Achile : un dispositif de décodage acoustico-phonétique et d'identification lexicale indépendant du locuteur à partir de modules mixtes

Alain Ghio

UNIVERSITE DE DROIT, D'ECONOMIE ET DES SCIENCES D'AIX - MARSEILLE III

FACULTE DE SCIENCES ET TECHNIQUES DE SAINT-JEROME

ACHILE : un dispositif de décodage ACoustico-pHonétique et d'Identification LExicale indépendant du locuteur à partir de modules mixtes

THESE effectuée par Alain Ghio dans le laboratoire CNRS ESA 6057 « Parole et Langage » à Aix-en-Provence.

Soutenance en vue d'obtenir le grade de DOCTEUR EN SCIENCES (spécialité Sciences et Technologies de l'Information

Soutenue le 21 novembre 1997 devant le jury composé de MM.

Jean-Paul Haton	**Professeur à l'Université de Nancy I Rapporteur**
Serge Huard	**Professeur à l'Université d'Aix-Marseille III**
Henri Méloni	**Professeur à l'Université d'Avignon**
Mario Rossi	**Professeur à l'Université d'Aix-Marseille I Directeur de Thèse**
Jean Véronis	**Maître de Conférence à l'Université d'Aix-Marseille I, Rapporteur**

RESUME

La reconnaissance de la parole est une activite dont le but est de faire identifier, par des machines, ce qui est dit par une personne. Le processus peut consister a reconnaitre des sons (decodage acoustico-phonetique), des mots (identification lexicale) ou des phrases. Un tel systeme est soit concu pour un seul utilisateur, soit pour differents locuteurs. Achile est un dispositif de decodage acoustico-phonetique et d'identification lexicle. Il permet la reconnaissance de mots isoles independamment du locuteur sans phase d'apprentissage, ni d'adaptation. Notre objectif est d'examiner jusqu'a quel point un modele a base de connaissances phonetiques est capable de decoder de facon automatique la structure phonique de la parole sans recourir aux methodes stochastiques. Le dispositif s'inspire, d'un point de vue fonctionnel, du traitement cognitif humain. La tache de reconnaissance est effectuee par repartition du travail et interaction d'une societe d'experts. Le signal de parole alimente tout d'abord les analyseurs de bas niveau. Pour cela, est utilisee, entre autre, une analyse spectrale fondee sur modele auditif qui tient compte de la notion de ponderation sonique et de bandes critiques. Les processus de bas-niveau transmettent leurs donnees a plusieurs modules de decodage fonctionnant en parallele (segmentation, reconnaissance globale et analytique). Les resultats sont ensuite transmis aux modules de haut-niveau qui agissent en utilisant des connaissances symboliques (representations phonologiques, acces lexical). Un moteur d'inferences se charge de prendre une decision finale en comparant les donnees d'un dictionnaire aux donnees decodees. L'evaluation du dispositif sur un lexique de 500 mots nous permet de quantifier la pertinence des connaissances, des analyses et des algorithmes employes. C'est aussi le moyen de pouvoir faire evoluer le dispositif en apportant des modifications.

Mots clés :

Systèmes homme-machine ; Reconnaissance automatique de la parole ;

Segmentation (linguistique) ; Acoustique – Analyse ; Phonétique acoustique ;

Phonologie – Informatique ; Vocabulaire – Identification

ABSTRACT

The aim of Speech Recognition is to identify with machines what a speaker is saying. This process can recognise sounds (acoustic-phonetic decoding), words (isolated-words recognition) or sentences. Engineers can build such a system only for a specified user or for different speakers. ACHILE is a system based on parallel-distributed processes for speaker-independent acoustic-phonetic decoding and words recognition. This is a speaker-independent isolated-words recognition system without learning and adaptation stage. We aim to examine to what extent a knowledge-based model can recognise segmental structure without stochastic modelling. The system proposed is inspired, in a functional way, by some features of human cognitive processing. This system is composed of a succession of demons who work on the pattern, each performing a different job. The speech signal first arrives at the low level analysis processes. A part of this analysis is realised with a spectral detection based on a perceptual model including frequencies weighting and critical bands analysis. Low-level detectors activate parallel distributed processes of decoding (segmentation, global and analytic recognition). Their results, then, are sent to the high-level processes, who act upon them using high level information (phonological rules, access to a dictionary...). Finally, a decision process selects the alternative that has the strongest evidence. The system has been tested on 500 words. It allows us to quantify the relevance of knowledge, analysis and algorithms used. It is also possible to change or add some parts which is important to improve the system.

REMERCIEMENTS

La page de remerciements

Est celle du soulagement

Car elle s'écrit finalement

A l'achèvement du document.

Pourtant,

Je suis resté un long moment

Devant ma feuille de papier blanc

Me demandant vraiment comment

Exprimer de vrais sentiments.

Parler d'abord de Pierre ou Jean

Me semblait manifestement

Injuste pour Anne ou Vincent.

Il fallait donc faire autrement.

Je décidai finalement

D'exprimer collectivement

Mes sincères remerciements

A mon Directeur tolérant,

A mon jury, mes enseignants,

A mes collègues, amis, parents.

NOTE SUR LES ILLUSTRATIONS

Dans ce document, les illustrations tirées d'ouvrages originaux sont accompagnées, dans la légende de la Figure, des références bibliographiques correspondantes. Les courbes représentant des signaux ou des résultats d'analyse ont été réalisées à partir du logiciel « Mes » développé au laboratoire « Parole et Langage » par Robert Espesser, ainsi qu'avec le logiciel « Phonédit » développé par la société S.Q.Lab, en la personne de Benoît Galindo.

AVANT PROPOS

Formé au métier d'ingénieur à l'Ecole Nationale Supérieure de Physique de Marseille, je dois mon initiation aux technologies vocales à John S. Mason, directeur du « Speech Processing Lab » de l'Université de Swansea (U.K.) dans laquelle je fus accueilli pour effectuer un stage sur la reconnaissance du locuteur. Pris de passion par cette discipline, je récidivai au cours de mon projet de fin d'études où me fus confié un programme de recherche et développement sur la commande vocale de matériel audio-visuel dans le service domotique du LEREA, l'ex laboratoire de R&D du groupe Thomson CE à Strasbourg. Une fois diplomé et envieux d'approfondir mes connaissances dans le domaine de la parole, j'obtins de la part de Mario Rossi, directeur du laboratoire « Parole et Langage », une bourse de docteur-ingénieur, que le CNRS alloue dans le but de former des ingénieurs aux métiers de la recherche. Ce travail m'a permis de présenter cette thèse. Ma formation de physicien, l'utilisation intensive de l'informatique et le contact permanent avec le milieu des linguistes-phonéticiens m'ont conduit au carrefour de trois disciplines, position extrêmement enrichissante du point de vue des connaissances, mais délicate. Position délicate car je me vois contraint de convaincre trois publics différents, ce qui relève du défi. C'est la raison pour laquelle je décidai de fournir, dans ce document, une première partie générale particulièrement fournie afin de permettre aux non-spécialistes de chaque domaine d'appréhender toutes les facettes de notre travail. Pour finir, j'aimerai citer une phrase de Régine André-Obrecht, chercheuse en traitement de la parole à l'IRIT (André-Obrecht, 1993, p.7):

> ***« la reconnaissance automatique de la parole est un sujet pluridisciplinaire (et comporte un grand nombre) de domaines que le chercheur en RAP ne peut ignorer. Comme il ne peut être spécialiste en chacun d'eux, ce chercheur doit être à l'écoute des autres chercheurs dans les matières qu'il ne maîtrise pas. »***

TABLE DES MATIERES

PARTIE -1-

PRESENTATION GENERALE

Introduction

Il fut un temps, pas si lointain, où l'ordinateur était un outil de spécialistes. L'emploi de l'informatique restait très spécifique et répondait à des tâches très précises: commandes de machines, télécommunications, calcul intensif... Pour réaliser ces travaux spécialisés, l'utilisateur communiquait avec son calculateur par d'incompréhensibles cartes perforées ou d'interminables lignes de commandes saisies sur un clavier. Ces moyens de communication sommaires convenaient parfaitement à une utilisation professionnelle mais rendaient l'ordinateur inaccessible à des utilisateurs novices ou occasionnels.

L'apparition et le développement massif de systèmes informatiques dans de nombreuses professions et dans la vie quotidienne ont obligé les ingénieurs à repenser cet état de fait: alors que, jusqu'à présent, l'homme se pliait aux exigences de la machine, il fallait maintenant que la machine s'adaptât à l'homme. Les systèmes Macintosh ont révolutionné la communication entre l'utilisateur et l'ordinateur en créant une interface plus adaptée, à partir d'icônes et de menus déroulants. Parallèlement, divers moyens de communication comme la souris, le crayon optique, le joystick, l'écran tactile, le son, ont permis de faciliter l'interaction avec les machines. Une étape a été franchie comme le prouve la multiplication des ordinateurs personnels dans les foyers. Mais ce développement reste cantonné à l'ordinateur individuel et touche très peu d'autres systèmes comme, par exemple, les bornes de renseignements, la commande de matériel audiovisuel, l'aide aux handicapés... Cela signifie que la communication homme-machine reste encore inadaptée car trop éloignée d'une communication simple et naturelle pour l'être humain.

De façon indéniable, la parole est le meilleur moyen d'expression que l'être humain possède. Aussi, de plus en plus, l'importance d'un dialogue homme-machine dans un langage oral le plus naturel possible devient grandissante au point de paraître maintenant indispensable à la mise en place de systèmes d'information accessibles à tous. Le mythe du « robot intelligent » parlant et comprenant fait surface. Cependant, le rêve ne doit pas faire oublier la réalité: l'utilisation de la parole dans la communication homme-machine doit faire l'objet d'une étude très précise pour évaluer si le contrôle vocal apporte véritablement un meilleur confort d'utilisation ou un accroissement des performances. Il s'agit de se poser la question suivante: « de la parole, quand et pour quoi faire ? ». Cette interrogation est nécessaire afin d'éviter de tomber dans la situation absurde où des équipes de recherche mettent au point des systèmes de synthèse et de reconnaissance de la parole puis cherchent, dans un second temps, des applications. Ces situations sont bien souvent décevantes car les réalisations se révèlent vite inadaptées à une utilisation réelle. Cette réflexion est d'autant plus importante dans la mesure où, aujourd'hui, aucune machine n'est encore apte à gérer et à comprendre un dialogue naturel: il faut donc se contenter de voix désagréables, de dialogues contraints de type question-réponse ou de menus dirigés. Il ne faut toutefois pas perdre espoir et penser qu'un jour viendra où nous pourrons dialoguer avec une machine en lui posant directement nos questions de façon spontanée. Tel est l'un des objectifs des industries de la langue et plus précisément du secteur de la reconnaissance automatique de la parole (R.A.P.).

I. LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

« Lorsqu'on considère un sujet nouveau, on a fréquemment tendance à, tout d'abord surestimer ce qui paraît déjà intéressant ou remarquable, et ensuite, par une sorte de réaction naturelle, à sous-estimer l'état réel de la situation quand nous découvrons que nos idées ont dépassé celles qui étaient réellement réalisables. »

Comtesse Ada Lovelace.

Plan du chapitre

RESUME

La Reconnaissance Automatique de la Parole (R.A.P.) consiste à identifier, par des moyens informatiques, ce qui est dit par un locuteur humain. C'est une activité prometteuse par ses multiples applications possibles: commande vocale de machines, saisie de données, sécurité, interface homme-machine, aide aux handicapés, apprentissage assisté par ordinateur... Ces technologies sont encore peu développées car l'opération de reconnaissance automatique de la parole par les machines s'avère difficile du fait des caractéristiques humaines du signal de parole. Les trois principales difficultés sont la diversité des informations (acoustique, lexique, syntaxe, émotions, appartenance sociale...), la continuité de l'information acoustique et la grande variabilité du signal de parole. Du fait de ces difficultés, la R.A.P. ne fonctionne actuellement que dans des situations contraintes où une partie des obstacles est supprimée: mots isolés, locuteur unique, élocution non spontanée...

L'opération de décodage s'effectue généralement par étapes. Après acquisition du signal de parole, il est nécessaire d'extraire une information pertinente, opération réalisée par calcul de paramètres acoustiques (énergie, coefficients spectraux, cepstraux...). Après une phase facultative d'analyse temporelle dite de « segmentation », l'étape d'identification proprement dite intervient. Pour cela, il existe diverses techniques: globales ou analytiques. Dans les méthodes globales, des processus algorithmiques opèrent une comparaison entre l'information extraite du signal à reconnaître et celle stockée dans des archives préalablement établies. Ces archives peuvent être des modèles (modèles de Markov), des réseaux (réseaux neuro-mimétiques) ou des prototypes. Une mesure de distance permet de proposer une solution correspondant à l'archive la plus proche. Dans les méthodes analytiques, le savoir d'expert est formalisé sous forme de règles qui agissent sur des informations a priori pertinentes. Actuellement, les techniques stochastiques (modèles de Markov) affichent des performances supérieures qui commencent à plafonner. Ces méthodes semblent insuffisantes pour résoudre, seules, la tâche complexe de décodage de la parole continue multilocuteurs.

Dans le but de trouver une alternative, notre objectif est d'étudier, mettre au point et réaliser un système de décodage acoustico-phonétique et d'identification lexicale, indépendant du locuteur, ceci afin d'examiner dans quelle mesure un modèle mixte à base de connaissances et de méthodes globales est capable de décoder de façon automatique la structure phonique de la parole sans recourir aux méthodes stochastiques. Notre effort s'est porté sur la recherche et l'utilisation de paramètres acoustiques pertinents et sur la mise en place d'une architecture originale où fonctionnent en parallèle divers processus de décodage.

I.1. Un secteur d'activité des industries de la langue

Les nombreuses recherches effectuées dans le domaine de la reconnaissance automatique de la parole sont intégrées dans une vaste branche d'activité scientifique et économique, celle des industries « de la langue ».

I.1.A. Les industries de la langue

D'après (Carré et al., 1991, p.10), le terme *industries de la langue* désigne «l'ensemble des activités qui visent à faire manipuler, interpréter ou générer par des machines le langage naturel écrit ou parlé par les humains». Parmi les activités technologiques développées dans ce cadre, on distingue généralement le traitement des langues naturelles - ou encore linguistique computationnelle - au secteur du *traitement de la parole* (Figure 1). Si le premier domaine aborde la langue dans son aspect exclusivement abstrait, le second prend en compte sa réalisation physique, et notamment les phénomènes acoustico-phonétiques qui caractérisent la parole. A l'aube des années 90, la part de marché des industries de la langue était estimée en France à environ 320 millions de francs dont un quart relevait du traitement de la parole (Source: ministère de la recherche et de la technologie). Ce marché reste encore très étroit si l'on considère que cette somme représente à peine le chiffre d'affaire d'une société de service en informatique d'environ 300 salariés (Carré et al., 1991).

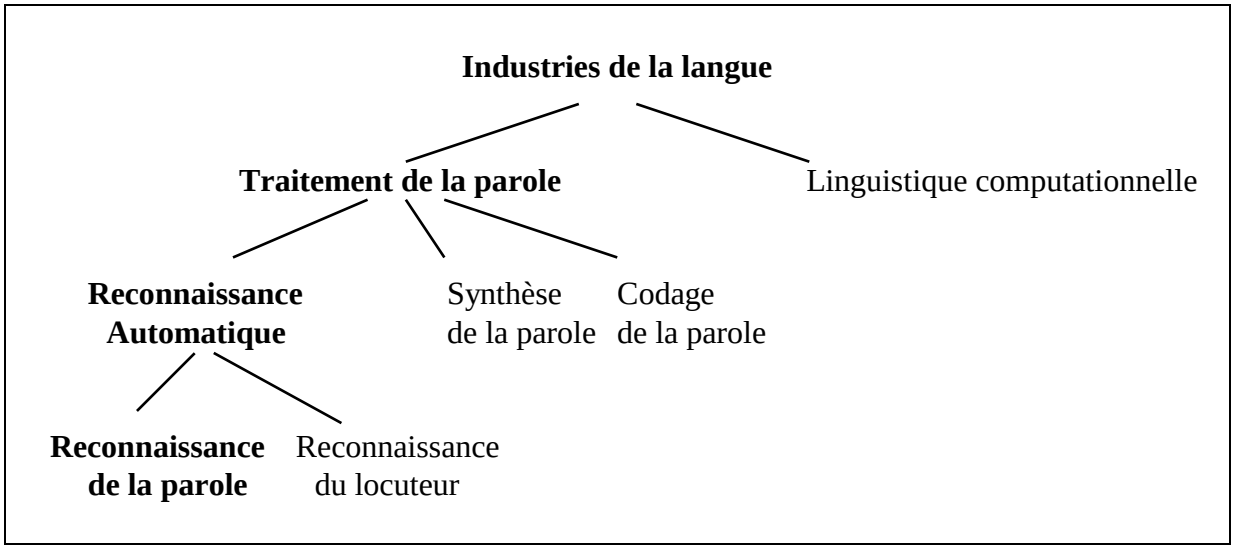


Figure 1: La place de la reconnaissance de la parole parmi les secteurs des industries de la langue

Les principaux débouchés de la linguistique computationnelle consistent à automatiser des traitements effectués sur la langue écrite: dictionnaire électronique, traduction, contraction de texte, vérification orthographique, correction grammaticale... L'interrogation de bases de données et de documents restent toutefois l'activité la plus importante.

Le domaine du traitement de la parole - ou encore technologies vocales - est lui aussi multi-sectoriel. On distingue ainsi trois activités majeures:

- la synthèse de la parole, qui consiste à créer artificiellement une voix.

- le codage et la compression de la parole, qui touchent surtout le domaine de la transmission.
- la *reconnaissance automatique*, qui concernent deux aspects distincts:
  - ⇒ la reconnaissance du locuteur, qui consiste à identifier ou authentifier une personne par sa voix.
  - ⇒ la reconnaissance automatique de la parole (R.A.P.), qui a pour but d'identifier ce qui est dit par un locuteur humain.

Dans notre cas, nous nous intéressons au dernier thème, celui de la reconnaissance automatique de la parole (R.A.P.).

I.1.B. Les applications de la reconnaissance automatique de la parole

Pour imaginer les applications possibles de la reconnaissance automatique de la parole (R.A.P.), penchons-nous tout d'abord sur les propriétés de la communication orale.

P1/ La parole est un moyen de *communication rapide* (2 à 4 mots/sec) par rapport à un clavier (1 mot/sec), à l'écriture manuscrite (0.4 mot/sec) ou à un menu à touches comme dans une interrogation à distance par téléphone (0.3 mot /sec) (Baudry, 1985).

P2/ La parole est un canal de *communication fonctionnant en parallèle* avec les sens naturels (vision, toucher...) et peut être ainsi utilisée en remplacement si ces derniers sont handicapés ou en complément si ils sont occupés.

P3/ L'émission et la réception du message oral sont *omnidirectionnels*: les personnes et systèmes concernés dans la communication ne sont pas obligatoirement proches et fixes. Ce n'est pas le cas, par exemple, dans un dialogue classique homme-ordinateur où l'utilisateur est fixé devant un clavier et un écran.

P4/ L'usage de la parole est *spontané et naturel*.

P5/ L'équipement terminal d'entrée pour la parole (le microphone) reste *simple* comparé à un clavier, un écran tactile, un ensemble de boutons poussoirs...

Du fait de ces propriétés, le champ des applications de la R.A.P. apparaît très vaste. Ci-dessous est donnée une liste non exhaustive de ces utilisations possibles.

- l'aide aux handicapés
  - ⇒ aide à l'apprentissage de la parole chez l'enfant sourd.
  - ⇒ aide à la lecture labiale.
  - ⇒ contrôle vocal de l'environnement (commande de prothèses, chaises roulantes...)
- la commande vocale de machines
  - ⇒ pour le grand-public (appareils ménagers, micro-ordinateurs, jouets...) du fait de la spontanéité de la communication orale.
  - ⇒ en milieu hostile (centrale nucléaire) ou obscur (chambres noires, salles closes...) du fait de l'omnidirectionnalité de la voix.
  - ⇒ en assistance au pilotage en avion pour permettre au pilote déjà sur-occupé de commander diverses fonctions.

⇒ en automobile pour actionner certaines fonctions gênantes comme les essuie-glaces, les phares, les vitres, l'autoradio...

- la saisie de données
  - ⇒ pour des transactions à distance par téléphone (domaine commercial, bancaire...)
  - ⇒ pour les tâches de tri quand les mains sont occupées (tri postal, tri de bagages...).
  - ⇒ pour les expérimentations quand les yeux sont occupés (observation au microscope) ou inopérants (observation d'optique nécessitant un faible éclairage).
  - ⇒ pour les relevés fastidieux (numéros d'immatriculation, cartographie).
  - ⇒ pour des relevés automatiques lors d'inspections et de contrôle de qualité.
- la sécurité
  - ⇒ pour l'accès en zone réglementée. Dans ce cas-là, un dispositif de vérification du locuteur peut être associé de façon efficace.
  - ⇒ applications militaires diverses.
- l'interface homme-machine dans le cas de systèmes complexes
  - ⇒ dans un centre de renseignements (réservations, horaires, achats, administration, météo, réseau routier...) surtout par l'intermédiaire du téléphone. Il est à remarquer que dans le cas de la France, l'existence du Minitel concurrence énormément la création de serveurs vocaux.
  - ⇒ la machine à dicter vocale (Bureautique, gestion...).
  - ⇒ le robot « intelligent ».
- apprentissage d'une langue assisté par ordinateur
  - ⇒ enseignement à distance.
  - ⇒ rééducation.

Certaines de ces applications sont encore du domaine de la science-fiction soit par manque de connaissances, soit parce que la technologie ne le permet pas encore. D'autres sont sur le point d'être commercialisées. La plupart sont du domaine de la recherche. L'interface homme-machine par la parole peut s'appliquer à partir du moment où il y a communication, c'est à dire très souvent. Les tendances actuelles consistent à associer un contrôle vocal à d'autres techniques de communication. Par exemple, en Bureautique, le clavier reste l'interface de saisie la plus efficace. Malgré tout, certaines commandes très fréquentes peuvent être contrôlées à la voix. En fait, le choix d'une application doit faire l'objet d'une étude très précise pour savoir si le contrôle vocal apporte véritablement un accroissement des performances ou un meilleur confort d'utilisation. L'exemple du crayon optique introduit dans la programmation des enregistrements des magnétoscopes puis rapidement abandonné est la preuve qu'une idée intéressante au départ peut s'avérer inadaptée en fin de compte.

I.2. Les principes de la reconnaissance automatique de la parole

S'il est vrai que, selon les objectifs choisis, les moyens mis en oeuvre pour réaliser un système de R.A.P. semblent différents, il existe tout de même un ensemble de points communs à toutes les méthodes, notamment sur le rôle du dispositif dans la chaîne de communication et dans les difficultés rencontrées.

I.2.A. Le rôle d'un système de reconnaissance automatique de la parole

Le rôle d'un système de R.A.P. est de décoder l'information transportée par le message vocal. Ce décodage permet alors de déclencher un affichage d'information, une sélection, une commande sur une machine... (Figure 2)

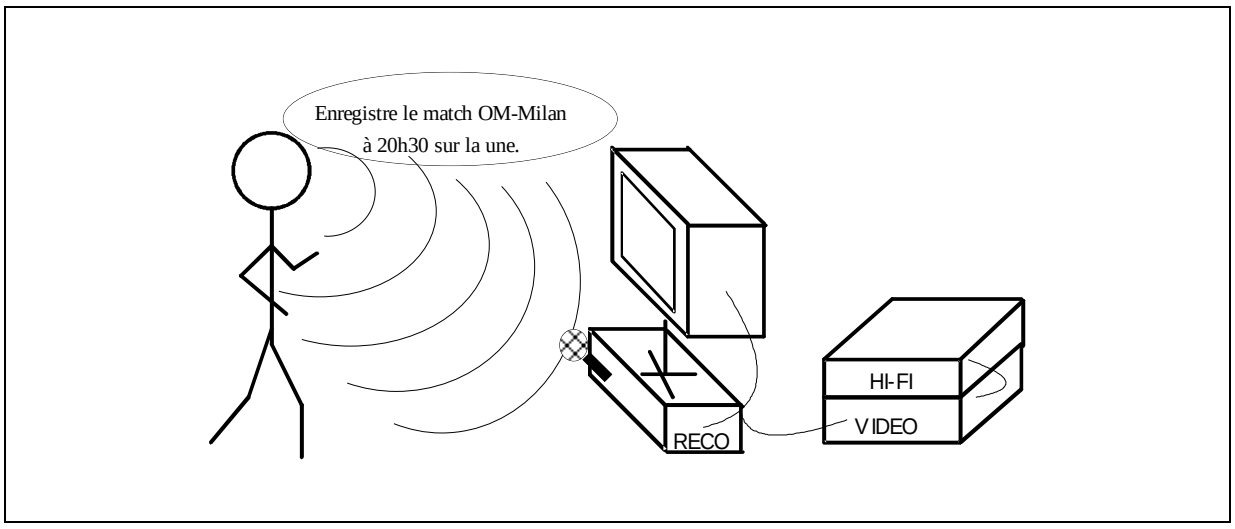


Figure 2: Exemple d'une utilisation de système de R.A.P. en domotique

I.2.B. La reconnaissance automatique de la parole dans la communication homme-machine

Dans le cadre d'une communication homme-machine, un système de décodage de la parole doit se substituer à un auditeur afin d'identifier et d'interpréter ce qui est dit par un locuteur. Le décodage doit alors être nécessairement couplé à un système d'affichage ou de synthèse vocale pour permettre un dialogue entre la machine et l'utilisateur. Avant d'aborder le problème spécifique de la communication homme-machine, présentons d'abord une vue plus générale de la communication orale entre deux humains. D'après (Pierrel, 1987, p.27), « la chaîne de communication orale peut être considérée en première approximation comme un double processus symétrique: d'un côté, l'émission ou la production, qui assure le passage de l'idée à la commande des nerfs moteurs du système phonatoire ; d'un autre côté, la perception qui, elle, permet de passer du signal acoustique de parole à l'idée » (Figure 3). Pour chaque fonction d'encodage et de décodage, on distingue un niveau périphérique dans lequel se manifestent des processus physiologiques (phonatoire ou auditif) et un niveau central de type cognitif. Il faut noter l'interaction qui existe entre l'émission et la perception, ce qui permet au locuteur de contrôler sa production. Certaines théories mettent en évidence l'interaction très forte entre production et perception: on parle alors de processus unique d'action/perception intégré sous forme de représentations sensori-motrices (Schwartz, 1995). La position extrême est représentée par la « théorie motrice » proposée par (Liberman, 1957 ; Liberman et al., 1967), qui fait l'hypothèse que les constituants de la parole sont perçus par référence à des commandes articulatoires aussi bien chez le locuteur que chez l'auditeur. Autrement dit,

l'information langagière serait codée en mémoire sous forme de représentations mentales de type articulatoire. L'auditeur utiliserait un processus d'inversion pour retrouver, à partir du signal acoustique, les commandes articulatoires associées au message sonore.

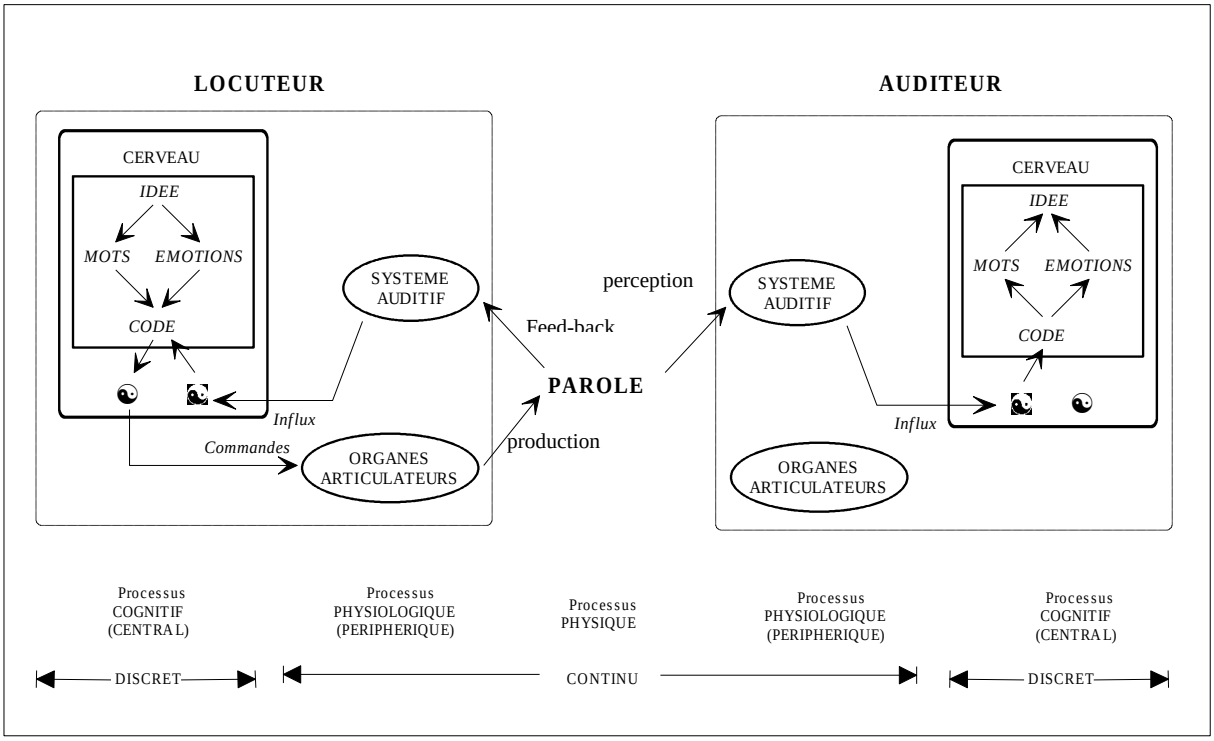


Figure 3: La chaîne de communication orale (description unidirectionnelle)

Si l'intégration production/ perception est généralement admise, les fondements de la théorie motrice restent extrêmement contestés. Au niveau cognitif, se distinguent deux types d'information: un support purement linguistique qui se manifeste par le choix et l'utilisation de mots, mais aussi un canal para-linguistique qui transporte les émotions, les intentions, les «sous-entendus» (Figure 3). Ceci explique l'écart qui existe parfois entre ce que le locuteur dit et ce qu'il veut dire.

I.2.C. Les questions préalables

Si, dans le cadre d'une communication homme-machine, un système de décodage automatique de la parole doit jouer le rôle de l'auditeur, deux questions se posent :

- Jusqu'où le décodage artificiel doit-il aller ?
- Doit-on imiter le fonctionnement humain pour réaliser un système de R.A.P.?

La première question consiste à se demander si le dispositif doit reconnaître les sons, les mots ou les idées. En principe, la seule chose importante est d'accéder aux idées. On parle alors de compréhension automatique de la parole (C.A.P.). Une telle tâche nécessite non seulement un bon décodage linguistique mais aussi une modélisation du monde réel, ce qui

s'avère extrêmement compliqué. Aussi, la solution consiste souvent à associer directement les idées à des mots: l'idée '🔔' = les mots 'cloche', 'sonnerie' ou 'alerte'. Cette simplification donne naissance à un sous-problème qui est celui de la reconnaissance automatique de la parole (R.A.P.). Cette réduction de la tâche se transforme parfois en malentendu: un très bon système de R.A.P. avec 98% de bon décodage peut devenir un très mauvais dispositif de C.A.P. si, par exemple, il confond "le" et "les" dans l'énoncé « détruit le fichier » et qu'il envoie alors une commande de destruction multiple de données au lieu d'un effacement simple ! Il faut garder à l'esprit que l'information n'est pas présente de façon homogène dans le discours oral. Dans notre exemple, la distinction « le/les » est fondamentale au niveau conceptuel. De plus, pour accéder aux idées, il n'est pas nécessaire de reconnaître de façon précise et exhaustive, l'ensemble de la chaîne parlée. Il suffit d'identifier certains sons, mots ou groupes de mots qui portent l'information. Enfin, il faut remarquer que la R.A.P. écarte la plupart du temps les informations paralinguistiques. On ne peut pas, par exemple, accéder à l'idée '🔔' par un cri d'alerte. Nous ne prétendons pas, dans notre travail, réaliser un dispositif de C.A.P. Notre investigation reste dans le domaine de la R.A.P. Toutefois, certains choix que nous avons effectués sont conditionnés par les réflexions présentées ci-dessus.

La seconde question qu'il faut se poser est la suivante: doit-on imiter le fonctionnement humain pour réaliser un système de R.A.P. ? Le meilleur système de R.A.P. que l'on connaisse reste incontestablement l'homme. Cela n'a rien d'étonnant par le fait que le langage est né d'une lente évolution au cours de laquelle le code linguistique s'est optimisé à l'appareil sensori-moteur dont dispose l'être humain. Compte tenu des performances humaines en matière de décodage de la parole, il peut sembler judicieux de vouloir non pas imiter, mais s'inspirer de son fonctionnement dans un dispositif artificiel. Les opposants à l'intrusion d'anthropomorphisme dans les techniques modernes proposent souvent l'image de l'avion qui vole très bien sans pour autant battre des ailes comme les oiseaux. On peut répondre à cet exemple en disant qu'effectivement, un avion ne bat pas des ailes, mais que son fuselage ressemble plus à un goéland qu'à un nénuphar et que les différences fonctionnelles entre un planeur et un aigle utilisant les courants ascendant pour se déplacer sont faibles. De même, ces personnes disent que les voitures se déplacent non pas sur des jambes mais sur des roues. On peut, là aussi, contredire cet argument en montrant qu'en milieu difficile comme un terrain mouvementé, un robot se déplace mieux sur six pattes qu'avec quatre roues. Il ne faut pas oublier que la parole est bien plus qu'un signal acoustique quelconque et qu'elle reste la manifestation physique d'un processus linguistique purement humain. On ne traite pas la parole comme on peut analyser les vibrations émises par un sous-marin. Dès 1971, Haton écrivait (Haton, 1971, p.77) « la plupart des chercheurs s'accordent maintenant à penser que le signal de parole est un signal à part, sur lequel les méthodes classiques de traitement ne donnent que des résultats moyens. La structure du message parlé, les relations entre les sons successifs en sont autant de facteurs importants spécifiques qui rendent difficile l'application des techniques générales de reconnaissance de formes pour une séquence de parole ». Les sons de parole ne sont pas dus au hasard. La prédiction des systèmes vocaliques des langues du monde montre que les unités phoniques de la communication orale sont nées d'une lente négociation entre perception et production (Boë et al., 1994). Tout au long de notre étude, nous avons essayé de ne pas perdre de vue cet aspect fondamental.

De nombreuses expériences montrent que le niveau périphérique auditif transcode, de façon non linéaire, l'information contenue dans le signal acoustique vers le nerf sensoriel de l'oreille (Figure 3). Autrement dit, lorsque l'être humain décode la parole au niveau cortical, le signal est passé par le système auditif qui effectue un filtrage. En R.A.P, les systèmes utilisent généralement le signal de parole brut comme matière première du décodage. Pourtant, il semble plus pertinent de se baser sur un signal filtré par le système auditif, ce filtrage physiologique pouvant être partiellement simulé sur un calculateur en tenant compte de phénomènes psycho-acoustiques. Cette démarche qui consiste à se rapprocher le plus possible du traitement humain n'est pas nouvelle (Caelen, 1979). Nous nous sommes orientés dans cette direction. Dans la phase de décodage proprement dite, il est aussi possible de s'inspirer du fonctionnement cognitif humain même si celui-ci est encore fort mal connu. Nous aborderons ce point dans un chapitre ultérieur.

I.2.D. Les difficultés en reconnaissance automatique de la parole

Plusieurs obstacles dus principalement à la nature humaine du message vocal font du traitement automatique de la parole un domaine complexe et loin d'être résolu. Ces difficultés proviennent des raisons suivantes :

- la diversité des informations
- le semi-continuum acoustique
- les phénomènes de variabilité

I.2.D.a. La diversité des informations

Le signal de parole transporte *plusieurs types d'information***:** les sons de parole (acoustique), les mots du vocabulaire, la grammaire, le sens du message, l'état émotionnel du locuteur, son identité, son accent, son appartenance sociale... A cela s'ajoutent des effets de contexte. Pour accéder à une information utile comme, par exemple, décoder ce qui est dit, il faut effectuer un tri en s'efforçant d'effacer les détails inutiles. Dans le cas de la reconnaissance de la parole multilocuteurs, le but est de s'affranchir des caractéristiques du locuteur, de la situation de communication... Cela entraîne une simplification de la tâche mais aussi une perte d'information qui peut être importante.

I.2.D.b. Le semi-continuum acoustique

La parole est un *semi-continuum* acoustique (Figure 4): lorsqu'une personne écoute parler quelqu'un, elle reçoit un flot semi-continu de sonorités ponctuées de pauses. Il n'existe pas de séparateurs clairs entre les unités phonétiques, lexicales... comparables aux blancs du langage écrit. Le problème de la segmentation reste donc très délicat.

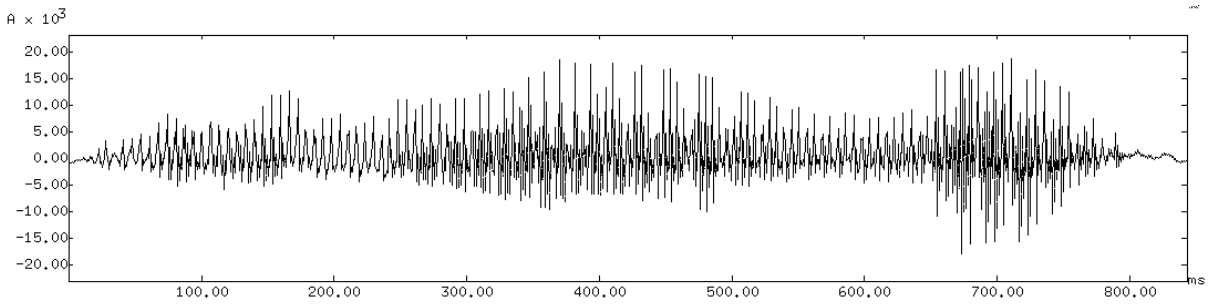


Figure 4: Signal acoustique sur l'énoncé "y en a un là ! "

I.2.D.c. Les phénomènes de variabilité

Il suffit de quelques observations pour constater que le signal de parole est extrêmement variable. Un même énoncé peut se réaliser sous des aspects très différents. Ces variations s'observent d'une part, entre locuteurs (Figure 5) mais aussi chez un même locuteur (Figure 6, p.13).

