveau,
- le contexte droit est inconnu à ce stade de l'analyse (analyse gauche-droite),
- les mesures effectuées sur un segment peuvent donner lieu à plusieurs interprétations,

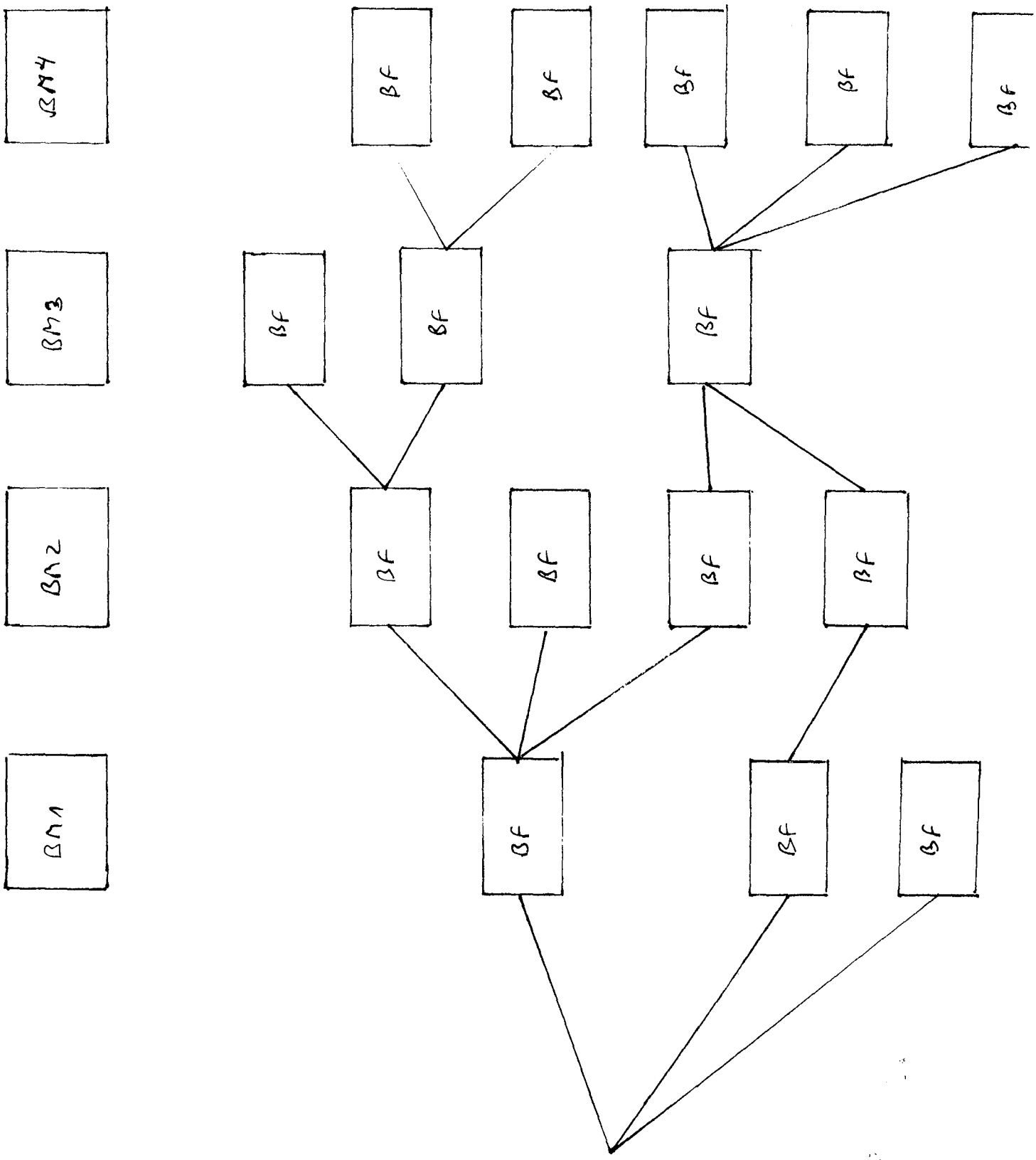
Par exemple: une analyse LPC va fournir 5 valeurs de formants, mais le deuxième ayant une forte largeur de bande, on peut penser qu'il correspond en fait à deux formants réels distincts, ce qui conduit à deux numérotations possibles pour les formants,

- les prémisses des règles ne sont pas booléennes .

Par exemple: à la question: le burst est-il faible? on peut répondre autrement que par l'affirmative ou la négative; ce raisonnement doit donc être approximatif,

- les seuils utilisés par l'expert sont toujours flous,
- le but du système consiste à ouvrir au maximum les listes de phonèmes fournies pour chaque segment pour que le phonème correct soit dans la liste, sans pour cela créer un trop fort indéterminisme pour les niveaux supérieurs,
- le moteur doit fonctionner aussi bien en chaînage avant qu'en chaînage arrière,
- le système doit pouvoir fournir une trace de son raisonnement.

Figure C.10 Exemple de treillis



## 2) Description statique

A partir des mesures acoustiques et des résultats des algorithmes de prétraitement relatif à chaque segment de l'énoncé, le moteur d'inférences construit progressivement un treillis de phonèmes représentatif de la phrase prononcée.

### a) Prétraitement et mesures

Pour chaque segment, nous disposons:

- du numéro de prélèvement de début de segment,
- du numéro de prélèvement de fin de segment,
- de la liste LPT des phonèmes trouvés par le prétraitement sur ce segment (dans le cas où aucune classe de phonèmes n'a pu être trouvée, la liste LPT contient tous les phonèmes du français),
- d'une base de mesures qui contient le nom et les résultats des procédures de traitement de signal effectuées sur ce segment.

### b) Le treillis phonétique (figure C.10)

Celui-ci est construit au fur et à mesure de l'avancement du raisonnement du moteur (gauche-droite). Le treillis se parcourt comme un arbre; chaque noeud comporte:

- le numéro du segment concerné Ns,

- une base de faits BF, cette base de faits est composée de triplets (I,V,L) où I est le nom d'un indice, V sa valeur et L une liste de couple (phonèmes, (pl_max,pl_min)). pl_max et pl_min représentent l'intervalle de plausibilité pour le phonème (intervalle initialisé à 0 0).

- la liste LHP des phonèmes hypothétisés sur le contexte droit,

- la liste des numéros des règles qui ont été activées sur ce noeud (liste initialisée à vide),

- la liste LR des phonèmes trouvés comme résultats (avec coefficients de vraisemblance).

Un noeud contient donc les résultats de l'analyse du segment Ns, son contexte gauche est constitué par la liste LR du noeud père de N. La liste des phonèmes pouvant succéder (contexte droit) au noeud N sera appelée LPH.

### 3) Fonctionnement du moteur

Rappelons que la syntaxe d'une règle s'écrit:

numero-de-règle (NR)
CONTEXTE_GAUCHE (liste-de-phonèmes) (LCG)
CONTEXTE_DROIT (liste-de-phonèmes) (LCD)
RESULTATS (liste-de-phonèmes) (LRR)
SI
conditions sur des mesures
ALORS
(liste de phonèmes conclusion) (LPC)

Nous indiquons entre parenthèses les noms que nous donnons dans la suite de ce chapitre aux différentes variables qui interviennent. Les listes suffixées par _PRED désignent les listes du père du noeud N, celles suffixées par _SUC désignent les liste du fils du noeud N.

Nous allons tout d'abord détailler le fonctionement du moteur dans le cas général.

#### a) Cas général

Nous disposons des informations suivantes pour traiter un segment donné:

- le résultat du prétraitement pour le segment suivant LPT_SUC (si le prétraitement n'a rien trouvé, la liste LPT contient l'ensemble des phonèmes du français).

- la liste des phonèmes hypothétisés par le segment précédent LHP_PRED.

**- Choix de la règle**

Pour qu'une règle soit activée, il faut qu'elle vérifie un ensemble de conditions:

- le numéro de la règle ne doit pas être contenu dans la liste des règles appliquées sur ce noeud,

- l'intersection de LPC et LPH_PRED doit être non vide (les phonèmes conclusion de la règle doivent être compatibles avec les phonèmes hypothétisés par le segment précédent),

- l'intersection de LCG et LR_PRED doit être non vide (le contexte gauche de la règle doit être compatible avec le résultat du noeud_père),

- l'intersection de LCD et LHP doit être non vide (le contexte droit de la règle doit être compatible avec le contexte supposé),

- l'intersection de LPC et LPT doit être non vide (résultats du prétraitement et conclusion de la règle),

- le minimum des plausibilités de chacune des prémisses doit être non nul (soit MAX leur maximum et MIN leur minimum).

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.

Illustrons cela sur un exemple:

Supposons que nous sommes à un noeud dont la liste LHP (contexte droit hypothétisé) 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 LHP du noeud en /d, t/ (intersection) et créer un frère dont la LCDS sera /p, b, k, g/ (soit la différence des deux).

- **application de la règle**

Si les conditions ci-dessus sont satisfaites, les règles sont activées:
* si la conclusion de la règle est une action, on exécute cette dernière;
* si la conclusion est une liste de phonèmes, le processus est le suivant:
pour tout indice acoustique i apparaissant dans les prémisses de la règle
pour tout phonème ph de la liste LPC (conclusion de la règle)
soit (pl_max,pl_min) l'intervalle de plausibilité
associé au phonème ph
mm = max (pl_max,MAX)
nn = min (pl_min,MIN)
on affecte au phonème ph dans la liste associée
à i la nouvelle valeur d'intervalle (mm,nn)

Exemple: soit la base de faits contenant uniquement
formant1 300 m(0,6 0,2) n(0,6 0,2) l(0,3 0,2)
formant2 2200 a(0,8 0.3)

et la règle (applicable):
SI
formant1 ^ (275 375) &
formant2 ^ (2000 3000)
ALORS [ i 1 m 1 1 1 ]

on obtient la nouvelle base de faits:

formant1 300 m (0,6 0,1) n(0,6 0,1) l(0,5 0,1) i(0,5 0,1)
formant2 2200 a(0.8 0.1) i(0,5 0,1) m(0,5 0,1)

- **classification des phonèmes**

Quand toutes les règles ont été appliquées sur un noeud, il faut classer les phonèmes par ordre de certitude. On cumule, pour chaque phonème, le nombre de fois où il atteint une plausibilité minimum de 0.5 (pour tous les indices). Celui qui atteint le plus grand score est placé en tête. En cas d'égalité, c'est celui qui atteint le plus grand maximum qui est retenu.

**b) Cas du premier segment d'une phrase**

Si le premier segment n'a pas été classifié durant la phase de prétraitement, on ne dispose d'aucune information pour commencer l'analyse. Dans ce cas, les règles sont utilisées en chaînage avant.

**c) Cas d'une double hypothèse de segmentation (figure C.11.)**

Lorsqu'une double hypothèse de segmentation est émise, c'est-à-dire soit par coupure d'un segment long lors du prétraitement (sous-segmentation), soit par application d'une règle de type action (sur-segmentation), une ligne de raisonnement est créée pour chaque hypothèse.
Par exemple: soit le segment P1-P2 (numéros des prélèvements de début et de fin) qui a été divisé en P1-P3, P3-P2. Une base de mesures est créée pour chacun des différents segments P1-P2, P1-P3 et P2-P3. Quand on aura analysé le segment dont les prélèvements sont immédiatement avant P1, chacune des solutions trouvées sera dupliquée. Chacun des noeuds aura 2 fils: un pour le segment P1-P2, un pour le segment P1-P3.

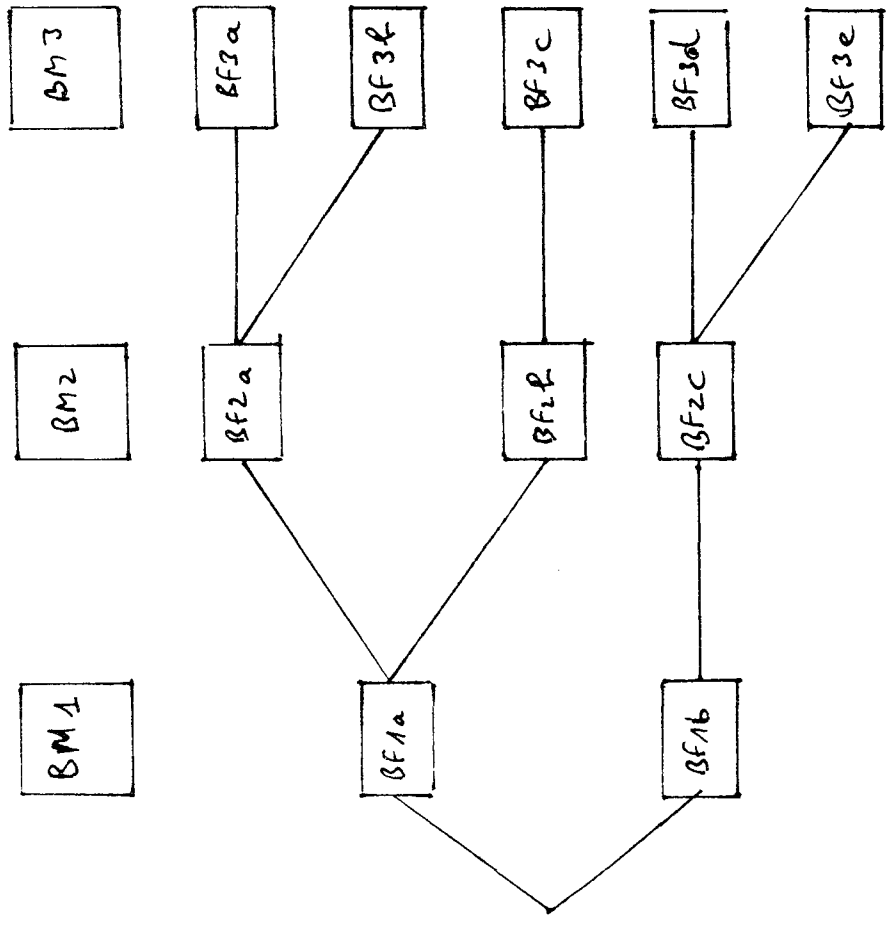
**d) Cas d'une double mesure (figure C.12.)**