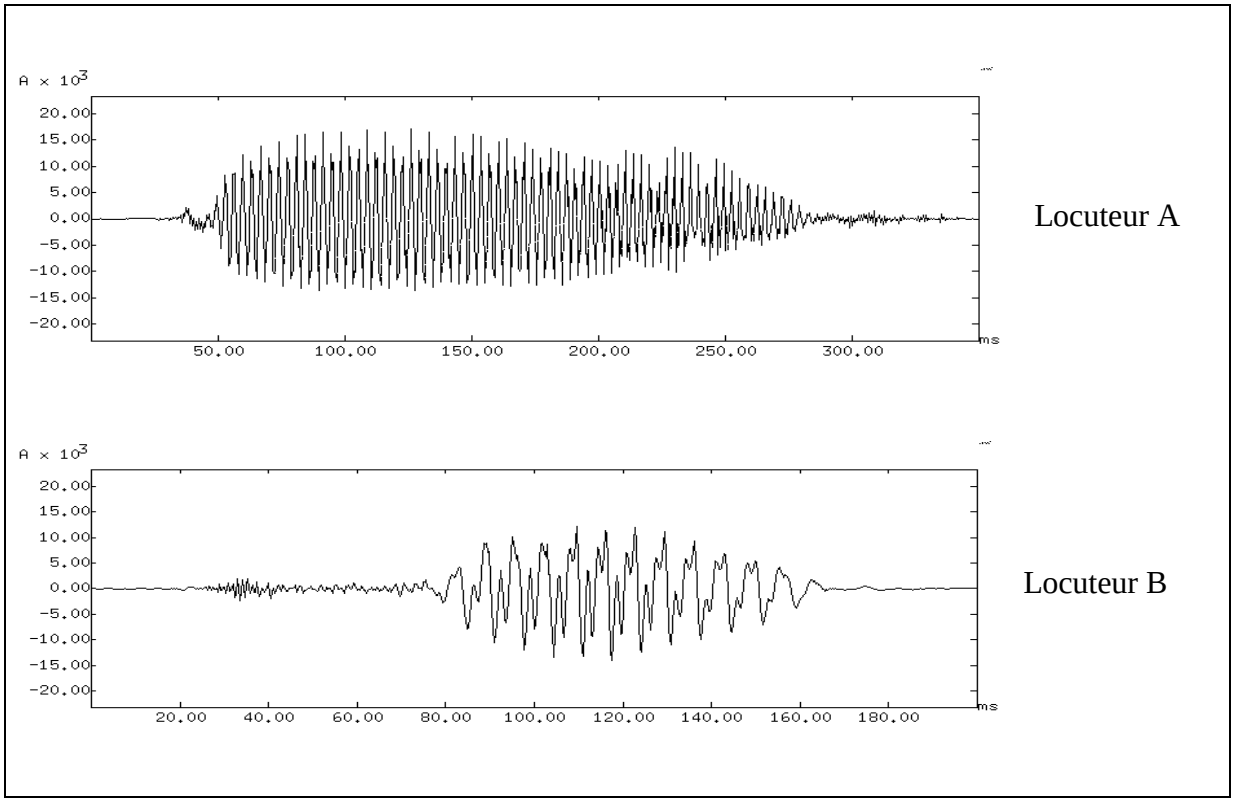


Figure 5: Exemple de variabilité interlocuteurs (signaux de parole sur l'énoncé "pire")

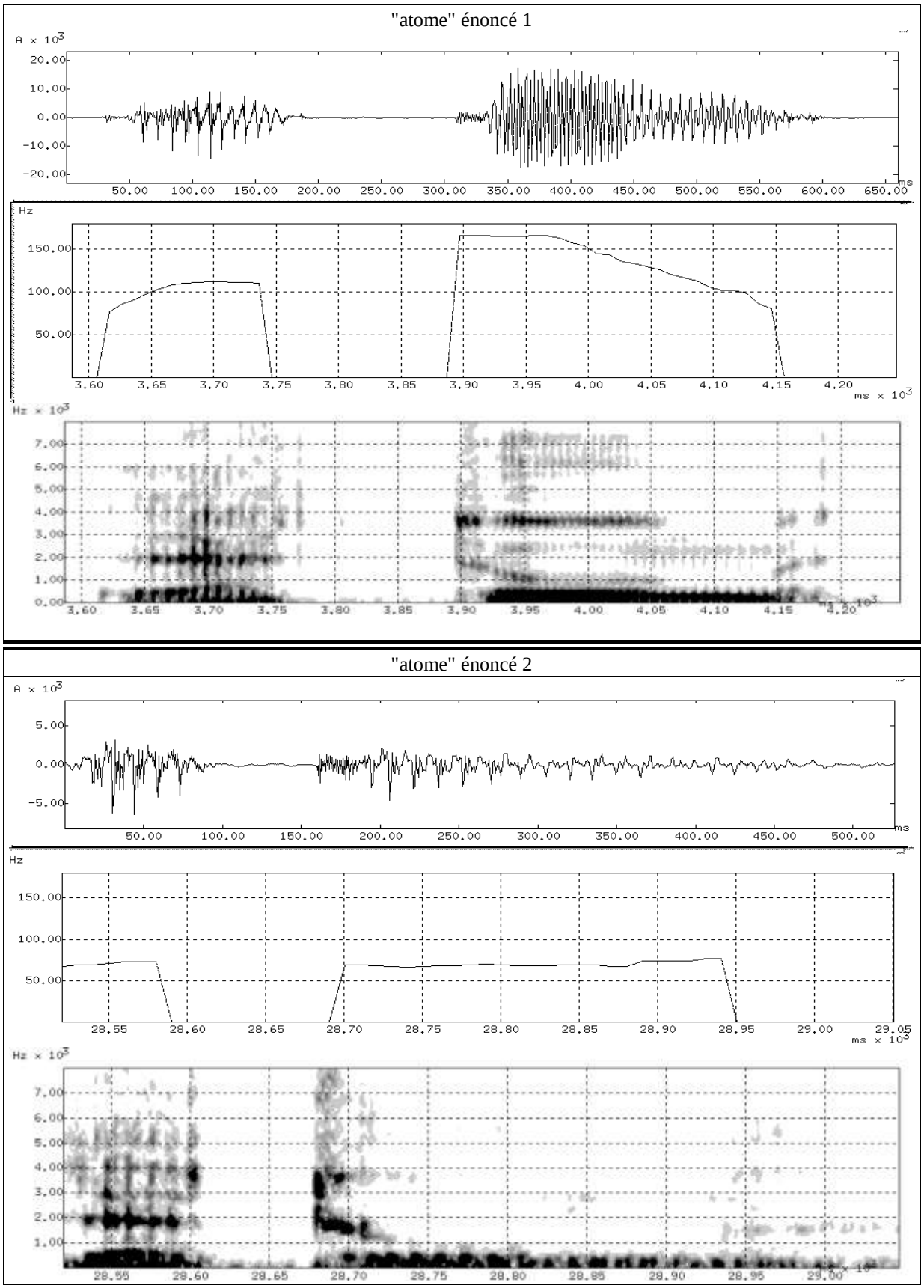


Figure 6: Exemple de variabilité intra-locuteur (signal de parole, fréquence fondamentale et spectrogramme pour deux énoncés du mot "atome" prononcé par le même locuteur dans deux contextes différents)

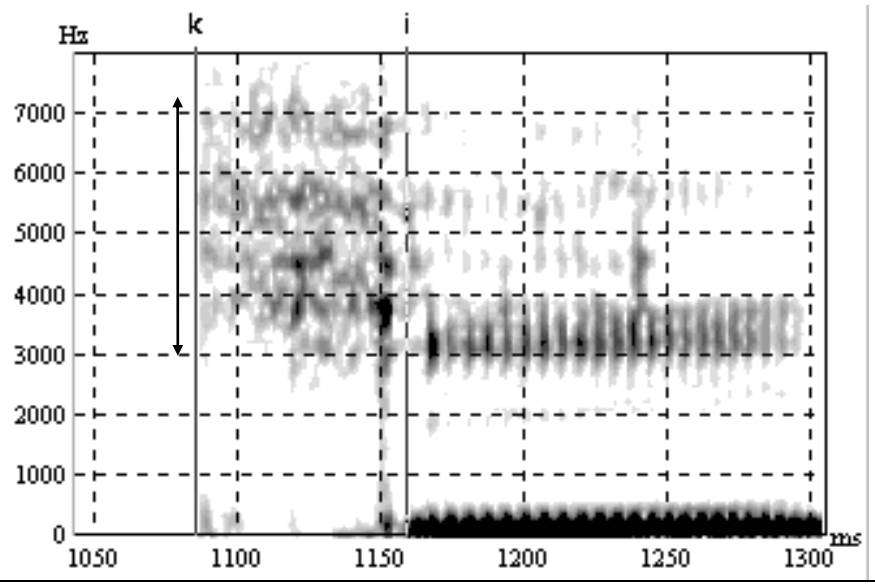
Bien souvent, le terme « variabilité » est utilisé pour combler une spécification insuffisante du contenu du signal de parole. Sous cette dénomination sont regroupés tous les phénomènes qui tendraient à écarter une production sonore d'une « norme » attendue, d'un invariant universel. Parler de « la » variabilité au singulier est une erreur; ce phénomène est multiple:

- la nature biologique de la parole entraîne une non-reproductibilité des phénomènes. Il s'agit d'imprécisions autour d'une valeur moyenne dues à l'imperfection du système phonatoire ou, autrement dit, des variations inhérentes à la définition de toute grandeur physique (flou empirique). On parle de **variabilité aléatoire**. Ces variations s'apparentant à un mouvement brownien, elles ne transportent aucune information pertinente et peuvent être réduites par des procédures de normalisation.
- les différences physiologiques entre les individus génèrent une **variabilité interindividuelle**. Les variations sont liées à l'originalité biologique de chaque personne et donc aux différences qui existent entre les appareils phonatoires de chaque individu. Il est tout de même possible de mettre en évidence des classes de locuteurs: enfant/adulte/personne âgée, homme/femme...
- les variations incessantes de la santé et de l'état psychologique d'un individu génèrent une **variabilité intra-individuelle**. Ainsi, un rhume ou une angoisse bouleversent la physiologie de l'individu et, par conséquent, sa production sonore. Une telle variabilité est, la plupart du temps, non contrôlable. Elle peut toutefois être réduite si le locuteur évolue en milieu protégé et si, les mesures acoustiques sont renouvelées plusieurs fois dans le temps, permettant de saisir une certaine invariance chez le locuteur.
- un même énoncé peut être intentionnellement réalisé différemment selon la situation d'élocution: colère, ironie, tristesse, joie... On parle de **variabilité paralinguistique**. Une telle variabilité est, la plupart du temps, voulue par le locuteur et peut être réduite dans le cadre d'un dialogue homme-machine contraint.
- les différences d'environnement entre les individus génèrent une **variabilité linguistique** que l'on peut scinder en deux:
  $\Rightarrow$ une variabilité sociolectale selon des critères sociologiques
  $\Rightarrow$ une variabilité dialectale qui se manifeste entre communautés linguistiques (ex: parler normand, parler méridional, parler québécois...).

  L'étude de ces différences de parler peuvent permettre de cerner les variations non pas par rapport à une norme, mais par rapport à une position « centrale ».
- la **variabilité contextuelle** reflète l'influence mutuelle qu'exercent les unités linguistiques entre elles. Elle se manifeste par le fait qu'un élément peut se réaliser de façons différentes selon les unités qui l'entourent. Ce type de variabilité est liée à la nature physique des instruments de la phonation. Les organes articulateurs sollicités dans la production de la parole possèdent une certaine inertie, un certain temps de réponse, une limite dans le déplacement... Chaque geste articulatoire mis en place pour produire un son est imprégné par les positions précédentes et par la configuration à venir (anticipation). Ces phénomènes sont connus sous le noms de coarticulation et de

coproduction. La Figure 7 illustre les variations concernant la composition spectrale de l'explosion de /k/, dont l'énergie se situe entre 3000Hz et 7000Hz en contexte _i et entre 500Hz et 4000Hz en contexte _lu. De telles variations contextuelles entraînent de grandes difficultés car la combinatoire des sons d'une langue peut être impressionnante. Toutefois, elles restent prévisibles et suivent une certaine régularité. Il est donc possible de s'en affranchir.

Spectrogramme de /ki/



Spectrogramme de /klu/

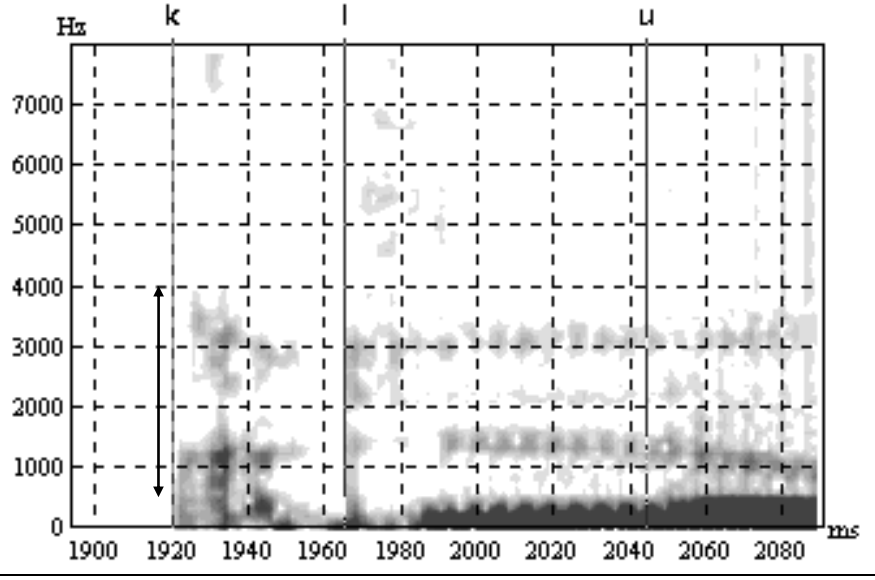


Figure 7: Exemple de variabilité contextuelle sur la composition spectrale de /k/. On remarque que la zone d'énergie sur l'explosion se situe entre 3000Hz et 7000Hz en contexte _i et entre 500Hz et 4000Hz en contexte _lu

Bilan sur les phénomènes de variabilité

Bien évidemment, toutes les sources de variabilité précédemment décrites opèrent simultanément et conjuguent leurs effets. Ainsi, la variabilité intra-locuteur (Figure 6, p.13) peut provenir de la variabilité aléatoire, intra-individuelle, paralinguistique et contextuelle.

Ces remarques sont importantes pour comprendre qu'il est impossible de traiter « la » variabilité globalement; chaque type nécessite une opération spécifique. Ainsi, Rossi distingue deux catégories fondamentalement différentes (Rossi, 1985):

- la variabilité aléatoire, de type chaotique, qui peut se traiter par des procédures de normalisation de bas niveau.
- les autres formes de variabilité, de type structurelle, qui sont théoriquement déterministes et porteuses d'information. Ainsi, la variabilité dialectale, qui se manifeste entre communautés linguistiques, peut être neutralisée au niveau des représentations phonologiques des éléments du dictionnaire...

Finalement, la principale difficulté face aux variabilités consiste à réaliser un système de décodage automatique qui sera apte à catégoriser de la même façon le [a] du mot "attend" prononcé isolément à Marseille par Monsieur X en colère et le [a] du mot "kaki" émis calmement par Mme Y, strasbourgeoise. Cette opération de décodage reste possible du fait de la grande redondance du signal de parole. En exploitant les différents niveaux d'analyse (phonétique, lexique...), il est possible de résoudre les nombreux problèmes d'ambiguïté et de réaliser ainsi un système robuste de reconnaissance automatique de la parole.

I.3. La réalisation de systèmes de reconnaissance automatique de la parole

Nous limitons la reconnaissance à une identification lexicale sans analyse syntaxique ou sémantique.

I.3.A. Les différentes phases dans la reconnaissance automatique de la parole

Le décodage de la parole depuis le signal acoustique jusqu'au(x) mot(s) se fait généralement par étapes successives (Figure 8).

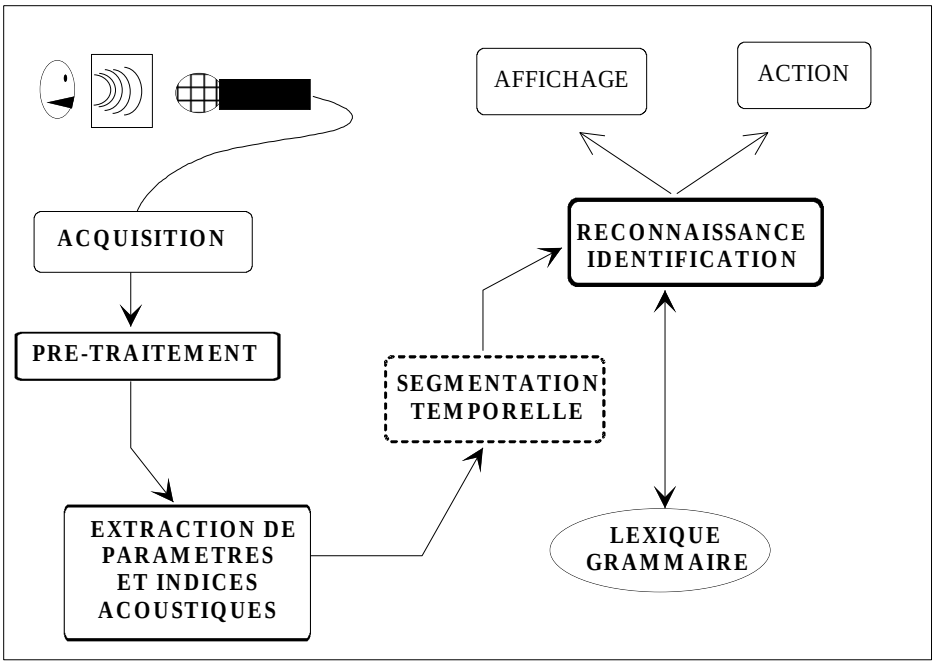


Figure 8: Les étapes de la reconnaissance automatique de la parole

I.3.A.a. L'acquisition

L'acquisition est une étape purement technique qui consiste à capter puis transformer le son de parole en un signal utilisable par un calculateur. Cette opération est aussi baptisée «numérisation».

I.3.A.b. Le pré-traitement

Une fois le signal numérisé, une étape de pré-traitement est souvent effectuée. Cela peut consister, par exemple, en la suppression par filtrage d'un bruit indésirable (ex: bruit du réseau électrique à 50 Hz). Il est aussi possible de vouloir accentuer les aigus pour équilibrer la composition spectrale en rehaussant les hautes fréquences (Figure 9).

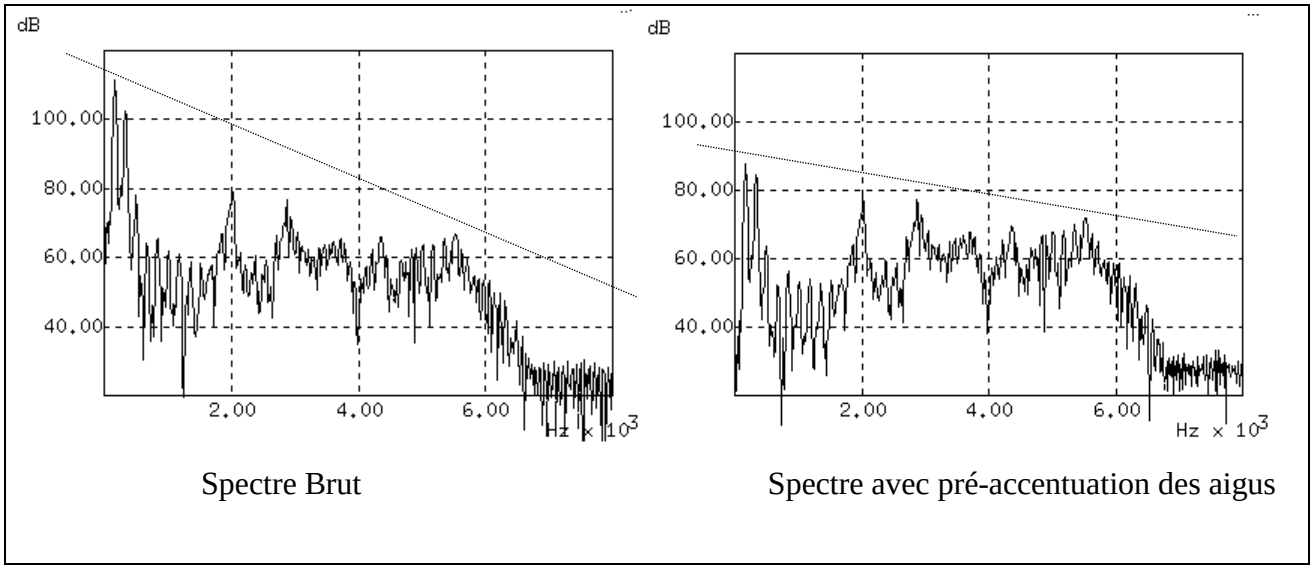


Figure 9: Exemple de pré-traitement sur le /i/ de l'énoncé "quitter" (accentuation des aigus avec un filtre de pré-emphase du type $h(z)=1-0.95\ z^{-1}$)

I.3.A.c. L'extraction de paramètres acoustiques

Le calcul de paramètres acoustiques a pour but d'extraire l'information utile contenue dans le signal de parole. Actuellement, les paramètres les plus utilisés en R.A.P. sont les *coefficients cepstraux obtenus sur un spectre de Fourier en échelle mel* (MFCC). Ils se calculent automatiquement à partir du signal de parole par une transformation mathématique complexe à laquelle on peut se référer dans (Davis & Mermelstein, 1980). D'autres types de paramètres analogues existent tels que les coefficients LPC, LPC-Cesptre, FFT-Cepstre... Ils sont directement utilisés comme source d'information acoustique dans les systèmes de reconnaissance utilisant une classification métrique ou des modèles statistiques. Généralement, ces paramètres plutôt abstraits sont choisis en fonction des performances qu'ils permettent d'obtenir. Des paramètres acoustiques plus directement liés à des caractéristiques physiques du signal peuvent aussi être utilisés. En guise d'exemple, citons l'énergie, la fréquence fondamentale, les représentations temps/fréquence de type spectrogramme, le taux

de passage par zéro, qui consiste à compter le nombre de fois par centiseconde où la courbe du signal coupe l'axe des zéros (Figure 10).

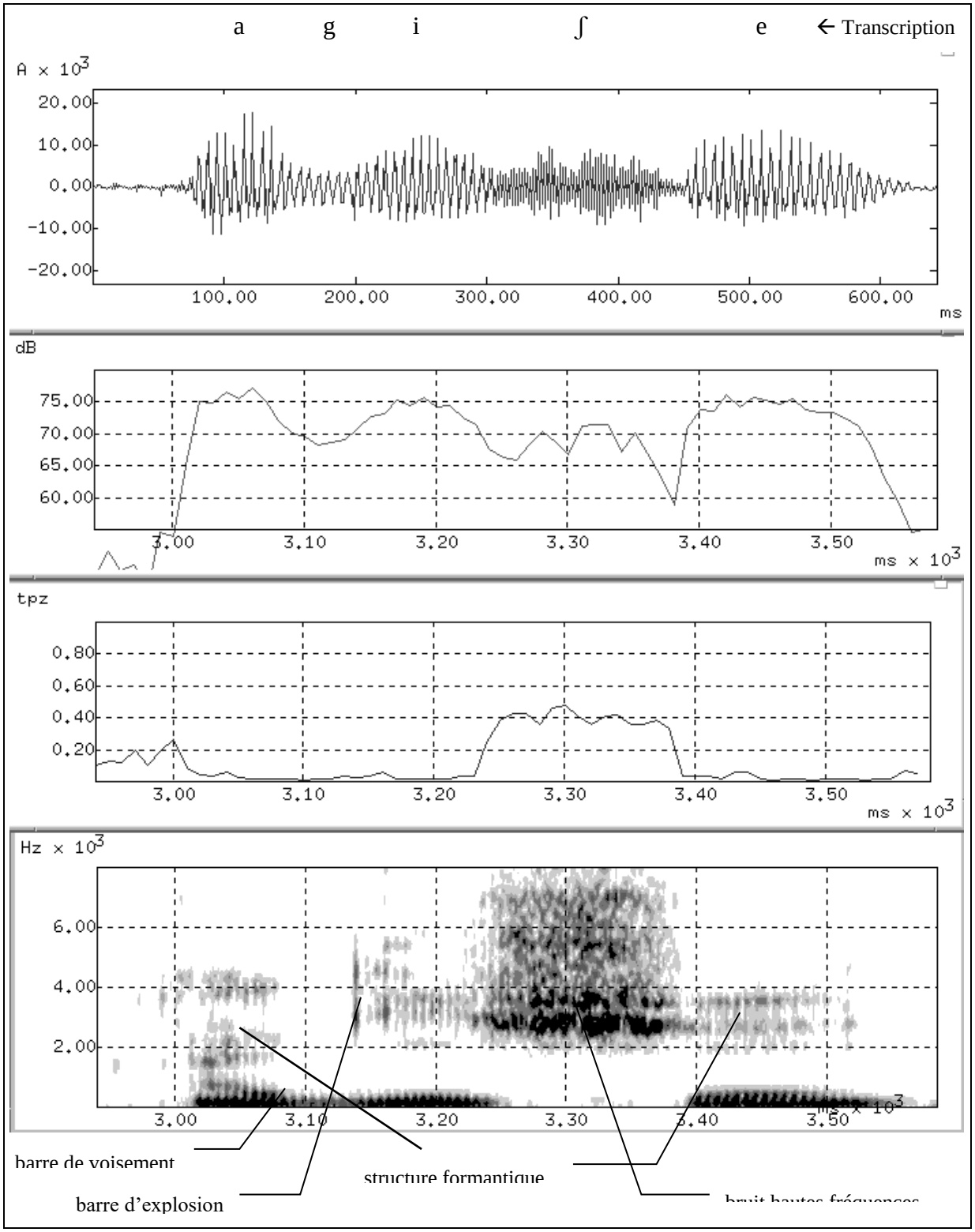


Figure 10: Exemple de paramètres acoustiques extraits de l'énoncé du mot "aguicher"

Sur la Figure 10, qui représente une visualisation de différents paramètres acoustiques extraits de l'énoncé "aguicher", on peut remarquer sur la courbe d'intensité que les fortes énergies correspondent aux voyelles. De même, sur la courbe du taux de passage par zéro, on se rend compte que les fortes valeurs sont relatives à la consonne /ʃ/ qui est une constrictive sourde possédant un important bruit de friction. Du spectrogramme, il est possible de déduire plusieurs informations: le caractère voisé des voyelles et de la consonne /g/ mise en évidence par la barre de voisement en basses fréquences, la structure formantique des voyelles, la barre d'explosion de l'occlusive /g/, le bruit hautes fréquences de la constrictive /ʃ/...

L'obtention de ces paramètres autorise le décodage de la parole. Les constrictives sourdes seront caractérisées par une faible intensité, un fort taux de passage par zéro, une forte contribution des hautes fréquences dans le spectre. Les occlusives voisées seront identifiées par une intensité moyenne, un faible taux de passage par zéro, une barre de voisement, une barre d'explosion... Cette information peut être utilisée directement dans les dispositifs de décodage par règles. Dans tous les cas, le problème essentiel réside dans le choix des paramètres acoustiques et des indices phonétiques (Rossi et al., 1983), certains étant plus pertinents et plus robustes que d'autres.

I.3.A.d. La segmentation

L'étape de segmentation consiste à analyser temporellement le continuum acoustique et repérer différents événements. Sur la Figure 11, on remarque 4 parties dans le signal de parole issu de l'énoncé "ski". On peut remarquer qu'un phonème peut être homogène acoustiquement (ex: [s], [i]) ou bien il peut être formé de segments hétérogènes (ex: [k] formé d'une tenue et d'une explosion). En fait, cet exemple est une simplification de la réalité. En effet, les sons de parole ne sont jamais homogènes. On distingue généralement une phase d'établissement, une phase stable et une phase de transition avec le phonème suivant.

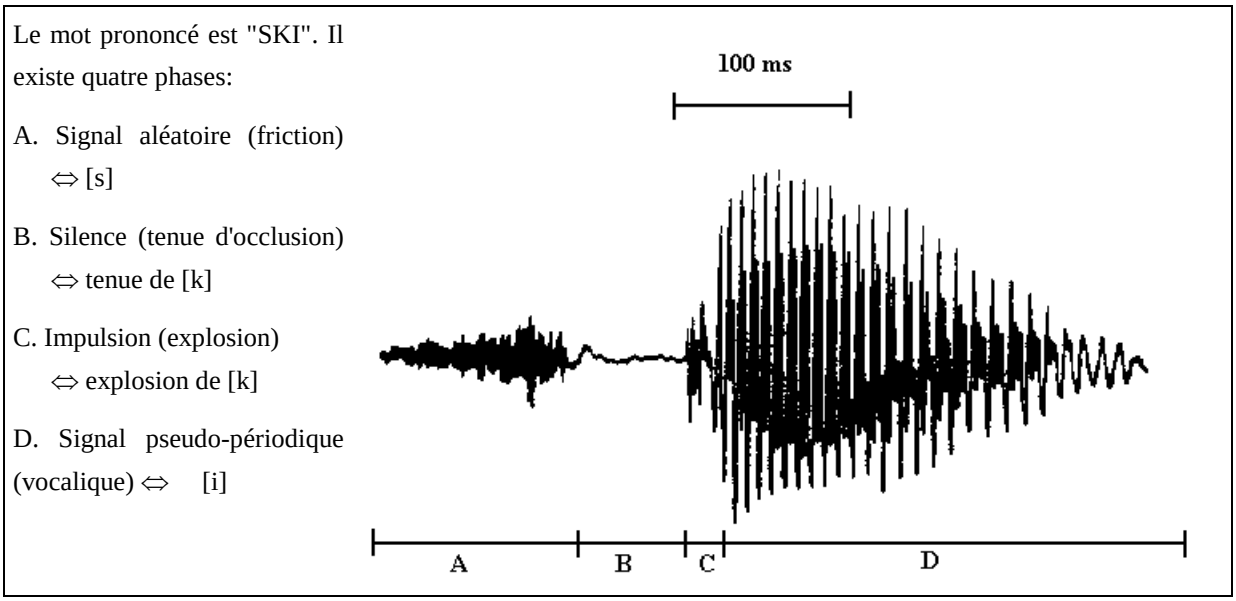


Figure 11: Segmentation du continuum acoustique en unités discrètes (Source: Calliope, 1989, p.258)

Le découpage de la parole en unités discrètes pose un problème. En effet, là où la linguistique met en évidence l'existence d'unités abstraites et discrètes telles que les phonèmes ou les mots, l'observation du signal acoustique ne laisse apparaître qu'un continuum (cf. § « Le semi-continuum acoustique » , p.11). Fant explique « sound segment boundaries should not be confused with phoneme boundaries. Several sounds of connected speech may carry information on one and the same phoneme, and there is overlapping in so far as one and the same sound segment carries information on several adjacent phonemes » (Fant, 1973, p.23). La confusion provient du fait que la description phonologique s'attache à représenter le contenu informatif du message alors que le signal de parole est le résultat de la mise en forme de ce message pour qu'il soit transmis et décodé. Si au niveau symbolique, la chaîne parlée est formée d'une séquence d'unités discrètes (Figure 12, a), la réalisation physique de cette chaîne fait intervenir des phénomènes physiologiques qui rendent opaques, voire inexistantes, les frontières entre les unités phoniques (Figure 12, d). Une telle représentation explique les différents niveaux de description de la parole:

a) représente la séquence idéale de phonèmes,

b) correspond à une séquence de segments sonores élémentaires,

c) illustre le fait qu'une ou plusieurs propriétés caractérisant un segment sonore peuvent s'étendre aux voisins,

d) met en évidence des fonctions continues exprimant un degré de corrélation entre les phonèmes et des événements particuliers présents dans l'onde sonore

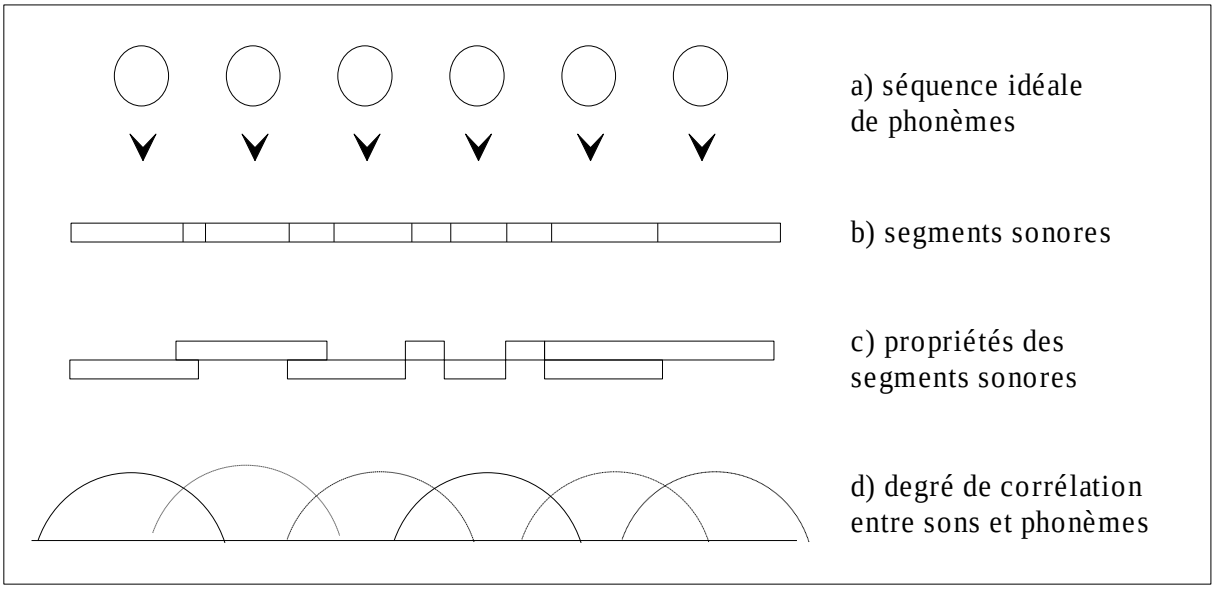


Figure 12: Schéma de Fant (Fant, 1973, p.22)

La Figure 13 illustre le passage d'une voyelle [∅] à une constrictive [ʃ]. Le spectrogramme laisse apparaître trois zones: à gauche une première zone nettement vocalique,

à droite une zone bruitée sans voisement correspondant à la fricative et, au milieu, une zone à la fois voisée et bruitée correspondant à la coarticulation entre les deux unités.

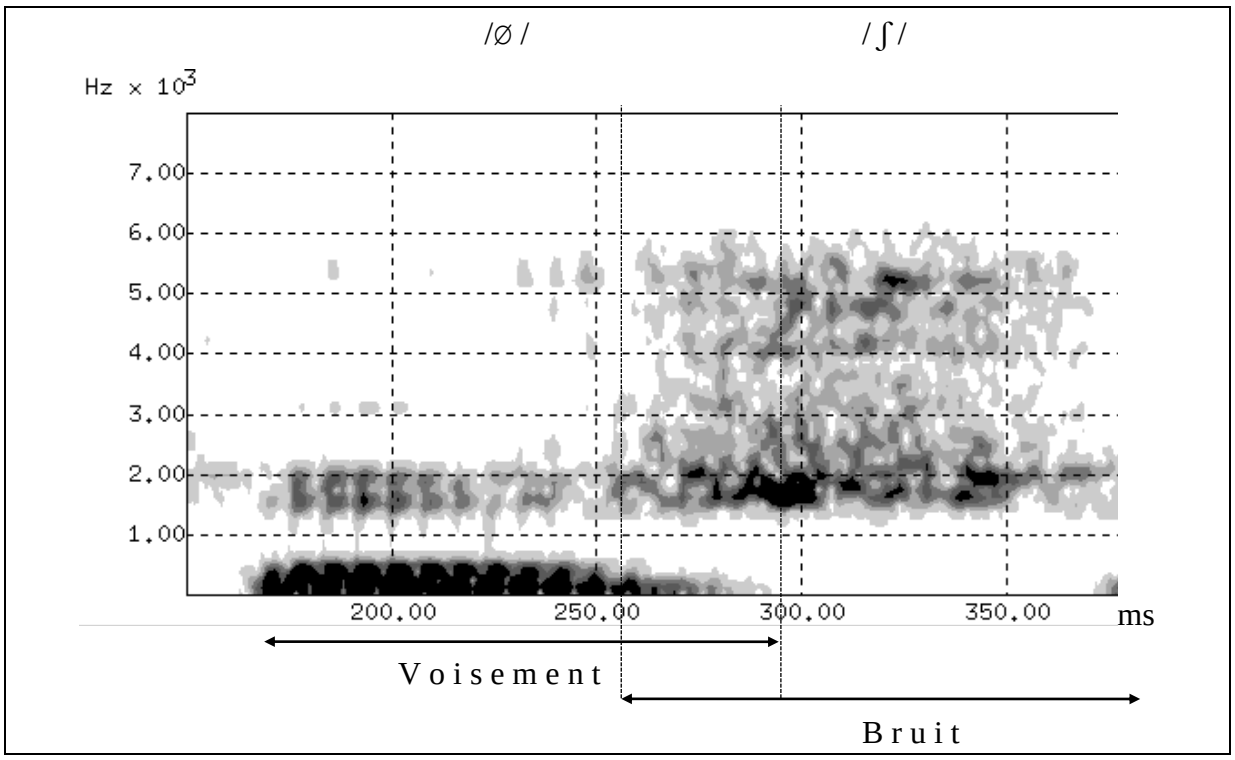


Figure 13: Asynchronie des discontinuités (spectrogramme de [ø ʃ])

Une des solutions, légitime, consiste à effectuer une segmentation, non par marquage des frontières, mais par repérage du « centre » de chaque unité phonique, celui-ci pouvant être vu comme le point de corrélation maximale (cf. Schéma de Fant). Cependant, un tel procédé peut s'avérer inadéquat dans le cas, par exemple, des occlusives sourdes où il est difficile de parler de « centre ».

En matière de segmentation automatique, le premier problème réside dans le choix d'une fonction ou plus généralement de critères qui mesurent les discontinuités du signal et de son spectre au cours du temps. Le second problème se pose pour le choix de l'unité de segmentation: doit-on couper en unités phonémiques (phonèmes, allophones), supra-phonémiques (mot, syllabe, diphones, polysons) ou sub-phonémiques (macro-classes, événements acoustico-phonétiques, traits, indices) ? Chacune de ces unités possède des avantages et des inconvénients: les phonèmes sont peu nombreux mais ils restent très sensibles au contexte, les mots absorbent les problèmes de coarticulation mais leur nombre peut être très étendu...

L'étape de segmentation, qui s'avère finalement délicate, est parfois évitée, comme dans le cas de la reconnaissance globale de mots isolés. Le dispositif identifie alors directement et globalement le mot après extraction des paramètres acoustiques. Dans les systèmes utilisant des chaînes de Markov avec modèles phonémiques, la segmentation s'obtient a posteriori, après identification des phonèmes. Il n'existe pas d'étape préliminaire de repérage temporel.

I.3.A.e. L'identification

La reconnaissance proprement dite, qui consiste à l'identification des unités de décodage, est une recherche paradigmatique, c'est à dire la sélection d'un élément parmi un lot d'unités en concurrence. Cela peut être, par exemple, un phonème parmi tous les phonèmes d'une langue, ou bien un mot dans un vocabulaire choisi... Le plus souvent, le système propose plusieurs solutions car de nombreuses ambiguïtés ne peuvent être levées sans faire appel aux niveaux linguistiques supérieurs. Ainsi, dans le cas d'une reconnaissance phonémique, on parle de treillis de phonèmes; seul un accès lexical permet de sélectionner la meilleure suite phonémique qui correspond à un mot existant dans le vocabulaire utilisé. Un très grand nombre de méthodes différentes peuvent être utilisées pour réaliser l'opération d'identification. Celle-ci apparaît plus ou moins complexe selon la tâche à accomplir: un seul ou plusieurs utilisateurs, vocabulaire réduit ou étendu...

I.3.B. Les différents systèmes de reconnaissance automatique de la parole

Face à la difficulté des problèmes rencontrés (essentiellement continuum et variabilité), il n'existe actuellement aucun système de R.A.P. valable pour n'importe quel utilisateur prononçant n'importe quoi, de n'importe quelle façon. Le problème est multidimensionnel et les marges de manoeuvre qui permettent de réduire chaque dimension du problème sont les suivantes:

1. nombre d'utilisateurs
2. mot isolé/parole continue
3. taille et difficulté du vocabulaire
4. milieu protégé/ bruité, parole de laboratoire/ spontanée
5. dialogue oral ou multimodal
6. reconnaissance / compréhension

La démarche actuelle consiste à se placer, tout d'abord, dans un cas simple pour limiter les difficultés. Le cas le plus favorable, celui de la reconnaissance mono-locuteur de mots isolés, permet d'éviter la variabilité inter-locuteurs et le continuum acoustique. La limitation du vocabulaire, du bruit ambiant et de la liberté d'énonciation améliore aussi grandement les performances. Au fur et à mesure que certains problèmes sont résolus (variabilité intra-locuteur, inter-locuteurs, milieu bruité...), les tolérances du dispositif peuvent être étendues. Les conditions de fonctionnement d'un système de R.A.P. apparaissent finalement importantes et donnent naissance à différentes catégories.

I.3.B.a. Système mono ou multilocuteurs

Dans le cas d'une reconnaissance mono-locuteur, le système est adapté à un simple utilisateur. La variabilité inter-locuteurs est ainsi supprimée. Généralement, il existe d'abord une phase d'apprentissage où l'utilisateur prononce un certain nombre d'énoncés. Chaque énoncé est renouvelé plusieurs fois afin de restreindre la variabilité intra-locuteur à des phénomènes aléatoires. Une fois ces informations stockées sous forme d'archives ou de modèles, le système fonctionne par comparaison entre un stimulus provenant de l'unique locuteur et la bibliothèque d'archives relative à cet utilisateur.

Dans le cas d'une reconnaissance multilocuteurs, le système doit fonctionner indépendamment de l'utilisateur. Si la méthode de décodage s'apparente à un système expert, les règles qui sont utilisées doivent être robustes. Si le système se sert de références sonores, celles-ci doivent être suffisamment représentatives pour toucher la plus grande population possible. Certains dispositifs fonctionnant avec des modèles de Markov effectuent ainsi un apprentissage sur plusieurs centaines de locuteurs.

Il existe aussi un type de dispositif mixte: mono-locuteur dans son principe, le système de décodage garde la possibilité de changer de bibliothèque d'archives selon le locuteur, rendant la reconnaissance multi-utilisateurs sans pouvoir être toutefois considéré comme multilocuteurs. Enfin, certains systèmes (Tubach et al., 1990 ; Gilles & Méloni, 1994) comportent une phase d'adaptation à l'utilisateur, au cours de laquelle des informations relatives au locuteur sont saisies. Ces informations sont ensuite utilisées de façon adaptée lors de la phase de reconnaissance proprement dite.

I.3.B.b. Reconnaissance d'énoncés isolés / parole continue

Dans un système de décodage d'énoncés isolés, les éléments sont considérés comme des unités insécables à reconnaître globalement. Il peut s'agir de phrases isolées ou plus fréquemment de mots isolés. Il existe des dispositifs, comme le logiciel Dragon, où les éléments du vocabulaire peuvent être juxtaposés mais ils doivent être prononcés en laissant apparaître des silences entre chaque unité. Autrement dit, il est possible de prononcer une phrase entière en laissant une pause entre chaque mot. Cette contrainte permet ainsi de réduire le problème du continuum acoustique. Ceci est nécessaire non seulement pour faciliter la séparation des unités lexicales, mais aussi pour supprimer les difficultés entraînées par les phénomènes de liaisons et de coarticulation entre la fin d'un mot et le début du suivant.

Dans un système de parole continue, la difficulté supplémentaire réside dans le repérage des unités lexicales à partir du flux phonémique. A titre d'exemple (El-Bèze, 1995, p.138), la suite phonétique /nuzavjõkaʃØtelãvØlɔpsyrlØʃã/ peut donner lieu à 3 millions de transcriptions orthographiques différentes si aucune information sémantique et syntaxique ne vient lever les ambiguïtés. Ainsi, /nu/ peut correspondre au pronom personnel "nous" ou à une forme conjuguée du verbe "nouer", /avjõ/ peut provenir du substantif "avion", du verbe "avoir"...

Il existe enfin un type de dispositif intermédiaire appelé « word-spotting » qui consiste à repérer des mots-clés dans la chaîne parlée. Ce type de fonctionnement s'appuie sur le fait légitime qu'il n'est pas nécessaire de décoder l'ensemble du message vocal pour obtenir l'information nécessaire.

I.3.B.c. Petit/grand lexique, vocabulaire facile/difficile

Autrefois, les performances des machines ne permettaient pas d'appréhender de grands vocabulaires, d'une part, à cause du manque de place pour stocker d'éventuels modèles ou références de mots, d'autre part, par manque de puissance de calcul pour comparer rapidement les unités à reconnaître et l'ensemble du vocabulaire. Aujourd'hui, grâce aux progrès technologiques, cet obstacle semble levé.

Un vocabulaire est qualifié de « difficile » quand certains mots du lexique sont très proches acoustiquement. Le cas extrême est celui des paires minimales où seul un trait relatif à un phonème permet la distinction de deux mots (ex: menthe/Nantes, four/sourd, pile/bile...). Bien que cela puisse se produire sur un petit vocabulaire, il est facile de deviner, de façon intuitive, que plus le lexique utilisé est étendu, plus les risques de confusion entre entités acoustiques voisines sont grands, rendant la tâche plus difficile. Une limitation plus ou moins draconienne du vocabulaire apparaît nécessaire pour obtenir des performances appréciables.

I.3.B.d. Parole en milieu protégé/difficile

Dans le cadre des travaux de recherche en traitement de la parole, la méthodologie courante consiste à utiliser des corpus d'énoncés enregistrés en laboratoire, c'est à dire en milieu non bruité avec des conditions d'élocution contrôlées. Si l'on travaille dans l'optique de réaliser un dispositif industriel, il faut bien garder à l'esprit que le traitement de la parole en situation réelle pose de nombreux problèmes supplémentaires. Ainsi, F. Néel* mentionne l'exemple d'un système de R.A.P. développé au LIMSI qui, en situation réelle (contrôle aérien), obtenait des performances 40 % inférieures à celles évaluées en laboratoire. De même, le CNET évalue à un facteur 3 l'accroissement du taux d'erreurs de reconnaissance en situation industrielle par rapport au laboratoire.

Pour espérer obtenir de bonnes performances en situation réelle, il est nécessaire d'utiliser, d'une part, des techniques de traitement du signal pour débruiter le signal. D'autre part, il est indispensable de faire appel aux études faites sur la parole spontanée afin de gérer des phénomènes tels que les données paraverbales (Kerbrat-Orecchioni, 1990), les pauses, les hésitations, les reprises (Goldman-Eisler,1968; Grosjean & Deschamps, 1975; Duez, 1987; Guaïtella, 1991).

I.3.B.e. Dialogue oral/multimodal

Traditionnellement, l'étude de la communication parlée se limite au canal acoustique. Or, il est bien connu que le canal visuel transporte énormément d'informations parfois plus importantes que celles du canal oral (Schwartz, 1995). Une approche multimodale consiste à utiliser cette information à plusieurs niveaux:

- d'une part au niveau segmental dans l'analyse du mouvement des lèvres pour accéder à la labialité. On peut citer, à titre d'exemple, le projet AMIBE (Montacié et al., 1995).
- d'autre part au niveau suprasegmental dans une optique interactionniste (signifiants proxémiques, posturaux, mimo-gestuels).

Evidemment, les dispositifs à analyse multi-canaux sont plus difficiles à mettre en oeuvre. Inversement, ils exploitent beaucoup mieux les sources d'information de la communication parlée et s'avèrent beaucoup plus robustes en milieu bruité.

* Ecole thématique "Fondements et perspectives en traitement automatique de la Parole", organisée par le GDR-PRC "Communication Homme-Machine", Marseille-Luminy, 1995

I.3.B.f. Reconnaissance de la parole / Traitement du Langage

Il existe différentes nuances dans le décodage de la parole. Le décodage acoustico-phonétique consiste à transcoder le signal vocal (physique continu) en phonèmes (unités abstraites discrètes du langage). Un accès lexical permet une transcription orthographique de la chaîne parlée. Pour cette opération, il est souvent nécessaire de faire intervenir des informations morphologiques, syntaxiques, sémantiques, touchant alors au Traitement Automatique des Langues Naturelles (TALN). Il faut toutefois souligner que la parole n'est pas simplement l'oralisation d'un texte écrit et que certaines techniques de TALN peuvent s'avérer inadaptée à la langue parlée. Signalons à ce sujet les études faites par le Groupe Aixois de Recherches en Syntaxe (G.A.R.S., 1977-1993) rebaptisé récemment Groupe d'Etude sur les Données Orales, qui met en évidence la spécificité de la syntaxe de l'oral. Quant aux systèmes de compréhension de la parole, ils restent encore peu développés compte tenu de la difficulté de la tâche, notamment dans la formalisation du monde réel de l'humain.

I.3.B.g. L'état actuel

Minimalistes au début (mots isolés, petit vocabulaire, mono-locuteur), les systèmes de R.A.P. deviennent de plus en plus complexes et intègrent de plus en plus de difficultés au fur et à mesure que les travaux de recherche progressent. L'accroissement des capacités des calculateurs contribuent aussi à pousser plus loin les performances. Un historique et un état de l'art sera présenté plus loin. Il reste à noter que, dans la réalisation des différents dispositifs décrits ci-dessus, diverses méthodes peuvent être utilisées selon les objectifs choisis.

I.3.C. Les différentes approches

Dans la réalisation des systèmes de reconnaissance, (Schwartz et al., 1988) distinguent deux types d'approches: les techniques globales et les méthodes analytiques.

I.3.C.a. Les approches globales

I.3.C.a.i. Le principe

Une approche globale consiste à considérer le message à identifier comme une forme insécable qu'il s'agit de reconnaître en lui attribuant une classe d'appartenance. Ainsi, dans le cas d'une reconnaissance de mots isolés, le dispositif cherche à reconnaître globalement tout le stimulus et à associer cette suite de sonorités à un mot, sans privilégier ou négliger une partie de cet ensemble, sans tenir compte de détails plus pertinents que d'autres. On parle aussi, sans connotation péjorative, d'analyse « aveugle » ou de modèle « d'ignorance ».

En général, il existe une phase préliminaire dite *d'apprentissage*. L'idée est de saisir des caractéristiques acoustiques du signal pour les stocker sous forme de références (méthodes métriques) ou de modèles (méthodes connexionnistes ou stochastiques). Dans la phase de reconnaissance proprement dite, le dispositif cherchera à mettre en relation les paramètres du stimulus avec ceux qui sont conservés en archives ou sous forme de modèles. Un système de R.A.P. nécessitant un apprentissage n'est pas forcément mono-locuteur. Il peut intégrer intelligemment des énoncés produits par plusieurs personnes pour former une ou plusieurs références communes (prototypes ou modèles). Cet apprentissage se fait une bonne fois pour

toute en laboratoire et l'archivage ainsi réalisé pourra servir, en théorie, à n'importe quel utilisateur.

I.3.C.a.ii. La notion d'espace métrique

Les méthodes globales nécessitent l'utilisation d'un espace métrique. Les axes de cet espace sont constitués par différents paramètres acoustiques. Ainsi, avec une analyse spectrale sur 3 bandes (basses, moyennes et hautes fréquences), il est possible de créer un espace à 3 dimensions où les coordonnées sur les axes sont relatives à l'énergie dans chaque canal. La Figure 14 illustre cette opération. A l'intérieur de cet espace métrique, deux points proches représentent le même phénomène acoustique, et en définitive, le même son.

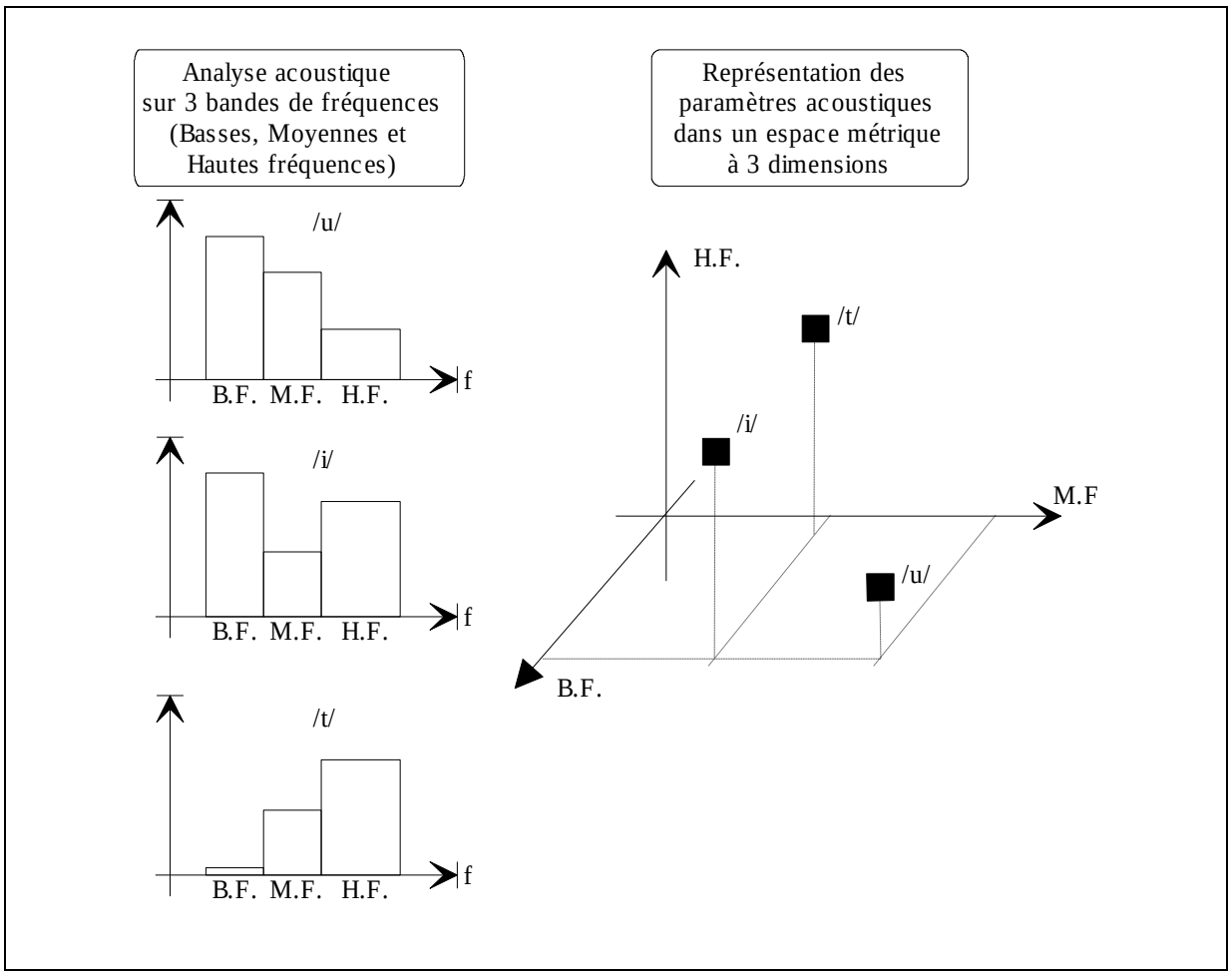


Figure 14: Exemple d'espace métrique tridimensionnel associé à des paramètres acoustiques issus d'une analyse à 3 bandes

Si le signal de parole était stable, un point suffirait à caractériser un phonème. Cette hypothèse n'étant pas vérifiée, la démarche généralement adoptée consiste à repérer la position des points en fonction du temps pour un énoncé précis. Le résultat obtenu est alors une matrice de points caractérisant cet énoncé-là. Le regroupement des matrices correspondant à chaque élément du vocabulaire permet la constitution du « dictionnaire » du lexique utilisé.

La phase de reconnaissance consiste ensuite à repérer de façon métrique (notion de distance et de proximité) un point issu des paramètres du stimulus avec ceux du dictionnaire. Le problème principal réside dans le choix des paramètres, de la distance, de l'alignement temporel et de la façon de tenir compte d'une utilisation multilocuteurs.

I.3.C.a.iii. Les approches statistiques

Dans une approche statistique, l'idée est « d'intégrer de façon aveugle et implicite des connaissances (sur la parole) par examen d'un corpus représentatif du langage envisagé, de façon à déterminer un ensemble de données statistiques » (Haton et al., 1991, p.14).

- Les **modèles de Markov** (Hidden Markov Model ou H.M.M.) ont été explorés depuis longtemps (Jelinek, 1976). Une bonne synthèse du sujet est disponible dans (Levinson et al., 1983). Ces techniques peuvent être considérées comme une extension des méthodes de comparaison de formes où la reconnaissance d'une série d'éléments consiste à trouver le chemin le plus probable dans la description des modèles de référence. Les références des unités de parole sont généralement stockées sous la forme de sources de Markov composées d'états et de transitions entre ces états. A chaque arc liant un état *i* à un état *j* est associée une probabilité de transition. Il existe aussi une probabilité de sortie représentant la probabilité d'émettre un symbole *k* pour la transition liant *i* à *j*. Les symboles *k* sont en fait des coefficients issus de la phase d'extraction de paramètres acoustiques (cf. § « L'extraction de paramètres acoustiques », p.17). Ce sont les données observables. L'apprentissage des caractéristiques du modèle est généralement réalisé par un algorithme dit de «Baum-Welch». La phase de reconnaissance proprement dite consiste à maximiser la probabilité d'être en présence d'un modèle connaissant la séquence d'observation. Cette maximisation passe par un algorithme dit «algorithme de Viterbi». Ce type de technique peut s'utiliser aussi bien pour modéliser la parole que le « langage » (El-Bèze, 1995).

- Le principe de base des **approches connexionnistes** (réseaux neuro-mimétiques ou R.N.M.) consiste à imiter le comportement du cerveau en simulant un grand nombre de neurones connectés entre eux en plusieurs couches. Les premières recherches en la matière remontent aux études sur le perceptron (Minsky & Papert, 1969). Une bonne introduction des approches neuro-mimétiques est disponible dans (Mac Clelland & Rumelhart, 1988) et (Mac Cord Nelson & Illingworth, 1991). En résumé, un réseau neuro-mimétique peut être considéré comme un dispositif qui:

  $\Rightarrow$ associe une classe de sortie à un objet d'entrée
  $\Rightarrow$ donne des résultats en fonction de données
  $\Rightarrow$ peut s'auto-organiser pour classer des données

L'analogie avec le fonctionnement humain est réelle par le fait qu'il existe toujours une phase d'apprentissage et une phase d'utilisation proprement dite pour une tâche spécifique. La connaissance n'est pas localisée mais est répartie sur l'ensemble des coefficients du réseau. Dans les dispositifs de type perceptron, la connaissance est intégrée en mettant en correspondance un stimulus avec une réponse symbolique (Figure 15). Appliquées à la parole,

ces techniques se heurtent à un problème spécifique dont il est difficile de rendre compte: l'aspect temporel. En effet, la parole est un signal dynamique et un système de reconnaissance de la parole doit être capable de capturer les propriétés temporelles du message vocal et non de se limiter à des prises de vues statiques. Pour cela, a été mis au point un type de réseau particulier: le Time Delay Neural Net (T.D.N.N.) (Waibel, 1987). La couche d'entrée du réseau est alimentée par des paramètres physiques extraits à partir du signal (coefficients MEL le plus souvent). L'un des axes de la matrice d'entrée est le temps. Généralement, chaque neurone d'une couche donnée est connecté à tous les neurones de la couche suivante. Ce n'est pas le cas pour le réseau TDNN, qui groupe, par exemple, trois trames, c'est à dire trois vecteurs de paramètres de la couche d'entrée, en une macro-fenêtre pour les connecter à un seul noeud de la couche suivante. Chacun de ces noeuds forme une représentation condensée de l'information présente dans cette partie du signal et non ailleurs. Le réseau décale ensuite la macro-fenêtre d'une trame et recommence l'opération.

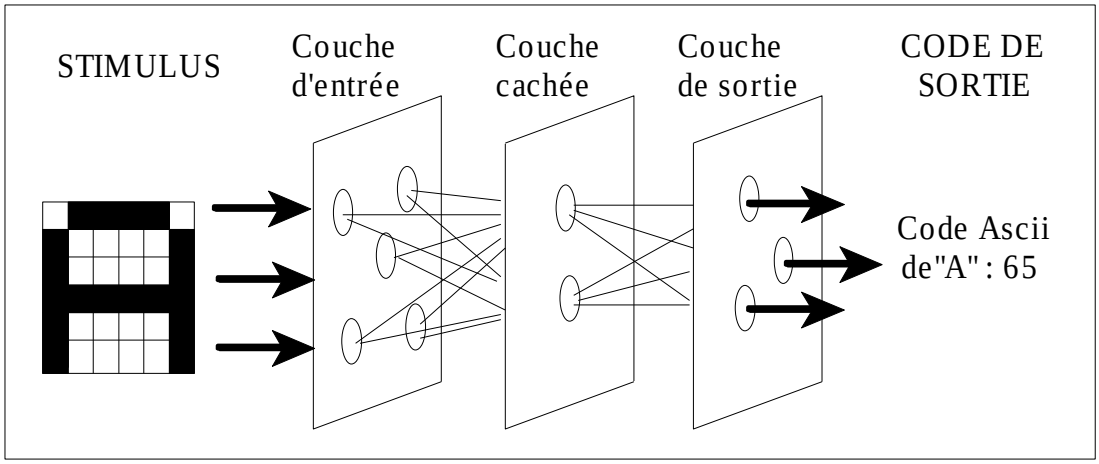


Figure 15: Mise en correspondance de deux objets par un réseau neuronal

bilan sur les méthodes statistiques

Les méthodes statistiques peuvent être considérées comme des classificateurs: les HMM modélisant des prototypes, les réseaux neuro-mimétiques proposant des frontières de catégories. Ces techniques semblent très efficaces car elles intègrent automatiquement des connaissances à condition que le corpus d'apprentissage contienne suffisamment de données pour réaliser une bonne modélisation. A titre d'exemple, certains modèles de langage sont entraînés sur des textes d'environ 250 millions de mots !!

Il faut toutefois signaler que ces méthodes se heurtent actuellement à un plafonnement des performances et toute amélioration apparaît extrêmement difficile. De plus, il existe une dégradation importante en dehors des conditions idéales. Enfin, l'adaptation à des conditions particulières s'avère difficile du fait de l'immense effort d'apprentissage à effectuer avec de nouvelles références.

I.3.C.b. Les approches analytiques

Dans les approches analytiques, l'idée est de formaliser explicitement des connaissances en matière de décodage de la parole. Ces dispositifs utilisent généralement un ensemble de règles et peuvent s'apparenter à des systèmes experts. Dans le cas de la parole, l'opération de décodage consiste à effectuer un repérage temporel des segments puis une identification des éléments en tenant compte des effets de contexte, en s'attardant sur des détails pertinents, c'est à dire en délaissant des phénomènes peu porteurs d'information... L'analyse est dirigée par la connaissance. Les règles doivent être robustes pour résister aux différentes sources de variabilité. Une bonne façon de procéder est d'utiliser la méthode des « îlots de confiance » où le décodage est réalisé en plusieurs passes. Dans un premier temps, le continuum acoustique est analysé grossièrement de gauche à droite. Pendant cette recherche, les parties où le décodage est le plus sûr sont repérées. En tenant compte des phénomènes de contexte, le processus de reconnaissance s'étend ensuite, en tache d'huile, de part et d'autre des îlots de confiance, jusqu'à ce que toutes les données acoustico-phonétiques soient utilisées.

L'approche analytique a d'abord été utilisée au MIT (Zue & Cole, 1979) puis poursuivie par différentes équipes (Haton & Lazrek, 1984; Stern et al., 1986; Carbonnel et al., 1986; Méloni & Bulot, 1986; Mizoguchi et al., 1986). L'obstacle majeur actuellement réside dans la formalisation du savoir de l'expert. Les modèles à base de connaissances doivent déterminer, prévoir et réaliser dans un système unifié tous les faits et lois qui gouvernent la parole. Ils doivent intégrer un nombre massif de relations entre les faits et les lois pour interpréter rapidement le message oral. La tâche relève du défi.

I.3.C.c. Les approches mixtes

L'image caricaturale du linguiste aux méthodes analytiques et de l'informaticien aux techniques stochastiques poursuit la communauté scientifique de la communication parlée depuis plusieurs années. Nous rappelons le fameux jugement d'un spécialiste: « Every time we fire a phonetician linguist, the performance of our system goes up »*. Pourtant, il semble que le fossé se comble entre les deux orientations scientifiques pour donner naissance à des approches mixtes qui tirent avantage de chacune des méthodes. Un exemple révélateur est celui du système de reconnaissance de lettres épelées développé au CRIN à Nancy: la première passe du décodage est effectuée à l'aide des modèles de Markov; les cas ambigus (ex: "P" /pe/ vs "T" /te/) sont traités analytiquement (étude précise de l'explosion) à l'aide de réseaux neuro-mimétiques, qui opèrent la discrimination. Un tableau récapitulant les points forts et points faibles de chacune des méthodes est fourni dans (De Leeuw & Caelen, 1994).

I.3.D. Les différentes stratégies

Dans la réalisation des systèmes de R.A.P., on distingue généralement les stratégies montantes et les stratégies descendantes.

* Jelinek F., IEEE ASSP workshop 1985, Arden House

- *Les stratégies montantes (bottom-up)*

  Une stratégie montante se fonde sur le décodage d'un niveau d'analyse inférieur pour émettre des hypothèses sur les niveaux supérieurs. Ainsi, à partir de paramètres physiques extraits sur le signal de parole (niveau acoustique), diverses hypothèses sur l'identification de segments phonémiques (niveau phonétique) sont proposées, puis sur les mots (niveau lexical)... La masse de données à traiter est ainsi réduite, au risque, cependant, de propager une erreur due à une mauvaise interprétation. Cette technique est la plus répandue.

- *Les stratégies descendantes (top-down)*

  Par opposition, dans une technique descendante, le système propose des hypothèses, connaissant certaines contraintes syntaxiques, lexicales ou phonétiques pour prédire des mots qui vont apparaître dans l'énoncé ; ensuite, le niveau inférieur (niveau acoustique) est consulté pour confirmer ou infirmer ces hypothèses.

- *Les doubles stratégies*

  La solution la plus judicieuse, mais difficile à mettre en oeuvre, consiste à effectuer un va-et-vient ascendant et descendant. Dans le système développé à Avignon (Gilles & Méloni., 1994), une première phase permet de dresser une cohorte de mots possibles à partir du décodage ascendant. La phase descendante écarte les hypothèses improbables en effectuant une analyse orientée sur le signal et permet de dégager de façon efficace les meilleurs candidats à la reconnaissance.

I.4. L'état des connaissances en reconnaissance automatique de la parole

Avant de prendre conscience de l'état des recherches et des réalisations actuelles en matière de reconnaissance automatique de la parole, faisons d'abord un rapide historique de ces travaux.

I.4.A. Historique des travaux effectués en reconnaissance automatique de la parole

La production de la parole par des moyens artificiels est un domaine que l'homme a abordé dès le siècle des Lumières (machine à parler du baron de Kempelen en 1779, têtes parlantes de l'abbé Mical en 1780). Si la synthèse vocale a connu très vite des réalisations, le problème dual, celui de la reconnaissance de la parole par des machines, a été abordé beaucoup plus tardivement et les progrès, moins rapides qu'en synthèse, n'ont suivi qu'au rythme du développement des moyens informatiques. Le descriptif ci-après est une synthèse de (Pierrel, 1987; Mariani, 1990; Haton et al., 1991). Il expose les grandes étapes des travaux effectués en reconnaissance automatique de la parole.

- *Dans les années 30*, R.J. Wensley construit le Televox, premier automate très rudimentaire exécutant quelques mouvements en fonction d'ordres téléphoniques.
- *Les années 50* voient l'apparition de dispositifs câblés de R.A.P. valables pour une dizaine de mots.

⇒ En 1952, Daves développe sur un circuit électronique câblé une reconnaissance mono-locuteur de dix chiffres qui aboutira plus tard sur un système acceptant plusieurs locuteurs (Duley & Balashek, 1958)

⇒ En 1956, Olson et Belar mettent au point la machine à écrire phonétique valable pour une dizaine de mots.

⇒ En 1958, Denes réalise un dispositif à deux étapes où une reconnaissance acoustique est affinée par l'emploi de contraintes linguistiques.

- *Au cours des années 60,* l'apparition des calculateurs numériques fournit de nouveaux moyens à la R.A.P. qui se tourne alors vers des méthodes où l'outil informatique devient essentiel.

⇒ En 1965, apparaît le premier dispositif de segmentation de la parole continue en phonèmes (Reddy, 1966).

⇒ En 1966, sont mis au point plusieurs systèmes multilocuteurs avec un vocabulaire d'une quarantaine de mots (King & Tunis, 1966 ; Cold, 1966)

⇒ En 1968, Alter et Reddy introduisent des connaissances linguistiques exploitées un an plus tard par Vicens qui réalise un dispositif utilisé par le robot manipulateur Hand-Eye-Ear.

⇒ En 1970, une équipe grenobloise (Tubach, 1970) met au point un système qui permet la dictée vocale d'un programme Algol prononcé mot à mot.

- *Les années 70* voient la création de multiples équipes de recherche en R.A.P. La nécessité de faire appel à des contraintes linguistiques dans le décodage automatique de phrases apparaît clairement, alors que la R.A.P. avait été jusque-là considérée comme un problème d'ingénierie. La fin de la décennie voit se terminer la première génération des systèmes commercialisés.

⇒ En 1971, le lancement du projet américain financé par l'A.R.P.A. (Advanced Research Projects Agency) avec un budget de 15 millions de dollars donne un véritable coup de fouet aux travaux en R.A.P.

⇒ En 1972, la compagnie Threshold commercialise pour un prix de 20 000 dollars le premier appareil de reconnaissance de mots - le VIP 100 - avec un vocabulaire de 32 mots et une utilisation mono-locuteur.

⇒ En 1976, la fin du projet A.R.P.A. (Klatt, 1977) voit la naissance de plusieurs systèmes acceptant la parole continue, un vocabulaire de 1000 mots, une grammaire artificielle adaptée à une tâche précise, avec plusieurs locuteurs coopératifs, dans une ambiance calme, avec un bon microphone. On retient les systèmes HWIM de Bolt Beranek & Newman, HARPY et HEARSAY de Carnegie Mellon University.

⇒ Dans le sillage d'A.R.P.A., de nombreux projets sont mis en route (De Mori, 1978). Les plus importants sont ceux d'IBM et du Standford Research Institute (Walker, 1978) aux Etats-Unis, ainsi que les travaux de la NEC (Nippon Electric Company) et de la NTT (Nippon Telegraph and Telephon) au Japon.

⇒ En 1978, VERBEX commercialise le premier système multi-locuteurs - le M1800 - pour un prix de 80.000 dollars. En même temps, la compagnie japonaise NEC

propose pour 60.000 dollars le premier dispositif reconnaissant des suites de mots enchaînés.

⇒ En 1979, Interstate met en vente le premier dispositif de reconnaissance à microprocesseur sur une carte de circuits imprimés, acceptant jusqu'à 100 mots pour un utilisateur (VRM, 1000 dollars).

- *Les années 80* mettent à profit les possibilités sans cesse croissantes de l'informatique dont les progrès exponentiels permettent d'acquérir des performances supérieures à faible coût. Les réalisations industrielles restent limitées.

  ⇒ La décennie est tout d'abord marquée par des techniques fondées sur la reconnaissance de formes, l'intelligence artificielle, les connaissances linguistiques, ce qui entraîne un rapprochement des différents domaines de recherche touchant à la parole. En même temps, les méthodes stochastiques commencent à prouver leur efficacité.

  ⇒ En France, la recherche en traitement automatique de la parole est relativement performante. En 1981, un groupe de recherche coordonné (le GRECO-PRC pôle Parole) mis en place par le Ministère de la Recherche et le C.N.R.S. a permis de regrouper les différents laboratoires de recherche français en traitement de la parole. Cette coordination a permis de créer une « culture commune » et d'effectuer un rapprochement entre recherche fondamentale et recherche appliquée. Un des premiers objectifs concrets a été la constitution d'une base de données lexicale (BDLEX) ainsi que d'une base de données des sons du français (BDSONS) absolument indispensables pour évaluer de façon objective les différents systèmes. Toutefois, les travaux de recherche pourtant fructueux n'ont abouti que sur un nombre restreint de réalisations industrielles.

I.4.B. Etat actuel en reconnaissance automatique de la parole

D'après (Haton et al., 1991), le traitement automatique de la parole est devenu une activité économique non négligeable comme le montre les efforts consentis par des grands groupes industriels tels que IBM, Texas Instruments, NEC, Hitachi, Matsushita, Toshiba, Philips, Siemens, Thomson... D'importants projets ont été mis en place un peu partout comme le programme D.A.R.P.A aux Etats-Unis, 5ème génération et ATR au Japon, ESPRIT en Europe. Il est à noter que le traitement automatique de la parole est fortement dépendant des innovations technologiques en matière de composants électroniques (vitesse des microprocesseurs et autres D.S.P. pour le calcul, capacité des mémoires pour le stockage des données), d'architectures (machines parallèles, processeurs symboliques) et des échanges de données (accès à des bases de données sonores, phonologiques, lexicales ou autres types de connaissances...).

Qu'en est-il de l'état des réalisations ?

- La reconnaissance de petits vocabulaires en mots isolés pour une utilisation mono-locuteur apparaît maintenant résolue à faible coût. On peut citer pour exemple la carte

«SoundBlaster» fonctionnant en temps réel sur un compatible PC. Malheureusement, ce type de reconnaissance s'avère vite limité.

- Le décodage multilocuteurs de petits vocabulaires semble lui aussi en voie de bon fonctionnement comme le prouve le système PHIL90 du CNET, robuste à travers le réseau téléphonique, avec détection de mots clés et rejet des mots inconnus. Fondé sur des modèles de Markov, il nécessite une collecte de données considérable (800 locuteurs) et l'utilisation d'un vocabulaire limité (50 mots). Intégré dans le serveur des «Baladins», le système est confronté à une situation réelle.
- Le décodage de grands vocabulaires contenant plusieurs milliers de mots isolés existe dans des dispositifs comme Tangora (IBM), Kurzweil, Dragon Systems. Toutefois, l'élocution reste contrainte par le fait qu'il faille laisser des silences entre les mots. Une telle restriction rend peu ergonomique un tel dispositif.
- Les systèmes de reconnaissance de parole continue semblent difficiles à mettre au point. Sur un vocabulaire limité et une tâche bien précise, des dispositifs tels que ceux réalisés par Fujitsu ou Verbex seront probablement utilisables dans quelques années. Une chose est sûre: le temps du dialogue naturel multilocuteurs entre l'homme et la machine se situe au delà des années 2000.

Bilan

Différentes réalisations ont mis en évidence la supériorité des performances obtenues à l'aide des méthodes stochastiques par rapport à toute autre approche. Néanmoins, de récentes interrogations de spécialistes de la R.A.P. indiquent que les performances des méthodes stochastiques appliquées aux coefficients MFCC (cf. § « L'extraction de paramètres acoustiques », p.17) commencent à plafonner malgré les progrès techniques (Bourlard, 1996). Il semble que le passage de la parole multilocuteurs de mots isolés à de la parole continue en conditions réelles d'utilisation ne pourra être réalisé uniquement sur la base de méthodes stochastiques. Par conséquent, il faudrait développer de nouvelles représentations du signal de parole et de nouvelles techniques de RAP capables d'intégrer simultanément diverses sources d'information.

I.5. Notre approche

Notre approche est une alternative aux méthodes stochastiques. Elle s'inscrit dans un courant scientifique actuel destiné à faire progresser la recherche fondamentale pour résoudre des problèmes apparemment insolubles dans un cadre purement stochastique. Notre objectif n'est pas de fournir un système industriel à cours terme, ni de rentrer dans une logique de performance coûte que coûte. Nous souhaitons garder un critère de contrôlabilité et d'explicabilité, c'est à dire la possibilité de donner une explication aux erreurs de décodage, ce qui permet de faire évoluer le dispositif. Notre objectif est d'examiner dans quelle mesure un modèle à base de connaissances est capable de décoder de façon automatique la structure phonique de la parole sans recourir aux méthodes stochastiques. Une telle approche comporte deux avantages non négligeables :

- servir de banc de test aux travaux effectués sur la communication parlée, notamment en confrontant aux données les connaissances et modèles proposés par les linguistes.
- permettre de faire évoluer un dispositif de décodage et autoriser ainsi un déblocage progressif des problèmes.

Dans ce but, notre effort s'est porté sur deux axes:

- la recherche et l'utilisation de paramètres acoustiques pertinents autres que les standards coefficients MFCC (Davis & Mermelstein, 1980) utilisés actuellement dans 90 % des systèmes de reconnaissance. A ce sujet-là, il convient de se poser une question: ces paramètres sont-ils réellement optimaux pour la reconnaissance de la parole sachant qu'ils fonctionnent très bien en identification du locuteur ? De tels résultats laissent penser que cette technique d'extraction d'informations est fortement imprégnée par le locuteur, ce qui est gênant dans un dispositif indépendant du locuteur.
- la mise en place d'une architecture originale où fonctionnent en parallèle divers processus de décodage .

Nous nous plaçons dans le cadre de reconnaissance de grand vocabulaire, ce qui nécessite une identification d'unités de type phonémique. Dans ce but, le système comporte un étage de décodage acoustico-phonétique. Le dispositif est construit indépendamment du locuteur et sans session d'adaptation. Nous limitons la reconnaissance à un accès lexical, sans analyse syntaxico-sémantique.

II. CONNAITRE LA PAROLE POUR MIEUX LA TRAITER

« Peut-on renoncer à comprendre quelque chose même en sachant qu'on y perdra tout. »

Bernard Faucon

Plan du chapitre

RESUME

La parole est la manifestation d'un processus de communication humain. Elle permet la transmission d'information par le canal acoustique mais aussi visuel. Un tel échange nécessite le partage de connaissances communes entre l'émetteur du message et le récepteur. Ce message peut être à la fois verbal, exprimé par des mots, ou para-verbal, exprimé par des émotions... La robustesse de la communication orale provient de la grande redondance du signal de parole. De nombreuses sources de connaissances apportent leur lot d'informations et permettent le décodage final du message: identification des sons, accès au lexique, adéquation syntaxique ou sémantique, interprétation pragmatique...

Dans le cadre de la théorie de l'information, le décodage de la parole s'apparente à un processus de compression du débit d'information (de 200 000 bits/s à 35 bits/s pour la reconnaissance de mots). La tâche de reconnaissance automatique de la parole est difficile et n'apparaît réalisable qu'en faisant collaborer toutes les sources d'informations. Le problème provient du manque de connaissances que nous possédons sur ces processus et la difficulté de les formaliser. La solution apparaîtra probablement quand notre savoir appréhendera mieux les aspects physiologiques, acoustiques et linguistiques de la communication orale.

II.1. A la recherche de l'information

Le rôle d'un système de R.A.P. est d'accéder à l'information présente dans un message oral. Cette recherche de l'information est loin d'être simple et immédiate. Plongeons-nous, pour cela, dans l'univers de la communication.

II.1.A. La communication

II.1.A.a. Généralités

De façon schématique, les processus de communication impliquent différents éléments (Figure 16):

- *l'émetteur* est l'entité à l'origine de la communication (ex: un livre, un locuteur...)
- le *récepteur* est l'entité à laquelle est destinée la communication (ex: lecteur, auditeur...)
- le *canal* est le moyen utilisé pour la communication (ex: texte, air libre, téléphone..)
- le *signal* est l'aspect matériel du message (ex: images, sons...)
- *l'information* est le sens associé aux signaux.
- un *bruit* est une perturbation troublant la communication.
- les *connaissances communes* sont un ensemble d'informations partagées par l'émetteur et le récepteur (clés du code)

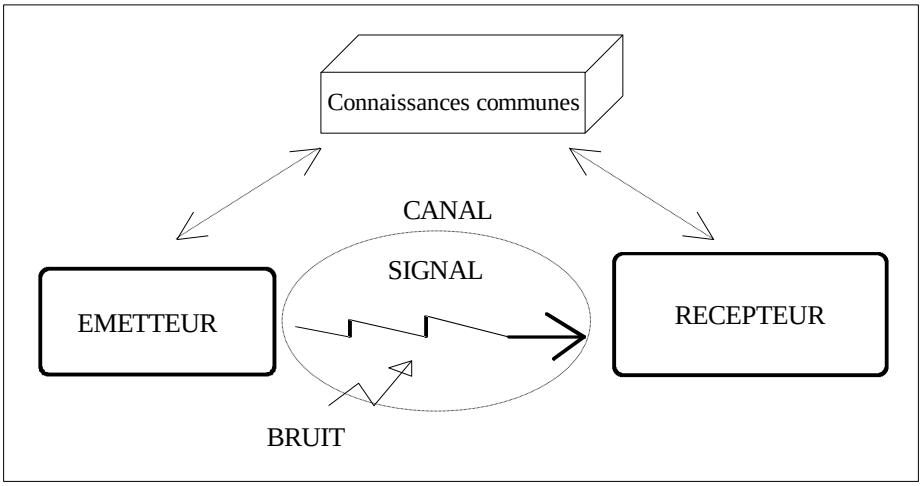


Figure 16: La chaîne de communication

Un processus de communication consiste à faire passer une idée (le *signifié*) à l'aide d'un moyen (le *signifiant*). Un *signe* est une convention associant un signifiant et un signifié (Saussure, 3e édition en 1968). Un *code* est à la fois un répertoire de signes (le *lexique*) et un ensemble de règles d'agencement (la *grammaire*). Une connaissance partagée du code est nécessaire entre les acteurs d'un processus de communication (Figure 16). La *sémiologie* est la science qui étudie les systèmes de signes (ex: signalisations, langages, ...). Au sens général, un *langage* est un *système sémiologique*. Il existe différents types de langages employés par l'être

humain: le langage visuel des sourds-muets ou des plongeurs, une communication par le toucher comme le braille, une communication sonore comme le langage tambouriné, le langage sifflé.

II.1.A.b. La communication parlée

La *langue* est un code. La *parole* est traditionnellement définie comme l'usage de la langue. Le *langage* est l'association de la langue (système abstrait) et de la parole (réalisation physique). En fait, comme le signale Saussure (Saussure, 3e édition en 1968, p.37), « ces deux objets sont étroitement liés et se supposent l'un l'autre: la langue est nécessaire pour que la parole soit intelligible et produise tous ses effets; mais celle-ci (la parole) est nécessaire pour que la langue s'établisse... Il y a donc interdépendance de la langue et la parole. » De plus, il existe un lien très fort entre langue et écriture. Aussi, bien souvent, la communication parlée est appréhendée uniquement comme un processus sonore verbal, c'est à dire l'expression de mots à travers le canal acoustique ou encore, l'oralisation d'un écrit. Pourtant, cette réduction est abusive pour les raisons suivantes:

- de nombreuses études montrent que les phénomènes vocaux para-verbaux sont essentiels dans la communication parlée, que ce soit d'un point de vue prosodique (accentuation, rythme) et paralinguistique (type de voix, émotions, ...).

- de nombreuses études montrent que le canal visuel est extrêmement important à la fois
  - ⇒ au niveau verbal dans la perception de la parole visible (Abry & Perrier, 1995)
  - ⇒ au niveau non-verbal dans la perception des gestes (Cosnier, 1982)

II.1.A.c. A la recherche de l'information sonore verbale

Nous avons conscience que considérer la parole uniquement comme un phénomène sonore verbal est une réduction. Pourtant, nous adopterons ce point de vue par la suite afin de limiter notre champ d'investigation. Sous le terme « information », sont souvent regroupées différentes définitions suivant le point de vue duquel on se place. Pourtant, que l'on soit linguiste ou physicien, cette notion reste la même, bien qu'on en parle en termes différents. Nous allons donc nous efforcer de regrouper ces points de vue pour travailler dans le même sens: rechercher l'information verbale présente dans le signal de parole.

II.1.B. Parole et information d'un point de vue qualitatif

Cette section aborde la notion d'information d'un point de vue linguistique. Bien que qualitative, cette description utilise facilement des notions mathématiques relatives à la géométrie et à la théorie des ensembles. Il est ainsi question de plan de description linguistique dans lequel se dessine un axe spatio-temporel (axe spatial gauche/droite dans le cas de message écrit en français, axe temporel dans le cas de message oral) le long duquel sont mises en concurrence différentes unités linguistiques.

II.1.B.a. Le plan de description linguistique

La relation *syntagmatique* est la présence successive d'unités élémentaires dans la chaîne de l'énoncé (ex: succession spatiale des syllabes, des mots dans l'écriture, succession temporelle des phonèmes dans la parole). Les relations *paradigmatiques* unissent les unités élémentaires (les *syntagmes*) qui peuvent faire l'objet d'un choix et qui, de ce fait, sont mutuellement exclusives en un point de la chaîne. On appelle *paradigme* ou classe de substitution un ensemble d'unités mutuellement exclusives à la même position (ex: verbes, noms, articles...). La Figure 17 fournit des exemples de ces relations. Les unités de même type nouent entre elles des relations de type paradigmatique et syntagmatique dans un même plan appelé *plan de description*, qui correspond à un niveau d'information. Il y a autant de plans de description que de types différents d'unités (plan phonémique, phonologique, lexical, syntaxique...).

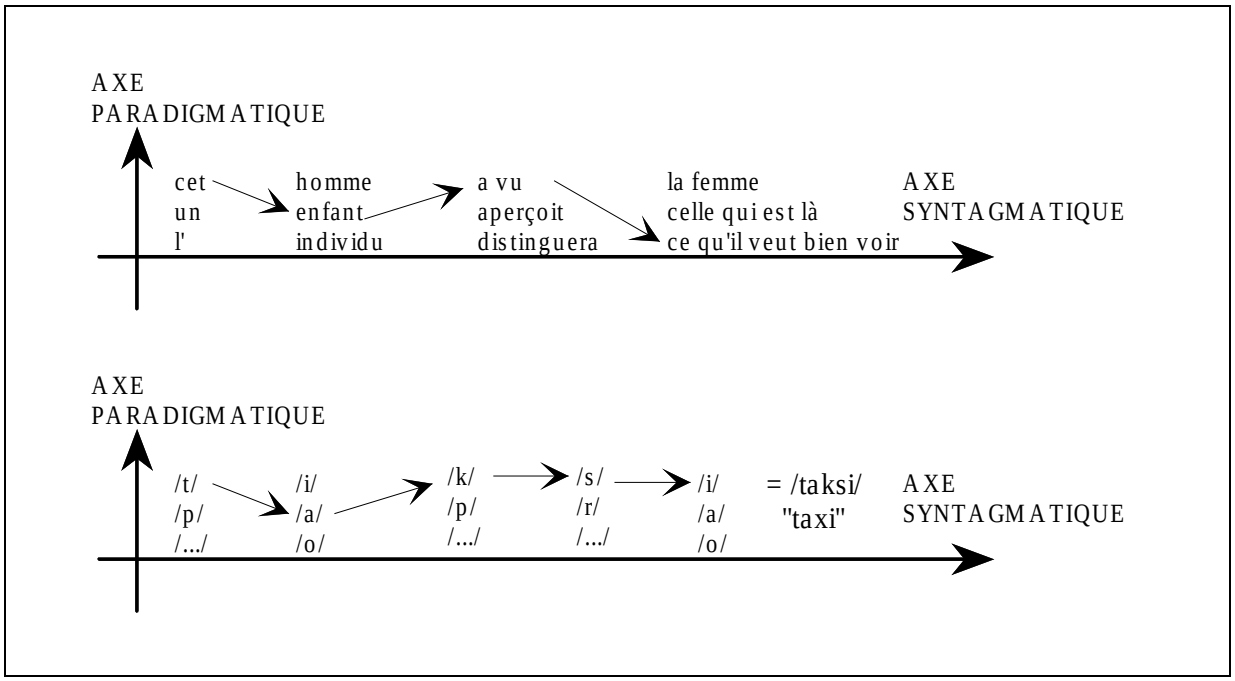


Figure 17: Exemples de relations syntagmatique et paradigmatique

Sur la Figure 18, les données observables (ex: signal acoustique) sont soumises à une première interprétation fournissant un premier plan de description (ex: traits phonétiques). Ensuite, les expressions linguistiques interprétées par {I} sont à nouveau interprétées fournissant un deuxième plan de description {II} (ex: phonèmes). Un modèle grammatical organisé en un nombre variable de plans de description est appelé *grammaire à niveaux multiples*. Les unités d'un niveau supérieur sont considérées comme des *unités virtuelles* ou idéalisées par rapport à celles qui sont issues des niveaux inférieurs, appelées *réalisations* ou *actualisations*. Les relations unissant des unités de même type sont appelées *relations homogènes*. Il existe, de même, des liens entre les différents niveaux.

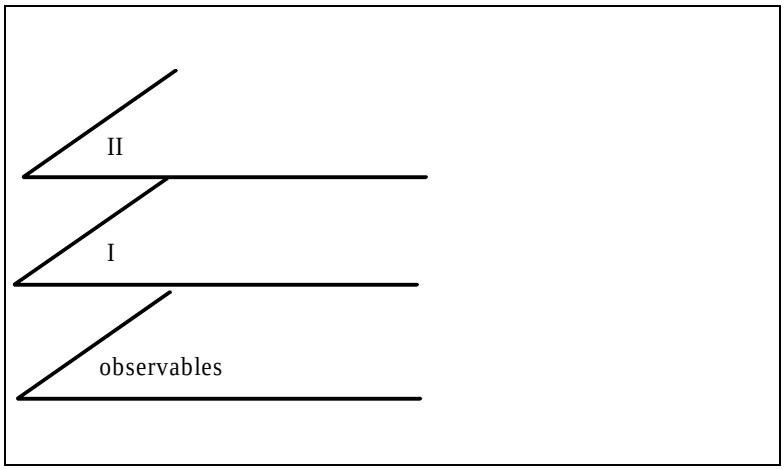


Figure 18: Grammaire à niveaux multiples.

II.1.B.b. Les relations inhomogènes

Les relations unissant des unités de type différent sont dites *inhomogènes*. Il existe deux sens dans ces relations comme le montre la Figure 19. La *lecture interprétative* décode les séquences acoustiques en significations (cas de la reconnaissance de la parole). Le *sens transformationnel* code des unités virtuelles en réalisations (cas de la synthèse vocale). Si à chaque unité d'un plan {I} correspond une et seulement une unité du plan {II} et inversement, les relations sont dites *biunivoques*. Malheureusement, les relations sont rarement biunivoques.

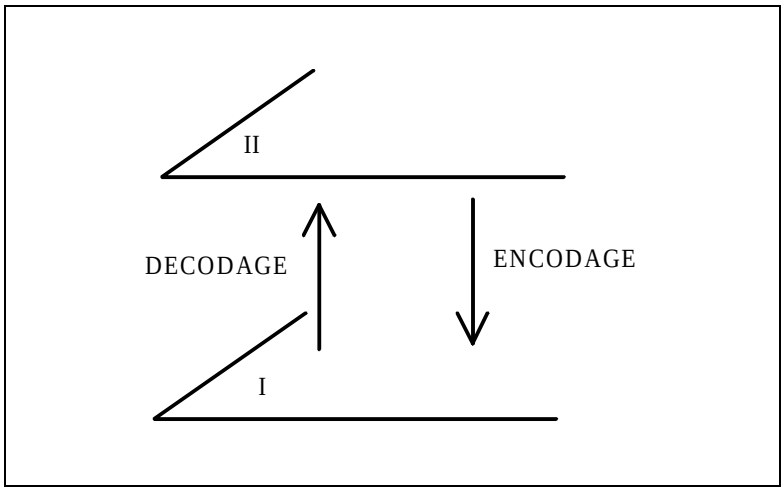


Figure 19: Encodage et décodage.

II.1.B.c. Les différents plans de description de la parole

Une description traditionnelle et simplificatrice des différents plans de description de la parole est proposée en Figure 20.