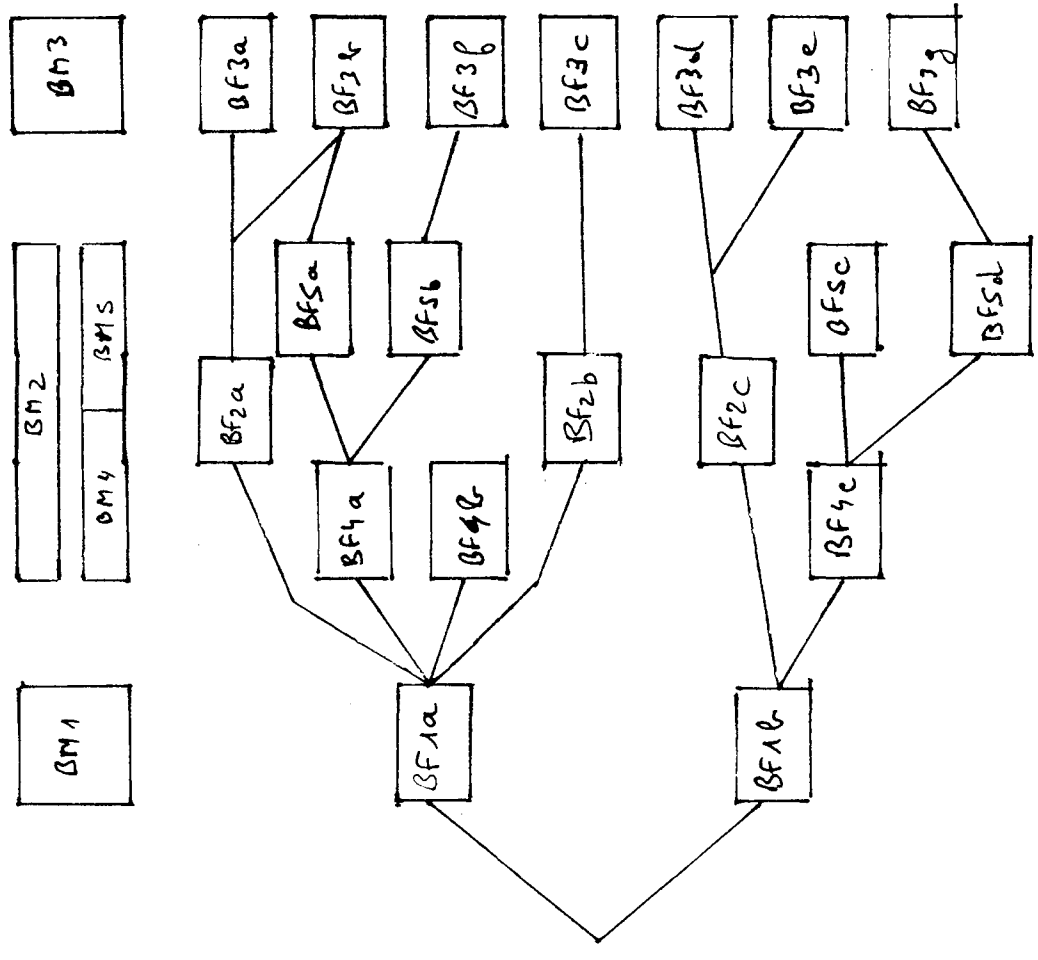
Lors de l'application d'une règle, si une procédure de calcul d'un indice fournit plusieurs résultats, par exemple numérotation ambigue de formants ou deux barres d'explosion trouvées pour une plosive, on duplique le noeud correspondant au segment étudié en autant d'exemplaires que de valeurs trouvées pour cet indice.

Treillis avant changement de la segmentation

Figure C.11

Treillis après changement de la segmentation

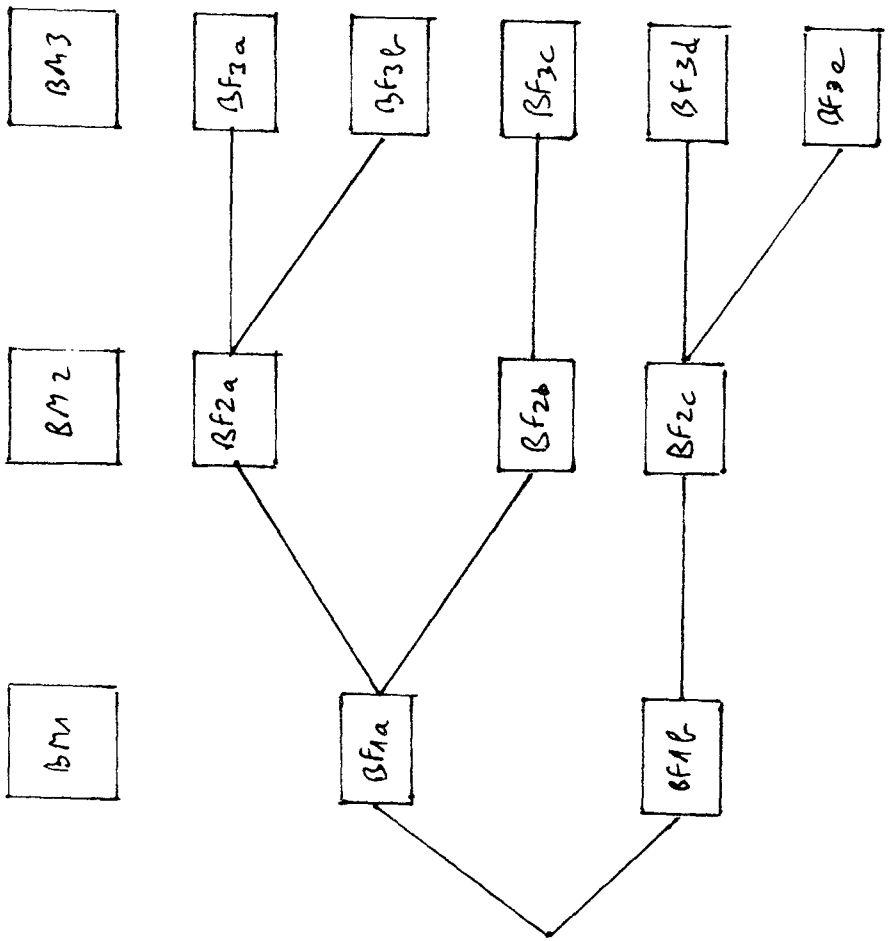


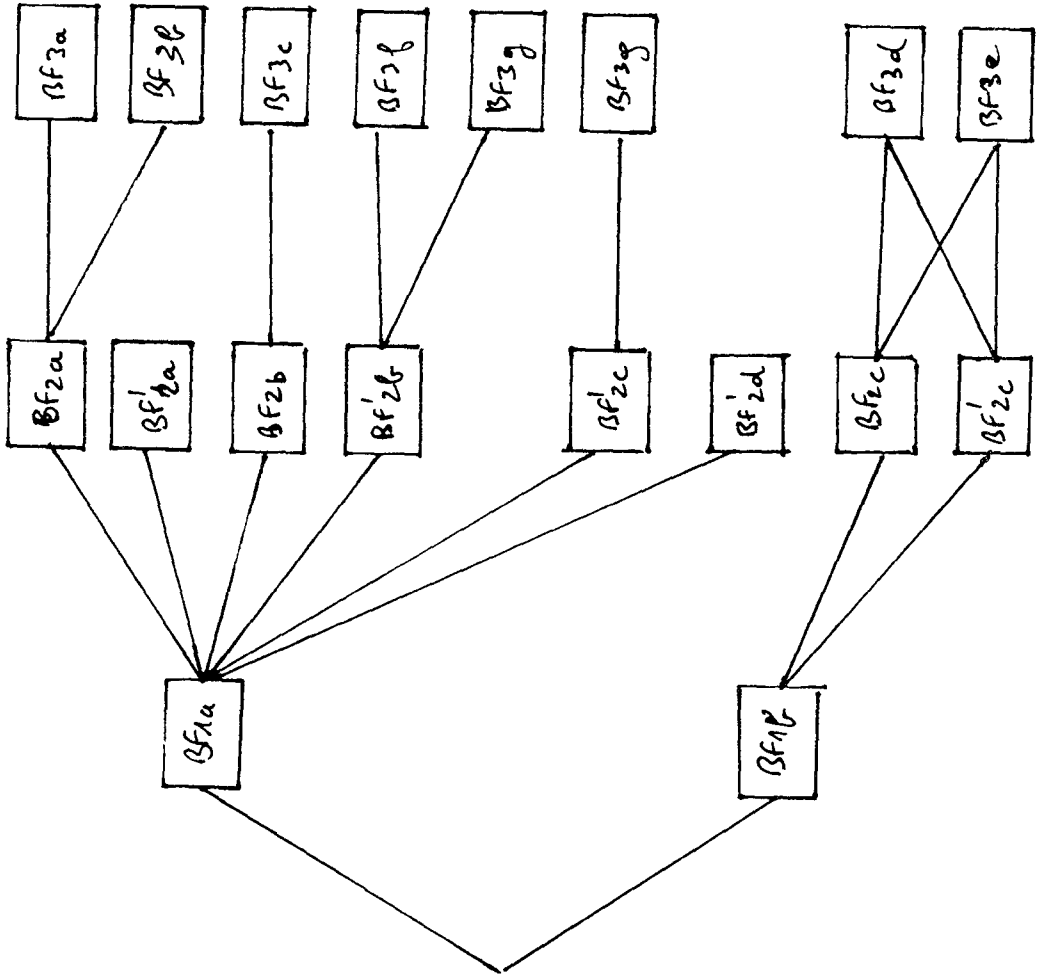



Treillis avant une double interprétation d'une mesure

Figure C.12

Treillis après une double interprétation d'une mesure



**c) Cas d'une règle de négation**

Pour une règle d'élimination de phonèmes, on affecte aux plausibilités trouvées pour les indices des coefficients négatifs. Le calcul des bornes de l'intervalle de plausibilité du phonème s'effectuera donc à l'aide de ces valeurs. Si un phonème possède une valeur MAX négative, cela signifie que ce phonème est définitivement éliminé. S'il a seulement son MIN négatif, cela signifie qu'au moins un indice infirme ce phonème.

**f) Exemple de fonctionnement**

Pour bien comprendre le fonctionnement du système, nous allons développer un exemple complet sur une phrase comprenant quatre segments et avec une base de règles réduite:

- **Base de règles:**

```
R1
CONTEXTE_DROIT [ y u oe ) w ch gh ui ]
  SI
       silence &
       présence burst_ACT &
       maximum-burst_ACT ^ (3000 4500)
   ALORS   [ t 1 ]

R2
CONTEXTE_DROIT [ i e ]
  SI
       silence &
       présence-burst_ACT &
       maximum-burst_ACT ^ (2500  3500)
  ALORS   [ k 1 ]

R3
CONTEXTE_GAUCHE  [ y u oe ) w ch gh ui ]
  SI
      friction-intense &
      limite-inf-friction_ACT ^ (2500 4000)
  ALORS  [s1 ]
```

```
R4
CONTEXTE_GAUCHE [ i e j nj an in eu ai a p t
                  k b d g f v z s m n R l ]
  SI
      friction-intense &
      limite-inf-friction_ACT ^ (2000  3500)
  ALORS  [ ch 1 ]

R5
CONTEXTE_GAUCHE [ i e ]
  SI
      friction-intense &
      limite-friction-montante
   ALORS  [ s-9  ch-9  gh-9  v-9  z-9  f-9 ]

R6
CONTEXTE_DROIT [ y u oe ) w ch gh ui ]
CONTEXTE_GAUCHE  [ i e j nj an in eu ai a p t
                  k b d g f v z s m n R l ]
   SI
       friction-intense &
       limite-friction-montante
   ALORS  [ s-9  ch-9  gh-9  v-9  z-9  f-9 ]

R7
   SI
       formant1_ACT ^ ( 583  734) &
       formant2_ACT ^ (1329 1432) &
       formant3_ACT ^ (2220 2637)
   ALORS [ a 1 ]

R8
   SI
       formant1_ACT ^ (170 280) &
       formant2_ACT >>(1700 1900)
   ALORS [ i 1  y 1 ]

R9
```

**bases de mesures**

Pour le segment 1:

LPT = [ i e ai a ) o u y an in on un oe ]
formant1 = 600
formant2 = 1400
formant3 = 2500

Pour le segment 2:

LPT = [ p t k b d g f v ]
maximum-burst = 3300
silence = vrai
presence-burst = vrai

Pour le segment 3:

LPT [ i e ai a ) o u y an in on un oe ]
formant1 = 300
formant2 = 1875

Pour le segment 4:

LPT = [ f s ch v z gh ]
limite-inf-friction = 3000
friction-intense = vrai
limite-friction-montant = vrai

Pour le segment 5:

LPT = [ i e ai a ) o u y an in on un oe ]
formant1 = 700
formant2 = 1400
formant3 = 2500

Après application de R7, R1, R2 et R8, on obtient un arbre à deux branches (cf figure C.13). On pourra noter qu'on n'a pas hypothétisé les branches /ti/ et /ku/.
Ensuite, les règles R3 et R4 s'appliquent et on obtient l'arbre de la figure C.14. Lors de l'application de la règle R5, le treillis se réduit à une seule branche a-t-y-s-a

## IV. Conclusion

Le moteur que nous avons développé correspond bien au cahier des charges que nous nous étions fixé. Il est maintenant opérationnel sur un Exormacs 68000.