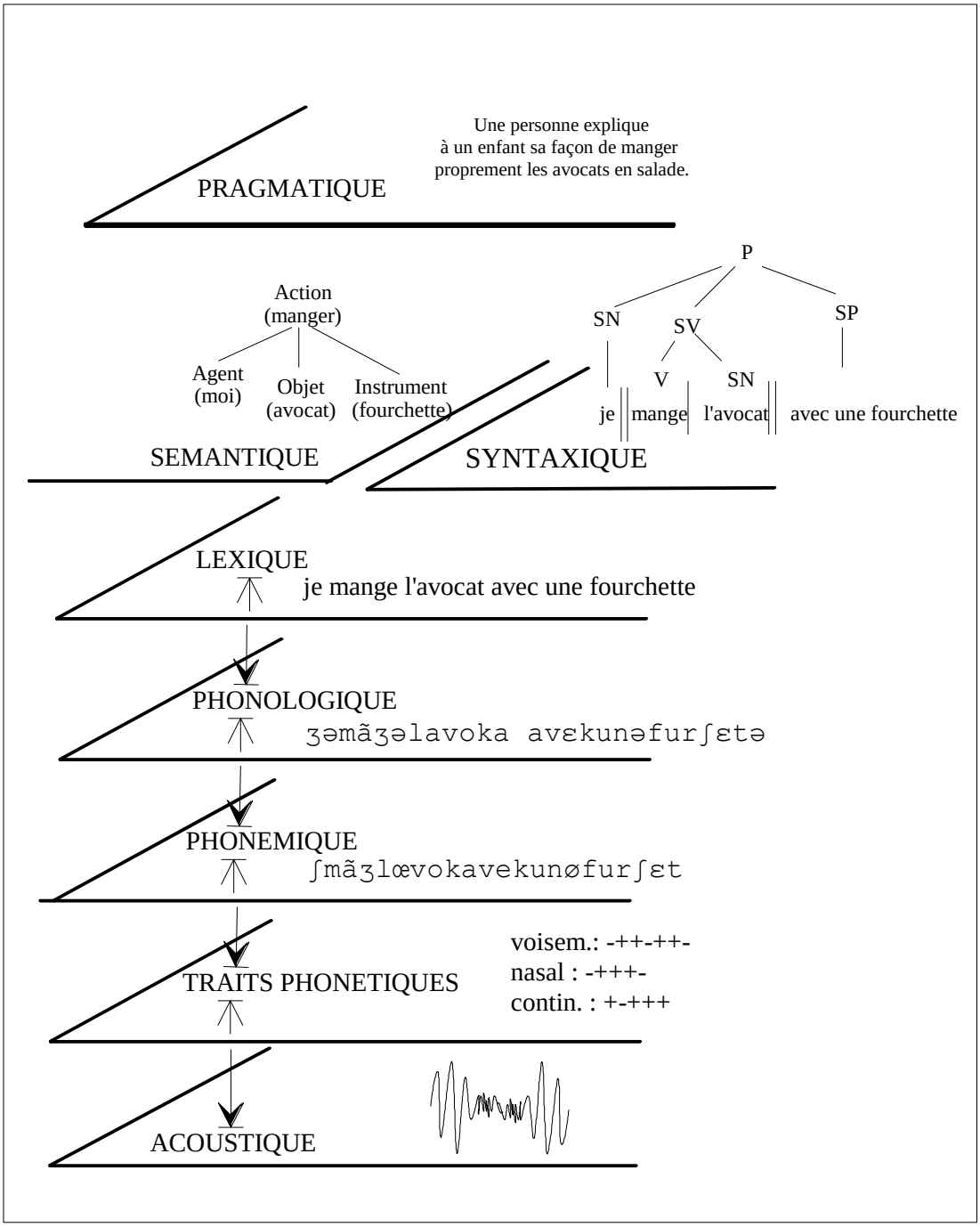


Figure 20: Exemple de différents plans de description (tiré du séminaire de doctorat, J. Véronis, 1993)

On distingue :

- l'*acoustique* a pour objet d'étude des sons en général.
- la *phonétique* a pour objet d'étude les sons de parole dans leur aspect physique.
- la *phonologie* a pour objet d'étude les sons de parole en fonction de leur aptitude à transmettre des contenus différents. Il s'agit d'un niveau abstrait.
- la *composante lexicale* étudie le mot en tant qu'élément d'un vocabulaire.

- la *syntaxe* explique la diversité et les contraintes des agencements des mots dans la phrase.
- la *sémantique* a pour objet l'étude de la signification du message en général.
- la *pragmatique* étudie le sens du message, c'est à dire l'interprétation que peut en faire l'auditeur dans un contexte particulier.

D'un point de vue linguistique, la reconnaissance de la parole consiste en une recherche paradigmatique d'unités le long d'un axe syntagmatique, ceci dans le cadre d'un plan de description précis. Ainsi, dans le plan phonético-phonologique, le décodage consiste à identifier des phonèmes parmi l'ensemble de tous les phonèmes. Dans le plan lexical, les divers mots du vocabulaire sont mis en concurrence...

Dans une stratégie de décodage ascendant, le passage d'un niveau à l'autre est dépendant d'une lecture interprétative du plan des réalisations vers les niveaux virtuels: des séquences acoustiques vers le plan phonémique, vers le plan lexical... En fait, il est possible de mettre en évidence quatre processus principaux (Pierrel, 1987):

- *le décodage acoustico-phonétique*

  Le décodage acoustico-phonétique transforme le continuum acoustique de parole en une suite discrète d'unités phonétiques symboliques (phonèmes, phones, diphones...). Il faut y ajouter des informations phonotactiques qui prennent en compte l'agencement des phonèmes entre eux (exemple: exclusion de séquences comme la suite de quatre consonnes consécutives) ainsi que des informations d'ordre phonologique qui tiennent compte des phénomènes de variabilité (phonèmes pouvant se prononcer de différentes façons, variations contextuelles, notion « d'accents » régionaux....) et des phénomènes d'altération des sons (coarticulation).

- *la composante lexicale*

  L'accès lexical cherche à regrouper les unités phonétiques pour former les mots du langage. Ici, se pose le problème de la segmentation lexicale qui n'apparaît pas explicitement dans le continuum acoustique. A ce niveau, doivent être inclus les phénomènes de liaisons, la gestion du schwa ('e' muet) et les règles morphologiques (notions de conjugaison, de préfixation, suffixation...).

- *la représentation de la phrase*

  Généralement, sont distinguées *l'analyse syntaxique*, qui essaie de déterminer une structure de la phrase et *l'analyse sémantique*, qui est liée à la signification des mots et des concepts sous-jacents. Cette étape peut fournir une représentation éventuellement incomplète, voire multiple, de la phrase stimulus. Le but ultime du décodage de la parole est plus de comprendre la signification du message vocal que de connaître de façon exhaustive tous les éléments de l'énoncé.

- *l'interprétation de l'énoncé*

  L'interprétation de l'énoncé est dépendante de l'aspect pragmatique, c'est à dire des informations relatives au contexte de la conversation (univers du dialogue). Elle permet de compléter et de valider les résultats de la reconnaissance de phrases puis, éventuellement, de provoquer une action correspondant à la réponse.

Dans le processus global de décodage de la parole, le rôle de la phonétique apparaît comme crucial. Celle-ci joue le rôle d'interface entre le monde physique continu et le domaine abstrait des représentations mentales de type symbolique. La phonétique est traditionnellement vue comme l'étude de la substance et la phonologie celle de la forme. Pourtant, la limite entre ces deux disciplines reste floue. L'exemple de « l'alphabet phonétique international » est particulièrement révélateur. L'appellation « phonétique » semble en contradiction avec la notion d'alphabet qui relève du domaine du symbole et donc de l'abstrait. Cela prouve peut-être qu'il n'y a pas vraiment lieu d'opposer les deux disciplines, la phonétique étant vue comme la face externe et la phonologie la face interne de cette interface entre matière préformée et représentations mentales.

Le passage du niveau physique vers les niveaux abstraits est le propre du décodage ascendant de la parole. Il s'apparente à la notion de compression de l'information en traitement du signal.

II.1.C. Parole et information d'un point de vue quantitatif

Ce paragraphe aborde la théorie de l'information de façon intuitive, sans rentrer dans de longues considérations mathématiques.

II.1.C.a. La notion d'information en traitement du signal

La mesure de l'information se résume à la réflexion suivante: soit un système pouvant prendre N états équiprobables (exemple: un dé à 8 faces), quelle quantité d'information doit être fournie pour connaître l'état du système ? La réponse possible est: la quantité d'information à fournir est égale au nombre minimum de questions à réponse binaire (oui ou non), nécessaire pour identifier l'état en question.

Par exemple, considérons un dé à 8 faces, et admettons que le tirage prenne la valeur 5. Pour connaître ce résultat, on peut demander tout d'abord: est-ce que le numéro est plus grand que 4 ? La réponse étant positive, on ne garde comme candidats que les chiffres compris entre 5 et 8, éliminant alors la moitié des choix possibles. En itérant le processus une deuxième fois (est-ce que le numéro est plus grand que 6 ? Réponse: non => on garde [5; 6]), une troisième fois (est-ce que le numéro est plus grand que 5 ? non => on trouve [5]). On obtient le résultat que l'on cherche en posant finalement 3 questions à réponses binaires qui fournissent une information suffisante pour identifier l'état du système.

Par définition, l'information fournie par une réponse binaire équiprobable est égale à un bit (unité d'information). Dans notre exemple, 3 bits d'information permettent de connaître l'état de notre dé à 8 faces. Du point de vue du dénombrement, 3 bits peuvent effectivement fournir 8 états ($2*2*2 = 2^3 = 8$). Autrement dit, 3 bits permettent de coder un système pouvant

prendre 8 valeurs. Il est possible de connaître le nombre de bits nécessaire pour définir parfaitement un système pouvant prendre *N* valeurs équiprobables. Il faut, pour cela, introduire la notion de logarithme à base 2 dont la propriété est la suivante: $Log_2(2^b) = b$.

Ainsi, dans l'exemple du dé à 8 faces, nous avons vu que l'information nécessaire au codage est de 3 bits. Or, $3\ bits = Log_2(2^3) = Log_2(8\ états)$.

De façon générale, l'information nécessaire à la définition d'une grandeur pouvant prendre *N* valeurs également probables est:

$$H\ en\ bits = Log_2(N\ états).$$

II.1.C.b. La notion de débit d'information

Le débit d'information représente la quantité d'information par unité de temps, nécessaire à la définition d'un phénomène. Dans le cas d'un signal numérique, le phénomène étudié est codé par un ou plusieurs paramètres. Diverses grandeurs interviennent:

- m le nombre de paramètres utiles pour le codage.
- F_p la fréquence de codage du p^{ème} paramètre.
- H_p le nombre de bits utilisés pour quantifier le p^{ème} paramètre
  $H = Log_2(N)$, N étant le nombre de niveaux de quantification

Le débit d'information s'exprime alors par:

$$D = \sum_{p=1}^{m} F_p \times H_p$$ où p est l'indice des paramètres.

Dans le cas fréquent où les fréquences de codage et le nombre de bits de quantification sont identiques pour tous les paramètres, le débit s'exprime finalement par:

$$D = F.m.H$$

II.1.C.c. Les différents débits d'information liés à la parole

Il est bien connu que le message vocal contient divers types d'informations: le message verbal sous différentes formes (acoustique, phonétique, lexicale), mais aussi des informations telles que le type de voix, l'intonation, le rythme... En conséquence, à la question de savoir quel débit d'information (en nombre de bits par seconde) est requis pour coder la parole, on ne peut répondre sans définir un *critère de fidélité* (Liénard, 1977). Ce débit ne sera pas le même s'il s'agit de coder toute l'information acoustique ou seulement un équivalent perceptif. Nous rejoignons ici la notion de plan de description entrevue précédemment.

Pour le codage numérique du signal de parole, le critère de fidélité est la ressemblance maximale entre l'onde originale et le signal codé. De façon classique, la numérisation s'effectue avec des échantillons codés sur H = 12bits (S/N=72dB) à une fréquence d'échantillonnage de F=16 kHz. Le débit s'exprime alors par: $D = F.m.H$ = 16 000 * 1 * 12 = 192 000

$$D_{physique} \approx 200\ 000\ bits/s$$

Si le critère de fidélité n'est plus la ressemblance maximale mais l'absence relative de différence perceptible entre le signal original et le signal codé, apparaît alors un niveau perceptif. L'état de l'art en matière de codage et synthèse vocale (rapport CCITT) montre qu'environ 16 coefficients codés chacun sur 128 niveaux et transmis toutes les 7 ms (143 fois par seconde) permettent de reconstituer un message vocal d'excellente intelligibilité et de naturel conservé. Le débit s'exprime alors, en prenant F = 143 Hz, m = 16 paramètres,
$H = Log_2(128) = 7$, par: $D = F.m.H = 143 * 16 * 7 = 16\,016$.

$$D_{perceptif} \approx 16\,000 \text{ bits/s}$$

Si le critère de fidélité devient uniquement l'équivalent phonétique entre les messages émis et reçus, moyennant les approximations suivantes pour le français (alphabet de N = 30 symboles phonétiques, d'occurrence équiprobable et prononcés à un rythme de F = 10 symboles par seconde), le débit s'exprime, en prenant F = 10 Hz, m = 1 symbole,
$H = Log_2(30) = 4.9$, par: $D = F.m.H = 10 * 1 * 4.9 = 49$.

$$D_{phonétique} \approx 50 \text{ bits/s}$$

A ce niveau, le locuteur et l'auditeur ne sont pas censés connaître les mots de la langue, mais uniquement les sons qu'elle utilise. Si le critère de fidélité devient l'équivalence des suites de mots émis et reçus, on peut alors, moyennant de nouvelles hypothèses simplificatrices (N = 20 000 mots dans la conversation courante, d'occurrence équiprobable, de vitesse d'élocution de F = 2.5 mots/s), mettre en évidence un débit de $D = F.m.H = 2.5 * 1 * Log_2(20\,000) = 2.5 * 14.3 = 35.7$

$$D_{lexical} \approx 35 \text{ bits/s}$$

Si, enfin, le critère de fidélité devient l'équivalence grammaticale des suites de mots émis et reçus, on peut alors, moyennant de nouvelles hypothèses simplificatrices (N = 40 catégories syntaxiques, d'occurrence équiprobable, de vitesse d'élocution de F = 2.5 mots/s), mettre en évidence un débit de $D = F.m.H = 2.5 * 1 * Log_2(40) = 2.5 * 5.3 = 13.3$

$$D_{syntaxique} \approx 13 \text{ bits/s}$$

II.1.C.d. La reconnaissance automatique de la parole: une compression de l'information

En se plaçant dans le cadre de la théorie de l'information, la reconnaissance automatique de la parole s'apparente à un processus de compression du débit d'information. En effet, l'opération de transcodage réduit le flux d'information de 200 000 bits/s à 50 bits/s dans le cas du décodage acoustico-phonétique, voire à 35 bits/s pour la reconnaissance lexicale. Ces résultats peuvent s'interpréter de deux façons:

- la parole contient un grand nombre d'informations diverses. Le processus de reconnaissance doit saisir celles qui sont pertinentes et délaisser les autres. La tâche est difficile: le facteur de compression est supérieur à 5000 pour l'identification lexicale.
- les informations sont très redondantes dans la parole. Il est judicieux d'exploiter cette redondance. Une solution consiste à passer du niveau acoustique concret (haut débit) à un niveau de représentation paramétrique du signal (ex: modélisation du spectre) de plus faible débit, puis à un niveau plus abstrait de codage (ex: indices acoustiques) suivi d'un

transcodage en étiquettes symboliques (ex: phonèmes...). Intuitivement, on comprend qu'une telle méthodologie à compression progressive permet d'utiliser les redondances de la parole et d'effectuer des vérifications à double sens entre chaque étape. Nous rejoignons ici la notion d'information sous son aspect linguistique décrite au § « Parole et information d'un point de vue qualitatif », p. 38.

Cette vision ascendante du décodage reste la plus généralement admise, peut-être à cause de sa relative simplicité de conception. Pourtant, elle ne doit pas faire oublier des points de vues plus dynamiques, dans une approche cognitive.

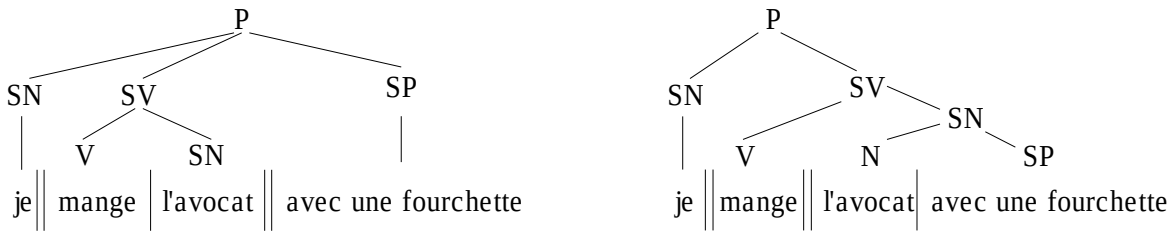
II.1.D. Parole et information d'un point de vue « cognitif »

Il est effectivement admis que le processus global de compréhension de la parole peut être divisé en un ensemble hiérarchisé de modules dont le rôle est, d'une part de mettre en correspondance le stimulus de parole avec des représentations internes des mots du langage, et d'autre part, d'extraire un message à partir de la chaîne lexicale constituée de plusieurs hypothèses construites par le processus d'accès lexical (Altmann, 1990). Cependant, plusieurs questions se posent: jusqu'à quel point les modules impliqués dans la chaîne de compréhension de la parole peuvent être considérés comme distincts (Fodor, 1983) ? De plus, jusqu'à quel degré ces modules interagissent-ils ? Ces questions restent actuellement sans réponse.

II.1.D.a. L'interaction des sources d'information

Il existe une très forte interaction entre chaque niveau de description; la recherche paradigmatique dans un plan de description ne peut pas se faire de façon isolée. Le passage du signal acoustique à la signification du message nécessite de faire coopérer au mieux les différentes sources d'informations. Chaque niveau apporte son lot de renseignements et doit collaborer au décodage complet. Ainsi, des connaissances sémantiques peuvent lever des ambiguïtés lexicales, des connaissances lexicales peuvent lever des ambiguïtés phonémiques, voire rectifier des erreurs... Revenons à l'exemple de la Figure 20, p.41 et imaginons les différentes ambiguïtés qui peuvent se créer à chaque niveau:

- au niveau phonologique, le décodage pourrait fournir différentes hypothèses /bãʒ/ , /mãʒ/ et /lãʒ/. Le niveau lexical nous permet d'exclure l'hypothèse / bãʒ/ car il n'existe aucun mot ayant cette réalisation phonétique.
- au niveau lexical, une confusion pourrait exister entre "lange" (du verbe "langer", 3e personne du singulier, indicatif, présent) et le substantif masculin "un lange". Toutefois, le niveau syntaxique nous permet d'exclure l'hypothèse du substantif par analyse grammaticale favorisant la présence d'un verbe après le pronom "je".
- au niveau syntaxique, les structures suivantes pourraient se trouver en compétition:



- Le niveau sémantique nous permet de favoriser la première structure où la fourchette, qui est un ustensile, joue un rôle instrumental à l'action de manger plutôt qu'un complément difficilement associable à l'avocat. Pour cela, il est possible d'utiliser des réseaux sémantiques plus ou moins complexes (Sabah, 1988) comme celui de la Figure 21. Le réseau proposé en exemple se lit de la façon suivante:

  ⇒ l'avocat (fruit) est une sorte de fruit, un fruit est une sorte d'aliment, un aliment est un objet de l'action de "manger" ; de plus, une fourchette est une sorte d'ustensile; un ustensile est un instrument de l'action de "manger".

  ⇒ une ambiguïté au niveau phonétique entre /mãʒ/ et /lãʒ/, issus des mots "mange" et "lange", eux-mêmes dérivés des verbes manger et langer, pourrait être levée par une étude sémantique. En effet, "la fourchette" peut difficilement se raccrocher à la notion de "langer", de même pour l'avocat. On peut ainsi se permettre de favoriser l'hypothèse /mãʒ/.

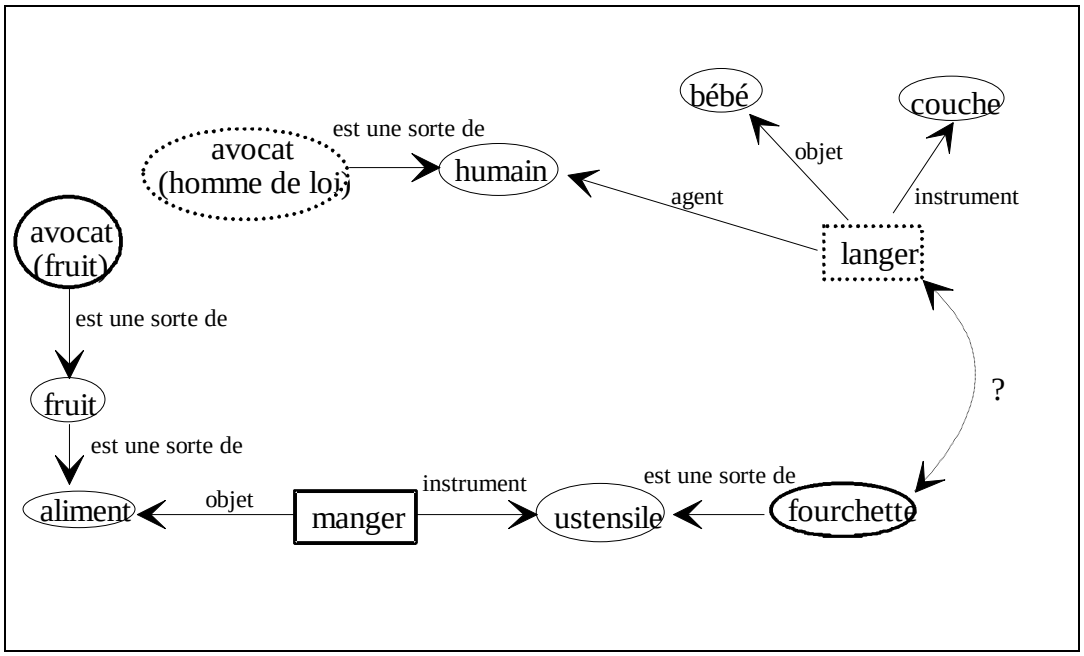


Figure 21: Exemple de réseau sémantique

Cet exemple est utile pour illustrer la complexité du message oral du fait de sa structure multiple. Cependant, cette redondance de l'information apparaît comme essentielle dans le processus de décodage. Seule la collaboration des différentes sources de connaissances permet la réalisation de systèmes performants de reconnaissance automatique de la parole.

II.1.D.b. L'organisation des sources d'information

La stratégie présentée précédemment laisse apparaître une hiérarchie entre les niveaux d'information (Figure 22a). Ceci reste une vision traditionnelle. Pourtant, rien ne s'oppose à proposer des modèles d'architecture non hiérarchisés où toutes les sources d'information interagissent de façon complètement libre (Figure 22b) ou contrôlée par un processus superviseur (Figure 22c). Ces points de vue peuvent être consultés dans (Caelen, 1995).

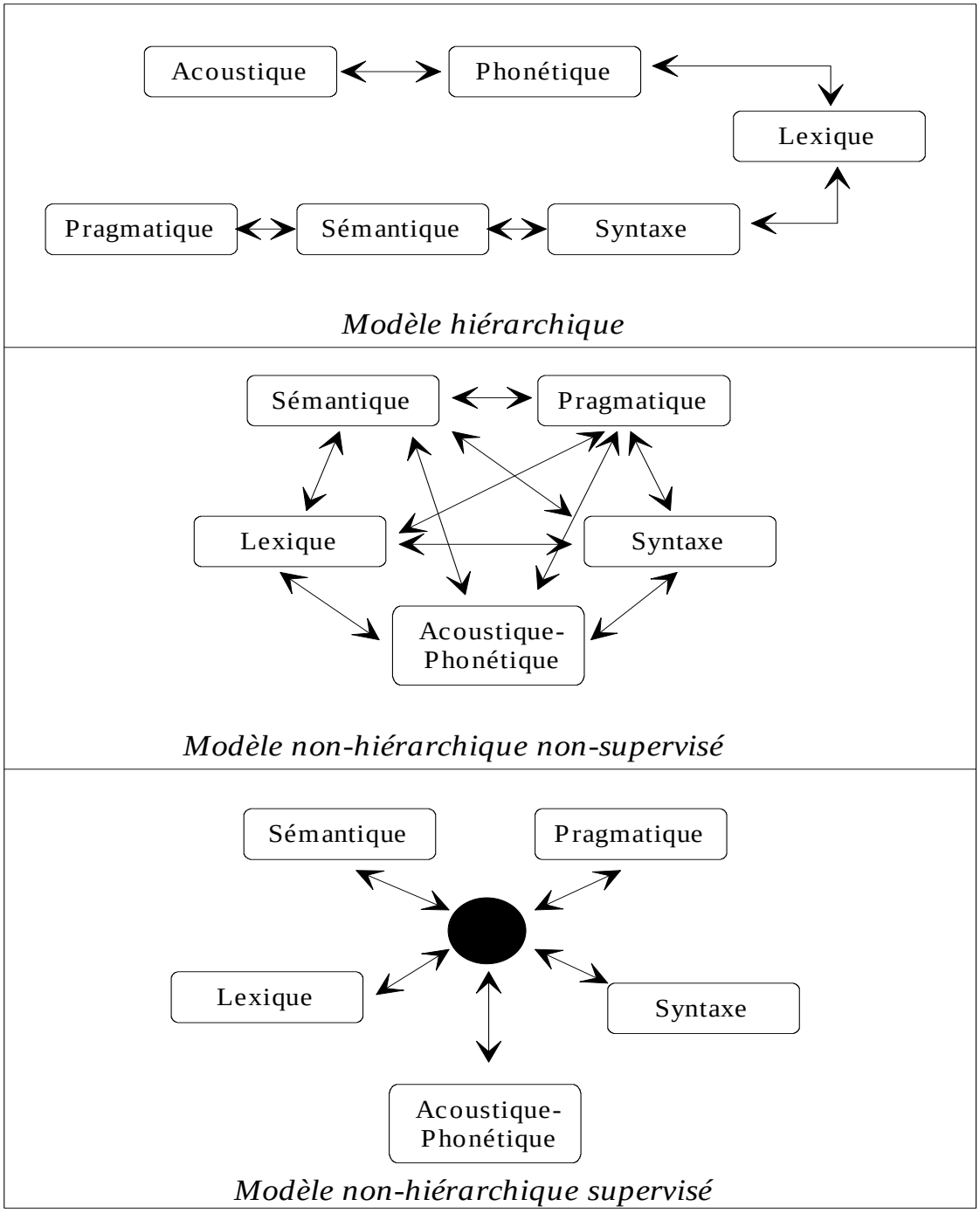


Figure 22: Organisation des sources de connaissances

On se rend compte que la notion d'information reste analogue que l'on soit linguiste ou physicien bien qu'on en parle en termes différents. Une chose est sûre: quelle que soit sa formation de base, une bonne connaissance des différents phénomènes relatifs à la parole est nécessaire pour mieux l'appréhender sous ses aspects physiologique, acoustique, sous forme de signal numérique et dans une perspective linguistique.

II.2. La physiologie de la parole

II.2.A. La production de la parole

La production de la parole est un mécanisme fort complexe encore incomplètement étudié. Les techniques d'investigation sont encore lourdes à mettre en place et restent peu répandues, ce qui explique la difficulté de faire avancer les connaissances dans le domaine. De façon simplifiée, la production de la parole peut être décrite par les phénomènes suivants (Figure 23).

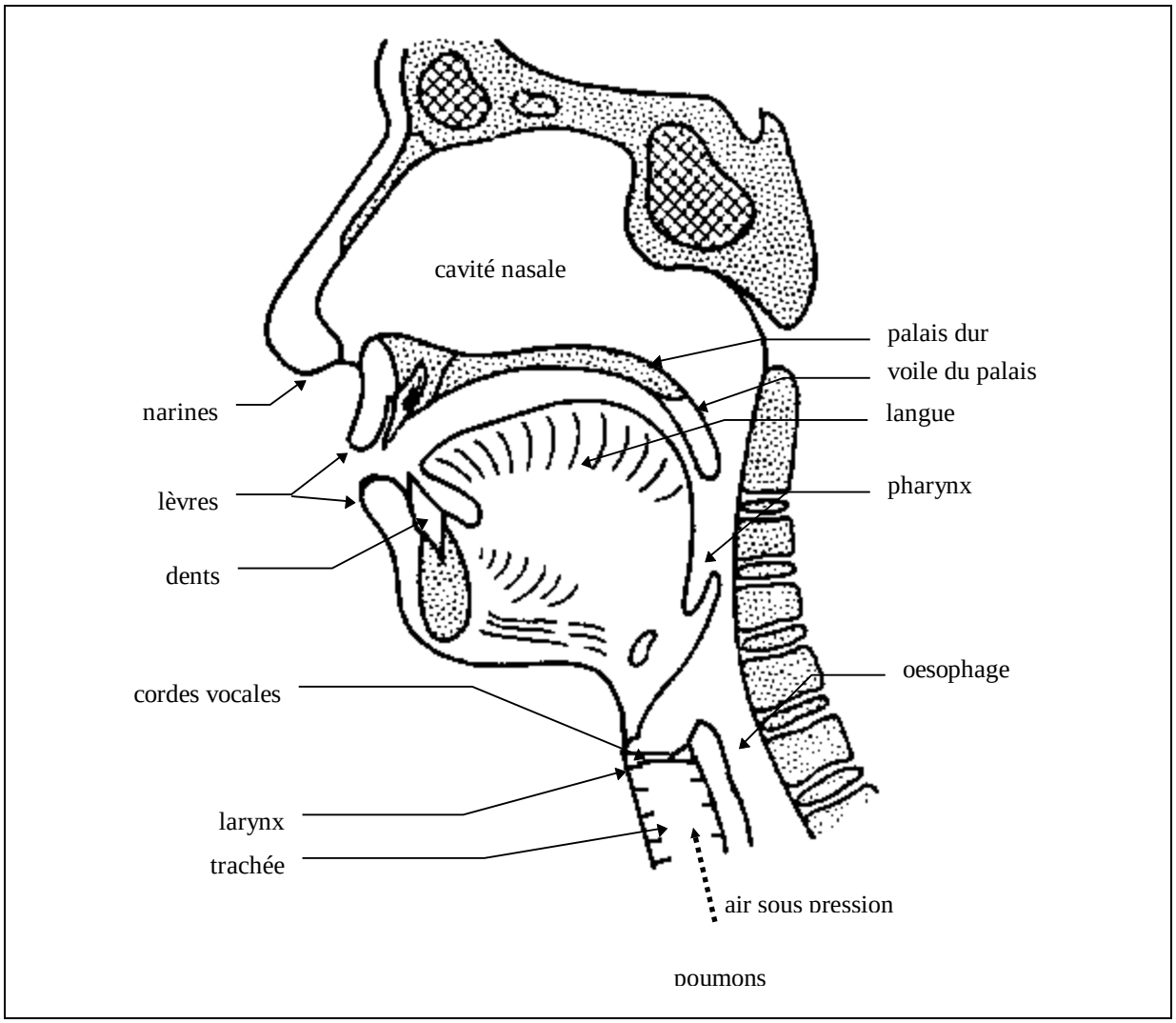


Figure 23: Anatomie de l'appareil vocal [contours anatomiques tirés de (Traissac, 1992)]

1/ Les poumons constituent la source d'énergie. Dans une phase d'expiration, ils créent une surpression qui a tendance à expulser de l'air par la trachée.

2a/ Si les cordes vocales sont « tendues », l'air mis sous pression par les poumons rencontre une résistance due à l'accolement des cordes. La glotte est l'espace libre compris entre les cordes vocales. La glotte est dite ouverte lorsque les cordes vocales sont écartées l'une de l'autre, permettant le passage de l'air. Par opposition, la fermeture de la glotte correspond à l'accolement des cordes vocales, formant alors un obstacle quasi-hermétique au passage de l'air (Figure 24, image 1). Le flux d'air ainsi obstrué crée alors une surpression croissante en amont des cordes vocales. Quand la force de pression sous-glottique dépasse la force de résistance des cordes vocales, l'air passe à travers la glotte (Figure 24, image 3), ce qui a pour effet de faire baisser cette pression: les cordes peuvent alors bloquer à nouveau le flux d'air (Figure 24, image 7) et le cycle recommence... Ce mouvement de vibration des cordes vocales peut se reproduire à grande vitesse (plusieurs centaines de fois par seconde). Le flux d'air continu qui arrive de la trachée est donc sectionné en « paquets » d'air, ce qui a pour résultat la génération d'un signal quasi-périodique (Figure 24).

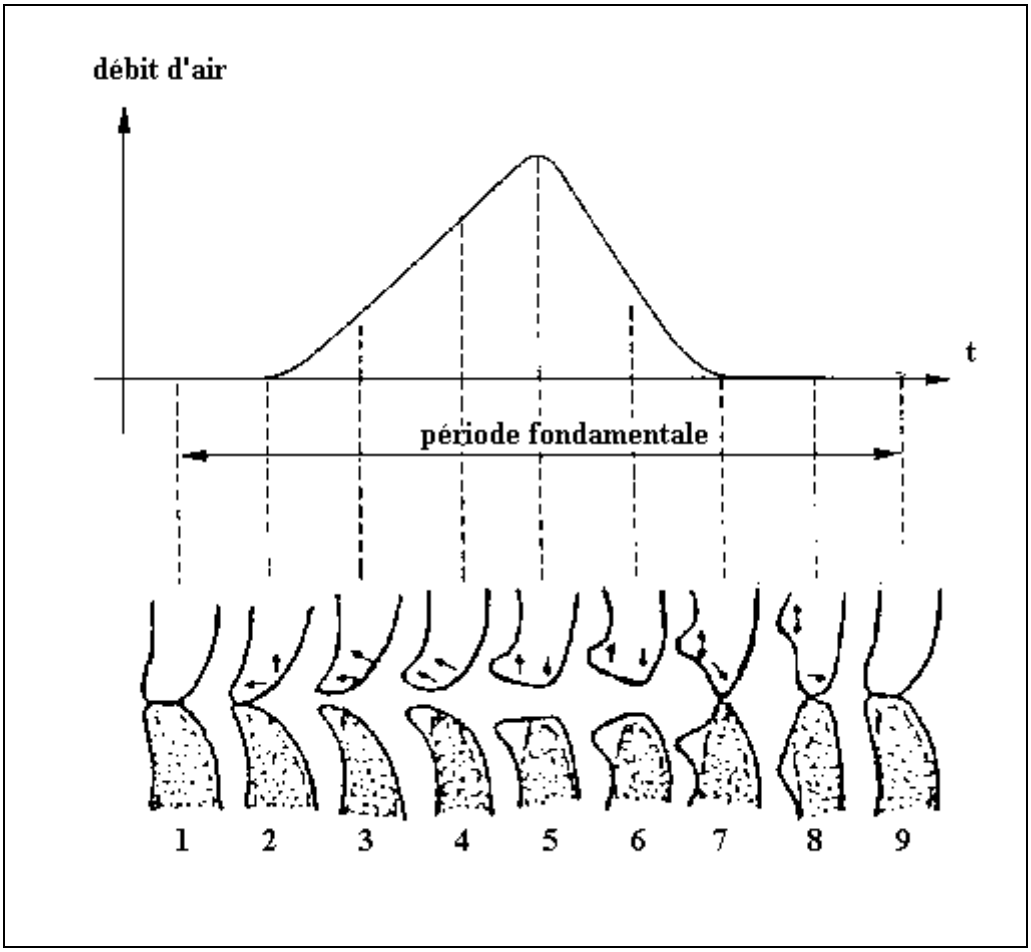


Figure 24: Configuration des cordes vocales et évolution du débit à travers la glotte (Source: Calliope, 1989, p.28)

Ce mécanisme est appelé voisement. Il intervient dans la production des voyelles et des consonnes sonores. Il peut être modélisé par un système à deux masses comme le proposent (Titze & Liang, 1994). D'un point de vue physiologique, il s'observe en posant la main sur la trachée; on y sent des vibrations.

2b/ Par opposition, si les cordes vocales sont « relâchées », l'air expulsé des poumons s'écoule librement dans le larynx. La phonation n'existe pas. Ce mécanisme intervient dans la respiration « normale » et dans la production des sons non voisés comme /p,t,k,f,s,j/.

3/ Que le voisement intervienne ou pas, l'air s'écoule ensuite dans la cavité pharyngale. Si l'une des sections du conduit vocal se rétrécit, l'écoulement devient turbulent et donne naissance à un bruit coloré. Ce mécanisme est présent dans la production des fricatives comme /f/ où la constriction a lieu au niveau des dents et des lèvres ou comme /ʃ/ où la constriction s'effectue au niveau du palais... Si la fermeture du conduit vocal est complète, on parle d'occlusion. Le flux d'air est alors stoppé momentanément. L'ouverture soudaine du conduit entraîne l'émission d'un signal « échelon », qui se traduit par une explosion. Ce mécanisme intervient dans la production des occlusives comme /p,t,k,b,d,g/.

4/ Si les cordes vocales vibrent et qu'une obstruction est mise en place dans le conduit vocal, les phénomènes de voisement et d'émission de bruit se produisent en même temps. Ce mécanisme est présent dans la production des consonnes voisées comme /v, z, m, n, b, g,.../

5/ Le signal acoustique généré, qu'il soit voisé et(ou) bruité, est modifié par les phénomènes de résonance dans les cavités pharyngale, buccale et nasale dont les formes varient en fonction de la position des organes articulateurs (langue, mâchoire) et du voile du palais qui ferme ou non la cavité nasale. C'est l'étape de la phonation et de la nasalisation.

6/ Le son, variable selon sa nature voisée, bruitée et selon la position des organes articulatoires, est émis à travers les lèvres et éventuellement le nez.

II.2.B. L'audition

Dans la perception de la parole, on distingue généralement deux phases: d'une part l'audition, où l'oreille transforme le signal acoustique en message nerveux et transmet l'information au cerveau par l'intermédiaire du nerf auditif ; d'autre part la cognition, où l'auditeur décode le message en interprétant les indices fournis à l'issue du pré-traitement auditif. Nous nous intéressons ici à la première phase. Pour cela, détaillons les organes périphériques de l'audition (Figure 25).

- *l'oreille externe*

  Le pavillon est le récepteur de l'onde acoustique. Le conduit auditif sert de guide d'ondes. L'ensemble est loin de se comporter de façon linéaire.

- *l'oreille moyenne*

  Le tympan est une membrane élastique étanche qui transforme l'onde acoustique en vibrations mécaniques. Sensible aux fréquences comprises entre 16 et 20 000

Hz, il ne possède pas vraiment de fréquence propre. Les osselets (marteau, enclume, étrier), fixés entre eux par des ligaments, forment une chaîne qui relie le tympan à la fenêtre ovale. Ils jouent le rôle d'amplificateur mécanique dont les caractéristiques varient selon la contraction des micro-muscles. La fenêtre ovale effectue le couplage entre le milieu aérien de l'oreille moyenne et le milieu aqueux de l'oreille interne.

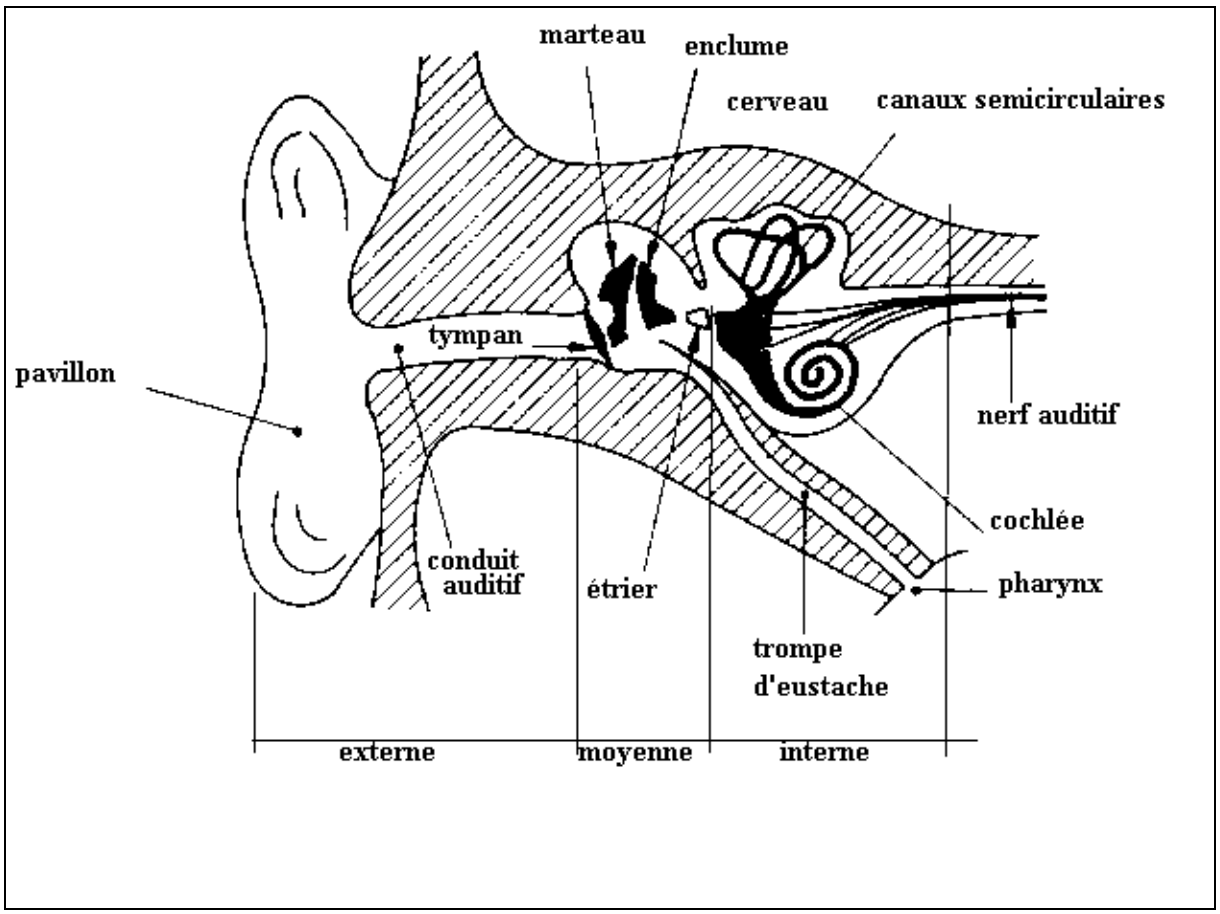


Figure 25: Anatomie de l'oreille (Source: Boite & Kunt, 1987, p.8)

- *l'oreille interne*

La cochlée est un tube de 3 cm de long enroulé en une spirale, d'où son autre nom de limaçon. Elle se comporte comme un résonateur: la position du maximum de vibration dans le tube cochléen dépend de la fréquence du son reçu. Elle contient la membrane basilaire qui supporte l'organe de Corti, une série de 20 000 cellules nerveuses dites ciliées. Le mouvement du fluide interne se propage dans la cochlée de façon sélective et, selon la position du maximum de vibration, les cellules excitées réagissent différemment, ce qui permet de juger la fréquence du son (Figure 26). Enfin, le nerf auditif transmet les impulsions nerveuses au cerveau.

Deux résultats essentiels sont à souligner:

- divers résultats (Zwicker & Feldkeller, 1981) montrent que l'oreille possède une meilleure résolution spectrale en basses fréquences (BF) qu'en hautes fréquences (HF). Cela se vérifie sur la Figure 26 où 5 mm de cochlée traitent 300 Hz de fréquences en BF (entre 0.3 kHz et 0.6 kHz) alors que sur la même longueur, les cellules ciliées traitent 4000 Hz de fréquences en HF (entre 4kHz et 8kHz).
- la sensation sonore dépend de l'intensité du son, mais aussi de sa fréquence ; un son pur de 50 dB de fréquence 1kHz est perçu aussi fort qu'un son de 70 dB émis à 60 Hz.

Nous reviendrons plus en détails sur ces points au § III.III.1, « Un modèle auditif », p.76.

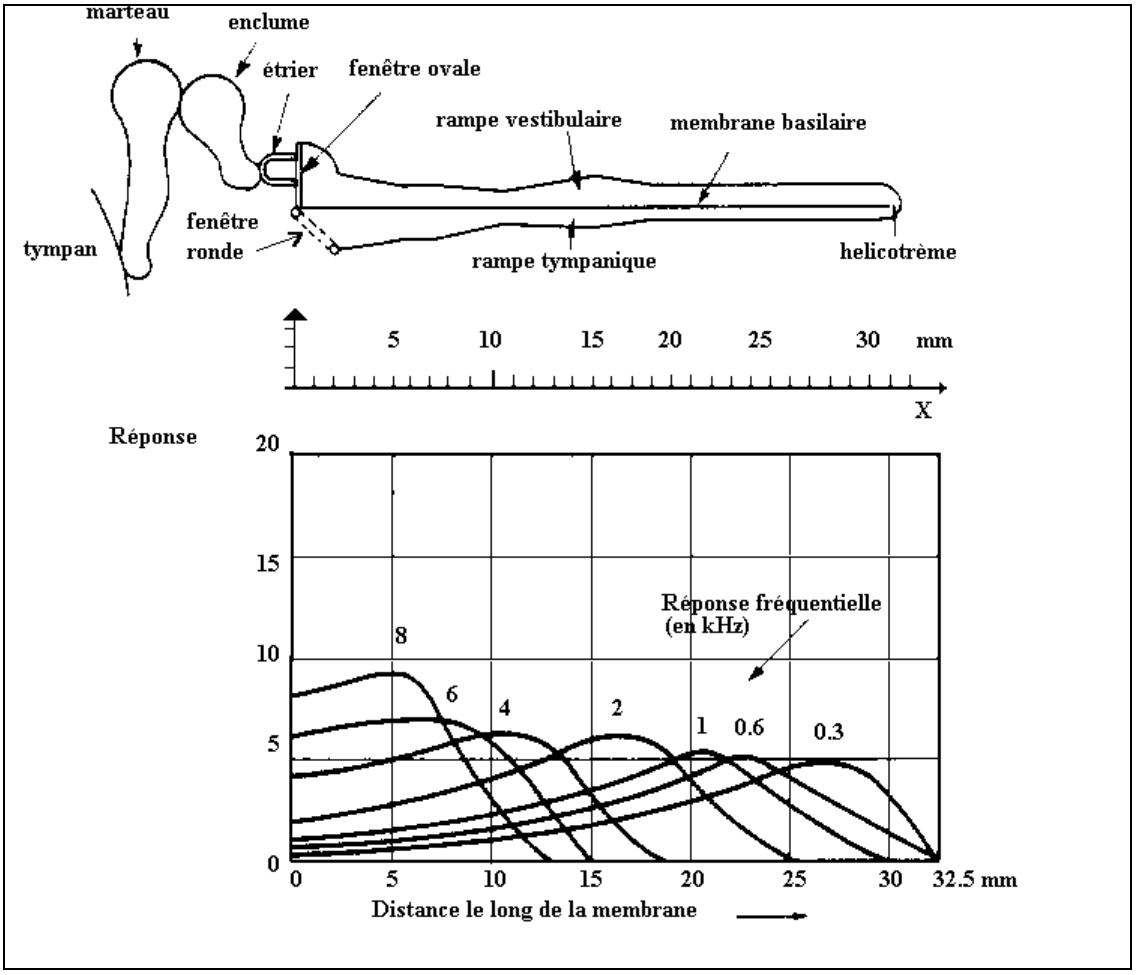


Figure 26: Section longitudinale de la cochlée et positions du maximum de sensibilité des cellules ciliées. (Source: Hassall et al., 1979, p.41)

II.3. La physique de la parole

II.3.A. L'aspect acoustique de la parole

D'un point de vue physique, la parole est un phénomène acoustique, de nature ondulatoire. Son signal est réel, continu, d'énergie finie, non stationnaire. Dans la production de la parole, les cordes vocales constituent une source sonore de signal quasi-périodique. On

appelle *fondamentale* - ou encore *Fo* - la fréquence de vibration des cordes vocales. Elle détermine la hauteur de la voix. Ci-dessous sont présentées quelques valeurs usuelles qui n'ont de sens que pour des sons voisés:

80 Hz < *Fo* < 200 Hz pour une voix masculine

150 Hz < *Fo* < 450 Hz pour une voix féminine

200 Hz < *Fo* < 600 Hz pour une voix d'enfant

Les turbulences créées par une obstruction éventuelle du conduit vocal constituent une source sonore de bruit coloré. Le conduit vocal joue le rôle de filtre dont les fréquences de résonance varient selon les positions des organes phonatoires. Tout ceci fournit la variété des sons de la parole. La Figure 27 illustre les phénomènes acoustiques mis en jeu dans le processus de production de la parole.

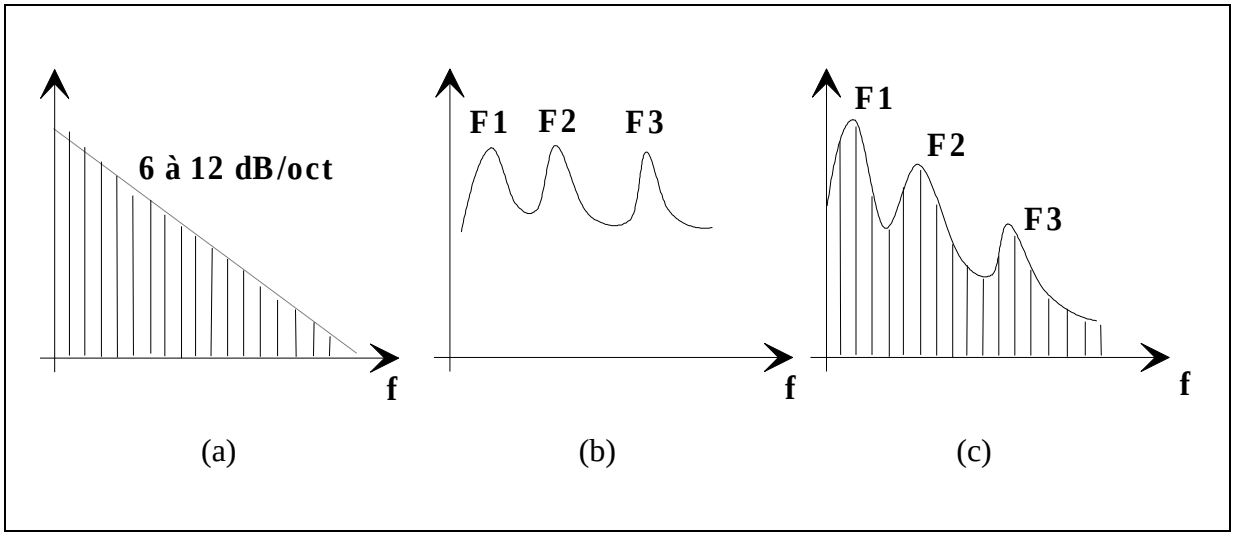


Figure 27: (a) Spectre de la source vocale, (b) fonction de transfert du conduit vocal, (c) spectre de la parole

La Figure 27a montre l'allure du spectre sonore au niveau du larynx. Il a une atténuation moyenne de 12dB/octave dans le cas des sons voisés ou de 6 dB/octave dans le cas des sons non voisés. La fonction de transfert du conduit vocal (Figure 27b) présente des pics correspondant aux fréquences de résonance des cavités du conduit vocal. Les lèvres agissent comme un filtre passe-haut dont la fréquence de coupure est proche de 6 kHz et dont l'atténuation en basses fréquences est d'environ 6 dB/octave. Finalement, en tenant compte de cet effet radiatif, le spectre de la voix (Figure 27c) possède un atténuation globale de 6dB/octave dans le cas des sons voisés. Dans le cas des sons non voisés, le résultat est un spectre à tendance plate. De plus, du fait des résonances, le spectre résultant (Figure 27c) possède différents maxima que l'on appelle formants dans le cas des voyelles. On les numérote par ordre croissant de fréquences (*F1<F2<F3<F4*). Chaque voyelle possède des formants spécifiques (Figure 28a, pour *F1* et *F2*). Les valeurs de ces pics ont été mesurées depuis longtemps (Perterson & Barney, 1952) et laissent apparaître une grande variabilité, ce qui se traduit par une dispersion. Autrement dit, la représentation d'une voyelle dans l'espace *F1-F2* ne se restreint pas à une valeur mais à un nuage de points (Figure 28b). Ces variations

sont dues essentiellement à la variabilité inter-locuteurs ainsi qu'aux effets contextuels (ex: le [a] de "kaki" est proche de /ɛ/).

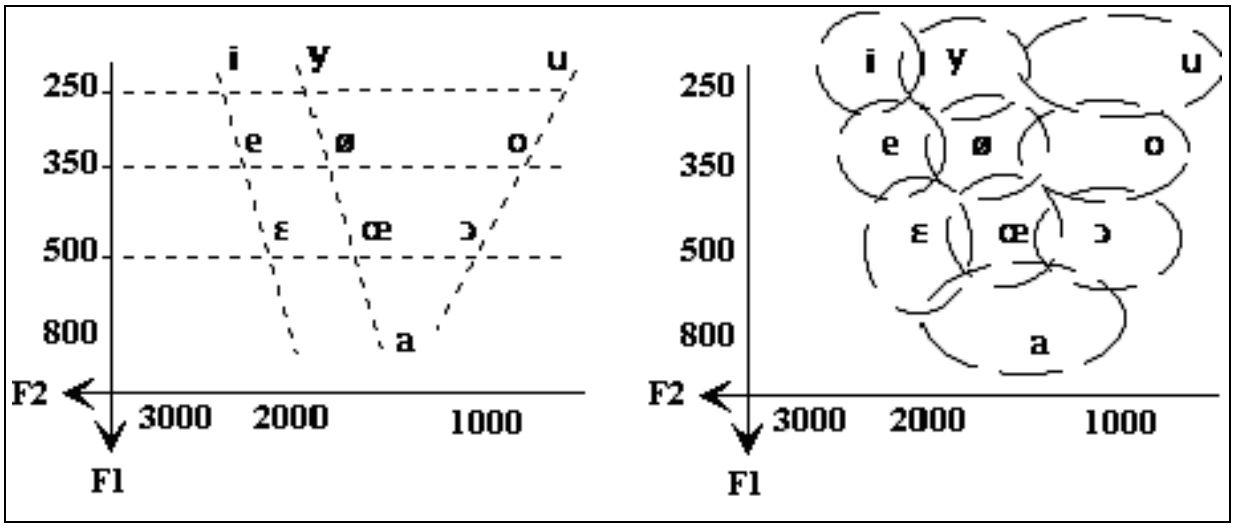


Figure 28: (a) Représentation biformantique et (b) champ de dispersion des voyelles orales du français (Source: Landercy & Renard, 1977, p.109)

Il faut remarquer que l'espace formantique n'est pas uniformément occupé. La zone /i, e, y/ est très peuplée contrairement à celle de /a/. Les voyelles du français sont réparties dans l'espace formantique non pas de façon linéaire mais logarithmique, un peu à l'image de l'espace perceptif (cf. Figure 26, p.53). Ceci nous renforce dans l'idée d'utiliser des caractéristiques perceptives dans le décodage de la parole comme nous l'avons précisé au paragraphe sur « Les questions préalables », p.9. Un tel procédé permet de mieux appréhender la répartition non uniforme des voyelles.

D'un point de vue temporel, le signal de parole est loin d'être stationnaire. Toutefois, sur des durées de l'ordre de la centiseconde (10 ms), on peut mettre en évidence des zones de stabilité comme le montre la Figure 29.

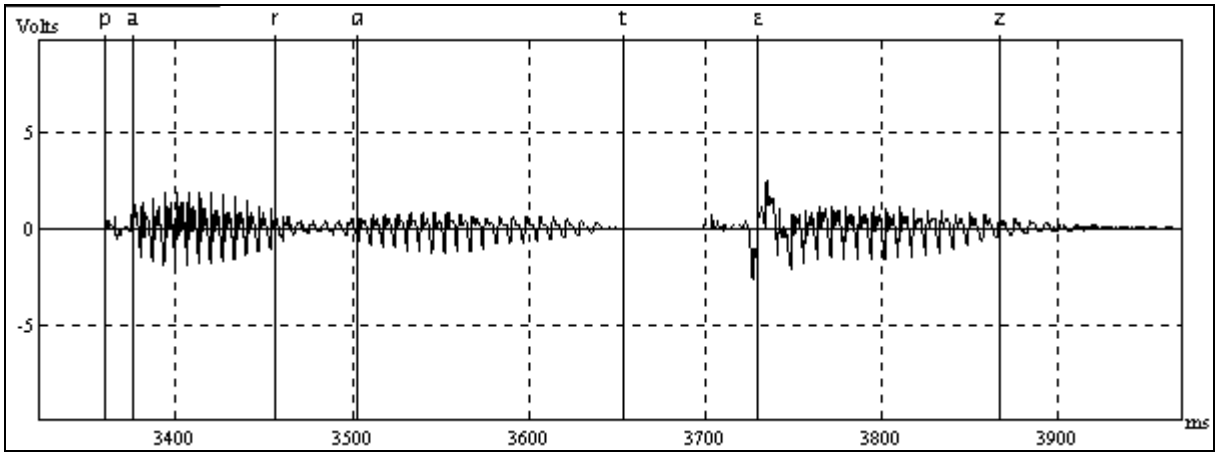


Figure 29: Allure du signal pour le mot "parenthèse"

Ces connaissances sur l'acoustique de la parole sont essentielles pour comprendre et réaliser un décodage du message vocal au moyen d'une méthode analytique qui s'attache à reconnaître justement des détails de nature spectrale, énergétique ou temporelle sur le signal de parole. De plus, ce savoir apparaît aussi comme indispensable pour se placer dans les meilleures conditions possibles pour saisir le son de parole.

II.3.B. L'acquisition de la parole

Ce paragraphe a pour but d'exposer les différents phénomènes physiques mis en jeu dans l'acquisition du signal de parole. Il permettra ainsi, au non-spécialiste, de les appréhender avec un peu plus de recul. Si à l'origine du traitement de la parole dans les années 50, l'instrumentation était analogique (spectrogrammes), la majorité des systèmes actuels utilisent l'électronique numérique (ordinateurs). Aussi, avant toute analyse, est-il nécessaire de capter et discrétiser le signal acoustique. On appelle cela la chaîne d'acquisition. Décrivons-la.

II.3.B.a. La saisie

Un *capteur* est un appareil qui permet de transformer un signal physique quelconque (optique, mécanique...) en un signal électrique. En théorie, il change seulement la nature physique du signal sans le détériorer. En pratique, il existe toujours une dégradation. Cette transformation est nécessaire car les signaux électriques sont faciles à mesurer. La saisie d'un son utilise un *microphone,* qui transforme les variations de pression de l'air en un courant ou une tension électrique. L'amplitude maximale du signal électrique en sortie du microphone est de quelques centaines de mV.

II.3.B.b. Le stockage

L'enregistrement sur bandes magnétiques permet de transformer un signal électrique en un signal magnétique par aimantation. En théorie, il change seulement la nature physique du signal sans modifier sa forme. En pratique, il existe aussi une distorsion. Cette transformation est utile car les signaux magnétiques sont durables et donc faciles à stocker contrairement aux signaux électriques, fugitifs.

II.3.B.c. La conversion analogique/numérique

Pour qu'un signal puisse être manipulé par un ordinateur, il faut qu'il remplisse deux conditions: qu'il soit quantifié (qu'il ne prenne que des valeurs calibrées) et qu'il soit discret (qu'on ne mesure sa valeur que de « temps en temps »). Le passage d'un signal électrique continu à un signal « informatique » utilise un convertisseur Analogique / Numérique. Cet appareil est conçu pour mesurer « de temps en temps » la valeur du signal électrique d'entrée et lui attribuer une valeur sur une échelle quantifiée (Figure 30). Cette valeur sera ensuite stockée dans une mémoire d'ordinateur avec les autres échantillons. Nous disposons ainsi d'un signal numérique qui pourra être utilisé indéfiniment sans distorsion. Pour cette opération, il faut remarquer que deux types de résolution entrent en jeu: une résolution en amplitude et une résolution temporelle.

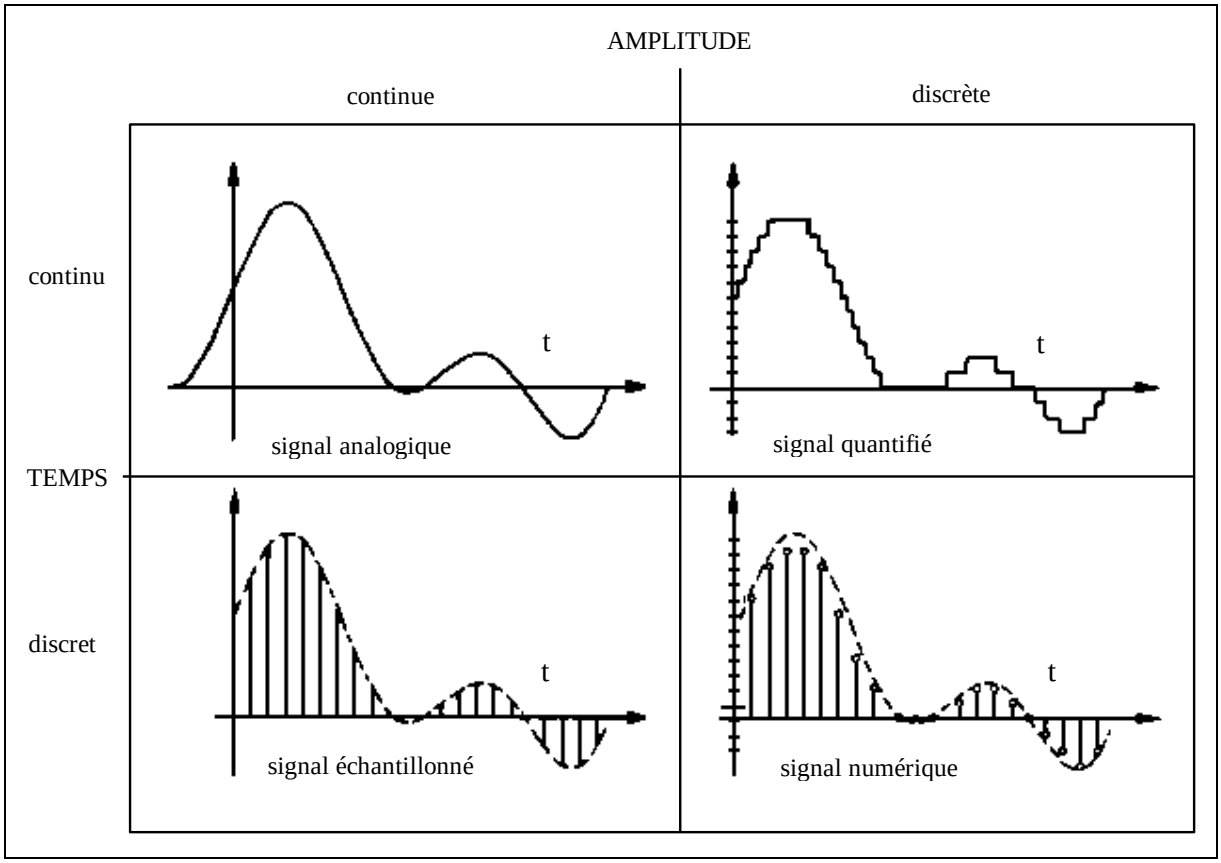


Figure 30: La numérisation du signal (Source : De Coulon, 1984, p.35)

- *la résolution en amplitude.*

  On appelle « nombre de bits » le nombre d'éléments binaires utilisés pour quantifier la valeur de l'échantillon. Si *N* est ce nombre, les échantillons pourront être codés selon 2^N niveaux différents. La dynamique de quantification est la valeur en décibels de ce nombre de niveaux et vaut: $dynamique\ (en\ dB) = 20 . Log_{10}(2^N) = 20 . N . Log_{10}(2)$ d'où

  $$dynamique\ (en\ dB) \approx 6.N$$

  Huit bits représentent un octet. Plus le nombre de bits utilisés est grand, meilleure sera la résolution mais plus grande devra être la mémoire de stockage. En général, pour le traitement de la parole, une quantification entre 8 et 16 bits est utilisée. Le Tableau 1 ci-après donne le nombre de niveaux possibles pour le codage et la dynamique de quantification en fonction du nombre de bits.

Tableau 1: Correspondance entre le nombre de bits et la dynamique escomptée

Nombre de Bits	Nombre de niveaux	Dynamique (en dB)
1	2	6
2	4	12
3	8	18
4	16	24
8	256	48
10	1024	60
12	4096	72
16	65536	96

- *la résolution temporelle.*

  On désigne par fréquence d'échantillonnage F_e le nombre de fois par seconde où une valeur du signal est mesurée. Cette valeur est appelée « échantillon ». L'inverse de F_e correspond à la période d'échantillonnage et est égale au temps séparant deux échantillons successifs. Elle est notée T_e ou Δt. Une valeur F_e = 16 000 Hertz (T_e = 0.0625ms) est commune pour la parole. Notons que le choix de F_e n'est pas anodin.

II.3.B.d. Les effets de l'échantillonnage

Soit un signal analogique x_a dont le spectre possède une fréquence maximale égale à $F_{max} = \frac{F_{\min}^{\max}}{2}$ (Figure 31, a). L'échantillonnage idéalisé (Figure 31, c&e) est réalisé en multipliant ce signal par une suite périodique d'impulsions de Dirac de période T_e (Figure 31, b&d). La conséquence est la périodisation du spectre X_a (démonstration au § « Quelques propriétés », p.89) , ce qui entraîne l'apparition de spectres secondaires séparés d'une période de $1/T_e = F_e$ (Figure 31, c&e, domaine spectral). Si les spectres secondaires recouvrent le domaine primaire du spectre, l'information spectrale est déformée (Figure 31, c, domaine spectral). Pour éviter ces recouvrements, il faut que le premier spectre secondaire soit à une distance spectrale supérieure à $2 * F_{max}$ du spectre primaire, c'est à dire qu'il faut que:

(Eq.1) $$F_e > 2.F_{max} \quad \Leftrightarrow F_e > F_{\min}^{\max}$$

Cette propriété est connue sous le nom de Théorème de SHANNON. La Figure 31 illustre parfaitement les effets de l'échantillonnage et le phénomène de recouvrement appelé aussi repliement de spectre (« aliasing » en anglais).

Pour réussir l'opération de numérisation du signal, quelques précautions pratiques sont à prendre.

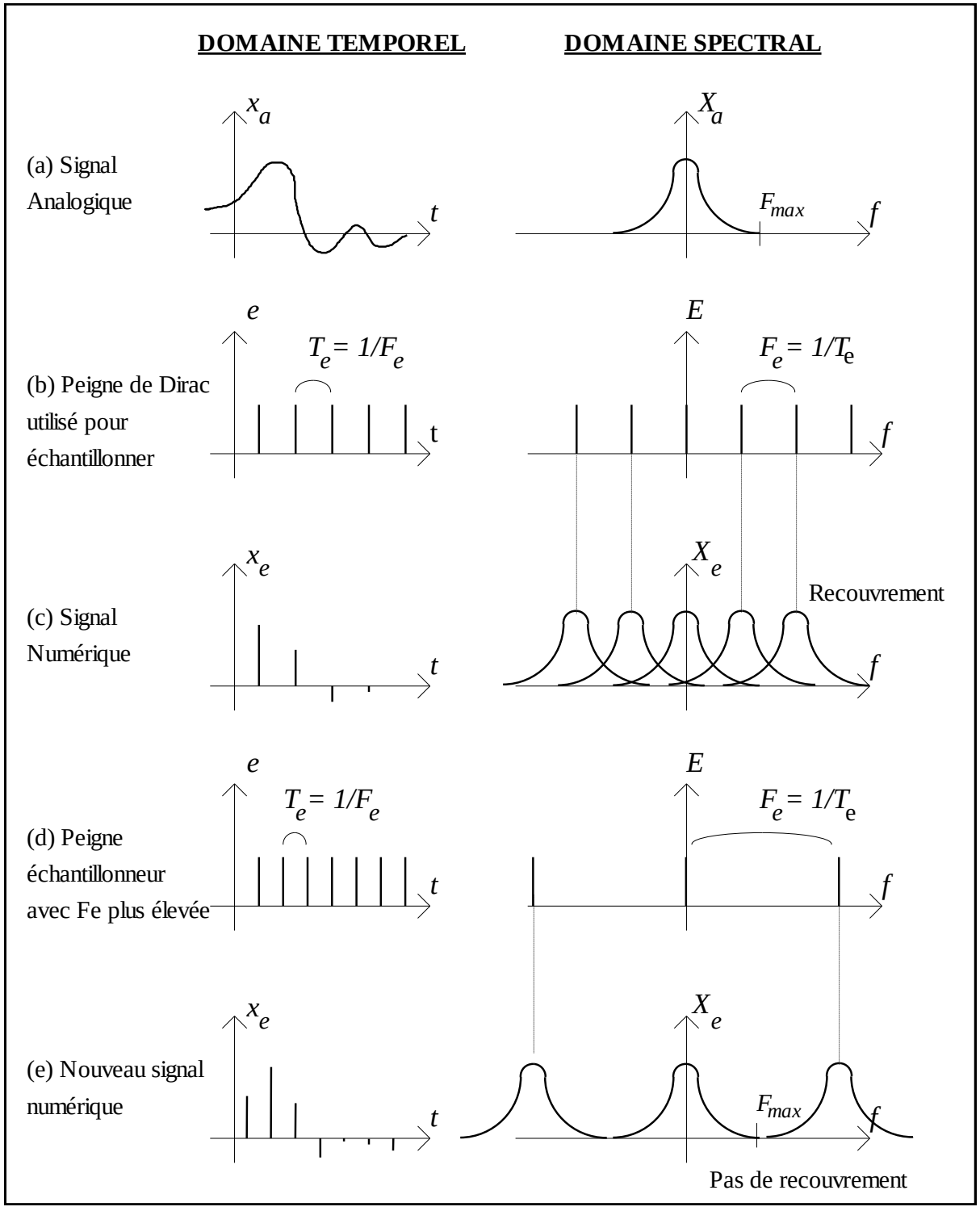


Figure 31: Les effets de l'échantillonnage (Source : Kunt, 1980)

II.3.B.e. Le conditionnement du signal électrique

La plage d'un Convertisseur Analogique / Numérique (C.A.N.) est l'amplitude autorisée en Volts que peut prendre le signal électrique d'entrée pour être convenablement converti. A une tension maximale analogique correspondra une valeur numérique maximale

dépendant de la résolution en bits du C.A.N. Les valeurs électriques étant généralement négatives et positives, on utilise aussi des valeurs numériques signées. Prenons l'exemple d'un C.A.N. possédant une plage d'entrée de +/- 2 V et une résolution de 8 bits. Les échantillons peuvent donc être codés selon 256 niveaux, c'est à dire dans l'intervalle [-128;128[. Cela signifie qu'un signal électrique de:

2 V	aura la valeur numérique de	127
1 V	" " "	63
-0.5 V	" " "	-32

Cette plage de +/- 2V est une grandeur courante. Or, le signal en sortie d'un microphone possède une valeur de quelques centaines de mV. Pour 100 mV d'amplitude crête à crête, seuls 5 % de la résolution du C.A.N. serait utilisés. Afin de remédier à cette insuffisance, une *amplification* du signal électrique est réalisée en sortie du microphone ou du magnétophone de telle façon qu'un son de puissance relativement forte produise, à l'entrée du C.A.N., une tension qui occupe 90 % de la plage utile. Par commodité, on utilise souvent un amplificateur à gain réglable pour ajuster le niveau de sortie du microphone ou du magnétophone.

La seconde opération de conditionnement du signal est un *filtrage passe-bas* afin d'éviter un recouvrement de spectre (« aliasing » en anglais) du fait de l'échantillonnage. Ce filtrage permet de réduire la largeur spectrale du signal tout en gardant la bande de fréquences utile. En cas de mauvais filtrage (bande passante trop grande par rapport à la fréquence d'échantillonnage), le signal numérique est parasité de façon irrémédiable. La Figure 32 représente le spectrogramme d'un signal de parole (mot prononcé "papi") échantillonné à 10kHz avec un filtrage passe-bas à 8kHz au lieu de 5kHz. Notons les masses énergétiques anormalement élevées en hautes fréquences: le [a], voyelle grave, possède un «faux-formant» vers 5000 Hz !

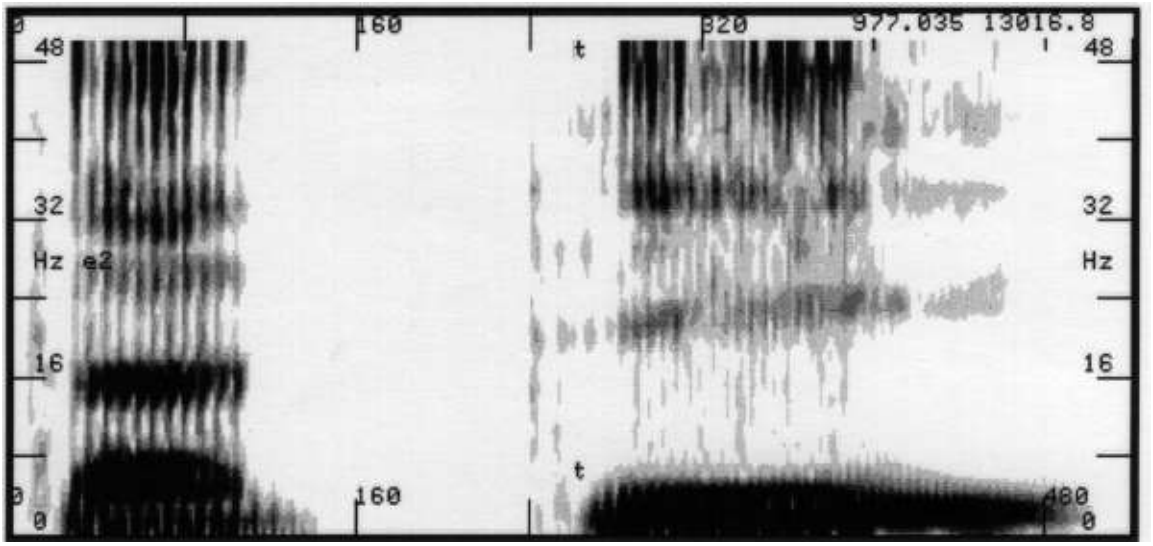


Figure 32: Exemple de repliement de spectre pour cause de mauvais filtrage

Ce type de problème ne se détecte pas facilement et peut avoir de graves conséquences. Par chance, les innovations technologiques permettent maintenant d'éviter ces erreurs de manipulation: le réglage du filtre s'effectue automatiquement à partir du choix de la fréquence d'échantillonnage. En traitement de la parole, la bande passante est généralement comprise entre 3 kHz et 10 kHz. En diminuant par filtrage la fréquence maximale du signal analogique de parole, qui peut monter au delà de 10 kHz, cela permet d'opter pour une fréquence d'échantillonnage inférieure à 20 kHz. Celle-ci est choisie habituellement au double de la bande passante par référence au théorème de SHANNON.

Une fois amplifié et convenablement filtré, le signal analogique est échantillonné et quantifié pour être finalement stocké en mémoire. Il peut être aussi enregistré sur un support optique ou magnétique. L'opération de restitution du son est possible à partir des échantillons de signal numérique grâce à un Convertisseur Numérique / Analogique (C.N.A.), un filtre, un amplificateur et un haut-parleur qui restitue le signal acoustique. Les deux avantages incomparables d'un signal numérique est d'être d'une part très peu dégradé par copie ou restitution et d'autre part manipulable par un calculateur numérique.

II.4. La phonétique et le traitement de la parole

II.4.A. Phonétique et phonologie

Traditionnellement, la *phonétique* est décrite comme la science qui « étudie les sons du langage dans leur réalisation concrète, indépendamment de leur fonction linguistique » (Dubois et al., 1973, p.373). On distingue généralement:

- la phonétique articulatoire qui s'attache aux mécanismes de la phonation
- la phonétique acoustique qui étudie les fondements physiques de la parole
- la phonétique perceptive qui s'intéresse à la réception et l'intégration des sons du langage
- la phonétique diachronique qui analyse l'évolution des sons du langage (historique)
- la phonétique corrective ou orthophonie qui se penche sur les moyens de corriger une « mauvaise » prononciation des sons

La *phonologie* « étudie les sons du langage du point de vue de leur fonction dans le système de communication linguistique. Elle étudie les éléments phoniques qui distinguent, dans une même langue, deux messages de sens différent, (...) et ceux qui permettent de reconnaître un même message à travers des réalisations individuelles différentes (...) Elle se différencie en cela de la phonétique, qui étudie les éléments phoniques indépendamment de leur fonction dans la communication » (Dubois et al., 1973, p.375). Ainsi, la différence phonique à l'initiale des mots "pain" et "bain" est d'ordre phonologique car elle permet une distinction de sens. Par contre, si une description de phonétique articulatoire fait la différence entre [ʁ], [r] et [x] qui représentent respectivement trois réalisations du "R" (le premier est le "R" grasseyé classique, le second est le "R" apical roulé, le troisième est le "R" guttural) une description phonologique en français ne présente qu'un seul "R" noté /R/ car toutes ces variantes sonores ne portent

aucune distinction de sens: [kaʁkasonə], [karkasonə] ou [kaxkasonə] représentent bel et bien la même ville de "Carcassonne".

La distinction entre phonétique et phonologie est en fait très délicate car il existe souvent un flou entre ces deux domaines. Comme le signalent (Dubois et al., 1994, p.361) « la phonétique ne peut faire abstraction du caractère social du langage » et donc de l'aspect communicationnel, « de même que la phonologie ne peut faire abstraction de la connaissance des sons concrets de la parole ». Le développement récent du concept de « phonologie articulatoire » (Browman & Goldstein, 1992) illustre le recouvrement des deux disciplines.

II.4.B. Le phonème

Le fonctionnement d'un code linguistique (cf. § II.II.1.II.1.A, « La communication », p.37) ne requiert qu'un nombre limité d'éléments distincts structurés en un *système phonologique.* Le *phonème* est la plus petite unité fonctionnelle distinctive de ce système. Une liste des phonèmes du français est donnée au Tableau 2.

Tableau 2: Les phonèmes du français

CONSONNES			
	[p] paie	[t] taie	[k] quai
	[b] baie	[d] dé	[g] gai
	[m] mais	[n] nez	[ñ] gagner
	[f] fait	[s] sait	[ʃ] chez
	[v] vais	[z] zéro	[ʒ] geai
	[w] ouais	[ɥ] huer	[j] yé-yé
	[l] lait	[R] raie	
VOYELLES			
	[i] lit	[y] lu	[u] loup
	[e] les	[ø] leu	[o] lot
	[ɛ] lait	[œ] leur	[ɔ] lotte
	[a] la	[ə] le	
	[ɛ̃] lin	[œ̃] l'un	
	[ã] lent	[ɔ̃] long	

Le phonème constitue en fait une unité symbolique et non observable. On l'obtient par interprétation de phénomènes particuliers observés. En effet, si en théorie, on juxtapose des phonèmes comme on enfile des perles à un collier, la réalisation physique d'une suite phonémique ne laisse apparaître qu'un continuum. De plus, si phonologiquement, chaque phonème est unique, d'un point de vue phonétique, un même phonème peut prendre de nombreuses apparences: on touche ici au problème de la non biunivocité qui existe entre le plan phonétique et le plan phonologique (cf. § « Les relations inhomogènes », p.40). Pour mieux apprécier cette différence, prenons l'exemple de l'écriture. Il n'existe qu'une première

lettre de l'alphabet (le graphème est une notion abstraite) mais celle-ci peut s'écrire sous un très grand nombre de formes dites allographies: a, A, ***a,*** a, A, **a**... La Figure 33 propose une classification des différences qui existent entre les sons du langage.

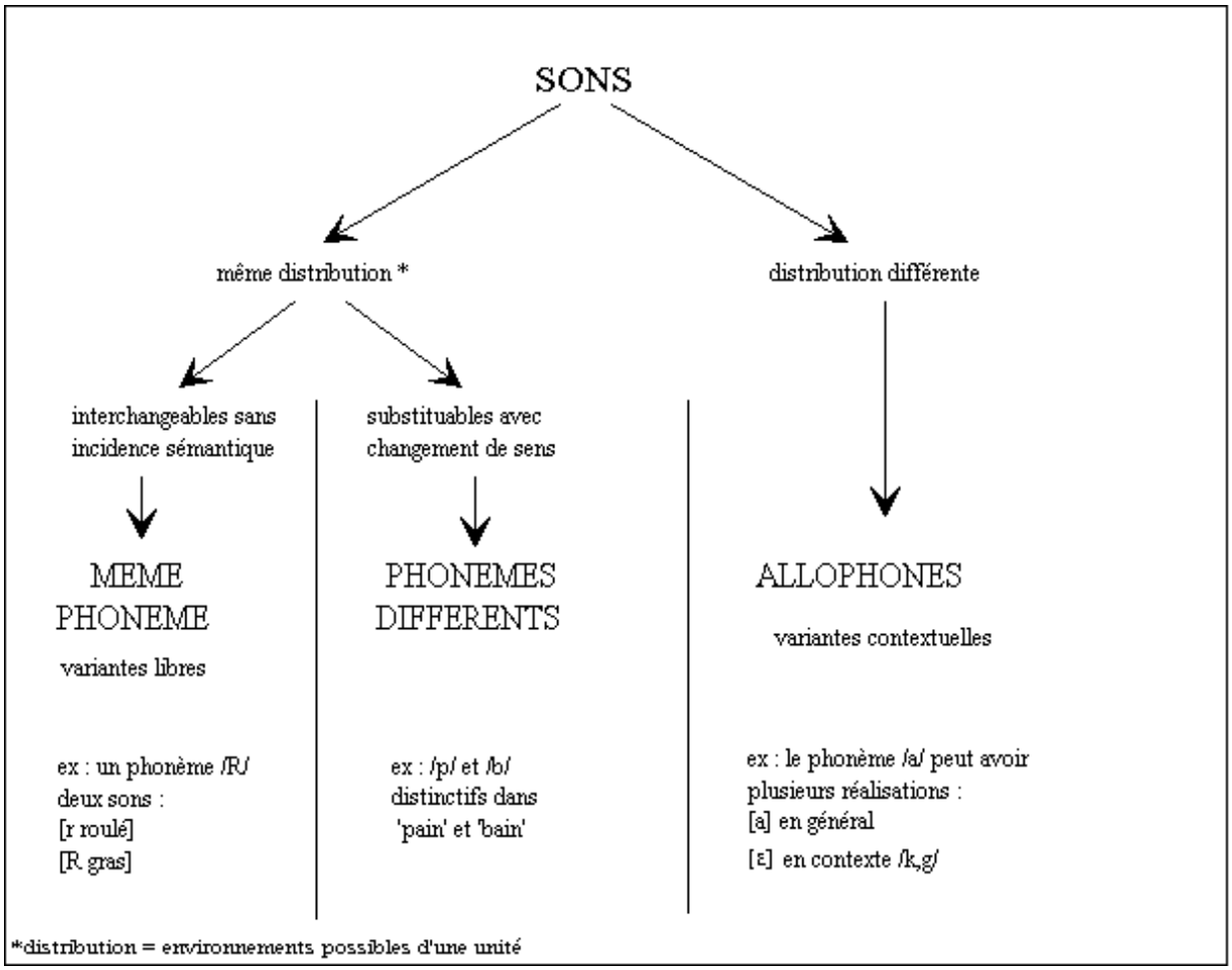


Figure 33: Les différences entre les sons du langage

Du fait de ces différences, plusieurs remarques sont à formuler:

- pour gérer le problème des variantes libres, il faut prévoir pour chaque phonème la possibilité de trouver plusieurs types de réalisations. Dans un système de reconnaissance analytique à stratégie montante, on cherchera à appliquer plusieurs jeux de règles pour identifier un même phonème. Dans un système stochastique, il faut prévoir dans le corpus d'apprentissage les différentes variantes libres de chaque phonème.
- pour gérer le problème des variantes contextuelles dans un décodage analytique fondé sur une stratégie montante, on cherchera à appliquer des jeux de règles dépendant du contexte. Dans l'optique d'une stratégie descendante, on cherchera à rectifier, a posteriori, le résultat du décodage une fois le contexte connu. Dans un système stochastique, il faut prévoir dans le corpus d'apprentissage les différents contextes possibles de chaque phonème.

Il faut mentionner l'existence de variantes dialectales qui nuancent encore ces distinctions. Prenons un exemple. /o/ et /ɔ/ sont contrastifs dans "saute" et "sotte " dans le cas du français « standard ». En revanche, en français méridional, ces deux mots sont prononcés sous la forme unique [sɔtə]. Cela ne signifie pas pour autant qu'il y ait neutralisation de [o] et [ɔ] en français méridional. La distinction /o/ et /ɔ/ existe, mais elle suit une loi de position, [o] étant présent en syllabe ouverte, [ɔ] en syllabe fermée. Le "o" de "saute" et "sotte" étant dans les deux cas en syllabe fermée, ces deux mots sont donc prononcés [sɔtə].

Cette remarque est importante car elle met en évidence les insuffisances souvent présentes dans les transcriptions phonologiques des éléments de dictionnaire. Ainsi, dans la base de données BD-Lex, "saute" est transcrit /sotə/, "sotte" est transcrit /sɔtə/. Pour s'adapter à une élocution méridionale, un aménagement est nécessaire, le décodage montant de [sɔtə] pouvant se rapporter aux deux mots. L'utilisation de règles phonologiques apparaît alors comme nécessaire.

II.4.C. L'utilisation de règles phonologiques

Les règles associant les phonèmes à leurs formes d'expression selon le contexte sont appelées *règles phonologiques*. Elles fonctionnent dans le sens transformationnel (cf. § « Les relations inhomogènes », p.40), c'est à dire que le « calcul » se fait de la façon suivante:

REGLE

phonème ⇒ séquence sonore

exemples de règles:

- voyelle devant fricative voisée entraîne allongement:

  / a / ⇒ [a:] comme dans "vase"

- assimilation consonantique régressive

  - assourdissement: / b / ⇒ [p] comme dans "absolu"

    / d / ⇒ [t] comme dans "méd(e)cin"

  - voisement: / k / ⇒ [g] comme dans "anecdote"

- dévoisement des consonnes en position finale

  ex: le mot "dix" se prononce /diz/ en position médiane (ex: j'ai dix ans)

  /dis/ en position finale (ex: le bus numéro dix)

Ces règles, fonctionnant dans le sens transformationnel, sont directement applicables pour la synthèse de la parole. Ce n'est malheureusement pas le cas pour la reconnaissance automatique. Toutefois, elles peuvent s'avérer utiles dans le sens descendant (du niveau abstrait des phonèmes vers le niveau acoustique) afin d'effectuer des vérifications ou de lever des ambiguïtés dues au contexte. Ainsi, si l'analyse acoustique montante identifie la syllabe §ap§.

Cela correspond-il simplement aux phonèmes /ap/ pris au sens phonologique ? Non. Car la règle d'assimilation régressive (assourdissement des occlusives voisées précédant une consonne sourde, voir plus haut) nous permet de penser que l'identification acoustique d'un [ap] peut conduire à deux hypothèses phonémiques: /ap/ ou /ab/. L'ambiguïté peut être levée par une étude du contexte, une analyse plus fine comme l'étude de la longueur de la voyelle ou par un accès lexical. Des détails sur l'utilisation de règles phonologiques en R.A.P. sont donnés dans (Rossi, 1980).