Le système actuel comporte environ 200 règles. Toutes les procédures d'extraction d'indices acoustiques n'ayant pas encore été programmées

Figure C.13

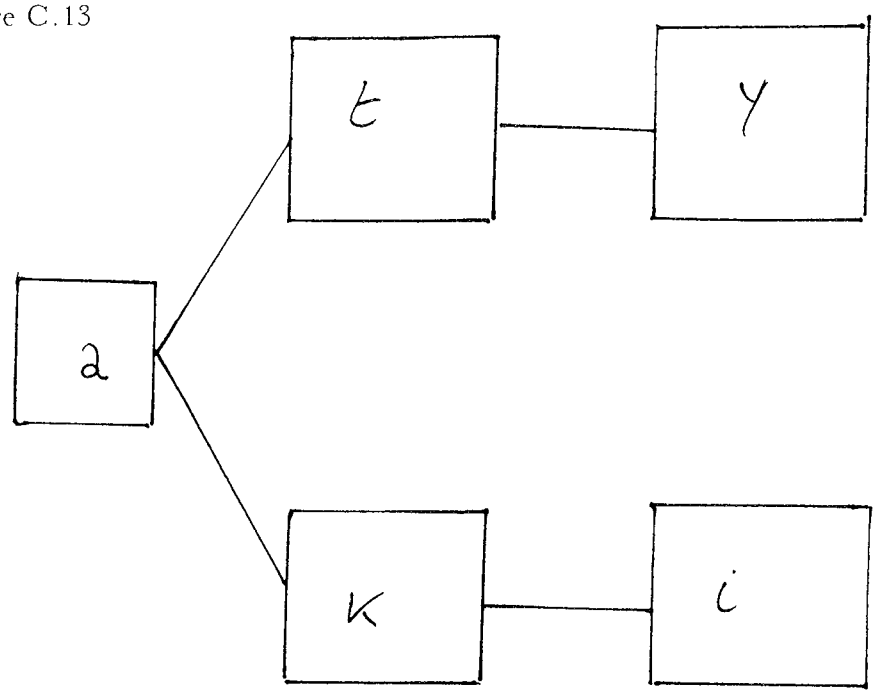
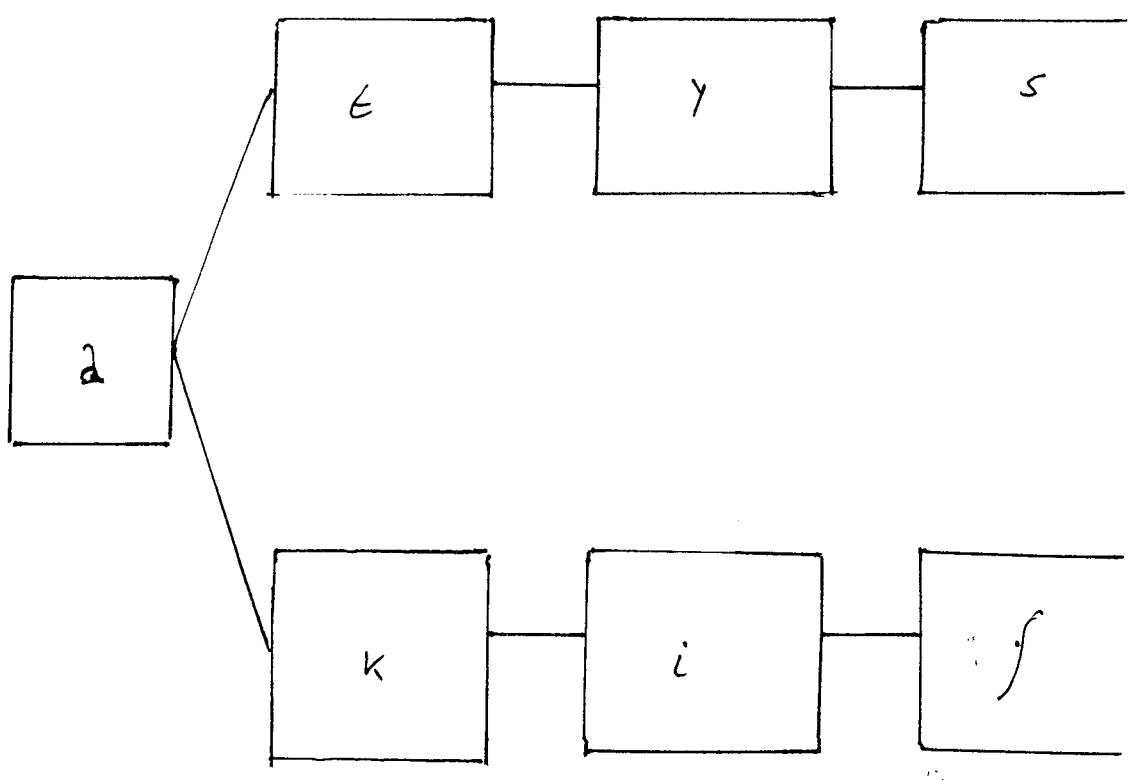



Figure C.14



lorsque le moteur d'inférences rencontre un tel indice dans la prémisse d'une règle, il demande de façon conversationnelle à l'utilisateur de donner la valeur de cet indice en analysant le spectrogramme.

# PARTIE D

# REALISATION ET RESULTATS

# INTRODUCTION

Dans cette partie, nous allons détailler les résultats de la reconnaissance. Tous les résultats ont été obtenus sur le corpus (cf C.1.III.) de parole continue en contexte multilocuteur. Dans un premier temps, l'expert a décodé ce corpus, puis il l'a manuellement étiqueté. Les pourcentages que nous donnerons dans cette partie sont tous fondés sur cet étiquetage manuel. Mais pour discuter des performances obtenues et des erreurs commises, nous les jugerons par rapport aux performances de l'expert.

Nous donnerons tout d'abord (chap 1) les scores des procédures de prétraitement: noyaux vocaliques, plosives et fricatives. Ensuite, dans le deuxième chapitre nous présenterons les résultats sur les sonantes. Enfin, nous indiquerons des taux de reconnaissance sur les plosives (chap 3).

# CHAPITRE 1

# LES RESULTATS DES PRETRAITEMENTS

## I. Les noyaux vocaliques

Le programme a été expliqué en détail dans la partie C.III. Taux de reconnaissance:

| 1ère passe | | | 2ème passe | | |
|---|---|---|---|---|---|
| Trouvés | Insertions | Omissions | Trouvés | Insertions | Omissions |
| 564 (97%) | 17 (3%) | 17 (3%) | 572 (98%) | 22 (<5%) | 9 (<2%) |

La figure D.1 montre le détail des erreurs de Novoca. ''O'' signifie omissions, ''I'' insertions, ''No'' le numéro de la phrase et ''NB'' le nombre de noyaux vocaliques présents dans la phrase. Les tests ont été effectués sur trois bandes de fréquence 250-1800, 250-2600, 250-2200. Pour chaque étude, on effectue deux passes: l'une à 4 dB, l'autre à 2 dB (valeur du seuil1). Les erreurs sont dues principalement à 6 raisons:

- omission de noyaux faibles situés à côté de consonnes très intenses et insertion de cette consonne en tant que noyau vocalique, (par exemple, dans ''toujours'' /ʒ/ plus intense que /u/ ),

- omission de /i/ presque complètement assourdi, soit en tête soit en fin d'énoncé,

- omission de phonèmes trop brefs (rythme très rapide),

- découpage de voyelles nasales en deux noyaux (importante fluctuation d'énergie au cours de la voyelle); ceci est compté comme une insertion,

**Figure D.1** : Résultats de Novoca (I: Insertions O: Omissions).

| | | 250 - 1800 Hz | | | | | | 250 - 2600 Hz | | | | | | 250 - 2200 Hz | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 4 dB | | | 2 dB | | | 4 dB | | | 2 dB | | | 4 dB | | | 2 dB | | |
| No | NB | O | I | | O | I | | O | I | | O | I | | O | I | | O | I | |
| 201 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 202 | 10 | 1 | 0 | ui | 1 | 0 | ui | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 203 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 204 | 8 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 205 | 9 | 1 | 1 | u gh | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 1 | u gh | 1 | 1 | u gh |
| 206 | 8 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 207 | 8 | 0 | 1 | dr | 0 | 0 | | 0 | 1 | dr | 0 | 2 | n dr | 0 | 1 | dr | 0 | 0 | |
| 208 | 5 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 209 | 14 | 1 | 0 | i | 1 | 0 | i | 1 | 0 | i | 1 | 0 | i | 1 | 0 | i | 1 | 0 | i |
| 210 | 5 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 141 | 14 | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r |
| 142 | 15 | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r |
| 143 | 11 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 144 | 9 | 0 | 1 | r | 0 | 1 | r | 0 | 0 | | 0 | 0 | | 0 | 1 | r | 0 | 1 | r |
| 145 | 10 | 2 | 1 | &i m | 1 | 2 | & Mn | 1 | 2 | &i m | 1 | 1 | & m | 2 | 1 | &i m | 1 | 1 | & m |
| 146 | 10 | 1 | 0 | e | 1 | 0 | e | 1 | 0 | e | 1 | 1 | e gh | 1 | 0 | e | 1 | 0 | e |
| 147 | 8 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 148 | 13 | 1 | 2 | o rr | 0 | 2 | u r | 0 | 2 | r r | 0 | 2 | u r | 0 | 2 | r r | 0 | 2 | u r |
| 149 | 10 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | |
| 150 | 13 | 0 | 1 | n | 0 | 3 | onrn | 0 | 2 | r n | 0 | 4 | nrrom | 0 | 2 | r n | 0 | 3 | onrn |
| 161 | 12 | 1 | 0 | y | 0 | 1 | m | 0 | 0 | | 0 | 1 | m | 0 | 0 | | 0 | 1 | m |
| 162 | 11 | 1 | 0 | on | 0 | 0 | | 1 | 0 | on | 0 | 1 | gh | 1 | 1 | ongh | 0 | 1 | gh |
| 163 | 11 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 1 | i |

Figure D.1 · suite

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 164 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 165 | 6 | 0 | 0 | | 0 | 1 | l | 0 | 0 | | 0 | 1 | l | 0 | 0 | | 0 | 1 | l | |
| 166 | 6 | 1 | 0 | e | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 167 | 10 | 2 | 0 | y e | 1 | 0 | y | 0 | 0 | | 0 | 0 | | 1 | 0 | e | 0 | 0 | | |
| 168 | 12 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 169 | 6 | 1 | 0 | i | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 170 | 10 | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | |
| 91 | 10 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 92 | 8 | 0 | 0 | | 0 | 0 | | 1 | 1 | gh y | 0 | 2 | gh m | 0 | 0 | | 0 | 0 | | |
| 93 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 1 | gh | 0 | 0 | | 0 | 0 | | |
| 94 | 10 | 0 | 1 | an | 0 | 1 | an | 0 | 1 | an | 0 | 1 | an | 0 | 1 | an | 0 | 1 | an | |
| 95 | 10 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 96 | 11 | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | 1 | 0 | u | |
| 97 | 10 | 0 | 0 | | 0 | 1 | m | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 1 | m | |
| 98 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 99 | 15 | 2 | 0 | i i | 1 | 0 | i | 2 | 0 | i i | 1 | 0 | i | 2 | 0 | i i | 1 | 0 | i | |
| 100 | 9 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 71 | 9 | 0 | 0 | | 0 | 1 | m | 0 | 0 | | 0 | 1 | gh | 0 | 0 | | 0 | 0 | | |
| 72 | 8 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 73 | 14 | 1 | 1 | r e | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | |
| 74 | 10 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 75 | 13 | 1 | 0 | y | 0 | 0 | | 0 | 0 | | 0 | 1 | z | 0 | 0 | | 0 | 0 | | |
| 76 | 10 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 77 | 8 | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | 0 | 0 | | |
| 78 | 11 | 1 | 0 | i | 1 | 1 | i r | 1 | 0 | i | 1 | 1 | i r | 1 | 0 | i | 1 | 1 | i r | |
| 79 | 13 | 0 | 0 | | 0 | 0 | | 1 | 0 | e | 1 | 1 | e gh | 0 | 0 | | 0 | 0 | | |

**Figure D.1** : suite

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 9 | 1 | 0 | u | 0 | 0 | | 1 | 0 | u | 0 | 0 | | 1 | 0 | u | 0 | 0 | | |
| 82 | 14 | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | 0 | 1 | r | |
| 83 | 10 | 1 | 1 | l e | 0 | 0 | | 0 | 1 | l | 0 | 1 | l | 0 | 1 | l | 0 | 1 | l | |
| 84 | 9 | 0 | 1 | l | 0 | 2 | & l | 0 | 1 | l | 0 | 1 | l | 0 | 1 | l | 0 | 1 | l | |
| 85 | 14 | 2 | 0 | i i | 1 | 0 | i | 1 | 0 | i | 1 | 0 | i | 2 | 0 | i i | 1 | 0 | i | |
| 87 | 13 | 1 | 0 | in | 0 | 0 | | 1 | 0 | in | 0 | 0 | | 1 | 0 | in | 0 | 0 | | |
| 88 | 15 | 0 | 0 | | 0 | 0 | | 0 | 0 | • . | 0 | 0 | | 0 | 0 | | 0 | 1 | n | |
| 89 | 10 | 1 | 0 | & | 0 | 0 | | 1 | 0 | & | 0 | 1 | gh | 1 | 1 | & gh | 0 | 1 | gh | |
| TOT | 581 | 26 | 15 | | 11 | 20 | | 16 | 15 | | 10 | 28 | | 17 | 17 | | 9 | 22 | | |
| Pourcent | | 4.5 | 2.6 | | 1.9 | 3.4 | | 2.7 | 2.6 | | 1.7 | 4.8 | | 2.9 | 2.9 | | 1.5 | 3.8 | | |

**Figure D.2** : Résultats sur les plosives (I: Insertions O: Omissions).

| No | Nb | 25 dB | | 27 dB | | 29 dB | | 31 dB | | 33 dB | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | O : | I | O : | I | O : | I | O : | I | O : | I |
| 201 | 6 | 1 : | 3 | 2 : | 2 | 2 : | 2 | 2 : | 0 | 2 : | 0 |
| 202 | 6 | 0 : | 1 | 1 : | 0 | 1 : | 0 | 2 : | 0 | 2 : | 0 |
| 203 | 3 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 204 | 3 | 0 : | 2 | 0 : | 2 | 0 : | 2 | 0 : | 2 | 0 : | 2 |
| 205 | 2 | 0 : | 3 | 0 : | 3 | 0 : | 3 | 0 : | 2 | 0 : | 2 |
| 206 | 2 | 0 : | 2 | 0 : | 2 | 0 : | 1 | 0 : | 1 | 0 : | 1 |
| 207 | 3 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 |
| 208 | 1 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 209 | 6 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 210 | 2 | 0 : | 0 | 0 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 |
| 141 | 4 | 0 : | 0 | 2 : | 0 | 3 : | 0 | 3 : | 0 | 3 : | 0 |
| 142 | 11 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 |
| 143 | 6 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 3 : | 0 |
| 144 | 4 | 0 : | 1 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 | 0 |
| 145 | 2 | 0 : | 3 | 0 : | 3 | 0 : | 2 | 0 : | 1 | 0 : | 0 |
| 146 | 2 | 0 : | 1 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 147 | 3 | 0 : | 1 | 0 : | 1 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 148 | 6 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 2 : | 0 |
| 149 | 4 | 1 : | 1 | 2 : | 0 | 2 : | 0 | 2 : | 0 | 2 : | 0 |
| 150 | 4 | 1 : | 0 | 1 : | 0 | 1 : | 0 | 2 : | 0 | 2 : | 0 |
| 91 | 4 | 0 : | 1 | 0 : | 1 | 0 : | 1 | 0 : | 0 | 0 : | 0 |
| 92 | 2 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 93 | 2 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |
| 94 | 2 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 | 0 : | 0 |

Figure D.2 : suite

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 95 | 4 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 | 2 | : | 0 |
| 96 | 5 | 0 | : | 1 | 0 | : | 1 | 0 | : | 1 | 1 | : | 0 | 2 | : | 0 |
| 97 | 5 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 1 | : | 0 |
| 98 | 5 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 2 | : | 0 | 2 | : | 0 |
| 99 | 7 | 1 | : | 0 | 2 | : | 0 | 2 | : | 0 | 2 | : | 0 | 3 | : | 0 |
| 100 | 4 | 0 | : | 1 | 0 | : | 1 | 0 | : | 1 | 0 | : | 0 | 0 | : | 0 |
| 161 | 6 | 0 | : | 1 | 1 | : | 1 | 1 | : | 1 | 1 | : | 1 | 1 | : | 0 |
| 162 | 4 | 0 | : | 0 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 |
| 163 | 5 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 | 1 | : | 0 |
| 164 | 5 | 1 | : | 2 | 1 | : | 2 | 1 | : | 1 | 1 | : | 0 | 1 | : | 0 |
| 165 | 4 | 0 | : | 1 | 0 | : | 1 | 0 | : | 1 | 0 | : | 0 | 1 | : | 0 |
| 166 | 2 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 |
| 167 | 3 | 1 | : | 0 | 1 | : | 0 | 2 | : | 0 | 2 | : | 0 | 2 | : | 0 |
| 168 | 5 | 0 | : | 2 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 1 | : | 0 |
| 169 | 4 | 0 | : | 1 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 |
| 170 | 4 | 2 | : | 0 | 2 | : | 0 | 2 | : | 0 | 2 | : | 0 | 2 | : | 0 |
| 71 | | | : | | | : | | | : | | | : | | | : | |
| 72 | 3 | 0 | : | 1 | 0 | : | 1 | 0 | : | 1 | 0 | : | 0 | 1 | : | 0 |
| 73 | 5 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 |
| 74 | 4 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 2 | : | 0 |
| 75 | 3 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 1 | : | 0 |
| 76 | 5 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 |
| 77 | 5 | 0 | : | 2 | 0 | : | 2 | 0 | : | 1 | 0 | : | 1 | 0 | : | 1 |
| 78 | 6 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 | 1 | : | 0 | 1 | : | 0 |
| 79 | 6 | 1 | : | 1 | 1 | : | 1 | 1 | : | 1 | 1 | : | 0 | 1 | : | 0 |
| 80 | 3 | 0 | : | 2 | 0 | : | 1 | 0 | : | 0 | 0 | : | 0 | 0 | : | 0 |

Figure D.2 : suite

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 82 | 5 | 0 | 1 | 0 | 1 | 0 | 1 | 2 | 1 | 2 | 1 |
| 83 | 3 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 84 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 85 | 5 | 0 | 1 | 1 | 1 | 1 | 0 | 2 | 0 | 3 | 0 |
| 87 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 88 | 7 | 3 | 0 | 3 | 0 | 3 | 0 | 3 | 0 | 3 | 0 |
| 89 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| total | 225 | 17 | 38 | 25 | 28 | 28 | 20 | 37 | 9 | 52 | 7 |

- insertion d'un noyau supplémentaire dans les groupes consonnantiques /bR/ /dR/ (remontée d'énergie après la plosive puis chute dans le /R/),

- insertion de quelques consonnes (/l/ /m/ /n/) très vocaliques (présence de formants très intenses qui les font ressembler à des voyelles).

## II. Les plosives

Les résultats complets sont donnés à la figure D.2. On a effectué 5 essais en faisant varier le paramètre déterminant (seuil absolu) de 25 à 33 dB. Plus ce seuil est élevé plus les insertions diminuent, mais plus les omissions augmentent. Un bon compromis semble 29 dB. Ce qui donne un pourcentage de reconnaissance de 87% avec un taux d'insertion de moins de 9%.

## III. Les fricatives

Notre corpus contenait 136 fricatives du type /z/ /ʒ/ /ʃ/ /s/. On dénombre 8 omissions et 9 insertions. Ce qui donne un taux de reconnaissance de 94% avec un taux d'insertions inférieur à 6%.
Dans les insertions ne sont pas comptabilisés:

- les fricatives /f/ et /v/,
- les /j/ et les barres d'explosion (qui présentent beaucoup d'énergie dans les hautes fréquences).

Le phonème responsable des trois quarts des insertions est le /R/; cela ne nous surprend pas, car nous savons que le /R/ ressemble à une fricative au contact d'une consonne sourde.

Parmi les 8 omissions, 7 sont dues à des /ʒ/. C'est la fricative qui pose le plus de problème, c'est celle qui présente le plus de pseudo-formants en basse fréquence. Le centre de gravité est abaissé et cette fricative se confond presque avec une voyelle (tout au moins pour sa partie basse fréquence).

# CHAPITRE 2

# LES RESULTATS SUR LES SONANTES

## I. But

Le but de cette phase consiste à discriminer trois groupes /R/ /l/ et /m,n/. C'est sur ces phonèmes que l'expert a obtenu ses plus faibles résultats. Nous avons donc essayé d'utiliser une approche différente d'un système à règles de production: une analyse statistique.

Afin de se rendre compte de la performance de cet algorithme, nous le testerons sur des phonèmes présentant des caractéristiques acoustiques voisines de celles des liquides et des nasales: /on/ /an/ /u/ /)/ et /a/.

### 1) Analyse factorielle discriminante

Elle utilise les propriétés des espaces vectoriels euclidiens pour décrire les individus et les variables. Cette technique est devenue d'une application courante grâce au développement de l'informatique et permet, entre autres, l'exploitation de grands fichiers de données. Parmi les principales applications de l'analyse factorielle discriminante, on peut citer:

- l'évaluation de la façon dont l'ensemble des variables quantitatives permet de reconstituer les groupes,
- la détermination des variables responsables de la discrimination,
- l'indication des individus non conformes au groupe auquel ils appartiennent.

L'application la plus intéressante pour notre étude sera la possibilité de classer les individus dont on connait les paramètres quantitatifs mais pas le groupe. Pour mieux lire et interpréter les résultats numériques obtenus, nous utiliserons des graphiques. L'analyse factorielle discrimante consiste, en fait, à déterminer une suite de combinaisons linéaires de variables

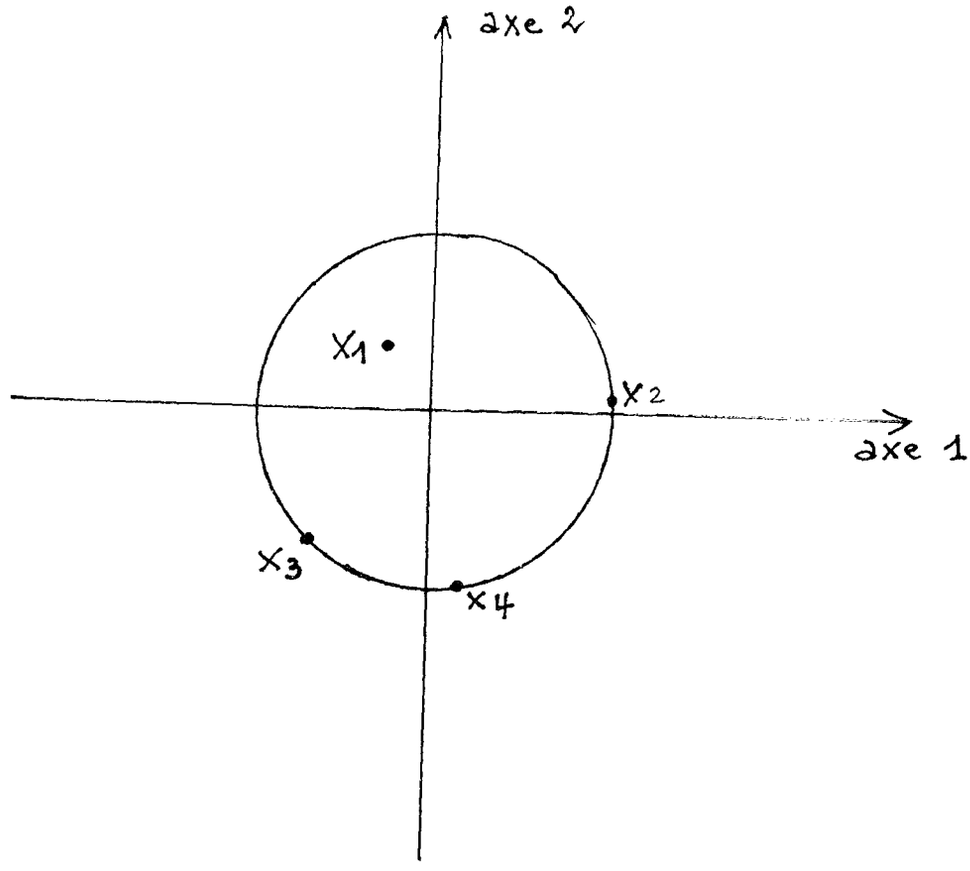



**Figure D.3 :**

| Expériences | Bandes | Larg. | Recouvr. | deb. | fin | Norm. | Eqie v |
|---|---|---|---|---|---|---|---|
| 1a | 13 | 200 | 0 | 200 | 2800 | non | non |
| 1b | 13 | 200 | 0 | 200 | 2800 | oui | non |
| 1c | 13 | 200 | 0 | 200 | 2800 | non | oui |
| 2 | 20 | 200 | 70 | 200 | 2870 | oui | non |
| 3 | 18 | 150 | 0 | 200 | 2900 | oui | non |
| 4 | 25 | 104 | 0 | 200 | 2800 | oui | non |

**Figure D.4 :**

quantitatives pour laquelle les centres de gravité de chacun des groupes sont les plus dispersés possibles. Les coefficients de ces combinaisons linéaires définissent une base orthonormée, au sens d'une métrique basée sur les covariances, du sous-espace engendré par les variables quantitatives. Chaque combinaison linéaire est appelée composante discriminante et l'axe qu'elle définit est dit axe discriminant. Pour connaître le groupe auquel un individu inconnu appartient, on calcule les distances le séparant des centres de gravité des différents groupes et il est alors classé comme appartenant au groupe dont la distance est minimale. Enfin, on représente graphiquement des projections sur différents plans définis par deux axes discriminants afin de voir comment les groupes se situent les uns par rapport aux autres, de déterminer quels sont les axes qui les discriminent le mieux et de mettre en évidence les individus mal affectés. Afin de juger de la qualité de la classification, on établira une matrice de confusion. L'étude du lien entre les variables quantitatives et la variable qualitative se fait à l'aide des composantes discriminantes. En effet, celles-ci ont été calculées pour réaliser la meilleure partition possible. Par conséquent, en calculant les coefficients de corrélation de chaque variable avec ses composantes, on obtient une information sur son apport à la discrimination: plus ce coefficient est proche de 1 pour une variable et une composante données, plus cette variable participe à la discrimination suivant l'axe correspondant à cette composante. Pour visualiser les résultats on peut utiliser un cercle de corrélation. Pour cela, on place les variables dans le plan des deux premières composantes principales; la distance d'une variable à l'origine des axes est proportionnelle à son écart type.

Dans l'exemple de la figure D.3, la variable X4 est fortement corrélée à la première composante principale tandis que la variable X2 l'est à la seconde. Pour résoudre numériquement ce calcul, il faut inventer une matrice carrée (P,P) dont le nombre d'éléments d'une ligne est égal au nombre de paramètres quantitatifs choisis. Cette opération devenant très coûteuse en temps de calcul, nous nous limiterons à une vingtaine de paramètres maximum.