II.4.D. Les traits phonétiques

II.4.D.a. La notion de trait pertinent

Toutes les propriétés de la parole ne sont pas forcément employées pour la transmission d'un message. Ainsi, le timbre de la voix, lié au locuteur, à son état physique ou émotif, le débit de parole, la hauteur de la voix... ne participent pas directement au transport d'information verbale. Par opposition, un *trait pertinent* est une propriété distinctive de la parole, qui permet à elle seule d'opposer deux phonèmes, c'est à dire deux unités d'information sémantique. Ainsi, le trait de voisement (vibration des cordes vocales) est distinctif pour opposer /b/, qui est voisé, et /p/, qui ne l'est pas, dans la transmission orale des mots "bain" [bɛ̃] et "pain" [pɛ̃]. D'un point de vue linguistique, il existe une description qui permet de classer les phonèmes par opposition mutuelle à l'aide de *faisceaux de traits* (Jakobson et al., 1951). Ainsi, indépendamment du contexte, le phonème /b/ peut être défini par l'ensemble {non syllabique - consonantique - non sonant - voisé - labial }, la voyelle /y/ par {syllabique - non consonantique - sonante - orale - labiale - fermée - antérieure}. Cette caractérisation par traits constitue un système minimal et pertinent. En général, la langue ne conserve que les oppositions phonologiques utiles, qui ont du rendement.

II.4.D.b. Le système des traits pertinents du français

Chaque phonème est défini de façon unique par la place qu'il occupe dans le système des traits pertinents. Il est à noter que la pertinence des traits varie suivant la langue et suivant les macro-classes de phonèmes. Ces traits peuvent être de différentes natures: articulatoires (labialité, antériorité, aperture...), acoustiques (acuité, voisement, compacité...) ou linguistique (syllabicité). Le Tableau 3 dresse une liste des principaux traits phonétiques avec leurs correspondances articulatoires, acoustiques et perceptives.

(Dell, 1985, p.63) définit le *trait de syllabicité* de la façon suivante: « sont syllabiques les sons qui peuvent à eux seuls constituer une syllabe ». Bien que cette définition puisse a priori poser des problèmes d'interprétation, il semble qu'elle permette de distinguer intuitivement les voyelles de tous les autres phonèmes. Ce trait ainsi que ceux relatifs au vocalisme et au consonantisme nous permettent de dresser une première classification des phonèmes du français en quatre grandes familles (Tableau 4).

Tableau 3: Les traits de la parole avec leurs corrélats
(d'après Landercy & Renard, 1977, p.169 ; Dell, 1985, pp.54-66)

Trait Articulatoire	Trait Acoustique	Corrélat Articulatoire	Corrélat Acoustique	Corrélat Perceptif
Consonantique	.	entrave au passage de l'air	vibration apériodique	bruit
Vocalique	Sonant	peu ou pas d'entrave au passage de l'air	structure formantique ou apparentée	timbre vocalique
Voisé	Sonore	vibration des cordes vocales	source périodique (fréquence fondamentale)	timbre quasi-musical mélodique
Nasal	.	abaissement du voile du palais	modification de l'amplitude relative des formants	son nasal
Antérieur	.	articulation antérieure de la langue	pour les voyelles accroissement de F2	timbre plus clair
Postérieur	.	articulation postérieure de la langue	pour les voyelles abaissement de F2	timbre plus sombre
Arrondi	.	arrondissement des lèvres	pour les voyelles abaissement de F2	timbre plus sombre
Ouvert	.	écartement des mâchoires	pour les voyelles, hausse de F1, rapprochement de F1 et F2	son plus intense
Labial	Grave (consonnes)	utilisation des lèvres	spectre favorable aux basses fréquences	timbre plus sombre
Dental	Aigu (consonnes)	articulation localisée au niveau des dents	spectre favorable aux hautes fréquences	timbre plus clair
Palatal	Compact (consonnes)	articulation localisée au niveau du palais	spectre favorable aux moyennes fréquences	timbre médian
Occlusion	Interrompu	fermeture + tenue + relâchement brusque	vibration apériodique impulsionnelle	explosion
Constriction	Continu	passage continu de l'air dans un rétrécissement du conduit vocal	vibration apériodique continue	friction

Tableau 4: Les macro-classes phonémiques du français (d'après Dell, 1985, p.294)

consonantique	non vocalique		obstruantes	Occlusives, Fricatives
	vocalique		consonnes nasales et liquides	Consonnes vocaliques
non consonantique		non syllabique	semi-voyelles	Consonnes vocaliques
		syllabique	voyelles	

II.4.D.c. Le système des voyelles et semi-voyelles du français

La catégorie des phonèmes non consonantiques du français regroupe les macro-classes des voyelles et des semi-voyelles. Elles sont toutes deux vocaliques (sonantes). La différence entre elles est assuré par le trait de syllabicité, les voyelles pouvant constituer à elles seules une syllabe, les semi-voyelles ne pouvant se le permettre. Pour éclairer cette définition, prenons un exemple (Dell, 1985, p.63). « Le son de timbre *u* qui apparaît dans la prononciation de *troua* [trua] compte pour une syllabe, mais pas celui qui apparaît dans *trois* [trwa]. Le premier, qui

est noté [u], est une sonante non-consonantique syllabique, autrement dit une voyelle. Le second, qui est noté [w], est une sonante non-consonantique non-syllabique. »

- *Les voyelles*

  Une approche phonologique permet d'isoler, en français, un système de 14 unités de type syllabique, vocalique, non consonantique. Rossi propose une classification des voyelles du français selon un arbre à cinq traits articulatoires (Figure 34).

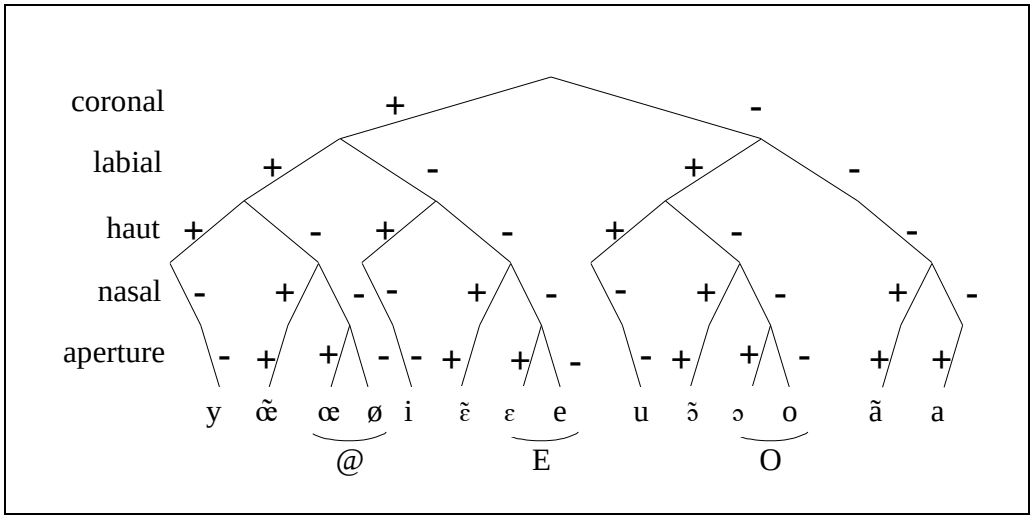


Figure 34: Classification en arbre des voyelles du français

  Cette classification est intéressante car elle permet de mieux comprendre les phénomènes d'archiphonémie. Ainsi, la neutralisation du trait d'aperture laisse apparaître 3 archiphonèmes (Figure 34) :

/@/	=	/ø/ ou /œ/
/E/	=	/e/ ou /ɛ/
/O/	=	/o/ ou /ɔ/

- *Les semi-voyelles (ou semi-consonnes)*

  Les *semi-voyelles* sont les phonèmes du type {non consonantique - vocalique - non syllabique}. Elles sont constituées par les phonèmes yod /j/, ué /ɥ/ et oué /w/. Les nomenclatures traditionnelles les appellent indifféremment glissante, semi-voyelle ou semi-consonne, dénominations révélant leur nature ambiguë. Toutefois, il semble que deux indices complémentaires permettent de valider leur existence ou en tous cas, de les distinguer des voyelles: l'intensité globale et la vitesse des transitions formantiques, une modification de la durée des transitions entraînant un changement de catégorisation phonétique (Liberman et al., 1956).

II.4.D.d. Le système consonantique du français

Le classement des consonnes en traits distinctifs posent plus de problèmes. Généralement, par le biais du trait vocalique, on distingue, d'une part les obstruantes pour lesquelles le flux d'air est contrarié partiellement (constrictives) ou complètement (occlusives), d'autre part, les consonnes vocaliques, moins « bruyantes » et plus « sonantes ». Examinons tout d'abord les obstruantes.

- *Les obstruantes*

  Les obstruantes sont les phonèmes du type {consonantique - non vocalique}. Parmi celles-ci, on distingue deux catégories différenciées entre elles par le mode d'articulation: mode occlusif (trait interrompu) ou constrictif (trait continu) appelé aussi fricatif (Tableau 5).

  Les occlusives sont caractérisées par la mise en place d'une obstruction totale du flux d'air en un point du conduit vocal. La phase de fermeture est appelée « tenue » de l'occlusion et peut être voisée ou non (cf. § « La production de la parole », p.49 et § « L'aspect acoustique de la parole », p.53). L'ouverture du conduit vocal se traduit par une explosion plus ou moins intense, longue et aiguë selon le lieu d'articulation et le contexte. Une phase de relâchement s'ensuit, pendant laquelle s'effectue l'enchaînement avec l'unité suivante. Notons que les occlusives présentent une organisation temporelle particulière, dans la mesure où, par nature, elles sont constituées de différentes phases (tenue et explosion). L'identification de ces unités est donc liée à ces deux propriétés. Toutefois, il faut noter que la tenue n'est pas nécessairement repérable, notamment si une occlusive non voisée est à l'initiale d'un énoncé ; le silence de la tenue est alors confondu avec la pause silencieuse. De même, l'explosion est parfois non identifiable car trop peu intense, surtout pour les labiales /p, b/.

  Les constrictives sont caractérisées par la mise en place d'une fermeture importante mais non totale en un point du conduit vocal. Cette constriction se traduit par la génération d'un bruit de friction, ce bruit pouvant être accompagné d'un voisement ou non (cf. § « La production de la parole », p.49 et § « L'aspect acoustique de la parole », p.53).

Tableau 5: Les traits articulatoires distinctifs des obstruantes du français.

		labiales	dentales	vélo-palatales
OCCLUSIVES (interrompues)	sourdes	/p/	/t/	/k/
	sonores (voisées)	/b/	/d/	/g/
CONSTRICTIVES (continues)	sourdes	/f/	/s/	/ʃ/
	sonores (voisées)	/v/	/z/	/ʒ/

- *Les nasales et les liquides*

La classe des phonèmes du type {consonantique - vocalique} se compose des nasales et des liquides. Les nasales peuvent, d'un point de vue articulatoire, s'assimiler à des obstruantes non continues (occlusives). Il existe, en effet, un point d'occlusion complète au niveau de la cavité buccale. Cependant, dans le cas des nasales, grâce à l'abaissement du voile du palais, le flux d'air s'écoule librement par la cavité nasale. La conséquence est l'existence d'une structure spectrale stable et continue, contrairement aux occlusives. Les sons produits sont donc du type {non syllabique - consonantique - vocalique - nasal}. Le Tableau 6 en donne les différentes versions selon le point où a lieu l'occlusion.

Tableau 6: Classement des consonnes nasales du français.

Labiale	Dentale	Palatale	Vélaire
/m/	/n/	/ñ/	/ ŋ /

Les consonnes /l/ et /R/ constituent la catégorie des liquides. Elles sont consonantiques, vocaliques et non nasales. Dans la production de /l/, la pointe de la langue bloque le fconduit vocal à l'arrière des incisives supérieures mais l'air contourne facilement la masse linguale par les côtés. On parle donc de liquide latérale. D'un point de vue phonologique, il n'existe, en français, qu'un et un seul phonème "r" que l'on note /R/, du fait de son unicité dans les oppositions distinctives. Pourtant, la production de /R/ est extrêmement variable (Straka, 1965). Les trois modes de réalisation du "R" généralement exposés pour le français sont le [ʁ] grasseyé, le [r] apical roulé et le [x] guttural. Ces différents types de réalisations phonétiques entraînent d'immenses difficultés dans le cadre d'un décodage automatique. Comme le propose Meunier dans (Meunier, 1994, p.64) « considérer /R/ comme une consonne phonétiquement polymorphe n'est pas forcément judicieux pour l'analyse de la variabilité. »

II.4.D.e. Trait articulatoire / trait acoustique

Généralement, les classements par traits ont la particularité d'être de type articulatoire: ouvert, fermé, labial, dental, nasal.... Il est toutefois possible de proposer un autre type de classification de type acoustique, fondée sur des caractéristiques physiques tels que aigu/grave, compact/diffus... Diverses expériences et constatations semblent donner une réalité à la notion de traits acoustiques organisés en un système cohérent (Rossi, 1977). Dans tous les cas, une telle description paraît beaucoup mieux adaptée au traitement automatique de la parole, où les données d'entrée sont d'origine acoustique. De plus, les passages entre les corrélats acoustiques et articulatoires sont parfois immédiats du fait de la correspondance partielle entre le plan articulatoire et le plan acoustique (Tableau 3, p.66). Ainsi, la notion d'antériorité pour les voyelles est propice à l'acuité, la labialisation se manifeste par l'abaissement du deuxième formant, un lieu d'articulation au niveau du palais se traduit par un

spectre compact... Le décodage peut donc s'effectuer par la recherche de traits non pas articulatoires mais acoustiques (Tableau 7).

Tableau 7a: Analyse binaire, en traits acoustiques, du système vocalique français (d'après Rossi)

Trait	a	i	u	o	e	y	ø	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃
extrême	+	+	+	-	-	+	-	-	-	-	+	-	-	-
diffus	-	+	+	+	+	+	+	-	-	-	-	-	-	-
aigu	-	+	-	-	+	+	+	+	-	+	-	+	-	+
bemolisé	-	-	+	+	-	+	+	-	+	+	-	-	+	+
nasal	-	-	-	-	-	-	-	-	-	-	+	+	+	+

Tableau 7b: Analyse binaire, en traits acoustiques, du système consonantique français (d'après Rossi)

Trait	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ñ	l	R	j	w
vocalique	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+
interrompu	+	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
compact	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+	-	+	-	+
aigu	-	+	+	-	+	+	-	+	+	-	+	+	-	+	+	+	+	+	+
voisé	-	-	-	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+

II.4.E. L'utilisation des traits en décodage acoustico-phonétique

Une analyse par traits permet de distinguer tous les phonèmes entre eux. L'identification d'une unité phonétique peut donc se ramener à une recherche de traits. Ainsi, pour identifier [p], il faut et il suffit d'identifier les traits {consonantique - non vocalique - interrompu - non voisé - labial}, pour reconnaître [i], il faut et il suffit de détecter les traits {syllabique - coronal - non labial - haut - non nasal - close}.Cette démarche peut s'avérer efficace dans la mesure où « l'analyse en traits constitue un degré d'abstraction qui permet d'éliminer certaines variations non significatives pour la reconnaissance, notamment certaines variations contextuelles » (Rossi et al., 1981, p.3).

L'erreur majeure, qui a souvent été commise dans le passé, est de croire que cette recherche est immédiate et que les traits sont visibles dans le signal de parole. Cette confusion est d'autant plus facile si l'on choisit une classification des phonèmes par traits acoustiques. La notion de trait est du domaine de l'abstrait; il s'agit d'une représentation formelle. D'après

Rossi (Rossi, 1981, p.8), « chaque trait est construit sur la base d'un faisceau d'indices en correspondance biunivoque avec certaines propriétés du signal, la propriété devenant indice par la valeur que lui assigne l'auditeur. Les propriétés correspondent elles-mêmes à une organisation particulière des trois paramètres du signal de parole: la fréquence, l'amplitude et le temps » (Figure 35).

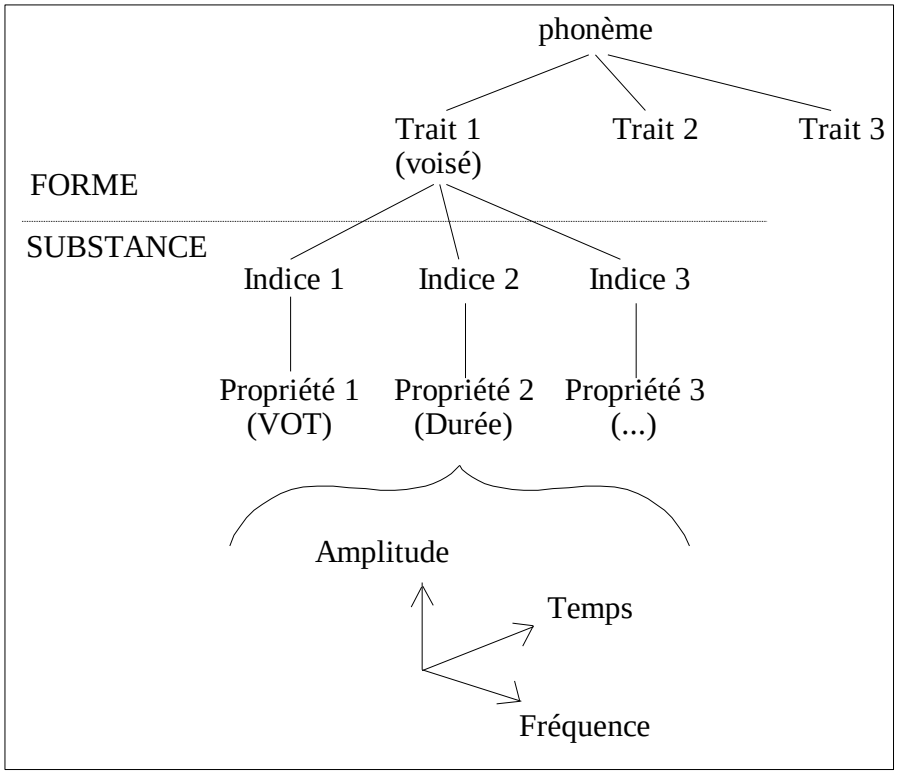


Figure 35: Traits, indices, propriétés et paramètres (d'après Rossi, 1981, p.8)

L'utilisation de traits en R.A.P. nécessite donc plusieurs étapes:

- la mise en oeuvre de détecteurs de propriétés à partir de paramètres acoustiques relatifs à la durée, l'amplitude et la répartition spectrale du signal de parole.
- les indices étant la composante abstraite de la substance, leur détermination reste immédiate à partir des propriétés.
- le regroupement d'indices en faisceau permet l'identification de traits.
- la combinaison des traits détectés autorise la caractérisation phonémique.

Ainsi, le trait de syllabicité, qui appartient au domaine de l'abstraction, peut être identifié à partir d'indices obtenus de propriétés relatives à l'intensité globale, à la durée des transitions et à la distribution (présence de consonnes adjacentes). La distinction entre voyelles et semi-voyelles est possible à condition que les détecteurs de propriétés soient robustes. Il faut remarquer que la notion de traits pertinents telle que la conçoivent (Jakobson et al., 1951) implique une organisation en un système économique et minimal. Elle exclut toute

redondance. Ainsi, dans le cas du français, la nasalité implique systématiquement une production voisée. Une classification phonologique des phonèmes du français ne mentionne donc jamais le trait de voisement en présence du trait de nasalité. Dans une optique minimaliste, cet « oubli » est justifié. Dans le cas du décodage acoustico-phonétique, la redondance doit être exploitée au maximum. Comme le font remarquer (Eskénazi & Liénard, 1981, p.58), « il nous semble imprudent de considérer certains traits comme auxiliaires ou redondants: a-t-on démontré qu'ils n'étaient pas utiles à la perception des phonèmes ? ». Ainsi, l'identification d'une nasale est dépendante non seulement de la recherche d'indices propres à la nasalité mais aussi de la présence du voisement. Cette redondance permet ainsi de confirmer ou d'infirmer certaines hypothèses.

PARTIE -2-

LES REALISATIONS

III. UNE ANALYSE SPECTRALE FONDEE SUR UN MODELE AUDITIF

Dans l'industrie, les informaticiens écrivent des programmes répondant au cahier des charges des utilisateurs. C'est impossible chez les scientifiques, qui doivent écrire eux-mêmes leurs codes. La première raison, évidente, est que pour écrire un programme scientifique il faut connaître la théorie qui le sous-entend.. « Il faut toujours modifier, revenir en arrière. En recherche, on ne connaît pas d'avance le cahier des charges » (Bernard Remaud)

« L'impossible preuve ». Extrait du « journal du CNRS », décembre 1992.

Plan du chapitre

RESUME

L'étape d'extraction de paramètres physiques est extrêmement importante car elle permet d'accéder à l'information utile contenue dans le signal de parole. L'analyse spectrale reste l'une des méthodes les plus efficaces. Le modèle auditif que nous avons développé s'inspire de résultats de psycho-acoustique relatifs à l'audition des sons, notamment à leur analyse fréquentielle par l'oreille. Cette démarche est justifiée par le fait que le décodage de la parole au niveau cortical s'effectue non pas sur le signal brut de parole, mais plutôt sur un signal filtré par le système auditif. Il se pourrait, en définitive, que l'information la plus pertinente pour l'identification d'un message oral soit celle qui serait passée par ce filtre auditif.

Deux propriétés principales de l'oreille ont été modélisées: la pondération sonique et l'intégration en bandes critiques. La pondération sonique consiste à corriger, de façon non-linéaire, les composantes spectrales du signal d'entrée. Ceci est en rapport avec les courbes d'isosonie, c'est à dire le fait qu'un son d'intensité 70 dB et de fréquence 50 Hz sera perçu à la même puissance sonore qu'un son de 50 dB produit à 1000 Hz. L'intégration en bandes critiques exploite le fait que l'oreille possède une résolution fréquentielle correcte en basses fréquences mais médiocre dans les hautes fréquences. Autrement dit, un écart de 100 Hz est pertinent dans les basses mais superflu dans les aigus.

Le résultat de l'analyse spectrale fondée sur un modèles auditif est une représentation temps-fréquence du signal d'entrée. Il s'agit, d'une part, d'un vocodeur baptisé « CritiVoc » et, d'autre part, d'un ensemble de coefficients dits « P.L.P. » (Prédiction Linéaire sur un spectre Perceptif) si l'analyse est poursuivie par une prédiction linéaire. Ces calculs sont finalement employés dans les différents modules de décodage.

III.1. Un modèle auditif

III.1.A. De l'intérêt de l'utilisation de modèles auditifs en reconnaissance automatique de la parole

Le décodage acoustico-phonétique débute par une extraction de paramètres physiques (cf. § « L'extraction de paramètres acoustiques », p.17). Cette étape est extrêmement importante car elle permet d'extraire l'information utile contenue dans le signal de parole. Une extraction médiocre pourrait handicaper grandement un système de décodage de la parole, tout comme une mauvaise mise au point ne permet pas d'obtenir une image intéressante en optique et traitement de l'image. L'extraction de paramètres acoustiques peut consister à calculer l'énergie, le taux de passage par zéro, la fréquence fondamentale... Toutefois, l'estimation de la répartition spectrale du signal de parole en fonction du temps reste l'une des méthodes les plus précises. C'est dans ce but qu'a été développé un type d'analyse fondée sur un modèle auditif.

Même si les systèmes de reconnaissance actuels semblent de plus en plus performants, l'homme reste bien supérieur à la machine en matière de décodage de la parole. Il paraît donc judicieux que des efforts soient faits pour étudier avec précision le comportement perceptif humain, et en particulier l'audition. Comme nous l'avons vu au paragraphe sur « Les questions préalables », p.9, le décodage de la parole au niveau cortical s'effectue non pas sur le signal brut de parole, mais plutôt sur un signal filtré par le système auditif. Il semblerait, en définitive, que l'information la plus pertinente pour l'identification d'un message oral soit celle qui serait passée par ce filtre auditif. Connaître en partie le fonctionnement de l'oreille permet de mieux gérer l'information transmise par le signal de parole. Ceci permet, d'une part, une meilleure assimilation (deux sons physiquement bien distincts mais perçus comme proches et devant être reconnus comme analogues). D'autre part, cela autorise une meilleure discrimination (deux sons globalement peu distincts d'un point de vue physique mais perçus comme différents et devant être reconnus comme distincts). On comprend dès lors l'intérêt de développer un système de traitement du signal modélisant au mieux le traitement auditif humain. En tenant compte des travaux de (Caelen, 1979; Teston, 1983; Yong & Mason, 1987; Hermansky, 1990; Junqua, 1990), nous nous sommes orientés dans cette direction.

III.1.B. La pondération sonique

La perception chez l'homme de la puissance sonore ne correspond pas directement à une mesure physique de l'intensité. La sensation sonore dépend, en effet, d'une part de l'intensité du signal mais aussi de sa fréquence. On parle alors d'intensité subjective.

III.1.B.a. Les courbes d'isosonie

La perception de la puissance sonore a été longuement étudiée avec des sons purs et des bruits de différents types par les psycho-acousticiens (Zwicker & Feldkeller, 1981). On peut mettre en évidence expérimentalement qu'un son pur de 50 dB émis à 2000 Hz paraîtra plus fort qu'un son de même puissance émis à 300 Hz. Des séries de courbes d'égales sensations sonores ont été établies à partir de nombreuses expériences. La Figure 36 fournit ces relevés qui ont été standardisés internationalement pour des sons purs écoutés dans des conditions bien précises. La consigne du test était d'ajuster l'intensité physique d'un son à

fréquence quelconque pour qu'il soit perçu à la même intensité subjective qu'un son émis à 1000 Hz. Chaque courbe est indicée par la puissance physique du son référence de 1000 Hz. En faisant varier cette puissance, on obtient un faisceau de courbes.

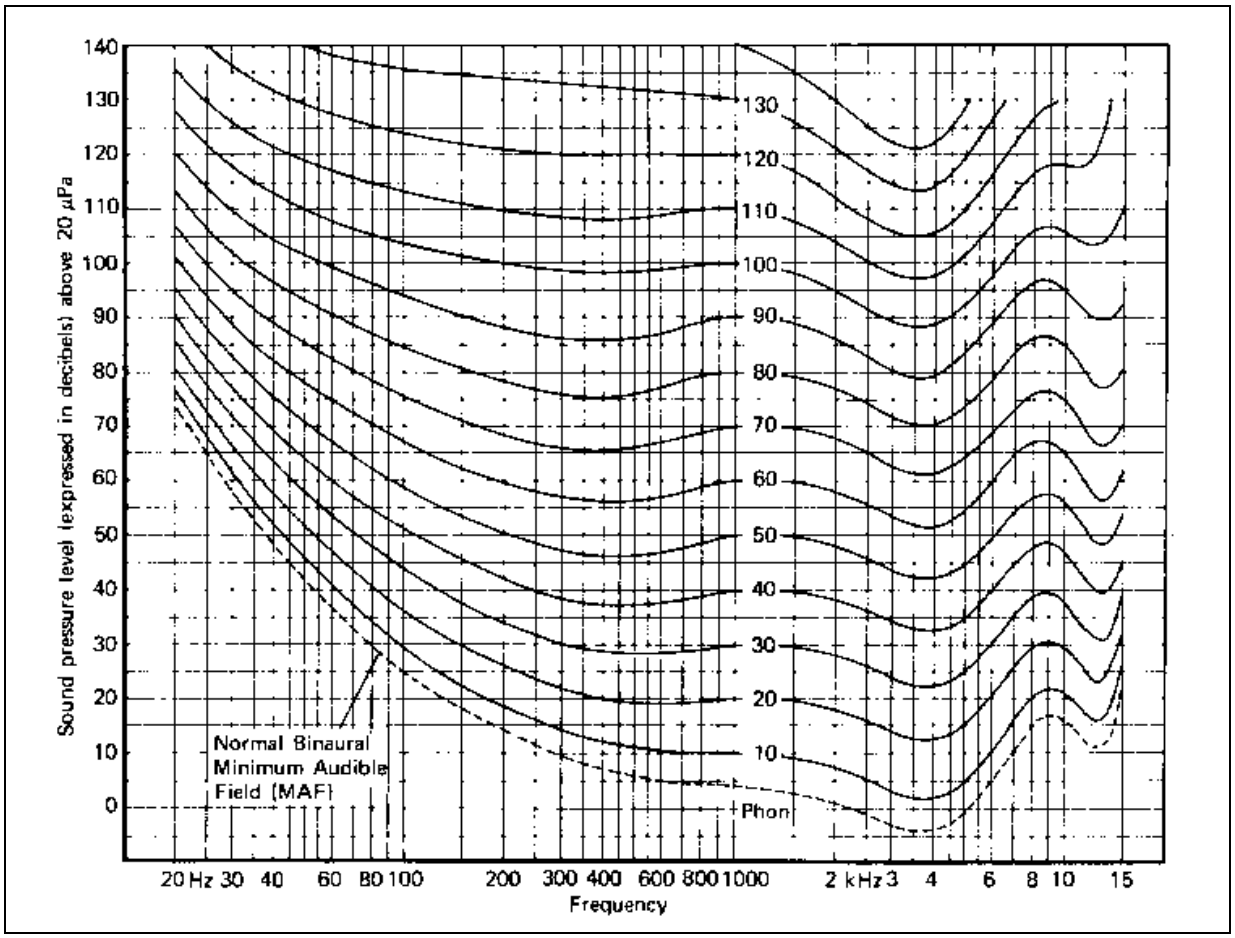


Figure 36: Courbes d'égale sensation sonore pour les sons purs (Source: Hassall et al., 1979)

Une unité d'intensité subjective est le *phone*. La Figure 36 rend compte du fait que pour percevoir un son de 50 phones, il faut l'émettre avec une puissance de 50 dB à 1000 Hz, 73 dB à 50 Hz, 42 dB à 4000 Hz. Une échelle proportionnelle à l'intensité subjective a été développée: l'échelle des *sones*. Son expression mathématique est la suivante:

$$S\ (\ en\ sones\ ) = 2^{(\frac{P\ -\ 40}{10})} \qquad \text{où } P \text{ est en phones}$$

La Figure 37 met en relation graphiquement le phone et le sone. On remarque qu'à phone constant, on se trouve aussi à sone constant. Le sone étant plus fréquemment utilisé, les courbes d'égale sensation sonore sont généralement appelées *courbes isosoniques.*

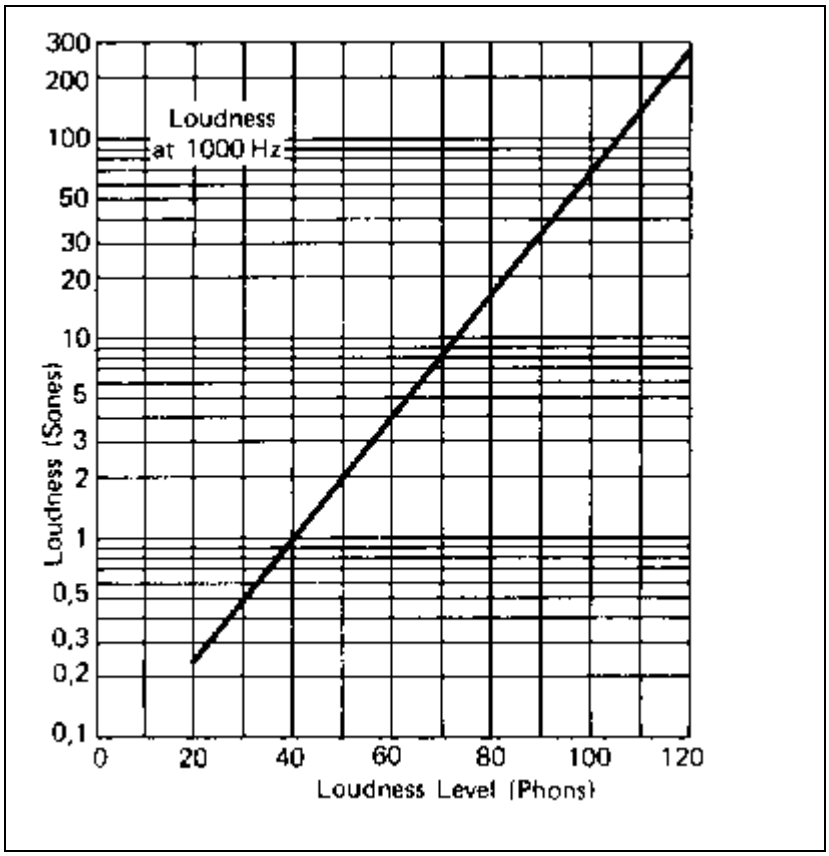


Figure 37: Relation entre l'intensité subjective exprimée en phones et en sones (Source: Hassall et al., 1979)

III.1.B.b. Les pondérations soniques

La réponse harmonique de l'oreille est loin d'être linéaire en fréquence pour des sons purs. L'une des conséquences de cette non-linéarité est une déformation de la répartition spectrale d'un son complexe à travers le filtre auditif. Afin de simuler ce pré-traitement, il apparaît intéressant de réaliser une correction de l'intensité en fonction de la fréquence avant d'aborder une phase de décodage. Pour effectuer cette correction, ont été réalisées *des courbes de pondération sonique* permettant de passer de l'intensité physique à l'intensité subjective. On a ainsi:

$$S(f) = E(f) + P(f)$$

où *f* est la fréquence
S est l'intensité perçue en dB
E est l'intensité réelle en dB
P est la pondération en dB

Plusieurs types de pondérations ont été proposés selon différentes conditions (Figure 38). La pondération A est relative à la courbe isosonique de 40 phones
B " " " " " " 70 "
C " " " " " " 100 "
D " " à la mesure de bruits tels que ceux des avions...
E est proposée par Stevens

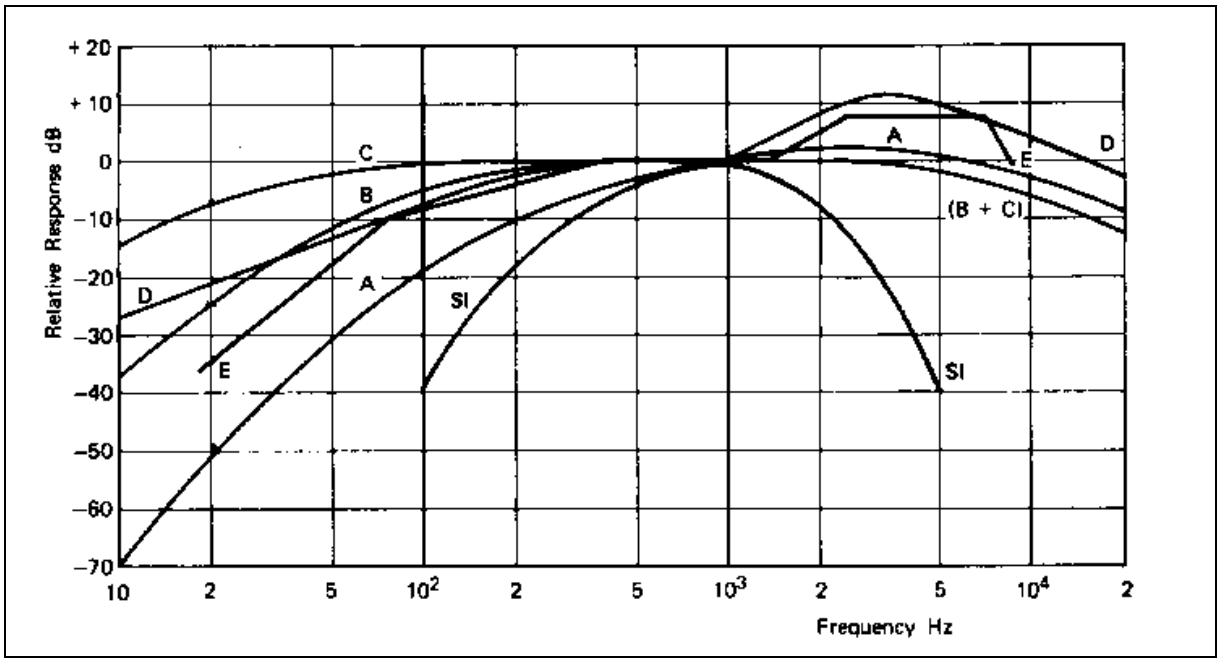


Figure 38: Courbes standardisées de pondération soniques (Source: Hassall et al., 1979)

Dans notre cas, nous avons retenu deux types de pondération: le type A et E. La courbe A standard est la plus utilisée du fait de sa simplicité mais a l'inconvénient d'être un peu trop normalisée. La courbe E de Stevens, plus complexe du fait de son lobe autour de 4000 Hz, tient mieux compte des phénomènes réels.

III.1.B.c. La modélisation des pondérations soniques

Pour modéliser une pondération sonique, deux méthodes peuvent être utilisées: un tableau de valeurs discrètes ou une expression analytique. Un tableau de valeurs discrètes permet de rendre compte de n'importe quelle courbe, aussi peu régulière soit elle. Toutefois, il ne permet pas de changer de domaine d'utilisation facilement. Aussi, avons-nous préféré utiliser une expression analytique obtenue à partir de relevés précis (Teston, 1983). La meilleure modélisation a été obtenue en exprimant la pondération en dB fonction du logarithme de la fréquence en Hz. Les expressions analytiques de ces courbes peuvent être réduites à des polynômes du type:

$$Pond\ (en\ dB) = a + b.Log(f) + c.[Log(f)]^2 + d.[Log(f)]^3 + ...$$

La pondération de type A est relativement régulière (Figure 39a). Néanmoins, elle présente un peu trop d'uniformité dans sa partie supérieure, au dessus de 2000 Hz. La pondération de type E (Figure 39b) a le mérite de relever, du fait de son lobe autour de 4000 Hz, les moyennes et hautes fréquences souvent acoustiquement faibles mais perceptivement présentes (le lobe est donc pertinent d'un point de vue perceptif). Toutefois, elle présente des imperfections en basses fréquences (trop faible atténuation) et en très hautes fréquences (trop forte atténuation). La solution que nous avons choisie consiste à adopter une pondération baptisée de type X dont le profil est celui de la courbe A en basses fréquences (0 à 2000 Hz), celui de la courbe E entre 2000 et 6000 Hz et une fonction de transfert unité au delà (Figure 39c et Figure 39d).

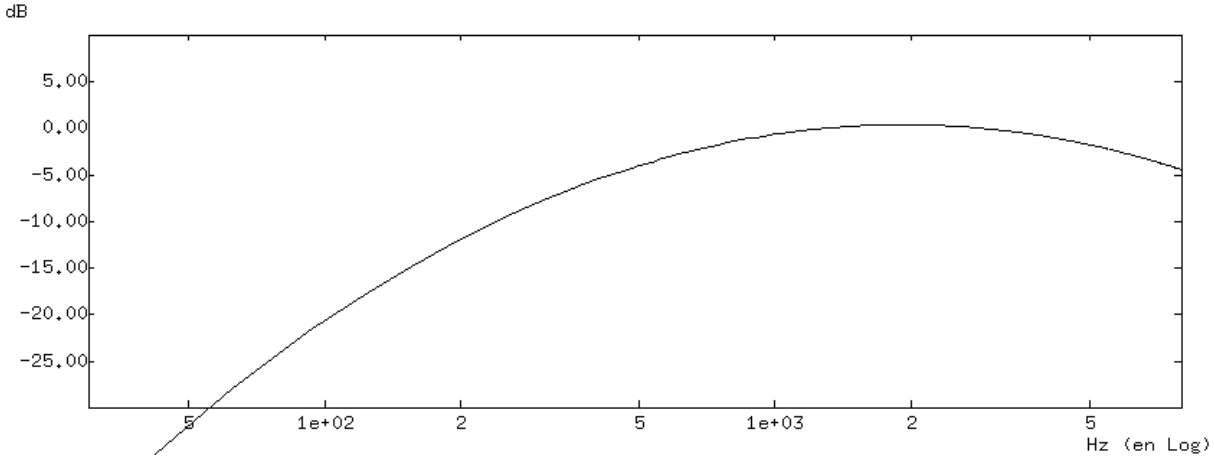


Figure 39a: Pondérations de type A (x en Hz avec progression logarithmique, y en dB)

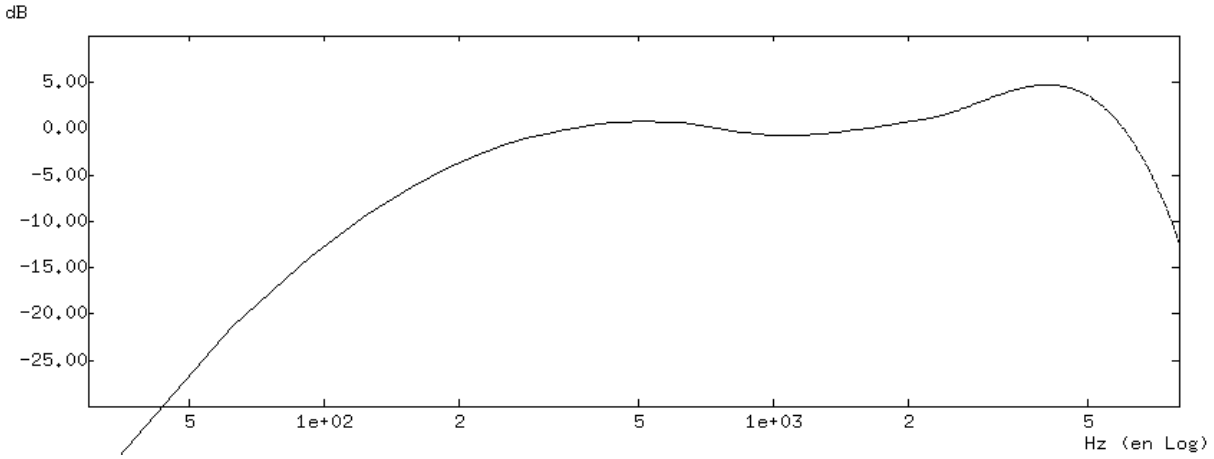


Figure 39b: Pondérations de type E (x en Hz avec progression logarithmique, y en dB)

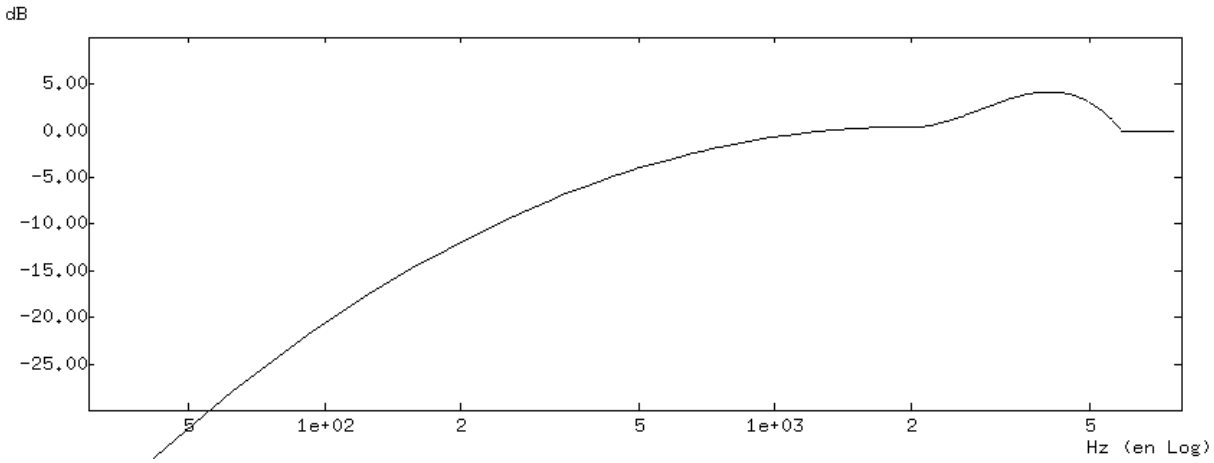


Figure 39c: Pondérations de type X (x en Hz avec progression logarithmique, y en dB)

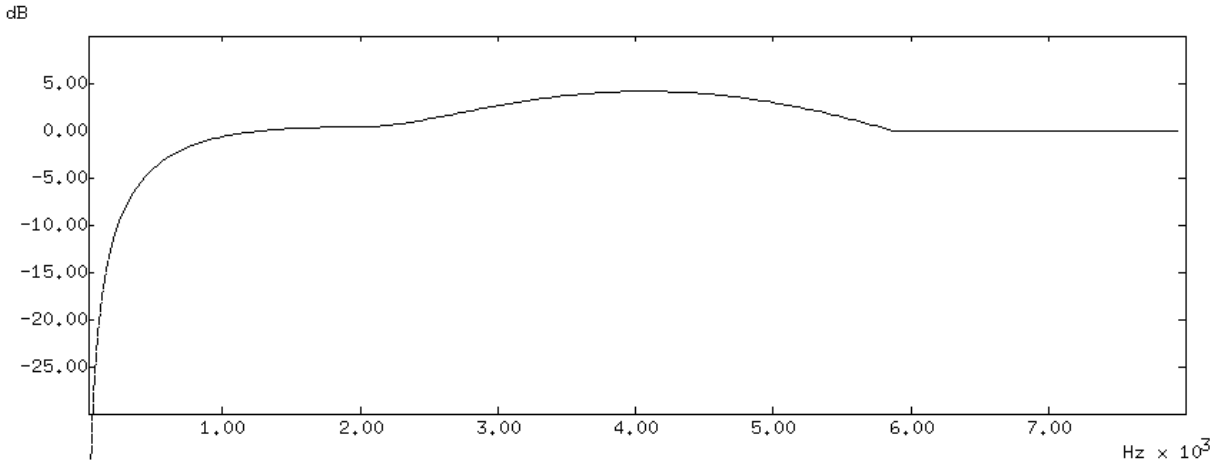


Figure 39d: Pondérations de type X (x en Hz avec progression linéaire, y en dB)

III.1.B.d. Le spectre subjectif

L'expression finale de la pondération (Figure 39d) est la suivante :

On pose $x = Log10(f \text{ en Hz})$

Pour f ∈ [0 ; 2000 Hz] $P_{dB}(f) = -136.963 + 83.623\ x - 12.725\ x^2$

Pour f ∈ [2000 ; 5880 Hz] $P_{dB}(f) = 10808.277 - 9418.1\ x + 2730.756\ x^2 - 263.407\ x^3$

Pour f ∈ [5880 ; 8000 Hz] $P_{dB}(f) = 0$

Le spectre subjectif est obtenu en pondérant les fréquences du spectre brut par l'une des courbes de pondération:

$S_{dB}(n) = E_{dB}(n) + P_{dB}(n)$ où *n* est la fréquence discrète du spectre numérique
S(n) est la valeur en dB à la fréquence *n* du spectre subjectif
E(n) est la valeur en dB à la fréquence *n* du spectre brut
P(n) est le facteur de pondération en dB à la fréquence *n*

L'objection que l'on peut formuler à une telle transformation est d'utiliser des résultats relatifs à des sons purs ou à bandes très étroites pour les appliquer à des sons complexes comme ceux de la parole. Plusieurs études ont permis de valider ce transfert de connaissances. Ainsi, à propos de l'évaluation des seuils différentiels de durée, Rossi affirme que «...nous pourrions utiliser en phonétique la plupart des résultats acquis dans les études de seuils à partir de sons purs.» (Rossi, 1972).