**a) Ensemble d'individus**

Pour appliquer l'analyse factorielle discriminante, nous avons choisi 525 individus (empruntés au corpus de parole continue du paragraphe C.2.) répartis en neuf groupes : /R/ l/ /m/ /n/ /on/ /an/ /u/ /a/ /)/. C'est la représentation spectrale, c'est-à-dire l'énergie dans certaines bandes de fréquence, qui a été choisie comme paramètre quantitatif. Pour cette étude, la bande 200-3000 Hz semble, d'après l'expert, être la plus discriminante. Pour des raisons de temps de calcul (matrice à inverser), nous avons découpé cette bande en vingt cinq parties au maximum. Pour

calculer l'énergie, nous avons utilisé trois méthodes:

- sans normalisation, sur tout le spectre,
- avec normalisation, sur tout le spectre,
- sans normalisation, sur le spectre visible.

Pour normaliser un spectre, nous additionnons une constante à toutes les valeurs de ce spectre. Cette constante est calculée en tenant compte de l'énergie de la voyelle adjacente. Le spectre visible est le spectre constitué uniquement de l'énergie visible sur les spectrogrammes numériques. Différentes expériences ont été réalisées conformément au tableau D.4.

Nous calculons les paramètres quantitatifs pour un phonème en procédant de la manière suivante:

- nous utilisons les limites du phonème qui ont été déterminées manuellement par l'expert (cf C.2.III.),
- nous éliminons les parties fortement instables et très coarticulées que constituent le premier quart et le dernier quart du phonème. En conséquence, nous moyennons les 50 % restant au centre. Si le phonème est constitué par moins de 6 prélèvements (soit moins de 3/100 ème de seconde environ), nous faisons la moyenne des trois prélèvements centraux.
- nous calculons l'énergie dans chaque bande en faisant la moyenne des canaux compris dans l'intervalle de fréquence.

**b) Résultats** (D5)

Comme on pouvait s'y attendre, lorsqu'on considère l'ensemble des neuf groupes, les résultats sont de qualité médiocre. Une approche de type statistique (aveugle) ne permet pas de différencier des phonèmes relativement proches acoustiquement, dans un cadre multilocuteur. Cela prouve les limites de cette méthode. Par contre, dans le cas où on sait que le phonème est soit une liquide soit une nasale (/R/ /l/ /m,n/), le pourcentage de réussite peut atteindre 85%, ce qui constitue un bon résultat. Cette méthode donne de meilleurs résultats lorsqu'on veut distinguer /R/ de /l/ (95 %).

**- influence de la normalisation**

On remarque que la normalisation améliore les résultats. En effet, la normalisation élimine, dans une large proportion, les disparités énergétiques pour un même phonème; donc le travail de l'analyse discriminante se trouve facilité pour former des groupes dont les centres de gravité soient les plus éloignés possibles.

- **influence de l'énergie visible**

L'utilisation de l'énergie visible dégrade fortement les performances. Nous espérions au contraire qu'en partant des mêmes données que l'expert, nous améliorerions les résultats. Mais, en procédant ainsi, on perd de l'information et de plus, deux spectres d'allure assez proche peuvent présenter des ''spectres d'énergie visible '' très différents; en effet, si tous les deux sont à la limite du visible, mais l'un est au dessus, l'autre juste en dessous (plus rien de visible).

- **influence du nombre de bandes et de leur recouvrement**

Plus le nombre de bandes est important, meilleurs sont les résultats. Cette constatation trouve une justification logique dans le fait que plus on augmente le nombre de bandes, plus on fournit d'informations au système. Mais cette croissance semble asymptotique; au-delà de vingt bandes, les résultats varient très peu, mais le temps de calcul s'accroît très fortement.
En revanche, le fait de prendre des bandes de même largeur mais qui se recouvrent, n'augmente pas de façon significative les résultats. Signalons que le recouvrement n'apporte pas d'éléments nouveaux, mais ne fait que dupliquer les informations spectrales.

- **erreurs et confusions**

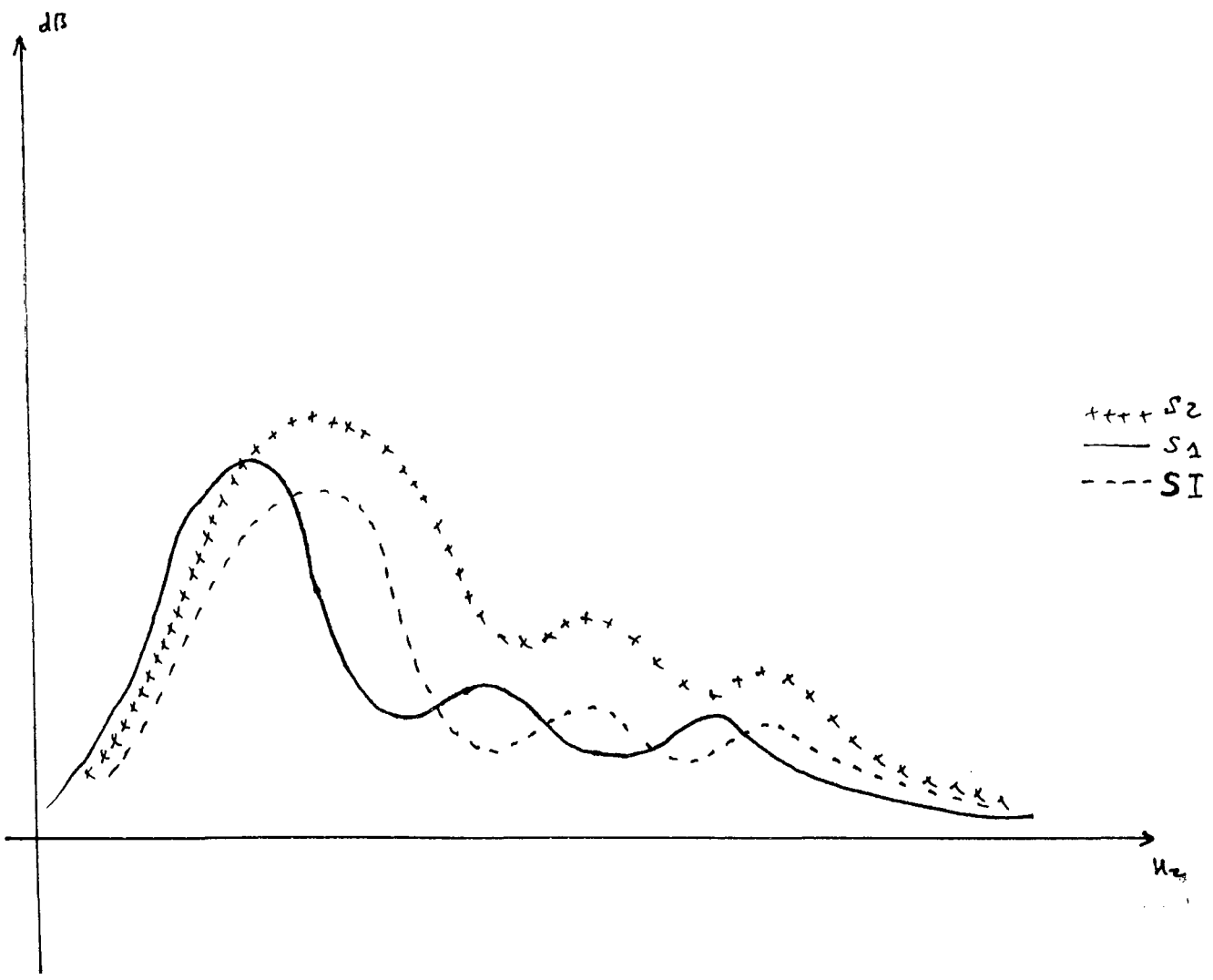
Si on détaille l'expérience 4 de l'étude No 1 (cf figure D.4), il parait intéressant de noter que les résultats sont homogènes, excepté pour le /)/. Précisons que dans le cas du /)/, c'est le petit nombre d'échantillons disponibles qui rend cette méthode inefficace. Dans le cas des voyelles, on peut constater que les groupes [ /on/ /an/ /a/ /u/ ] et [ /R/ /m/ /n/ /l/ ] ne provoquent que 42 erreurs (sur 507 individus) soit 8.3 % d'erreur. Ce ne sont pas les voyelles qui constituent la principale source d'erreurs. Les sons /m/ /n/ sont pratiquement indiscernables: pour ces deux sons, les informations contextuelles semblent indispensables pour les différencier . Les /l/ sont souvent confondus avec les /m/ ou /n/, près d'un quart des /l/ sont classés /m/ ou /n/. L'étude [ /l/ /R/ /m,n/ ] semble montrer qu'on peut séparer ces 3 groupes avec un pourcentage correct (plus de 85%).
Nous allons essayer de définir les principales bandes de fréquence qui sont utiles dans la discrimination de ces trois groupes. Sur le cercle des corrélations on remarque que les variables 5 à 8 et 13 à 17 favorisent la discrimination entre les groupes /l/ et /m,n/. Ce sont donc les bandes 800-1400 Hz et 2000-2750 Hz qui permettent de distinguer ces deux groupes. Pour les groupes /R/ et /m,n/ ce sont les variables 3, 9, 10, 11, c'est-à-dire les bandes 500-650 Hz et 1400-1850 Hz qui interviennent dans leur discrimination.

| Etudes | Expériences | 1a | 1b | 1c | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|
| r,l,m,n,o,on,an,)',a | | 56.5% | 59.5% | 50.2% | 62.2% | 63.2% | 64.4% |
| r,l,<m,n>,<on,an> | | ..... | 80.2% | ..... | 82.7% | 82.7% | 84.1% |
| r,l,<m,n> | | 82.3% | 82.3% | 80.3% | 84.1% | 83.4% | 85.5% |
| r,l,<on,an> | | 85.6% | 90.5% | 77.6% | 89.7% | 92.1% | 92.0% |
| l,<m,n> | | 88.1% | 88.1% | 85.6% | 89.3% | 88.8% | 90.4% |
| r,<on,an> | | 87.6% | 94.4% | 81.5% | 94.9% | 96.1% | 95.5% |
| r,l | | 94.7% | 93.6% | 88.3% | 94.7% | 95.2% | 95.2% |

**Figure D.5 :**

| Etudes | |
|---|---|
| r,l,e,n,u,on,an,j,a | 58.6% |
| r,l,<r,n>,<on,an> | 71.8% |
| r,l,<e,n> | 71.0% |
| r,l,<on,an> | 80.2% |
| l,<e,n> | 78.1% |
| r,<on,an> | 88.8% |
| r,l | 83.6% |

**Figure D.6 :**



## 2) Approche globale: modèles de référence

Nous voulions comparer la méthode précédente (analyse factorielle discriminante) avec une méthode classique par modèle de référence. Pour obtenir ces modèles, nous avons calculé un spectre moyen. Pour un phonème inconnu, on calcule sa distance avec chacun des modèles de référence. Il est affecté à la classe du modèle pour lequel la distance est minimum. La distance est calculée sur les 64 premiers canaux (0-3000 Hz): distance de Hamming.

Les résultats obtenus (cf figure D.6.) pour les neuf groupes sont proches de ceux de l'analyse factorielle discriminante. Quand le nombre de groupes à discriminer est faible, les résultats sont nettement moins bons (perte de l'ordre de 10%). En effet, l'analyse discriminante permet de trouver une combinaison linéaire des différentes bandes de fréquence propres à chacun des cas envisagés. Mais cette approche ne tient pas du tout compte de la forme propre de chaque phonème. Un exemple permet d'illustrer ce point: sur la figure D.6., nous avons représenté trois spectres de phonèmes; SR1 et SR2 sont deux spectres de référence des phonèmes PH1 et PH2. Pour l'ordinateur, le spectre inconnu SI est plus proche en distance euclidienne du phonème PH1. Mais pour l'expert, il est plus proche du phonème PH2. En effet, l'expert observe l'emplacement des formants et des creux dans les spectres, alors que la machine compare des spectres de façon globale sans connaître les informations pertinentes pour l'oreille (position et non pas intensité relative des formants).

# RESULTATS DU MOTEUR D'INFERENCES SUR LES PLOSIVES

## I. Conditions d'obtention des résultats

Pour obtenir les résultats sur les plosives, nous nous sommes servis de la segmentation manuelle fournie par l'expert. Nous ne voulions pas que les résultats soient perturbés par les erreurs dûes aux erreurs du prétraitement. De plus, nous ne nous sommes pas servi des procédures d'extraction des indices: suivi de formants et analyse du burst.
Nous désirerions seulement savoir si les règles étaient correctes. Quand les performances seront jugées satisfaisantes, nous utiliserons ler résultats du prétraitement et des procedures d'extraction d'indices.

## II. Résultats

Sur 20 phrases du corpus, nous avons obtenus dans 65 % des cas le bon phonème en tête. C'est bien sûr moins bien que les performances obtenues par l'expert, mais c'est un résultat qui pourra être amélioré car nous allons analyser les erreurs et, avec l'aide de l'expert, modifier ou ajouter des règles. Il semble, à première vue, que certains indices soient déterminants pour l'expert et il faudra en tenir compte dans les pondérations: burs-concentré et friction-importante. Si un indice clair est présent dans le signal, l'expert y attache une importance capitale.

L'occlusive la mieux reconnue, c'est le /t/. Cela correspond assez bien avec les taux de reconnaissance de l'expert (87%). Le phonème /d/ semble au contraire posser beaucoup plus de problème (burst moins clair).