III.1.C. Les bandes critiques

III.1.C.a. La non linéarité de la résolution fréquentielle de l'oreille

Il a été remarqué expérimentalement (Zwicker & Feldkeller, 1981) que l'oreille possède *une bonne résolution spectrale en basses fréquences (BF), mais médiocre en hautes fréquences (HF)*. Cette propriété peut être simulée en effectuant une analyse spectrale par bandes dont la largeur est proportionnelle à la fréquence. En effet, si l'on choisit des bandes étroites en BF, on obtiendra une bonne discrimination fréquentielle. Inversement, le choix de bandes larges en HF nous donne une faible résolution spectrale. Par commodité a été construit un espace fréquentiel adapté à cette situation.

III.1.C.b. Le domaine auditif

Le *Bark* est l'unité fréquentielle de l'espace auditif. Il tient compte de la non-linéarité du phénomène de perception des sons dans le domaine des fréquences. Ainsi, à un écart de fréquences fixe en Bark correspond une différence de fréquences hertziennes faible pour les BF et élevée pour les HF. La correspondance entre le domaine des Hertz et celui des Barks est empirique (Figure 40).

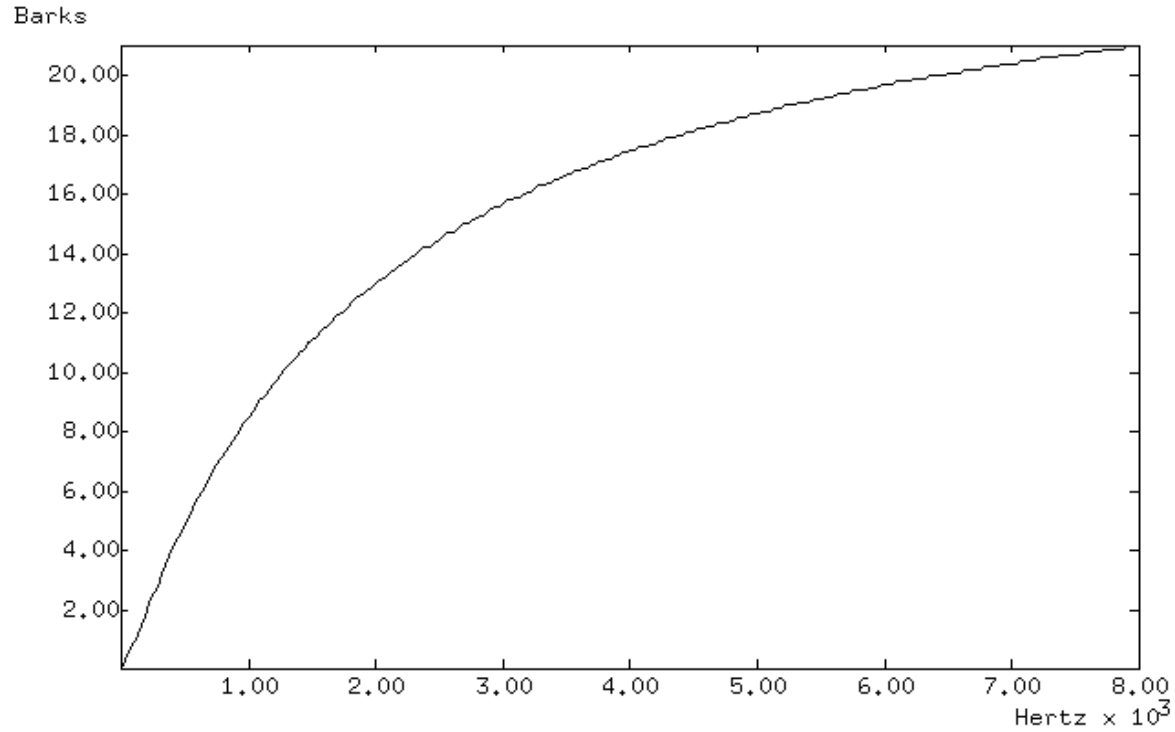


Figure 40: La relation Hertz/Bark

Il existe plusieurs relations analytiques. La plus connue est celle de Zwicker (Zwicker & Therhart, 1980).

$$z\ (\textit{en barks}) = 13 \cdot \arctan(0{,}76 \cdot F) + 3{,}5 \cdot \arctan\left(\frac{F}{7{,}5}\right)^2 \qquad \text{avec } F \text{ en kHz}$$

Il existe aussi une relation logarithmique (Junqua, 1990):

$$z\ (en\ barks) = 6 . \ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right) \quad \text{avec } f \text{ en Hz}$$

Nous avons toutefois préféré celle de Traunmuller plus simple et tout aussi valable (dans Calliope, 1989).

$$z\ (en\ barks) = 26{,}81 . \frac{f}{1960 + f} - 0{,}53 \qquad \text{avec } f \text{ en Hz}$$

Le fait que la courbe de Traunmuller ne passe pas, à tort, par l'origine peut être compensé par une relation linéaire assurant la continuité et le passage de la courbe par l'origine:

$$z\ (en\ barks) = \frac{f}{102{.}44} \qquad \text{pour } f \leq 200 \text{ Hz}$$

La relation inverse donnant les Hz en fonction des Barks est la suivante:

pour $z < 2$ *barks* $\qquad f\ (en\ Hz) = 102{,}44 . z$

pour $z \geq 2$ *barks* $\qquad f\ (en\ Hz) = \frac{1960 . B}{26{,}81 - B} \qquad$ où $B = z + 0{,}53$

Un autre type de transformation fréquemment utilisée est le *mel*. La fonction est approximativement linéaire jusqu'à 1 kHz puis logarithmique au dessus. Il existe différentes expressions analytiques:

$$m\ (en\ mels) = 2595 . \log_{10}\left(1 + \frac{f}{700}\right) \quad \text{avec } f \text{ en Hz} \qquad \text{(O'Shaughnessy, 1990)}$$

$$m\ (en\ mels) = 1000 . \log_2 (1 + f) \quad \text{avec } f \text{ en kHz} \qquad \text{(Furui, 1989)}$$

Nous avons choisi d'utiliser le *Bark* pour sa réalité physiologique, plutôt que le *mel* qui représente une échelle non linéaire arbitraire.

III.1.C.c. L'intégration en bandes critiques

Les bandes critiques sont des zones fréquentielles distantes d'un *Bark*. Elles ont été mesurées expérimentalement par des tests psycho-acoustiques. Les fréquences centrales des bandes sont régulièrement espacées tous les Barks dans le domaine auditif. La conséquence de cette propriété est une répartition non uniforme des bandes dans le domaine linéaire des Hertz (Figure 41 ; Tableau 8). Les bandes sont serrées et étroites en basses fréquences, éloignées et larges en hautes fréquences, à l'image de la perception auditive (cf. Figure 26, p.53). Une autre caractéristique des bandes critiques est de couvrir et intégrer une bande large de fréquences qui se chevauchent entre les bandes, ce qui génère un effet dit « de masque » nettement visible sur la Figure 41.

Tableau 8: Bandes critiques psycho-acoustiques (Source: Hassall et al., 1979)

Bande critique (Bark)	1	2	3	4	5	6	7	8
Fréquence centrale (Hz)	50	150	250	350	450	570	700	840
Largeur de bande (Hz)	100	100	100	100	110	120	140	150
Bande critique (Bark)	9	10	11	12	13	14	15	16
Fréquence centrale (Hz)	1000	1170	1370	1600	1850	2150	2500	2900
Largeur de bande (Hz)	160	190	210	240	280	320	380	450
Bande critique (Bark)	17	18	19	20	21	22	23	24
Fréquence centrale (Hz)	3400	4000	4800	5800	7000	8500	10500	13500
Largeur de bande (Hz)	550	700	900	1100	1300	1800	2500	3500

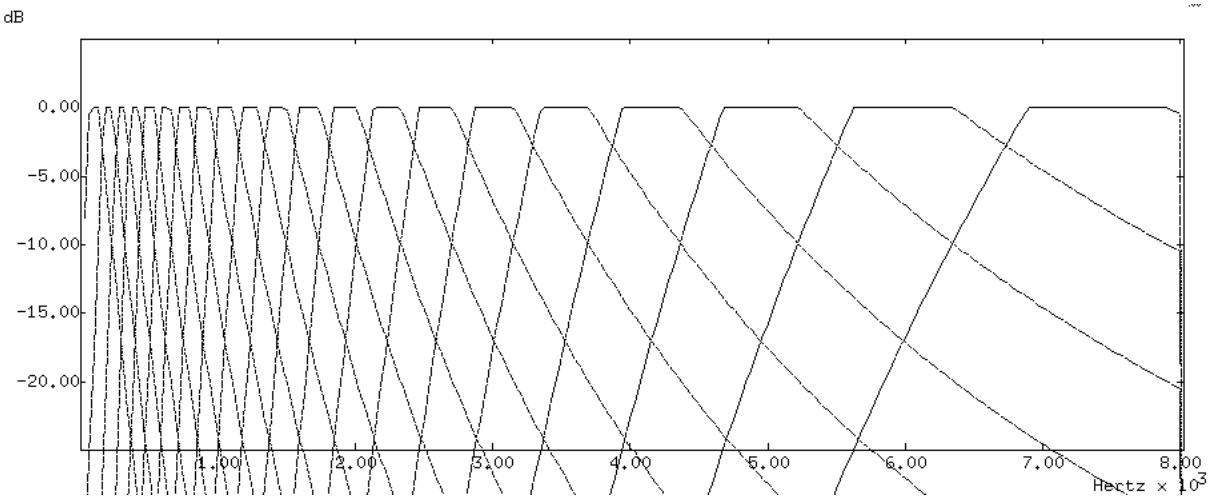
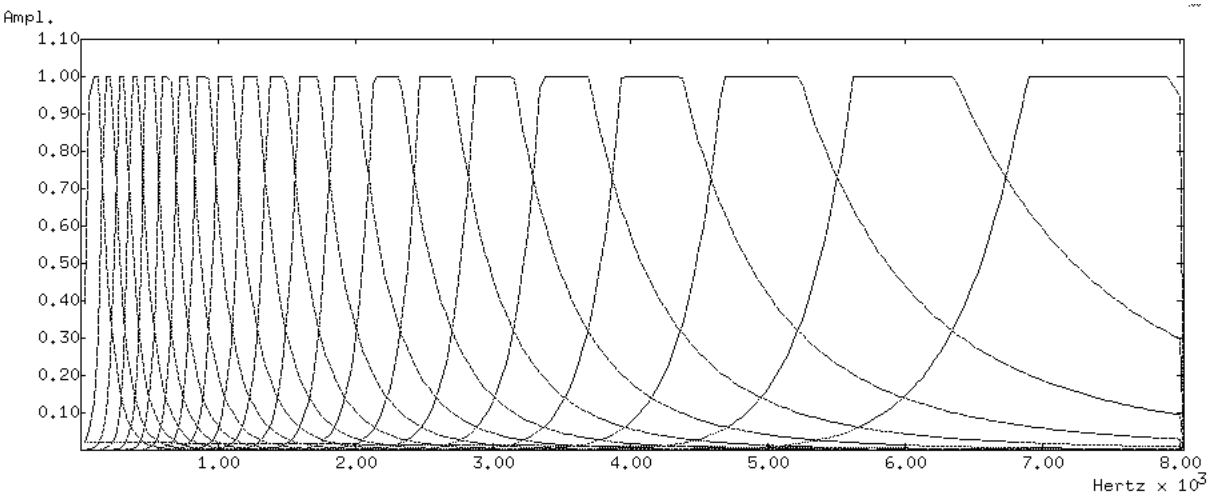


Figure 41a: Bandes critiques en dB



41b: Bandes critiques en linéaire

La production langagière ayant lentement évolué de façon à optimiser l'ensemble sensori-moteur (cf. « Les questions préalables, p.9), il est admis que le système auditif est relativement bien « réglé » pour l'extraction d'information phonétique pertinente. Une preuve évidente est la relative finesse de la discrimination fréquentielle en basses fréquences, autorisant la variété formantique, tout en réduisant la nuance en hautes fréquences, ce qui permet de restreindre les effets de la variabilité non pertinente des bruits aigus comme ceux des fricatives. Dans notre quête de la recherche d'information, nous avons simulé cet effet d'intégration en bandes critiques. L'originalité de cette opération réside dans le choix non aveugle de la place et de la largeur des bandes. Les fréquences centrales des bandes sont régulièrement espacées tous les Barks dans le domaine auditif. Ainsi, la fréquence centrale de la bande *K* est: $zc_k = k - 0.5$ *(en barks)*

Pour simuler l'effet de masque, nous avons construit les bandes comme volontairement asymétriques avec une pente de +25dB/Bark dans la partie ascendante et de -10dB/Bark dans la partie descendante (Hermansky et al., 1985). Ces bandes théoriques à allure triangulaire (Figure 42a) possédant une trop faible bande passante, nous avons opté pour une amplification de leur fonction de transfert d'un gain de 4.5 dB suivie ensuite d'une saturation à la valeur maximale de 0 dB. Le résultat de cette opération sur une bande est illustré en Figure 42b. Nous obtenons ainsi des bandes de largeur correcte par rapport à celles qui sont proposés par (Hassall et al., 1979).

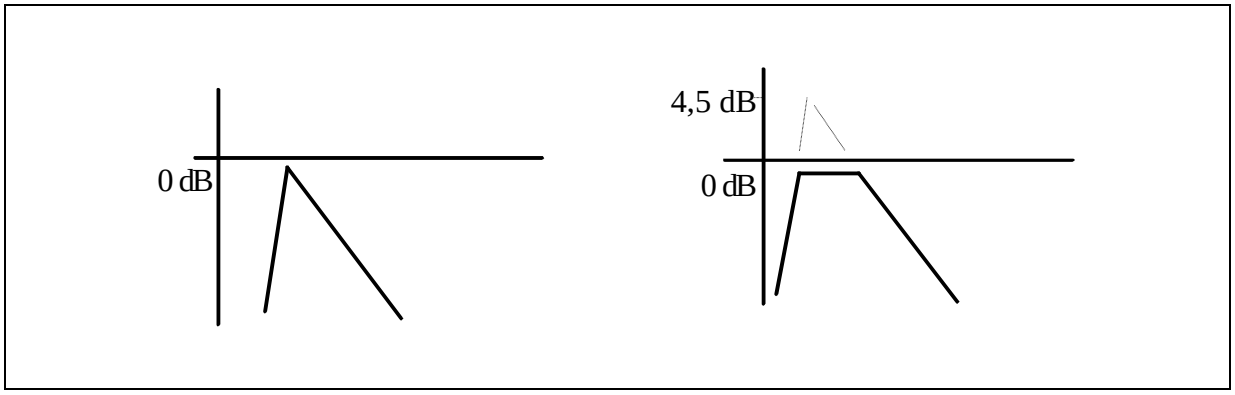


Figure 42a.: Bande triangulaire *42b: Bande élargie*

Il faut bien garder à l'esprit que les choix adoptés sont liés à l'étude expérimentale du système auditif humain et que la modélisation mathématique est une approximation analytique de ces phénomènes physiques. Le Tableau 8, page 84, indique les valeurs des bandes critiques mesurées à partir d'expérimentations sur des sujets humains. En les comparant aux résultats relatifs au modèle (Tableau 9), on se rend compte que la modélisation est relativement fidèle. L'équivalent graphique de ce tableau est la Figure 41, p.84.

Le spectre à bandes critiques est obtenu en intégrant le spectre brut sur chacune des bandes. L'énergie dans la bande *K* est alors:

$E_k = \sum_{i=0}^{N} C_k(i).P(i)$ où $P()$ est le spectre brut

C_k est la fonction de pondération de la bande k

Le résultat de l'analyse par bandes critiques (Figure 43) est finalement un vecteur à *m* composantes énergétiques où *m* est le nombre de bandes critiques qui dépend de la bande passante utilisée (19 canaux pour une bande passante de 5000 Hz, 21 pour 8000 Hz*)*. Ce traitement, inspiré de caractéristiques auditives, peut être utilisée dans le cadre d'une analyse spectrale par vocodeur.

Tableau 9: Bandes critiques du modèle auditif

k **(en barks)**	***zc*$_k$** **(en barks)**	***fc*$_k$** **(en Hz)**	***BP*$_k$** **(en Hz)**	***domaine*** **(en Hz)**	
1	0.5	51	104	0	100
2	1.5	154	100	100	200
3	2.5	250	96	200	300
4	3.5	347	104	300	400
5	4.5	453	115	395	510
6	5.5	569	127	505	635
7	6.5	697	140	630	770
8	7.5	838	155	760	920
9	8.5	995	174	910	1085
10	9.5	1172	195	1075	1270
11	10.5	1370	221	1260	1480
12	11.5	1595	253	1470	1720
13	12.5	1853	291	1710	2000
14	13.5	2152	340	1990	2320
15	14.5	2500	401	2300	2700
16	15.5	2915	481	2675	3155
17	16.5	3413	587	3120	3710
18	17.5	4025	733	3660	4390
19	18.5	4794	941	4320	5265
20	19.5	5790	1252	5160	6415
21	20.5	7131	1748	6260	8000
22	21.5	9033	2638	7715	10350
23	22.5	11941	4368	9760	14125
24	23.5	16942	8633	12625	21260

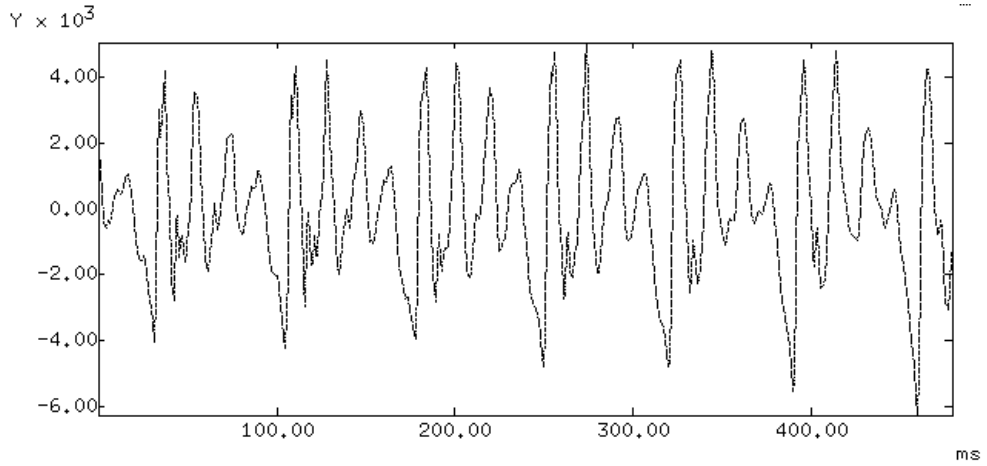


Figure 43a: Signal temporel de /a/ dans "anniversaire", locutrice

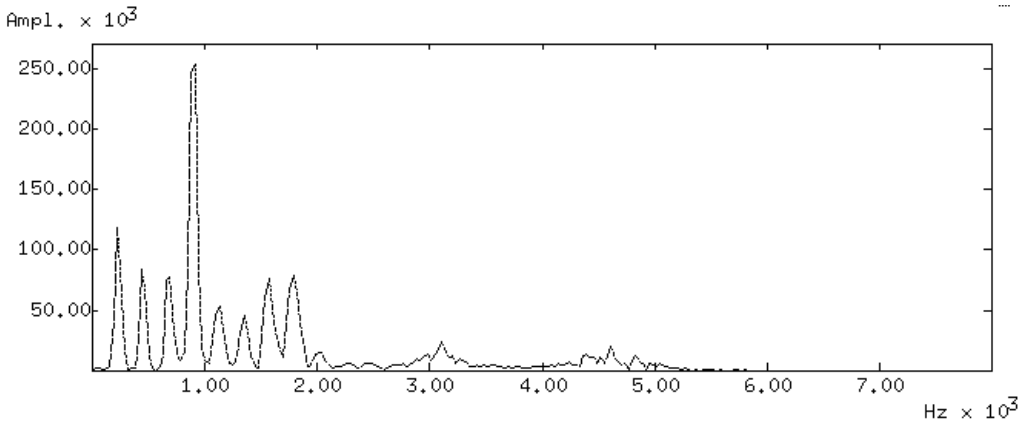


Figure 43b: Spectre du signal

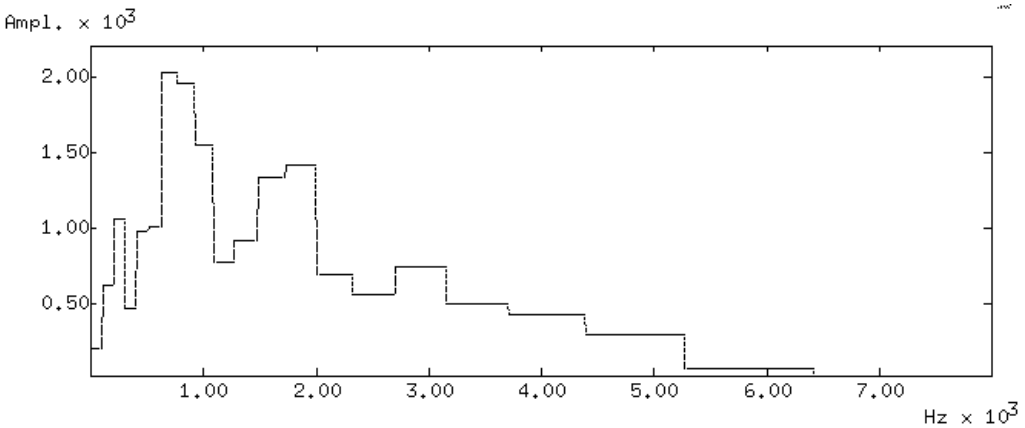


Figure 43c: Analyse en bandes critiques

III.2. L'analyse spectrale par vocodeur

III.2.A. De l'utilité du vocodeur

Un vocodeur est un dispositif qui construit une représentation temps-fréquence du signal de parole. Cette estimation de la structure spectrale s'effectue à l'aide de filtres, en découpant la bande passante en sous-bandes dans laquelle est évaluée l'intensité du signal. Le regroupement en bandes de fréquences entraîne une réduction du débit d'information. Cela permet aussi de s'affranchir, en partie, de la variabilité aléatoire. Les premières réalisations remontent à (Dudley, 1959) avec du matériel analogique. Depuis, d'autres systèmes ont vu le jour et, avec le développement de l'informatique, des dispositifs numériques sont apparus comme, par exemple, au CNET (Génin, 1976). Le jeu de caractéristiques d'un vocodeur sont importantes. Il est nécessaire de faire un choix sur le nombre de bandes, leur répartition, leur bande passante, leur forme, ainsi que sur la courbe de sensibilité de l'ensemble des bandes. Dans le cadre d'un traitement numérique du signal, il existe deux méthodes pour effectuer une analyse par vocodeur:

- réaliser un filtrage temporel par convolution discrète entre le signal et la réponse impulsionnelle de filtres passe-bande de type F.I.R. (Finite Impulse Response). C'est la technique suivie par (Giraud, 1991). Des exemples d'implémentation sont donnés dans (Rabiner, 1979) et (Mac Clellan, 1979).
- calculer un spectre de Fourier et sommer les raies du module suivant diverses bandes de fréquences.

Nous avons choisi la seconde solution car elle semble mieux adaptée pour simuler la forme compliquée des bandes critiques (Figure 41, p.84).

III.2.B. La Transformée de Fourier

III.2.B.a. Définitions et notations

La Transformée de Fourier est une transformation mathématique qui consiste à projeter une fonction dans un espace, celui des fréquences, dont la base est constituée de fonctions orthogonales, en l'occurrence de sinusoïdes. Autrement dit, elle permet de changer de représentation et d'exprimer une fonction de forme compliquée (un signal quelconque) en une combinaison linéaire de fonctions élémentaires de forme simple (des ondes sinusoïdales).

Il est possible de calculer une Transformée de Fourier pour un signal analogique *x(t)* (temps continu) ou pour un signal numérique *x(n)* (temps discret).

Traditionnellement, on note la Transformation de Fourier de la façon suivante :

- pour un signal analogique (*t* en secondes)

(Eq.2) $$x(t) \xrightarrow{T.F.} X(f) = \int_{-\infty}^{+\infty} x(t).e^{-2j\pi ft}\,dt$$ où *f* est la fréquence en Hz

- pour un signal numérique* ,

(Eq.3) $$x(n) \xrightarrow{T.F.} X(\nu) = \sum_{n=-\infty}^{+\infty} x(n).e^{-2j\pi\nu n}$$ où ν est la fréquence réduite**

X(ν), qui est le spectre du signal x, est une fonction complexe qui peut s'exprimer en fonction de sa partie réelle et imaginaire :

$$X(\nu) = \mathrm{Re}[X(\nu)] + j\ \mathrm{Im}[X(\nu)]$$

Dans le cas d'un signal réel x, ce qui est le cas de la parole, les parties réelle et imaginaire sont respectivement données par :

(Eq.4) $$\mathrm{Re}[X(\nu)] = \sum_{n=-\infty}^{\infty} x(n).\cos(2\pi\nu n) \quad \text{et} \quad \mathrm{Im}[X(\nu)] = -\sum_{n=-\infty}^{\infty} x(n).\sin(2\pi\nu n)$$

On peut aussi exprimer le spectre sous forme de module et argument :

$$X(\nu) = |X(\nu)|\ e^{j.\arg[X(\nu)]}$$

où $|X(\nu)| = \sqrt{(\mathrm{Re}[X(\nu)])^2 + (\mathrm{Im}[X(\nu)])^2}$ est appelé spectre d'amplitude

$\arg[X(\nu)] = \arctan\left(\frac{\mathrm{Im}[X(\nu)]}{\mathrm{Re}[X(\nu)]}\right)$ est appelé spectre de phase

Il faut signaler que le terme $|X(\nu)|^2$ est appelé spectre d'énergie.

III.2.B.b. Quelques propriétés

- La périodicité

D'après (Eq.3), on peut écrire $X(\nu+1) = \sum_{n=-\infty}^{+\infty} x(n).e^{-2j\pi(\nu+1)n} = \sum_{n=-\infty}^{+\infty} x(n).e^{-2j\pi\nu n}.e^{-2j\pi n}$ ($X(\nu)$; $e^{-2j\pi n} = 1$)

D'où l'égalité $X(\nu+1) = X(\nu)$, ce qui signifie que la Transformée de Fourier(TF) d'un signal numérique est une fonction périodique en ν de période 1*** . Cette propriété de périodicité autorise donc la restriction de l'étude de $X(\nu)$ à un intervalle unité. On utilise généralement l'intervalle [-½ ; ½], appelé intervalle principal.

---

* Le remplacement de la variable continue *t* par une variable discrète *n* peut s'écrire de la façon suivante : $t = n.\Delta t$ où Δt est la période d'échantillonnage ($\Delta t = 1/F_{ech}$). Pour décrire le signal x, nous utiliserons par la suite la notation simplifié $x(n) = x(n\,\Delta t)$

** La comparaison des notations de (Eq.2) et (Eq.3) permet d'établir une relation entre *f*(en Hz), *t*(en s.), *n*(sans unité) et ν. Cette relation est la suivante : $f.t = \nu.n$. D'après les indications de la note précédente qui précise que $t = n.\Delta t \Leftrightarrow t = n / F_{ech} \Leftrightarrow n = t.F_{ech}$, on obtient finalement $f.t = \nu.t.F_{ech}$ d'où la relation :
$\nu = \frac{f}{F_{ech}}$ qui correspond à une grandeur sans unité d'où son nom de fréquence réduite.

*** Si on reprend l'égalité $\nu = f / F_{ech}$, on montre ainsi que la TF est une fonction périodique en *f* de période F_{ech} .On démontre là les résultats présentés au § « Les effets de l'échantillonnage », p.58

- La symétrie

Les relations (Eq.4) concernant un signal réel (cas de la parole) permettent de mettre en évidence le fait que :

$Re[X(-\nu)] = Re[X(\nu)]$ fonction paire ⇔ symétrie par rapport à l'axe $\nu = 0$

$Im[X(-\nu)] = - Im[X(\nu)]$ fonction impaire ⇔ symétrie par rapport à l'origine

Ces propriétés de symétrie autorisent donc la restriction de l'étude de X(ν) à un intervalle où ν est positive. On utilise finalement l'intervalle [0 ; ½]* .

III.2.B.c. Problèmes pratiques liés au calcul de la Transformée de Fourier

Il existe deux difficultés pratiques associées au calcul de la relation (Eq.3). La première est due à la nécessité d'un nombre infini d'échantillons du signal $\underline{x}$, phénomène qu'il est impossible de traiter concrètement. Le deuxième obstacle est relatif à l'aspect continu de la fréquence ν. Pour obtenir une représentation spectrale complète, il est nécessaire de calculer la valeur $X(\nu)$ en n'importe quel point $\nu \in [0 ; ½]$. Cette grandeur étant continue, cette opération est impossible à effectuer dans un système de traitement numérique.

La solution à ces problèmes peut paraître simple: il suffit de limiter la durée du signal et de discrétiser la fréquence. La mise en oeuvre et l'interprétation des résultats après ces choix nécessitent une étude minutieuse et détaillée afin d'éviter une utilisation erronée.

III.2.C. La Transformée de Fourier à Court Terme

III.2.C.a. Principe et notations

Nous avons vu que le calcul de la TF d'un signal $\underline{x}$ à durée illimitée est impossible d'un point de vue pratique. Une manière brutale de surmonter cet obstacle est de limiter temporellement le signal en le tronquant sur un intervalle de L échantillons centrés autour de l'échantillon $x(n_0)$. Cet intervalle est appelé *fenêtre temporelle d'analyse.* Nous la notons $\underline{w}$. Il s'agit d'une fonction de pondération dont les valeurs $w(m)$ sont nulles en dehors d'un intervalle $[-L/2 ; L/2]$. La version à durée limitée du signal $\underline{x}$ est notée $\underline{x}\big|_{n_0}^{L}$. Il s'obtient par le calcul suivant:

(Eq.5) $\underline{x}\big|_{n_0}^{L}(n) = x(n).w(n- n_0)$ où $w(m) = 0$ pour $|m| > L/2$

Dans ce cas là, l'équation (Eq.3) devient $X\big|_{n_0}^{L}(\nu) = \sum_{n=-\infty}^{+\infty} x(n).w(n-n_0).e^{-2j\pi\nu n}$

Sachant que $w(n- n_0)$ est nul pour $|n- n_0| > L/2$, on obtient finalement une TF qui se calcule sur un nombre fini de L échantillons. C'est pourquoi nous appellerons cette transformation la Transformée de Fourier à Court Terme notée :

* Si on reprend l'égalité $\nu = f / F_{ech}$, on montre ainsi pourquoi les études spectrales sur la parole, qui est un signal réel, s'effectue sur l'intervalle f ∈ [0 ; F_{ech} / 2].

(Eq.6) $$X|_{n_0}^{L}(\nu) = \sum_{n=n_0-L/2}^{n_0+L/2} x(n).w(n-n_0).e^{-2j\pi\nu n}$$

Notre objectif est atteint. Il reste à présent à étudier la relation qu'il existe entre le spectre obtenu $X|_{n_0}^{L}(\nu)$ et le spectre recherché $X(\nu)$. Autrement dit, étudions les répercutions spectrales du fenêtrage.

III.2.C.b. Répercutions spectrales dues au fenêtrage du signal

Nous savons que $TF\{w(n-n_0)\} = TF\{w(n)\}.e^{-2j\pi\nu n_0}$ et qu'une multiplication dans le domaine temporel entraîne une convolution dans le domaine spectral (Kunt, 1980). La correspondance spectrale de la relation (Eq.5) est donc :

$$X|_{n_0}^{L}(\nu) = X(\nu) * \left[W(\nu).e^{-2j\pi\nu n_0}\right] \quad \Leftrightarrow$$

$$X|_{n_0}^{L}(\nu) = \int_{-\infty}^{+\infty} W(\phi).e^{-2j\pi\phi n_0} X(\nu-\phi).d\phi$$

où
- ν est la fréquence réduite
- $*$ est le signe de convolution
- $X|_{n_0}^{L}$ est la TF du signal $\underline{x}|_{n_0}^{L}$
- $X(\nu)$ est la TF du signal $\underline{x}$
- $W(\nu)$ est la TF du signal $\underline{w}$

Pour simplifier les équations, plaçons nous en $n_0 = 0$. Nous occultons ainsi les phénomènes de phase sans perturber le raisonnement axé essentiellement sur des problèmes d'amplitude. La dernière relation devient alors :

$$X|_{n_0}^{L}(\nu) = \int_{-\infty}^{+\infty} W(\phi).X(\nu-\phi).d\phi$$

Cette relation montre que les valeurs $X|_{n_0}^{L}(\nu)$ représentent une version « lissée » et approximative de $X(\nu)$. La justesse de l'estimation est dépendante de la fonction $W(\nu)$, appelée fenêtre spectrale. $X|_{n_0}^{L}(\nu)$ reproduira d'autant plus fidèlement les propriétés de $X(\nu)$ que $W(\nu)$ sera plus proche d'une impulsion de Dirac. En pratique, $\underline{w}$ devra être un filtre passe-bas à bande étroite.

Prenons l'exemple d'une (co)sinusoïde $\underline{x}$ tronquée par une fenêtre rectangulaire $\underline{w}$ (Figure 44).

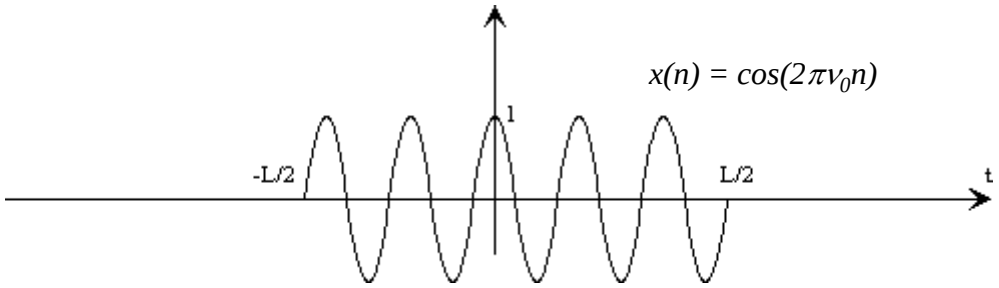


Figure 44: Sinusoïde fenêtrée

(Harris, 1976) fournit les relations suivantes :

signal	amplitude de la TF (sur l'intervalle principal)
$x(n) = cos(2\pi\nu_0 n)$	$X(\nu) = [\delta(\nu - \nu_0) + \delta(\nu + \nu_0)] / 2$ *
$w(n) = 1$ pour $\lvert n\rvert \leq L/2$ 0 ailleurs	$W(\nu) = \frac{\sin(\pi\nu L)}{\sin(\pi\nu)}$

La fonction $\frac{\sin(\pi\nu L)}{\sin(\pi\nu)}$ est connue sous le nom de fonction de Dirichlet. Il faut remarquer que la valeur de cette expression en $\nu = 0$ peut s'exprimer grâce aux séries de Taylor qui montrent que $\frac{\sin(Nx)}{\sin(x)}$ converge vers N lorsque x tend vers 0. La fonction de Dirichlet vaut donc L pour $\nu = 0$. Nous l'avons représentée autour de l'intervalle principal ($\nu \in$ [-½ ; ½]), ceci pour différentes valeurs de L (Figure 45).

Le spectre résultant de la sinusoïde fenêtrée, qui est issu de la convolution des TF de la sinusoïde infini par celle de la fonction rectangle, est finalement :

$$X|_0^L(\nu) = \frac{1}{2}.\left(\frac{\sin\left(\pi(\nu-\nu_0)L\right)}{\sin\left(\pi(\nu-\nu_0)\right)} + \frac{\sin\left(\pi(\nu+\nu_0)L\right)}{\sin\left(\pi(\nu+\nu_0)\right)}\right)$$

Une représentation graphique de cette fonction est fournie en Figure 46, p.94. L'exemple montre que les valeurs $X|_{n_0}^L(\nu)$ représentent une version approximative de $X(\nu)$. On peut remarquer que l'estimation est d'autant plus juste que la durée d'observation est longue.