## Conclusion

Arrivé à ce stade, l'expert va jouer un rôle très actif pour modifier la base de connaissances du système. Au besoin, nous ajouterons des stratégies nouvelles pour tenir compte des remarques de l'expert.

# CONCLUSION

# CONCLUSION

L'objectif de cette thèse était l'amélioration du décodage acoustico-phonétique pour la parole continue en contexte multilocuteur. Pour réaliser ce but, nous avons formalisé la compétence d'un phonéticien, expert en lecture de spectrogrammes, dont les performances sont nettement supérieures aux systèmes automatiques actuels.
L'analyse de l'expert permet d'enrichir les procédures de décodage:

- en augmentant le nombre et la qualité des informations acoustiques pertinentes relatives à la segmentation et à l'identification phonétique,
- en mettant en évidence des stratégies d'analyse efficaces.

Pour recueillir cette expertise, nous avons étudié le comportement de l'expert au cours du décodage de 50 phrases (parole continue, multilocuteur, rythme rapide d'élocution) phonétiquement équilibrées. Nous avons formalisé les connaissances nécessaires:

- sous forme procédurale (procédures de prétraitement),
- sous forme déclarative (règles de production).

Pour exploiter cette expertise, nous avons développé notre propre moteur d'inférences dont les principales caractéristiques sont:

- de fonctionner en chaînage avant et arrière,
- de pouvoir suivre plusieurs lignes de raisonnement,
- de ne pas séparer la phase de segmentation de celle d'identification,
- de pouvoir envisager plusieurs segmentations et de les poursuivre parallèlement,
- d'être facilement modifiable (au niveau des règles) par un expert,
- de fournir un véritable treillis phonétique (et pas simplement

une liste de phonèmes par segment),
- de prendre en compte l'incertain introduit lors de l'utilisation par l'expert de seuils de tolérance flous.

Bien que notre système soit maintenant opérationnel sur un Exormacx 68000, il reste encore un certain nombre de règles à écrire ou à modifier (sur les sonantes), seulement 200 règles étant implantées actuellement. De plus, les prémisses de ces règles contiennent des indices dont les procédures d'extraction n'ont pas encore toutes été programmées. Dans la version actuelle du système, quand on active une règle faisant appel à un indice non calculable automatiquement, le moteur pose la question à l'utilisateur qui y répond de façon conversationnelle. Autrement, la procédure de détection d'indices est activée.

Ensuite, il faudra interfacer ce module de décodage acoustico-phonétique avec le système de dialogue homme-machine qui est actuellement développé au CRIN dans le cadre du projet Dialogue (cf Carbonell et al, AFCET 85). Dans le système, le processeur de décodage acoustico-phonétique serait activé dans le cadre d'une analyse ascendante et également pour valider les hypothèses en provenance des niveaux supérieurs (approche descendante).

Le système expert à règles de production ne constitue pas pour nous la forme définitive d'un décodeur acoustico-phonétique. Ce formalisme permet d'acquérir, de façon incrémentale, l'expertise et autorise un expert du domaine à modifier lui-même la base de connaissances. Mais le décodage acoustico-phonétique exige, pour des applications industrielles ou grand public, une réponse en un temps très court, ne pouvant excéder quelques secondes. Pour cette raison, Djoudi Mahieddine (actuellement étudiant en D.E.A.) a pour tâche de compiler cette base de connaissances.

Enfin, nous étudions, avec un autre étudiant de DEA arabophone, les problèmes posés par l'extension d'APHODEX à la reconnaissance phonétique de l'arabe (cette langue présente en effet des problèmes spécifiques: pharyngalisation, gémination, nécessitant un raisonnement contextuel élaboré).

Ultérieurement, nous pensons mêler une approche déclarative et une approche par formes de référence, formes stylisées à base de traits ou spectres de référence dans différents contextes, ces spectres pouvant être générés à la demande, à l'aide d'un modèle du conduit vocal dont on modifierait la longueur ou la forme de l'onde glottale (analyse-synthèse) ou obtenus par analyse stochastique (chaînes de Markov).

# ANNEXES

```
* Les lignes qui ont une '*' comme premier caractère sont des commentaires
* Regles sur les plosives
*
*-------------------------------------------------------------------------------
*-------------------------------------------------------------------------------
* regles sur /p/
*-------------------------------------------------------------------------------
* regle generale burst et friction faible

R201
   SI
      presence-silence_ACT &
      presence-burst_ACT &
      burst-faible_ACT &
      friction-faible_ACT

   ALORS
      [ p 1 b 1 ]
FIN
*-------------------------------------------------------------------------------
* burst renforce sous F2

R202
     CONTEXTE_DROIT [ i e ai in y ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ ( formant2_SUC , formant2_SUC * 9 / 10 )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* burst renforce sous F3

R203
     CONTEXTE_DROIT [ i e y ai in ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ ( formant3_SUC , formant3_SUC * 9 / 10 )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* burst renforce sous F4

R2031
     CONTEXTE_DROIT [ y ai in ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ ( formant4_SUC , formant4_SUC * 9 / 10 )
```

```
   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* burst contine entre 1500 et 4000 Hz

R204
     CONTEXTE_DROIT [ ai in ]
   SI
     presence-burst_ACT &
     burst-continu-1500-4000

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* burst renforce entre F2 et F3

R205
     CONTEXTE_DROIT [ ai in a an o on u & oe un eu ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ ( formant2_SUC , formant3_SUC )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* burst renforce entre F3 et F4

R206
     CONTEXTE_DROIT [ ai in ) o on u & oe un eu ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ ( formant3_SUC , formant4_SUC )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
* renforcement vers F1

R207
     CONTEXTE_DROIT [ eu ]
   SI
     presence-burst_ACT &
     renforcement-burst ^ ( formant1_SUC - 100 , formant1_SUC + 100 )

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------------
```

```
* renforcement vers F2

R207
     CONTEXTE_DROIT [ y eu ]
   SI
     presence-burst_ACT &
     renforcement-burst ^ ( formant2_SUC - 100 , formant2_SUC + 100 )

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
* renforcement vers F3

R208
     CONTEXTE_DROIT [ a an y eu ]
   SI
     presence-burst_ACT &
     renforcement-burst ^ ( formant3_SUC - 100 , formant3_SUC + 100 )

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
* renforcement vers F4

R208
     CONTEXTE_DROIT [ y eu ]
   SI
     presence-burst_ACT &
     renforcement-burst ^ ( formant3_SUC - 100 , formant3_SUC + 100 )

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
* renforcement vers 4500 Hz

R208
     CONTEXTE_DROIT [ a an ]
   SI
     presence-burst_ACT &
     renforcement-burst ^ ( 4400 , 4600 )

  ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
* faible entre F2 et F3

R209
     CONTEXTE_DROIT [ ) ]
```

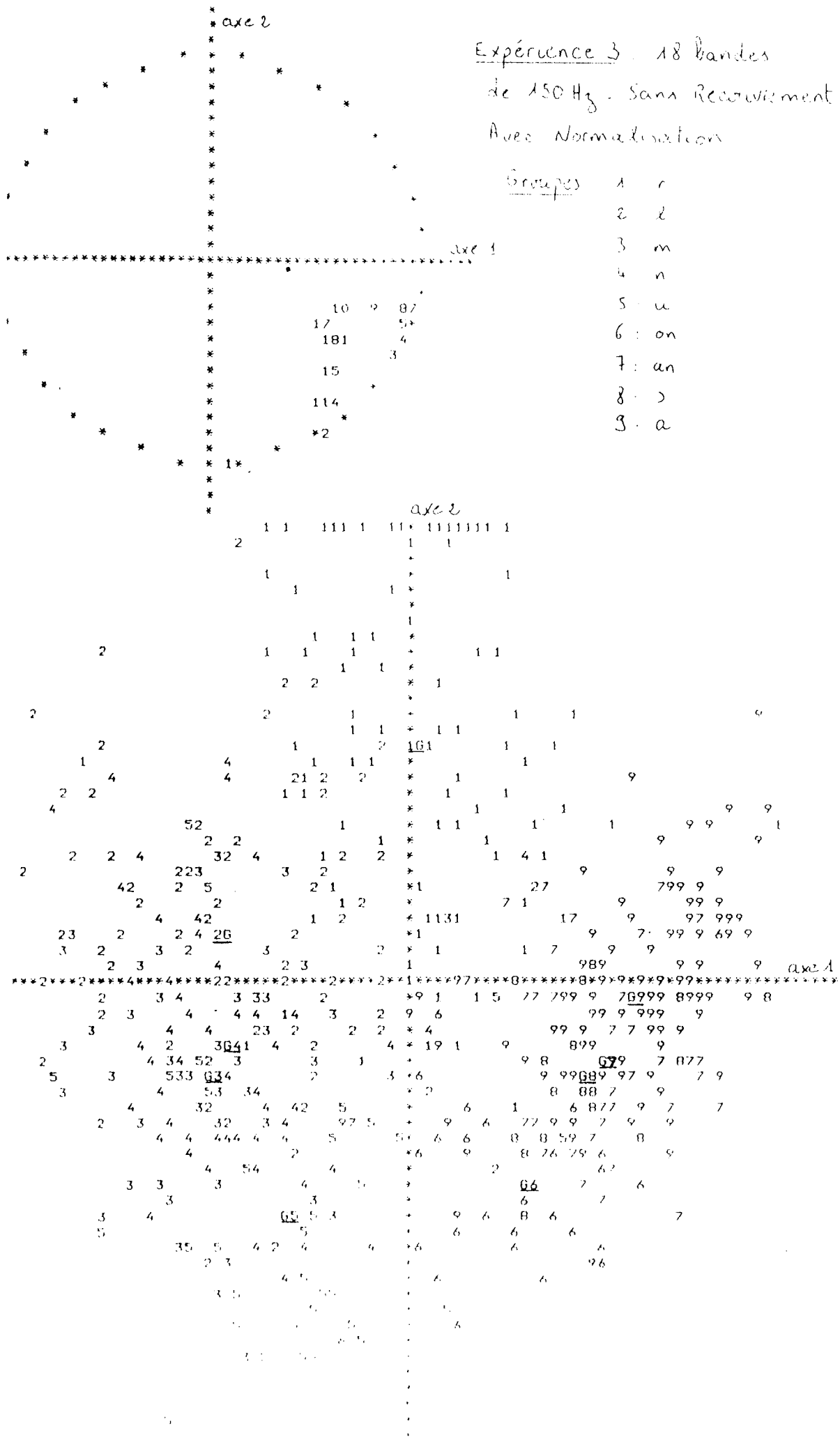
```
   SI
     presence-burst_ACT &
     renforcement-faible-burst_ACT ^ ( formant2_SUC , formant3_SUC )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
*
R214
     CONTEXTE_DROIT [ oe un ]
   SI
     presence-burst &
     renforcement-faible-burst_ACT ^ ( formant4_SUC , formant4_SUC - 100 )

   ALORS [ p 1 ]

FIN

*-------------------------------------------------------------------------
* faible sous F1

R210
     CONTEXTE_DROIT [ ) o on ]
   SI
     presence-burst_ACT &
     renforcement-faible-burst_ACT << ( formant1_SUC , formant1_SUC - 100 )

   ALORS [ p 1 ]

FIN
*-------------------------------------------------------------------------
* tres faible au dessus de 5000 Hz

R211
     CONTEXTE_DROIT [ o on ]
   SI
     presence-burst_ACT &
     rien-visible-sup-5000

   ALORS [ p 1 ]
FIN
*-------------------------------------------------------------------------
* burst continu entre 1500 et 3500 Hz

R212
     CONTEXTE_DROIT [ u ]
   SI
     presence-burst_ACT &
     burst-continu-1500-3500

   ALORS [ p 1 ]
```

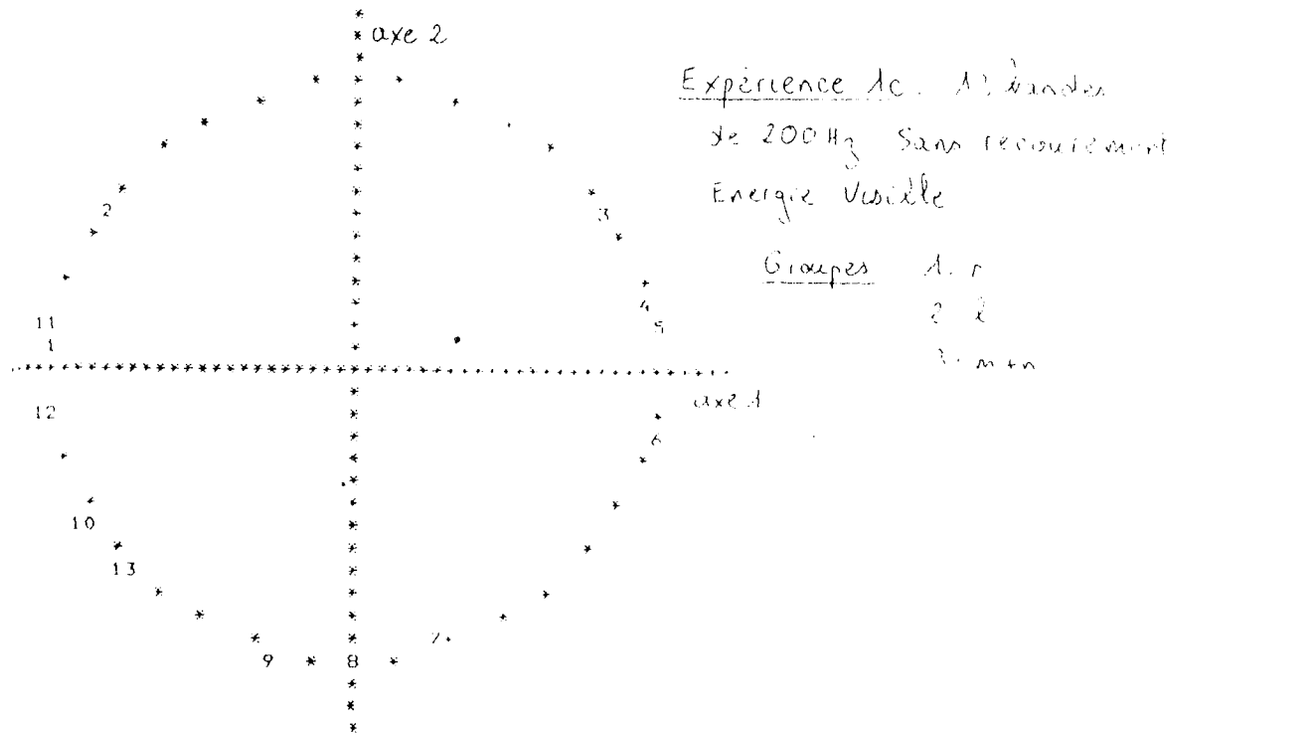
```
FIN
*-------------------------------------------------------------------------------
* burst continu entre 1500 et 4500 Hz

R213
     CONTEXTE_DROIT [ y ]
   SI
     presence-burst_ACT &
     burst-continu-1500-4500

   ALORS [ p 1 ]

FIN

*-------------------------------------------------------------------------------
*
*-------------------------------------------------------------------------------
* transitions
*-------------------------------------------------------------------------------
* transitions F1 montant

R300

    CONTEXTE_DROIT [ a ) ]
  SI
    silence_ACT &
    presence-transition &
    trans-F1-montant

 ALORS [ p 1 ]

FIN

*-------------------------------------------------------------------------------
* transitions F2 montant

R300

    CONTEXTE_DROIT [ ai a y ]
  SI
    silence_ACT &
    presence-transition &
    trans-F2-montant

 ALORS [ p 1 ]

FIN

*-------------------------------------------------------------------------------
* transitions F3 montant

R300
```

```
    CONTEXTE_DROIT [ i e ai a ]
  SI
    silence_ACT &
    presence-transition &
    trans-F3-montant

 ALORS [ p l ]

FIN

*-------------------------------------------------------------------------------
* regle de voisement

R333
    CONTEXTE_DROIT [ d v b z g gh ]
  SI
    VRAI

  ALORS [ p t k -9 ]


*-------------------------------------------------------------------------------
*-------------------------------------------------------------------------------
* regles sur /b/
*-------------------------------------------------------------------------------
* renforcement vers 2.5 KHz
*-------------------------------------------------------------------------------
* burst continu entre 1500 et 4500 Hz

R300
     CONTEXTE_DROIT [ i ]
   SI
     presence-burst_ACT &
     renforcement-burst_ACT ^ [ 2500*9/10, 2500*11/10 ]

   ALORS [ b l ]

FIN

*-------------------------------------------------------------------------------
* renforcement entre F3 et F4

R301
    CONTEXTE_DROIT [ i a ]
  SI
    presence-burst &
    renforcement-burst ^ ( formant3_SUC formant4_SUC )

  ALORS [ b l ]
FIN

*-------------------------------------------------------------------------------
```

```
* renforcement vers 2KHz

R302
    CONTEXTE_DROIT [ i ]
  SI
    presence-burst &
    renforcement-burst ^ ( 2000*9/10 , 2000*11/10 )

  ALORS [ b l ]

FIN
*-------------------------------------------------------------------------------
* renforcement vers F2

R302
    CONTEXTE_DROIT [ e a ) o on u y eu ]
  SI
    presence-burst &
    renforcement-burst ^ ( formant2_ACT * 9 / 10 , formant2_ACT * 11 / 10)

  ALORS [ b l ]

FIN
*-------------------------------------------------------------------------------
* renforcement vers F3

R302
    CONTEXTE_DROIT [ e a ) o on y ]
  SI
    presence-burst &
    renforcement-burst ^ ( formant3_ACT * 9 / 10 , formant3_ACT * 11 / 10)

  ALORS [ b l ]

FIN
*-------------------------------------------------------------------------------
* renforcement vers F4

R302
    CONTEXTE_DROIT [ y ]
  SI
    presence-burst &
    renforcement-burst ^ ( formant4_ACT * 9 / 10 , formant4_ACT * 11 / 10)

  ALORS [ b l ]

FIN

*-------------------------------------------------------------------------------
* renforcement vers F2

*-------------------------------------------------------------------------------
```

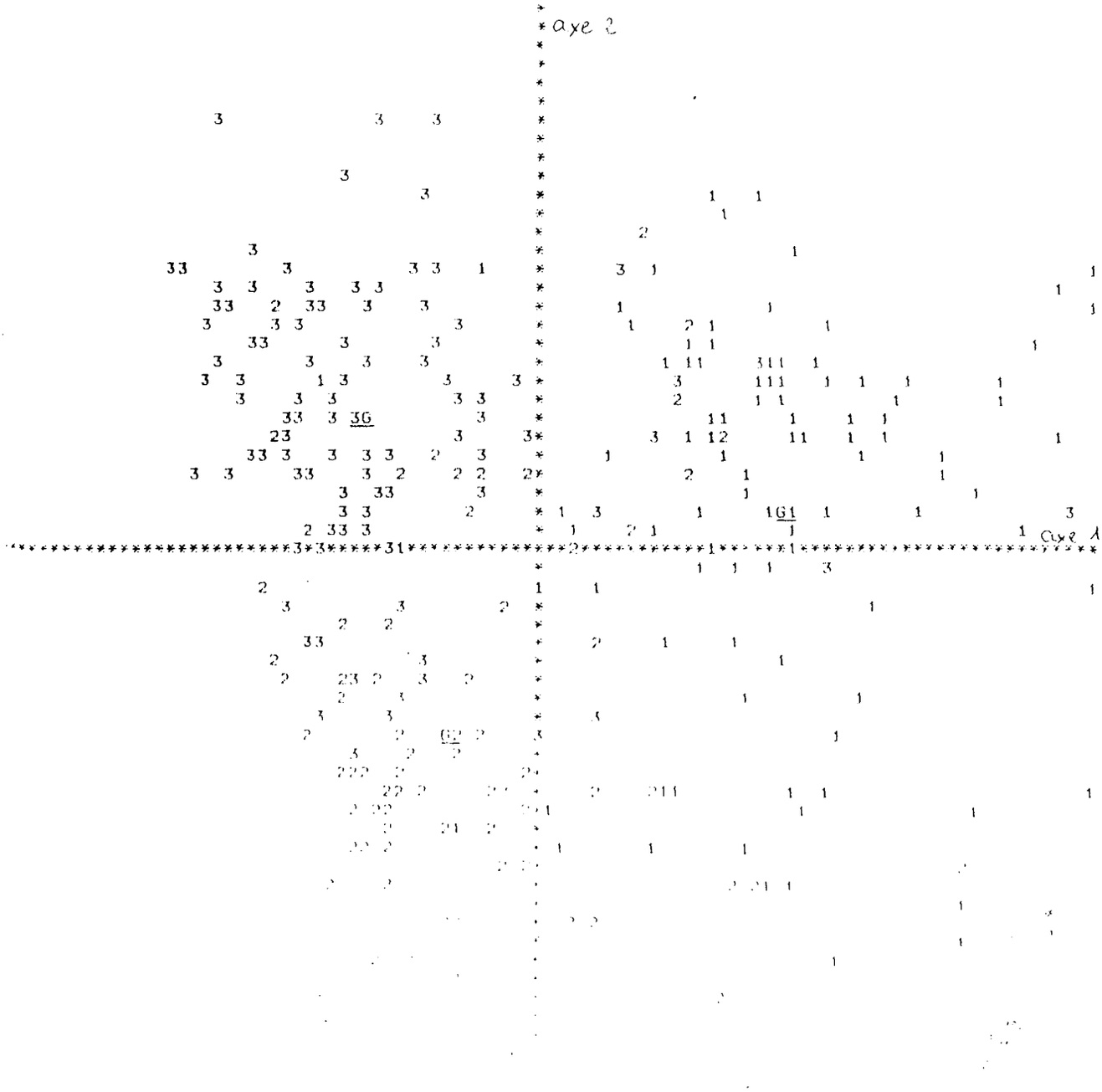
```
*-------------------------------------------------------------------------------
* regles sur /l/
*-------------------------------------------------------------------------------
* regles sur Fl

R301
      CONTEXTE_DROIT [ i e o u y ]
   SI
      concentration-energie_ACT ^ ( 220 280 )

   ALORS
      [ l ]
FIN

*-------------------------------------------------------------------------------
* regles sur Fl

R302
      CONTEXTE_DROIT [ ai ) eu oe un & ]
   SI
      concentration-energie_ACT ^ ( 270 330 )

   ALORS
      [ l ]
FIN
*-------------------------------------------------------------------------------
* regles sur Fl

R303
      CONTEXTE_DROIT [ a ]
   SI
      concentration-energie_ACT ^ ( 320 380 )

   ALORS
      [ l ]
FIN
```


axe 2
axe 1
Expérience 3 : 18 bandes
de 150 Hz . Sans Recouvrement
Avec Normalisation
Groupes
1 : r
2 : l
3 : m
4 : n
5 : u
6 : on
7 : an
8 : ɔ
9 : a
axe 2
axe 1




Expérience 1c. 12 bandes de 200 Hz Sans recouvrement
Energie Visible

Groupes 1. r
2 2
3. m+n



Phrases équilibrées

No 71 : Il s'est réfugié dans ma chambre.
No 72 : Le troupeau s'abreuvait au ruisseau.
No 73 : Le client s'attend à ce que vous fassiez une réduction.
No 75 : Une rançon est exigée par les ravisseurs.
No 76 : Ainsi cette comédie est en un acte.
No 77 : Papa aime mon vin quand il est bon.
No 78 : Le ciel est tout noir, il va tomber des cordes.
No 79 : On dit que l'essor de ce village est important.
No 82 : Le chevrier a corné pour rassembler son troupeau.
No 84 : L'oie est dans sa main, son coeur bat et saute.
No 85 : Une rivière dessinait des méandres dans sa prairie.
No 88 : Je me suis entretenu avec l'institutrice de ma jeune soeur.
No 89 : Quand le soleil se lève, je sors de mon lit.
No 91 : L'été tout le monde se mettait aux fenêtres.
No 92 : Le cocher a fouetté sa jument.
No 93 : Je rends souvent visite à mon oncle.
No 94 : Ma soirée se passera sans incident.
No 95 : La police veut les papiers du chauffeur.
No 96 : Jean, quant à lui, est très grand pour son âge.
No 97 : Le microscope qui est sur pied, est le mien.
No 98 : Le jardin entoure un petit lac.
No 99 : Il a broyé du noir depuis la perte de son ami.
No 100: Le forçat s'est évadé du bagne.
No 141: On entend les gazouillis d'un oiseau dans le jardin.
No 142: La barque du pêcheur a été emportée par une tempête.
No 143: Ce livre provient de la bibliothèque.
No 144: J'en conclus qu'il n'y a rien à voir.
No 145: Le mal s'envenime faute de soins.
No 146: je suis sûr que vous connaissez ces noms.
No 147: Il s'arrêtait tout l'été ici.
No 148: Voilà toujours deux choux pour le repas de midi.
No 149: Les manches de son manteau sont décousues.
No 150: Ce vaisseau parcourt les mers à travers le monde.
No 161: Une société de musique va bientôt défiler.
No 162: Le juge veut prolonger l'interrogatoire.
No 163: Ici, ma mère a acheté des coupons de tissu.
No 164: Pierre cogne par derrière comme un sourd.
No 165: La pluie ne fait pas le beau temps.
No 166: Sans fleurs, la maison est triste.
No 167: Elle a vraiment toujours des doigts menus.

No 168: Ce boucher n'a encore plus de lard à l'étalage.
No 169: Confie moi à quoi tu penses.
No 170: Ce dont nous discutons vous laisse rêveur.
No 201: Est-ce que le conducteur arrête l'auto?
No 202: C'est toujours comme ça depuis dix ans, tu sais.
No 203: Ce cheval ne veut pas marcher au pas.
No 204: La bière est moins forte que le rhum.
No 205: Ici, il fait toujours très froid en hiver.
No 206: J'aime Sylvie quand elle est mignonne.
No 207: Diane ne reviendra pas avant lundi.
No 208: Aimez vous ce dessin?
No 209: J'ai déjà lu la réponse qu'il m'a envoyée par la poste.
No 210: Mes gants sont usés.

# BIBLIOGRAPHIE


[Bellman 57] R. Bellman : ´Dynamic Programming´, Princeton, N.F. :Princeton, Univ. Press, 1957.

[Boe 78] Boe J.L. ´Anatomie et physiologie de la phonation, une introduction´

[Boe 80] Boe J.L., Abry C. & Corsi P. ´Les problèmes de normalisation interlocuteurs. Mèthodes d'ajustement aux limites´, 11 èmes JEP, Strasbourg, 1980

[Boyer 84] A. Boyer, ´Etude d'algorithmes de reconnaissance de mots enchainès´, Rapport de D.E.A., Universitè Nancy I, 1984.

[Bridle 82] J.S. Bridle, MD. Brown, R.M. Chamberlain : ´An algorithm for connected word recognition´ Proc 1982 IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, France, pp. 899-902, May 1982.

[Callec 82] Callec A., Monne S., Querre M., Travarain O. & Mercier G. ´Automatic segmentation of phonetic units and training in the KEAL speech recognition system´, IEEE ICASSP 82, Paris, pp 2000-2003, 1982

[Carbonell 82] N. Carbonell, J.P. Haton, F. Lonchamp, J.M. Pierrel : ´Elaboration experimentale d'indices prosodiques pour la reconnaissance; Application a l'analyse syntaxico-sèmantique dans le système Myrtille II´, Sèminaire GRECO-GALF ´Prosodie et reconnaissance automatique de la parole´, Aix, 1982.

[Carbonell 84] N. Carbonell, M.O. Cordier, D. Fohr, J.P. Haton, F. Lonchamp, J.M. Pierrel : ´Acquisition et formalisation du raisonnement dans un système-expert de lecture de spectrogrammes vocaux´, Colloque ARC, Orsay 1984.

[Carbonell 84] N. Carbonell, D. Fohr, J.P. Haton, F. Lonchamp, J.M. Pierrel : ´An Expert System for the Automatic Reading of French Spectrograms', IEEE ICASSP 84, San Diego, March 84.

[Carbonell 83] Carbonell N., Haton J.P., Lonchamp F. & J.M. Pierrel ´Elaboration d'un système expert pour le dècodage phonètique de la parole´, Speech Communication, No 2-3, pp 231-233, 1983

[Carbonell 85] Carbonell N., Damestoy J.P., Fohr D., Haton J.P., Lonchamp F. & Pierrel J.M. ´Techniques d'intelligence artificielle en dècodage acoustico-phonètique´, 14 èmes JEP, Paris, pp 299-303, 1985

[Carre 84] Carrè R., Descout R., Eskènazi M., Mariani J. & Rossi M. ´The french language database : defining, planning and recording a large database´, IEEE ICASSP 84, San Diego, 1984

[Carton 74] Carton F. ´Introduction à la phonètique du français´, Collection ETUDES No 303, Bordas, pp 22-25, 1974

[Charpillet 85] F. Charpillet, J.P Haton, J.M Pierrel : ''Un système de reconnaissance de parole continue pour la saisie de textes lus'' 5ème congrès AFCET R.F.I.A., Grenoble, novembre 1985.

[Colmerauer 83] A. Colmerauer, H. Kanoui, M. Van Caneghem : ´Prolog, bases thèriques et dèveloppements actuels´, TSI, vol. 2, No 4, 1983, pp 271-311

[Combescure 81] P. Combescure : ´Vingt listes de dix phrases phonètiquement èquilibrèes´, Revue d'acoustique, 14, n 56, 1981.

[Damestoy 85] J.P. Damestoy, J.P. Haton : ´A Frame Language for the Control of Phonetic Decoding in Continus Speech Recognition´, IEEE ICASSP 85, Tampa, Florida, March 85.

[Delattre 68] Delattre P. 'From acoustic cues to distinctive features', Phonetica, Vol. 18, pp 198-230, 1968

[De Mori 76] De Mori R., Laface P. & Piccolo E. 'Automatic detection and description of syllabic features in continuous speech', IEEE Trans. ASSP October, pp 365-379, 1976

[De Mori 83] R. De Mori : 'Extraction of Acoustic Cues Using a Grammar of Frames', Speech Communication, n 2-3, 1983.

[Di Martino 84] J. Di Matino : 'Contribution à la reconnaissance globale de la parole : Mots isolès et mots enchaînès', Thèse de docteur ingènieur en informatique, Universitè de Nancy I, avril 1984.

[Feigenbaum 71] E. Feigenbaum, 'On generality and problem solving' Machine Intelligence, 6.

[Fohr 85] Fohr D., Haton J.P., Lonchamp F. & Sauter L. 'Mèthodes de segmentation syllabique en reconnaissance de la parole', 14 èmes Journèes d 'Etude sur la Parole, Paris, pp 164-167, 1985

[Fohr 85] N. Carbonell, J.P. Damestoy, D. Fohr, J.P Haton, F. Lonchamp, J.M. Pierrel : 'Techniques d'intelligence artificielle en dècodage acoustico-phonètique' 14 ième Journèes d'Etudes sur la Parole, Paris, juin 85, pp 299

[Fant 73] Fant G. 'Speech sounds and features', MIT Press, pp 32-83, 1973

[Fujimura 62] Fujimura O. 'Analysis of nasal consonants', JASA 34, pp 1865-1875, 1962

[Fujimura 75] Fujimura O. 'The syllabe as a unit of speech recognition', IEEE Trans. ASSP-23, pp 82-87, February 1975

[Ganascia 85] Ganascia J.G. 'La conception des systèmes experts', La Recherche, No 170, pp 1142-1151, 1985

[Gillet 84] Gillet & All 'SERAC : un système expert en reconnaissance acoustico-phonètique', 4 ème congrès AFCET-RFIA, Paris, 1984

[Green 84] Green P.D. & Wood A.R. 'Knowledge-based speech understanding : towards a representation approach', Proc. ECAI, 1984

[Gubrynowicz 82] Gubrynowicz R. 'Application de la thèorie des sous-ensembles flous à l'analyse et la reconnaissance automatique de la parole', Rapport DT/LAA/TSS/RCP/101, CNET Lannion, 1982

[Haton-79] Haton J.P. & Lièнard J.S. 'La reconnaissance de la parole', La Recherche, No 99, 1979

[Haton 74] J.P. Haton : 'Une mèthode dynamique de comparaison de chaîne de symboles de longueurs diffèrentes : application à la recherche lexicale', CRAS, sèrie A, 278, pp 1527-1530.

[Haton 85] J.P. Haton : 'Techniques d'intelligence artificielle en comprèhension de la parole et vision : ètat des recherches' T.S.I.

[Itakura 75] F. Itakura : 'Minimum Production Residual Principle Applied to Speech Recognition', IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 67-72, February 1975.

[Jelinek 82] F. Jelinek, R.L. Mercer, L.R. Bahl : 'Continuous Speech Recognition: statistical methods' C.S.R. group, IBM T.J. Watson Research Center, Yorktown Heigths NY 10598

[Johanssen 83] Johanssen J., Mac Allister J., Michaele T. & Ross S. 'A speech spectrogram expert', IEEE ICASSP 83, Boston, 1983

[Johnson 84] Johnson S.R., Connoly J.H. & Edmonds E.A. 'Spectrogram analysis : a knowledge-based approach to automatic speech recognition', Proc. of Expert Systems 84, CUP, 1984

[Kaufman-73] Kaufman A. 'Thèorie des sous-ensembles flous', Masson, Paris, 1973 à 1977

[Laface 80] Laface P. & De Mori R. 'Use of fuzzy algorithms for phonetic and phonemic labelling of continuous speech', IEEE Trans. PAMI, Vol. 2, pp 136-148, 1980

[Lauriere 82] Laurière J.L. 'Reprèsentation et utilisation des connaissances', TSI, No 1, pp 25-42, 1982

[Lazrek 83] M. Lazrek : ´Decodage acoustico-phonètique en comprèhension automatique de la parole continue´, Thèse de 3 ième cycle, Universitè de Nancy I, 1983.

[Lazrek 84] M. Lazrek, J.P Haton : ´Segmentation et identification des phonèmes dans un système de reconnaissance automatique de la parole continue´, Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle,Paris, janvier 1984, pp 5.

[Lemoine 85] Lemoine E. ´Recherche de paramètres discriminatoires pour les consonnes L, R, M et N dans un ènoncè de parole continue´, Rapport de DEA, Universitè de Nancy 1, 1985

[Lesser 75] V.R. Lesser et al. : ´Organisation of the HEARSAY II Speech Understanding System´ IEEE Trans. ASSP, 23, p 11-23, 1975

[Levinson 83] S.E. Levinson, L. R. Rabiner, M. M. Sondhi : ´Speaker Independent Isolated Digit Recognition Using Hidden Markov Models´, Proc 1983, IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, pp. 1049-1052, April 1983.

[Lonchamp 85] F. Lonchamp ´Reading Spectrograms : the view of the expert´

[Lonchamp 85] F. Lonchamp ´Les sons du français, analyse acoustique desciptive´

[Lowerre-76] Lowerre B.T. ´The HARPY speech recognition system´, PH. D. Thesis, Carnegie Melon University, Pittsburg PA, 1976

[Lowerre-77] Lowerre B.T. ´Dynamic speaker adaptation in the HARPY speech recognition system´, IEEE ICASSP-77, Hartford, pp 788-790, 1977

[Mariani 81] Mariani J. ´Reconnaissance de la parole continue par diphones´, Actes du sèminaire GRECO-GALF ´Processus d'encodage et de dècodage phonètiques´, Toulouse, 1981

[Mariani 81] J. Mariani : ´ESOPE : Un Système de comphrèhension de la parole continue´ Thèse de doctorat ès Sciences, Universitè d'Orsay

[Meloni 82] Meloni H. ´Contribution à la recherche sur la reconnaissance automatique de la parole continue´, Thèse de docteur d'ètat es sciences, Universitè d'Aix-Marseille 2, 1982

[Memmi 84] Memmi D., Eskènazi M., Mariani J. & Stern P. ´SONEX : système expert en lecture de spectrogrammes´, Rapport ATP Intelligence artificielle, 1984

[Mercier 82] Mercier G. ´Acoustic-phonetic decoding and adaptation in continuous speech recognition´, in ´Automatic Speech Analysis and Recognition´, J.P. Haton Ed., Reidel, pp 69-99, 1982

[Mermelstein-75a] Mermelstein P. ´A phonetic-context controlled strategy for segmentation and phonetic labelling of speech´, IEEE Trans. ASSP-23, pp 79-82, February 1975

[Mermelstein-75b] Mermelstein P. ´Automatic segmentation of speech into syllabic units´, JASA 58, pp 880-883, 1975

[Myers 81] C. Myers, L.R. Rabiner : ´A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition´ IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP_29, n[ 2, April 1981.

[Nakagawa 83] S. Nakagawa : ´A Connected spoken word Recognition Method by o(n) Dynamic Programming Pattern Matching Algorithm´, Proc 1983, IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, pp. 296-299, April 1983.

[Neel 83] Neel F., Eskènazi M. & Mariani J. ´Cadrage automatique pour la constitution de dictionnaires d'entitès phonètiques´, Speech Communication, No 2, pp 193-195, 1983

[Perennou 82] G. Perennou, G. Caelen : ´Utilisation de la prosodie pour la reconnaissance de la parole continue dictèe´, Seminaire ´prosodie et reconnaissance automatique de la parole´, Aix, 1982.

[Perennou 81] Perennou G. & Decalmes M. ´Le dècodage au niveau phonologique dans ARIAL II´, Actes du sèminaire GRECO-GALF ´Processus d'encodage et de dècodage phonètiques´, Toulouse, 1981

[Pierrel 75] J.M. Pierrel : ´Contribution à la compréhension automatique du discours continu´. Thèse de 3ème cycle, Université de Nancy I, 1975.

[Pierrel 81] J.M. Pierrel : ´Etude et mise en oeuvre de contraintes liguistiques en comprèhension automatique du discours continu´. Thèse d'état, Université de Nancy I, 1981.

[Pister 84] C. Pister : ´ Adaptation au locuteur par apprentissage automatique : application à un système de reconnaissance automatique de la parole´, Doctorat de 3 ième cycle en informatique, Université de Nancy I, 1984.

[Rabiner-76] Rabiner L.R., Cheng M.J., Rosenberg A.E. & Mac Gonegal C.A. ´Comparative performance stydy of several pitch detection algorithms´, IEEE Trans. ASSP-24, vol 1, pp 399-417, October 1976

[Reddy 73] D.R. Reddy, L.D. Erman, R.D. Neely : ´A Model and a System for Machine Recognition of Speech´ IEEE Trans. A.U., 21, p 229-238, 1973

[Ruske 81] Ruske G. & Schotola T. ´The efficiency of demi-syllabes segmentation in the recognition of spoken words´, IEEE ICASSP 81, Atlanta, pp 971-974, 1981

[Ruske 82] G. Ruske, T. Schotola : ´ The Efficiency of Demisyllable Segmentation in the Recognition of Spoken Words´, in ´Automatic Speech Analysis and Recognition, J.P. Haton ed, Reidel 1982

[Sakoe 71] H. Sakoe, S. Chiba : ´A Dynamic Programming Approach to Continuous Speech Recognition´ in Proc.Int.Congr.Acoust., Budapest, Hungary, 1971.

[Sakoe 78] H. Sakoe, S. Chiba : ´Dynamic Programming Algorithm Optimisation for Spoken Word Recognition´, IEEE Transactions on Acoustics, Speech, and Signal Processing, February 1978.

[Sauter 84] L. Sauter : ´RAPACE : un système de reconnaissance analytique de parole continue´, Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, Paris, janvier 1984, pp 89.

[Shirai-78] Shirai K. & Honda H. ´Feature extraction for speech recognition based on articulary model´, Proc. 4th ICJPR, pp 1064-1068, 1978

[Stern 85] Stern P.E., Eskènazi M. & Memmi D. ´Elaboration d'un système expert en lecture de sonagrammes´, 14 èmes JEP, Paris, pp 295-298, 1985

[Van Den Berg 70] Jw. Van Den Berg ´Mechanisn of the larynx and the laryngeal Vibrations´ Manual of Phonetics, Malberg ed., 278-308, North Holland Publishing Co, Amsterdam London.

[Wakita-77] Wakita H. ´Normalization of vowels by vocal tract length and its application to vowel identification´, IEEE Trans. ASSP-25, No 2, pp 183-192, April 1977

[Zadeh-78] Zadeh L.A. ´Fuzzy sets as a basis for a theory of possibility´, Fuzzy sets and systems, No 1, pp 3-28, 1978

[Zue 82] Zue V. ´Acoutic phonetic knowledge representation : implications from spectrogram reading experiments´, in ´Automatic Speech Analysis and Recognition´, J.P. Haton Ed., D. Reidel, 1982

NOM DE L'ETUDIANT : *FOHR Dominique*

NATURE DE LA THESE : *Doctorat de l'Université de NANCY I en Informatique*

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 10 [illegible] 1993 n° 360

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

R. MAINARD