* le signe δ représente la fonction Dirac. δ(x) = 1 pour x = 0
0 partout ailleurs

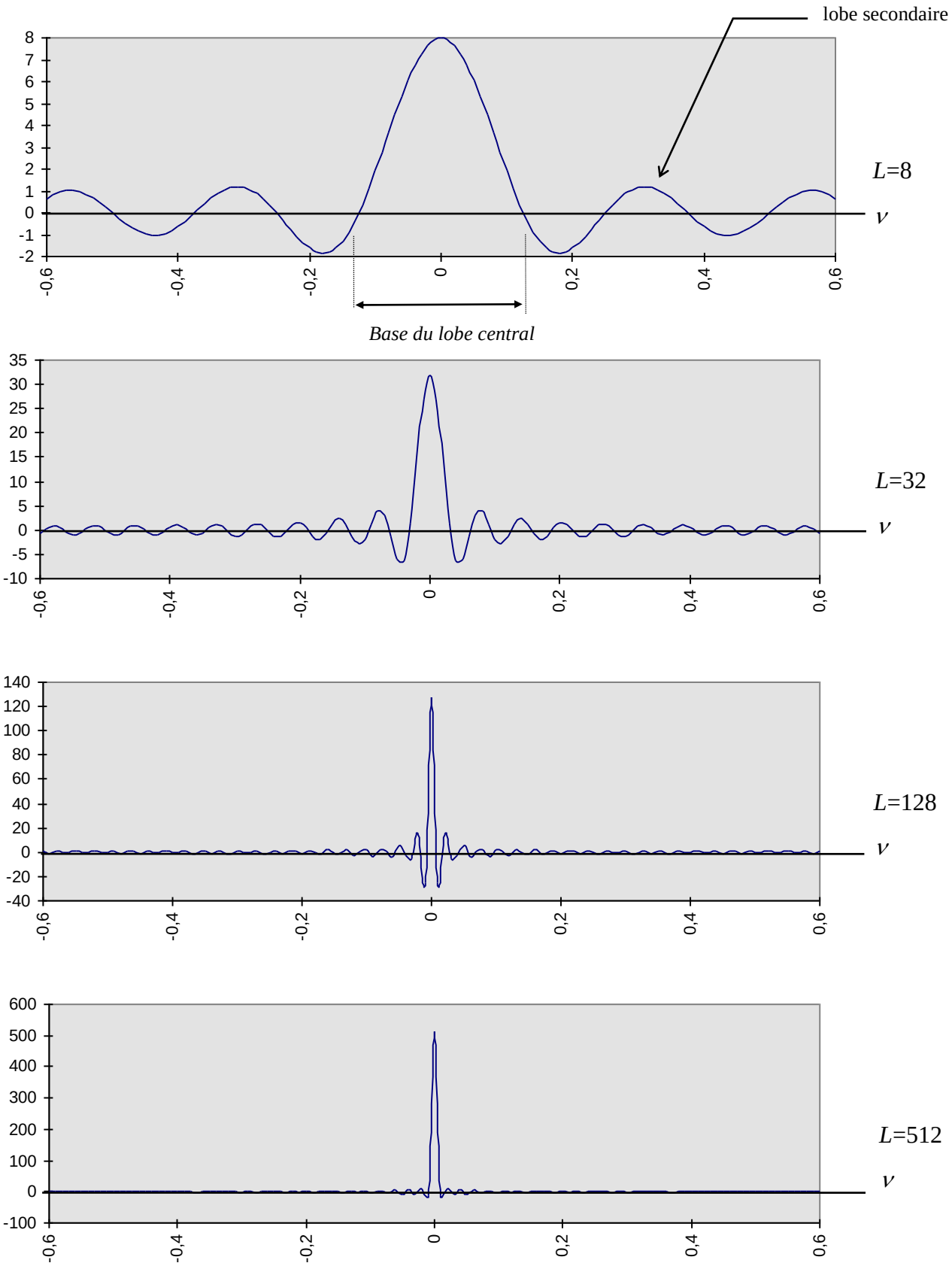


Figure 45: Fonction de Dirichlet pour différentes valeurs de L (abscisses en fréquence réduite, ordonnées en amplitude linéaire)

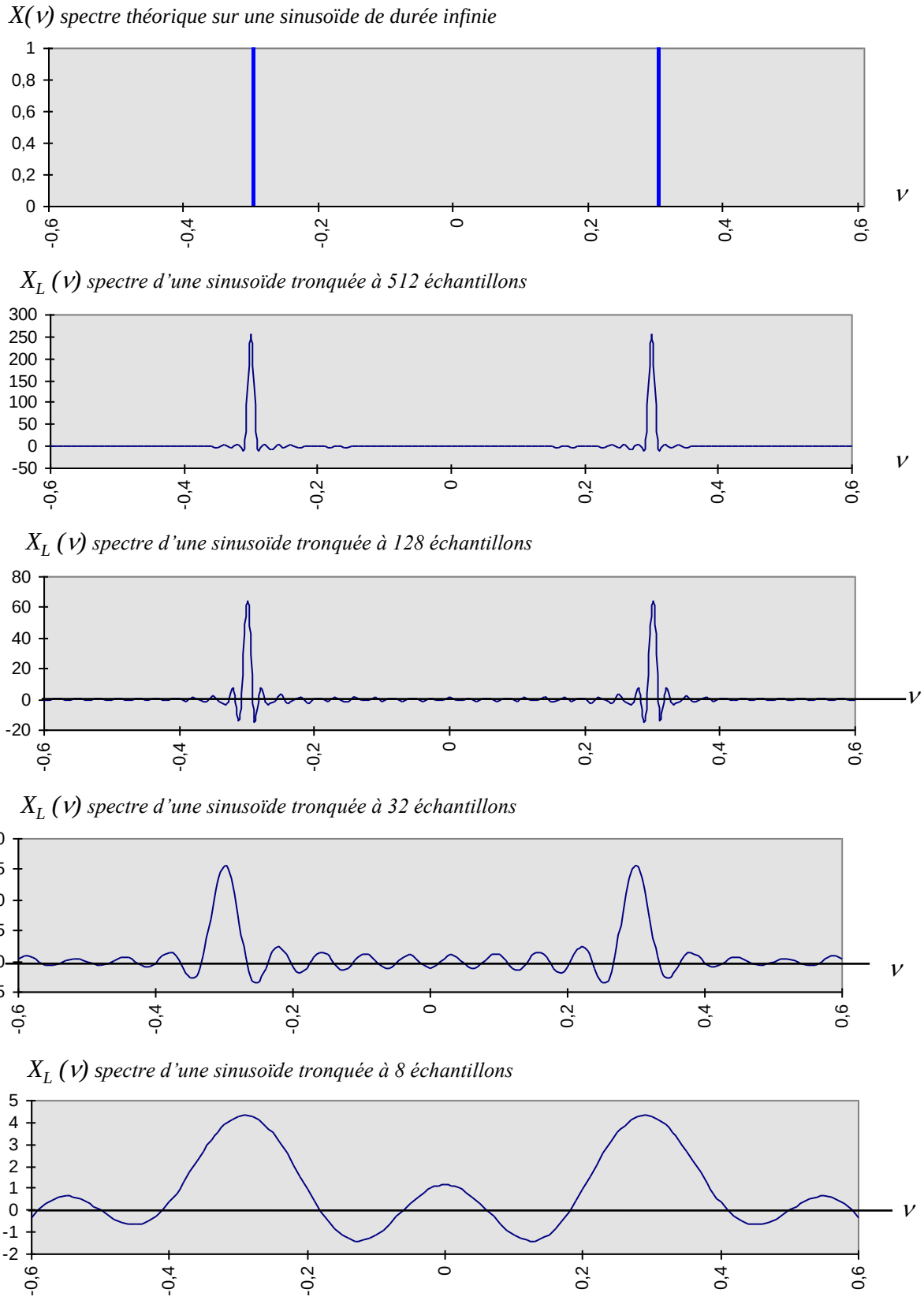


Figure 46: Résultats de l'analyse spectrale d'une sinusoïde (ν_0= 0.3) en fonction de la longueur de la fenêtre d'observation (abscisses en fréquence réduite, ordonnées en amplitude linéaire)

III.2.C.c. Longueur de fenêtre et résolution spectrale

Nous appelons résolution spectrale la capacité d'une analyse à distinguer deux fréquences distinctes. Reprenons l'exemple de la sinusoïde tronquée. Soient deux sinusoïdes mélangées x_1 et x_2 d'amplitudes et de phase égales mais de fréquences différentes ν_1= 0.3 et ν_2= 0.33. L'analyse spectrale est effectuée sur ce signal tronqué à L échantillons. Le résultat de cette observation est présenté à la Figure 47. Contrairement à la figure précédente, nous n'avons représenté que la partie où $\nu > 0$, la partie gauche étant symétrique (cf. § « Quelques propriétés », p.89). On peut remarquer qu'il existe une longueur critique L_c de la fenêtre d'observation au dessous de laquelle les deux sinusoïdes ne sont plus séparées. Graphiquement, il semble que $L_c \approx 60$. La fusion des informations fréquentielles est due à B, la largeur de la base du lobe central* de la fenêtre spectrale $W(\nu)$. Pour que l'analyse spectrale puisse distinguer deux fréquences séparées de $\Delta\nu$, il faut que :

(Eq.7) $B \leq \Delta\nu$

Or, la demi-largeur de ce lobe central est égale à la première valeur de ν pour laquelle $W(\nu)$ s'annule. Dans le cas de la fenêtre rectangle, $W(\nu) = \frac{\sin(\pi\nu L)}{\sin(\pi\nu)}$. La première valeur de ν où $W(\nu) = 0$ correspond à l'équation :

$\sin(\pi\nu L) = 0 \quad \Rightarrow \pi\nu L = \pi \quad \Rightarrow \nu = 1/L$ qui représente la demi-largeur.

La largeur complète de la base de la fenêtre spectrale est donc :

(Eq.8) $B = 2/L$.

La condition (Eq.7) devient :

$B \leq \Delta\nu \quad \Leftrightarrow \quad 2/L \leq \Delta\nu \quad \Leftrightarrow \quad L \geq 2 / \Delta\nu$

Dans notre exemple, $\Delta\nu = \nu_2 - \nu_1 = 0.03$. Nous en déduisons $L_c = 2 / 0.03 \approx 67$ points. Ce seuil critique est nettement visible sur la Figure 47.

Cette condition est nécessaire mais pas suffisante. En effet, un deuxième facteur peut venir brouiller l'estimation spectrale. Il s'agit des oscillations des lobes secondaires (Figure 45, p.93). Reprenons l'exemple des deux sinusoïdes x_1 et x_2 de fréquences différentes ν_1= 0.3 et ν_2= 0.33, mais cette fois d'amplitudes différentes avec $A_2 = A_1 / 5$. Les résultats de l'analyse spectrale d'un tel signal sont présentés à la Figure 48, p.97. On voit que, même pour une longueur de fenêtre L=70, c'est à dire au dessus du seuil critique calculé ci-avant, la contribution spectrale de la sinusoïde x_2 est quasiment confondue avec les lobes secondaires de la contribution spectrale de la sinusoïde x_1 . Elle apparaît donc difficilement détectable.

* ce terme est défini dans la Figure 45, p.93

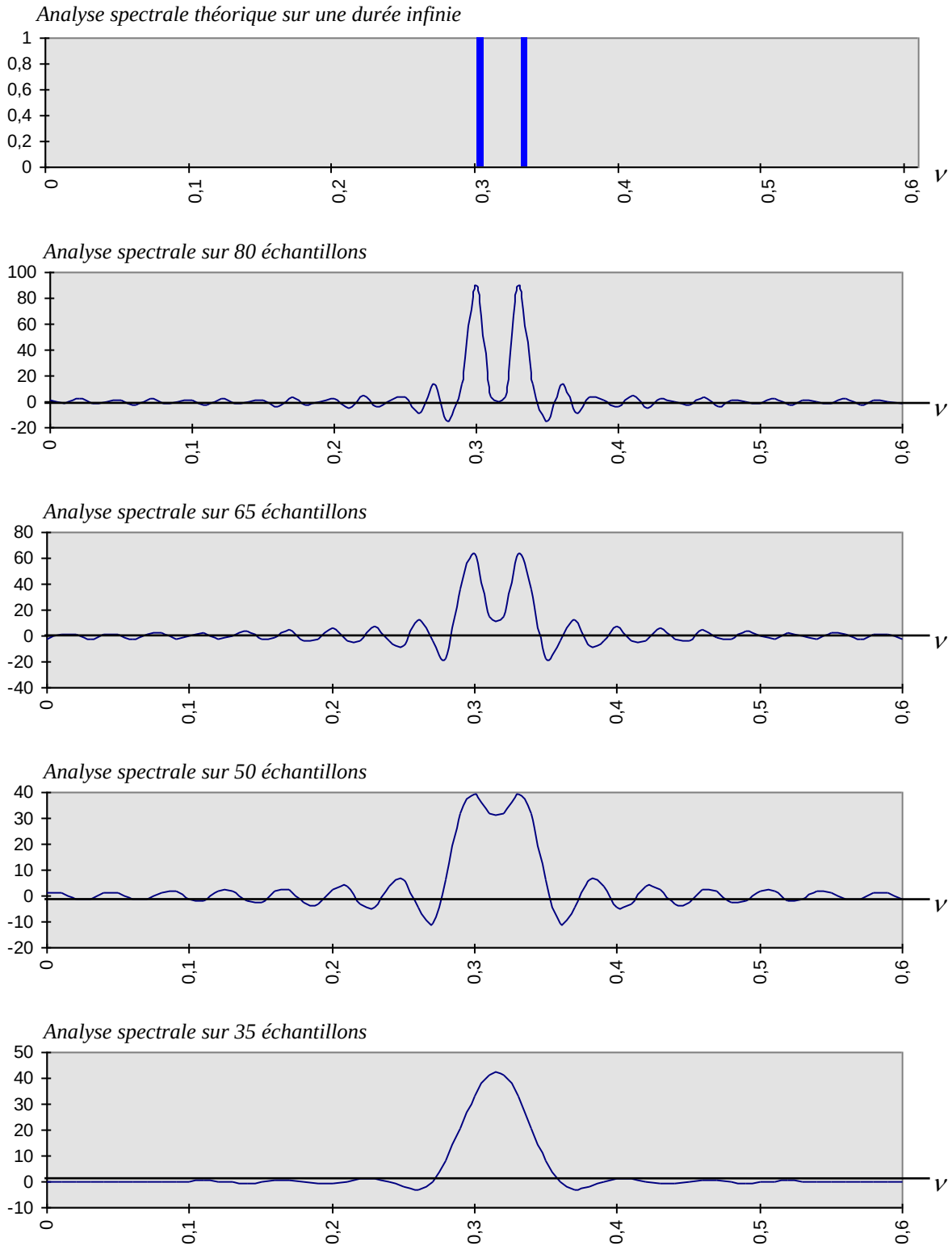


Figure 47: Analyse spectrale de deux sinusoïdes en fonction de fenêtres d'observation de longueurs différentes (abscisses en fréquence réduite, ordonnées en amplitude linéaire). On remarque la dégradation de la résolution spectrale avec les fenêtres de courte durée.

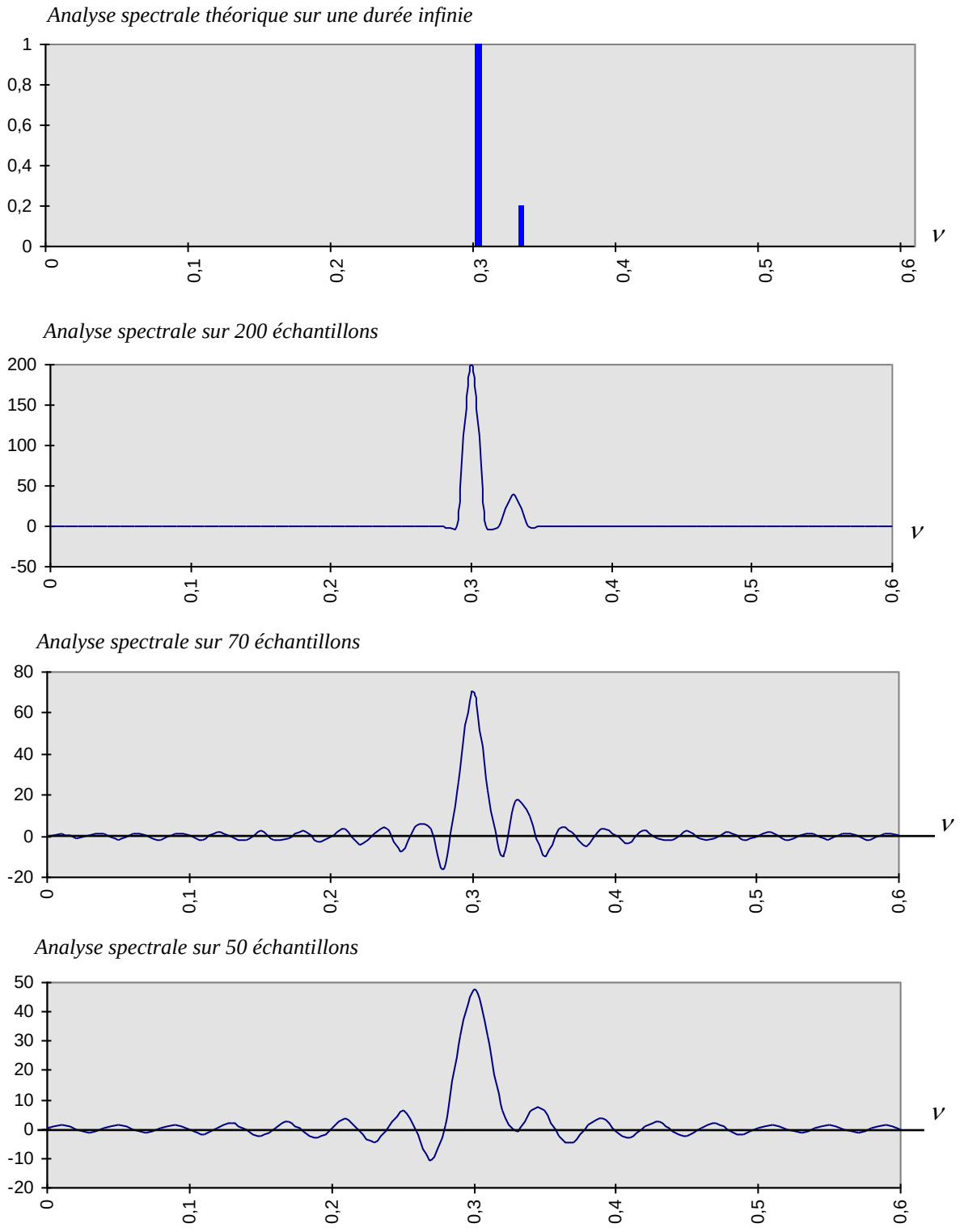


Figure 48: Résultat spectral de deux sinusoïdes d'amplitudes différentes analysées avec des longueurs de fenêtres d'observation différentes (abscisses en fréquence réduite, ordonnées en amplitude linéaire). On met en évidence la perturbation qu'apportent les lobes secondaires.

III.2.C.d. Choix de la durée d'observation

Se pose maintenant le problème du choix de la durée d'observation. La parole étant loin d'être un signal stationnaire, deux conditions essentielles sont à respecter:

- la durée d'analyse doit être suffisamment longue pour obtenir une estimation spectrale correcte. Nous avons vu précédemment les conséquences désastreuses d'une fenêtre d'observation trop courte.
- la durée d'analyse ne doit pas être trop longue, sous peine de réaliser des transformées de Fourier sur des signaux non stationnaires n'ayant aucune homogénéité (le spectre obtenu est alors peu pertinent).

Dans le domaine de la parole, des durées de *5 à 100 ms* sont généralement utilisées. Dans le cas d'un échantillonnage à 16000 Hz, qui est une valeur classique, une fenêtre de

5 ms	inclut	80	échantillons,	ce qui permet une résolution* optimale de	400 Hz
10 ms	"	160	"	"	200 Hz
20 ms	"	320	"	"	100 Hz
30 ms	"	480	"	"	67 Hz
40 ms	"	640	"	"	50 Hz
50 ms	"	800	"	"	40 Hz
75 ms	"	1200	"	"	27 Hz
100 ms	"	1600	"	"	20 Hz

La durée devra être choisie selon les besoins de l'analyse. Une étude portant sur la fréquence fondamentale nécessite une fenêtre d'au moins 30 ms ($\Leftrightarrow \Delta f = 67Hz$), ce qui rend possible la distinction des harmoniques de f0. On parle alors d'analyse en bandes étroites. Par contre, si l'observation doit se focaliser sur des changements rapides dans le temps, c'est à dire qu'une fenêtre de 10 ms est utilisée, la résolution ne sera plus que de 200 Hz. On parle ainsi d'analyse en bandes larges. Une telle analyse met alors plus en évidence l'aspect formantique que l'aspect harmonique de la parole. Dans notre travail, nous avons généralement choisi des fenêtres de 20 ms, ce qui permet un bon compromis entre précision temporelle et résolution spectrale.

III.2.C.e. Rôles et propriétés des fenêtres non rectangulaires

III.2.C.e.i. Position du problème

Rappelons tout d'abord que le fenêtrage est imposé par le fait qu'on ne puisse faire des calculs numériques que sur des signaux à durée limitée, et que limiter à L le nombre d'échantillons du signal revient à multiplier le signal par une fenêtre rectangulaire. Or, cette multiplication dans le domaine temporel entraîne une convolution au niveau spectral qui se traduit par une distorsion du spectre du signal (Figure 46, p.94). L'origine de ces déformations spectrales réside essentiellement dans des problèmes de discontinuité en bordure des fenêtres d'analyse (effets de tranchant). Le but des fenêtres d'analyse non rectangulaires est de réduire l'effet de discontinuité. Pour cela, il faut que l'ordre de dérivabilité de ces fenêtres soit le plus important possible sur les bords, ce qui est réalisé en rendant ces dérivées nulles ou proches de

* avec une fenêtre rectangulaire, $\Delta\nu = 2/L$. Or , $\Delta\nu = \Delta f(en\ Hz) / Fech(en\ Hz) \Rightarrow \Delta f(en\ Hz) = 2.Fech(en\ Hz)/L$

zéro. Cela a pour effet de faire tendre lentement le signal fenêtré vers 0 sur les bords de la fenêtre. Dans le domaine temporel, les fenêtres d'analyse sont caractérisées par leurs *L* coefficients non nuls appelés pondérations de fenêtre. Dans le domaine spectral, la forme idéale d'une fenêtre d'analyse est celle d'une impulsion de Dirac (cf. § « Répercutions spectrales dues au fenêtrage du signal », p.91). La forme réelle est celle d'une filtre passe bas à bande étroite. Tout comme dans le cas de la fenêtre rectangulaire (Figure 45, p.93), la partie passante de ce filtre est appelée *lobe principal* de la fenêtre. Le spectre présente des ondulations (*lobes secondaires)* qui doivent être le plus bas possible

III.2.C.e.ii. Les caractéristiques des fenêtres d'analyse

Les caractéristiques spectrales d'une fenêtre d'analyse sont les suivantes:

- la bande-passante du lobe principal
- la hauteur des lobes secondaires, donnée en dB par rapport au lobe principal.

A longueur de fenêtre *L* fixé, le théorème de Parceval (conservation de l'énergie de la fenêtre) met en évidence *un compromis entre la largeur du lobe principal et la hauteur des lobes secondaires de la fenêtre*, c'est-à-dire que si on diminue la largeur du lobe principal, on augmente la hauteur des lobes secondaires et inversement (Kunt, 1980). L'autre remarque importante que l'on puisse faire sur ces caractéristiques est que cette largeur est inversement proportionnelle à la longueur *L* de la fenêtre. Ceci est lié au fait qu'*une extension dans le domaine temporel correspond à une compression dans le domaine fréquentiel*, et vice versa. De nombreuses fenêtres d'analyse ont été proposées (Kunt, 1980). Nous présentons celles qui sont les plus utilisées.

III.2.C.e.iii. Les fenêtres classiques

L'expression analytique des fenêtres suivantes est valable pour $|n| \le \frac{L}{2}$ (Figure 49; Source: Kunt, 1980, pp. 119-120). Partout ailleurs, la pondération est nulle.

- **Fenêtre rectangulaire** *w (n) = 1*

C'est la fenêtre qui présente la meilleure résolution spectrale, mais aussi les lobes secondaires les plus élevés.

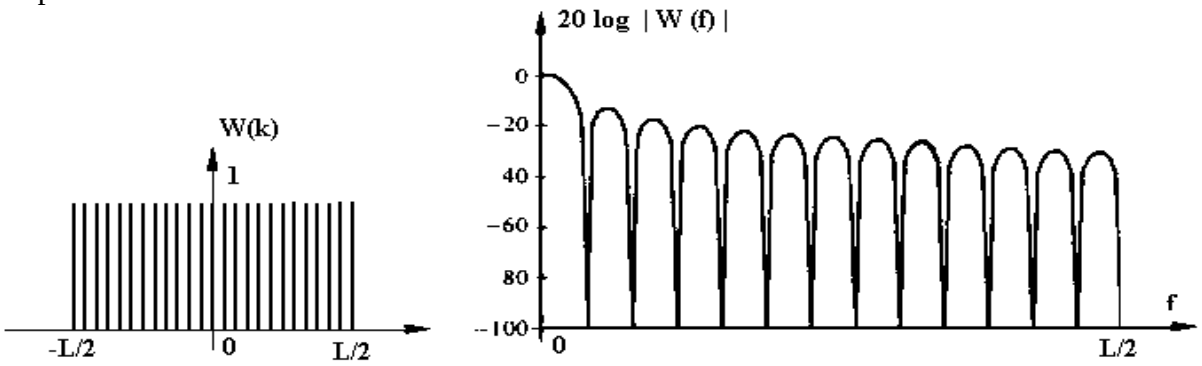


Figure 49a: Fenêtre rectangulaire

- **Fenêtres de Blackman** $w(n) = a_0 + 2.\sum_{l=1}^{p} a_1 . \cos\left(\frac{2\pi nl}{L}\right)$ avec $a_0 + 2.\sum_{l=1}^{p} a_1 = 1$

Elles sont appelées, dans certains cas, Blackman-Harris d'ordre p. On a ainsi:

	Blackman	Blackman-Harris 2	Blackman-Harris 3
p	2	2	3
a_0	0.42	0.42323	0.35875
a_1	0.25	0.49755	0.48829
a_2	0.04	0.07922	0.14128
a_3	0	0	0.01168

Ces fenêtres offrent des lobes secondaires extrêmement bas: -67 dB pour la BH2, -92 dB pour la BH3, ceci étant bien sûr obtenu au détriment de la résolution.

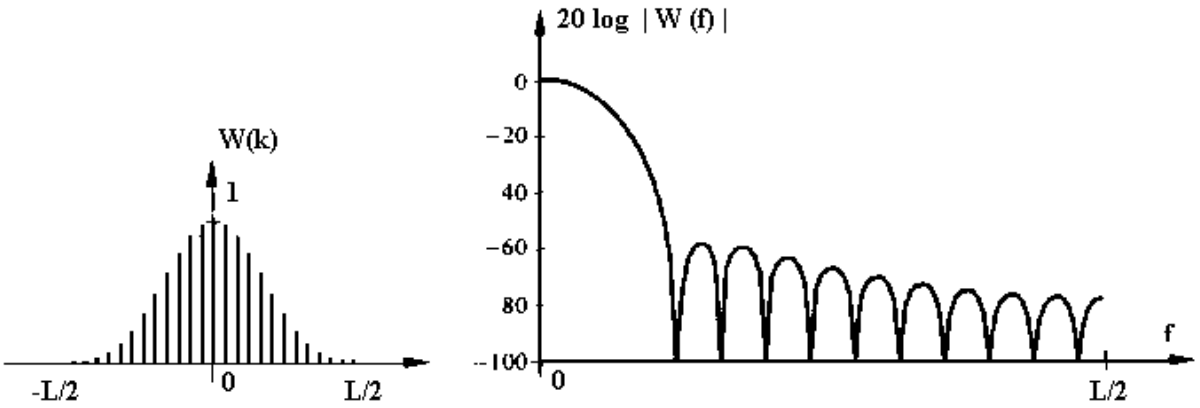


Figure 49b: Fenêtre de Blackman

- **Fenêtres de Hanning / Hamming** $w(n) = \alpha + (1-\alpha).\cos\left(\frac{2\pi n}{L}\right)$

Pour $\alpha = 0.5$, on obtient la fenêtre de Hanning dont les pondérations aux bords sont nulles. Pour $\alpha = 0.54$, on obtient la fenêtre de Hamming qui possède un piédestal. Elle présente un bon compromis entre la largeur du lobe principal et la hauteur des lobes secondaires. Elle a en outre la particularité d'avoir le premier lobe secondaire très fortement atténué.

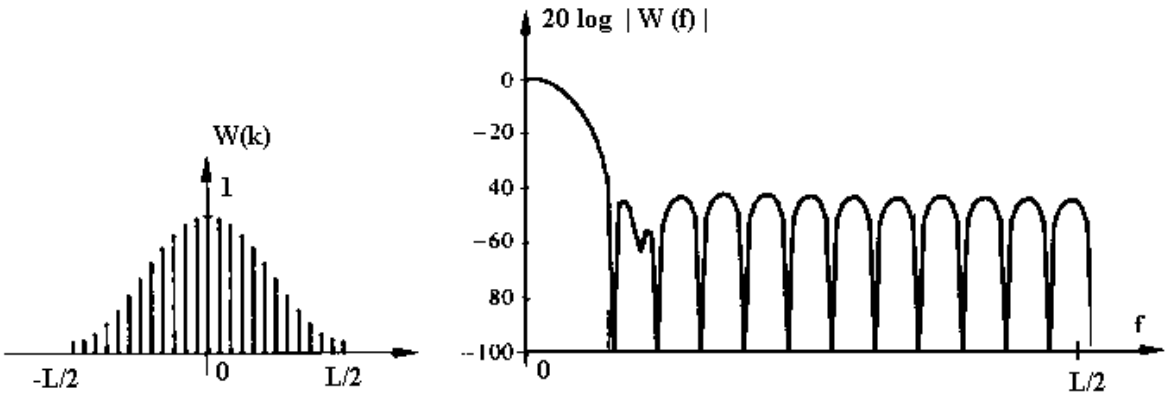


Figure 49c: Fenêtre de Hamming

- **Fenêtres de Kaiser-Bessel** $$w(n) = \frac{I_0\left[\beta L.\sqrt{1-\left(\frac{n}{L/2}\right)^2}\right]}{I_0[\beta L]}$$

I_0 représente la fonction de Bessel modifiée de première espèce d'ordre zéro et β est le paramètre caractérisant l'échange d'énergie entre le lobe central et les lobes secondaires. Pour de meilleures performances (Kunt, 1980), les valeurs du produit βL doivent être choisies dans l'intervalle [4 ; 9]. Le critère de construction de cette fenêtre est de maximiser l'énergie à l'intérieur de la bande de fréquences utile. L'avantage est de pouvoir régler à volonté la largeur du pic central en augmentant β.

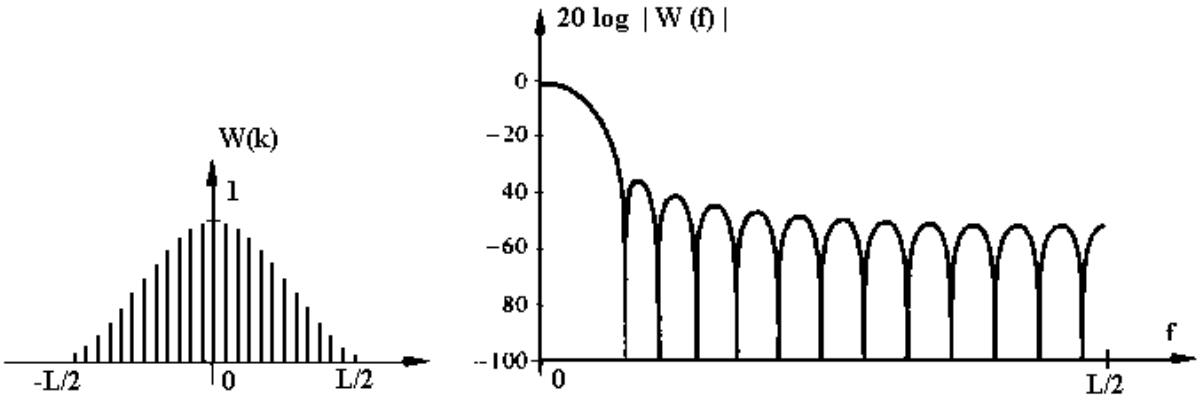


Figure 49d: Fenêtre de Kaiser-Bessel pour bL= 4.5

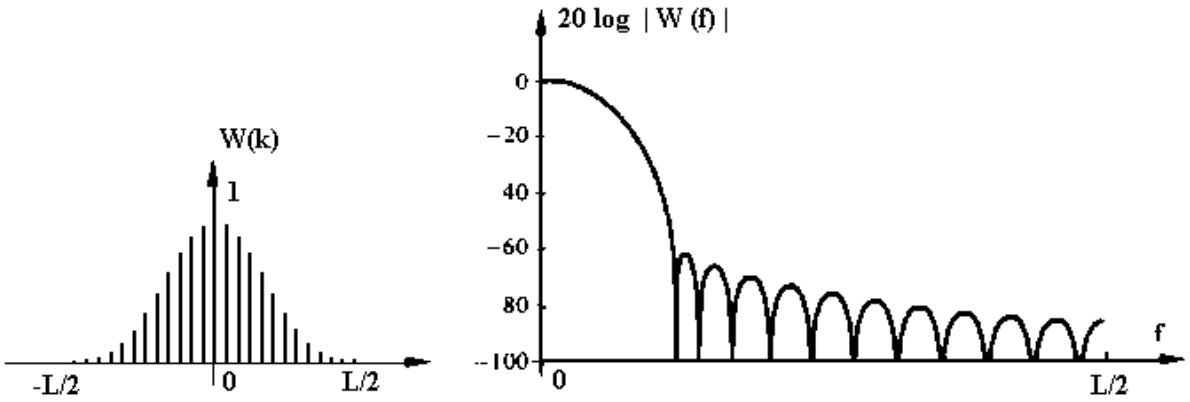


Figure 49e: Fenêtre de Kaiser-Bessel pour bL= 8.5

Pour notre utilisation, nous avons testé ces différentes fenêtres dont les caractéristiques sont résumées dans le Tableau 10. (Kunt, 1980, p.108) montre que le rapport des amplitudes du lobe central et du premier lobe secondaire de $W(\nu)$ varie très peu en fonction de L. Il s'exprime en décibels de la façon suivante :

$$\lambda_{fen} = 20.\log_{10}\left|\frac{W_{fen}(\nu_{\sec ondaire})}{W_{fen}(0)}\right|$$ où $\nu_{secondaire}$ est la fréquence réduite du 1er lobe secondaire

Le calcul de la bande passante suppose le repérage de ν_1 et ν_2 pour lesquelles $W(\nu_1) = W(\nu_2) = W_{ref}$ où W_{ref} est un niveau de référence en dessus duquel le filtre est déclaré « passant » et en dessous duquel il est dit « non passant ». La bande passante est alors

$$\Delta\nu = \nu_2 - \nu_1$$

Le calcul de la bande passante des fenêtres spectrales nous a posé un problème car les études sur le sujet proposent des résultats contradictoires (Figure 50):

- (Kunt, 1980) définit la bande passante comme la largeur de la base du lobe principal (Figure 45. p.93). Elle consiste à repérer ν_1 et ν_2 pour lesquelles $W(\nu_1) = W(\nu_2) = 0$. Cette définition nous semble peu justifiée car elle se fonde sur une mesure de niveau $W_{ref} = 0$, qui est une équation aux solutions multiples. De plus, elle semble en contradiction avec la notion de bande « passante » et « non passante » expliquée ci-avant. En effet, des parties de spectre, comme les lobes secondaires, sont déclarées « non passantes » car en dehors de la bande passante alors que leur valeur $W(\nu_{lobe\ secondaire})$ est au dessus de W_{ref}.
- (Harris, 1976) utilise la notion électronique de la bande passante qui consiste à repérer ν_1 et ν_2 pour lesquelles $W(\nu_1) = W(\nu_2) = Wmax$ - 3dB. Sur une représentation en mesure linéaire, W_{ref} se mesure alors de la façon suivante :

  $$W_{ref}(en\ dB) = Wmax - 3\text{dB} \Leftrightarrow W_{ref}(en\ linéaire) = 0{,}707 \cdot Wmax$$

  Cette mesure est en adéquation avec la notion de bande « passante » et « non passante » expliquée ci-avant. Cependant, un tel critère apparaît trop peu contraignant compte tenu de la pente douce du filtre.
- Nous avons choisi un compromis entre les deux propositions. Nous définissons la bande passante comme $\Delta\nu = \nu_2 - \nu_1$ où ν_2 et ν_1 sont repérées pour $W(\nu_1) = W(\nu_2) = Wmax$ - 6dB

  Sur une représentation en mesure linéaire, W_{ref} se mesure alors de la façon suivante :

  $$W_{ref}(en\ dB) = Wmax - 6\text{dB} \Leftrightarrow W_{ref}(en\ linéaire) = 0{,}5 \cdot Wmax$$

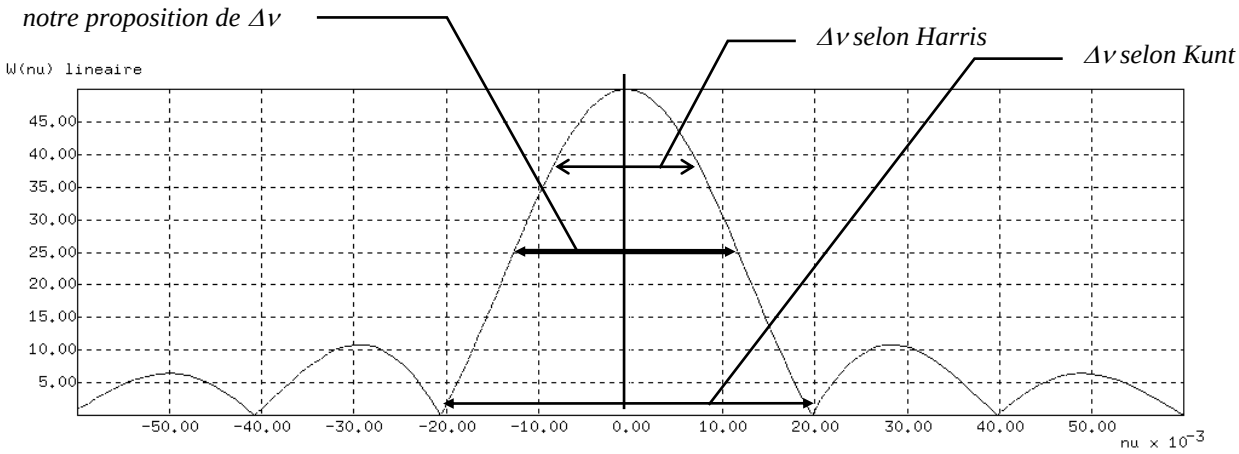


Figure 50: Mesure de la bande passante de la fenêtre spectrale rectangulaire pour L = 50 selon différents critères (abscisses en fréquence réduite, ordonnées en amplitude linéaire).

Tableau 10: Caractéristiques des fenêtres d'analyse classiques

FENETRE	Hauteur du lobe secondaire (en dB)	Bande passante du lobe principal (exprimée en fréquence réduite)
Rectangle	- 13	1.2 / L
Hamming	- 43	1.8 / L
Blackman	- 60	2.3 / L
Blackman-Harris 2	- 67	2.3 / L
Blackman-Harris 3	- 92	2.7 / L
Kaiser βL=4.5	- 40	1.8 / L
Kaiser βL=8.5	- 63	2.4 / L

Γ

Dans tous les cas, le calcul montre que la bande passante est inversement proportionnelle à L, caractéristique que nous connaissions déjà (cf. § « Les caractéristiques des fenêtres d'analyse », p.99). De plus, elle met en évidence un facteur Γ dépendant du type de fenêtre. Finalement, nous pouvons exprimer la bande passante réduite sous la forme :

(Eq.9) $$bp_{réduite} = \frac{\Gamma}{L}$$

La fenêtre qui jouit de la meilleure résolution spectrale est la fenêtre rectangulaire mais celle-ci possède aussi des lobes secondaires intenses. Inversement, Blackman-Harris 3 a un fort taux de réjection des lobes secondaires (-92dB) mais détient la moins bonne résolution. Nous avons finalement choisi la fenêtre de Hamming du fait de son bon compromis entre la hauteur des lobes secondaires et la bande passante du lobe principal. C'est d'ailleurs la fenêtre d'analyse préconisée par (Mac Aulay & Quatieri, 1986) dans leur méthode d'analyse-synthèse de la parole.

III.2.C.f. Le pas de décalage de la fenêtre d'analyse

III.2.C.f.i. Position du problème

L'utilisation de fenêtres d'analyse est un moyen de limiter artificiellement la durée du signal étudié afin de rendre possible une analyse spectrale sur calculateur. Nous avons étudié les effets de ce fenêtrage et notamment l'influence de la longueur et forme de la fenêtre. Il ne faut pas oublier qu'une telle analyse reste locale à la partie de signal observé, d'où son nom de Transformée de Fourier à Court Terme. Pour obtenir une analyse spectrale à long terme, il faut répéter l'analyse par TFCT en décalant la fenêtre d'observation le long du signal. La question est de savoir *de combien d'échantillons doit-on décaler la fenêtre de découpage pour faire une analyse spectrale convenable ?*

III.2.C.f.ii. Solution du problème

Par construction, les fenêtres w sont symétriques, c'est à dire que $w(n-n_0) = w(n_0-n)$. On peut donc écrire (Eq.6) , p.91, sous la forme :

$$X|_{n_0}^{L}(\nu) = \sum_{n=n_0-L/2}^{n_0+L/2} x(n).w(n_0-n).e^{-2j\pi\nu n}$$

Cette fonction est dépendante de n_0 et de ν.En fixant ν et en considérant n_0 variable, on peut réécrire cette relation sous la forme :

$$X|_{\nu}^{L}(n_0) = \sum_{n=n_0-L/2}^{n_0+L/2} x(n).w(n_0-n).e^{-2j\pi\nu n}$$

qui n'est rien d'autre qu'un produit de convolution :

$$X|_{\nu}(n_0) = w(n) * \left[x(n).e^{-2j\pi n\nu}\right] \qquad \text{où * est le signe de convolution}$$

ce qui peut s'interpréter comme le filtrage linéaire d'un signal x par un filtre de réponse impulsionnelle w. Rappelons que w est un filtre passe-bas à bande étroite (cf § « Répercutions spectrales dues au fenêtrage du signal », p.91). Soit $F_{\min}^{\max}$ la bande passante de ce filtre. $X|_{\nu}(n_0)$ issu du filtrage par w, est alors lui aussi limité à la bande passante $F_{\min}^{\max}$. Par conséquent, d'après le Théorème de Shannon (Eq.1, p.58), $X|_{\nu}(n_0)$ devra être échantillonné à une vitesse d'au moins $F_{\min}^{\max}$ fois par seconde pour éviter le phénomène de recouvrement, ce qui se traduit par la relation :

(Eq.10) $F_{fen} \geq F_{\min}^{\max}$ où F_{fen} est la fréquence de fenêtrage exprimée en Hz

Explicitons cette relation. Remarquons d'abord que $F_{\min}^{\max}$, la bande passante exprimée en Hz, est liée à la bande passante $bp_{réduite}$ de l'équation (Eq.9), p.103, par la relation

$$F_{\min}^{\max} = F_{ech}.bp_{réduite} = F_{ech}.\frac{\Gamma}{L} \qquad \text{où } F_{ech} \text{ est la fréquence d'échantillonnage}$$

Considérons ensuite la grandeur P_{fen} défini comme le pas de décalage de la fenêtre exprimé en échantillons. De façon triviale,

$$P_{fen} = \frac{F_{ech}}{F_{fen}}$$

La relation (Eq.10) devient alors:

$$F_{fen} \geq F_{\min}^{\max} \Leftrightarrow \frac{F_{ech}}{P_{fen}} \geq \frac{F_{ech}.\Gamma}{L} \Leftrightarrow \frac{1}{P_{fen}} \geq \frac{\Gamma}{L}$$

ce qui donne comme résultat final: $P_{fen} \leq \frac{L}{\Gamma}$

III.2.C.f.iii. Résultats

Prenons L=100, ce qui nous permet d'exprimer le pas de décalage en %. A partir des résultats du Tableau 10, p.103, nous en déduisons les résultats :

Tableau 11: Décalage conseillé des fenêtres d'analyse

Type fenêtre	Γ	P_{fen}
Rectangle	1.2	83%
Hamming	1.8	55%
Kaiser βL=4.5	1.8	55%
Blackman	2.3	44%

Le chiffre de 55% pour la fenêtre de Hamming signifie qu'en toute rigueur, *il faut décaler des fenêtres d'analyse de Hamming d'environ la moitié de leur longueur totale pour respecter les conditions de Shannon.*

Il est facile de comprendre qu'un recouvrement des fenêtres soit nécessaire dans le cas de l'utilisation de fenêtres de Hamming, Kaiser, Blackman... où une partie du signal en bord de fenêtre est écrasée par les pondérations. Décaler, par exemple, d'une demi-fenêtre permet de saisir dans la fenêtre suivante la partie écrasée dans la fenêtre présente.

III.2.C.g. Conclusion concernant la Transformée de Fourier à Court Terme

Si nous avons choisi d'étudier ces principes de traitement du signal, c'est pour examiner les fondements théoriques des outils mis au point dans cette étude et pour asseoir sur des bases solides des pratiques souvent intuitives. Il semble que ces pratiques intuitives, comme le fait d'effectuer un décalage de 50% pour des fenêtres de Hamming soient en cohérence avec les résultats présentés dans les paragraphes précédents.

III.2.D. Les Transformée de Fourier Discrète (T.F.D.)

Nous avons vu que pour obtenir une représentation spectrale classique d'un signal numérique, il est nécessaire de calculer la valeur $X(\nu)$ en n'importe quel point $\nu \in [0 ; ½]$ (cf. § « Problèmes pratiques liés au calcul de la Transformée de Fourier », p.90). Cette grandeur étant continue, cette opération est impossible à effectuer dans un système de traitement numérique. La Transformée de Fourier Discrète (TFD) est la solution au problème de l'aspect continu de ν. Présentons les modifications à apporter à la transformation de Fourier (Eq.3, p.89) pour obtenir sa version discrète. Le remplacement de la variable continue ν par une variable discrète k peut s'écrire de la façon suivante :

$\nu = k.\Delta\nu$ où $\Delta\nu$ est le pas de discrétisation de l'axe des fréquences.

Il s'agit là d'un échantillonnage de l'espace fréquentiel à ne pas confondre avec l'échantillonnage temporel du signal. Les fréquences discrètes $\nu_k = k.\Delta\nu$ sont les fréquences observables. Comme $X(\nu)$ est périodique de période unité (cf. § « Quelques propriétés », p.89), il est possible de diviser cet intervalle en N incréments et par conséquent,

(Eq.11) $\Delta\nu = 1/N$

Si l'intervalle choisi est [-½ ; ½], les *N* fréquences discrètes ν_k que nous noterons dorénavant k sont égales à -N/2, -N/2 + 1,... La TFD s'exprime finalement par :

(Eq.12) $x(n) \xrightarrow{T.F.D.} X(k) = X(\upsilon = \frac{k}{N}) = \sum_n x(n).e^{-2j\pi\frac{k}{N}n}$ où k est une grandeur fréquentielle discrète

La TFD a les mêmes propriétés que la TF continue, c'est à dire qu'elle peut être vue comme un simple changement de base du vecteur signal $\underline{x}$ en un vecteur $\underline{X}$ par la transformation suivante $\underline{X} = [M] \cdot \underline{x}$ où $[M] = \{M_{nk}\}$ est la matrice de changement de base que l'on appelle base harmonique. Elle se note :

$$M_{nk} = e^{2j\pi\frac{n}{N}k}$$

La relation inverse qui permet de reconstituer le signal à partir du spectre discret est la suivante:

$$X(k) \xrightarrow{T.F.D.^{-1}} x(n) = \frac{1}{N}.\sum_k X(k).e^{2j\pi\frac{n}{N}k}$$

III.2.E. Le Transformée de Fourier Rapide ou Fast Fourier Transform (F.F.T.)

En traitement du signal, la T.F.D. est très souvent utilisée. Aussi, de nombreux algorithmes de calcul ont été mis au point pour réduire le temps de calcul. On parle alors de *Transformée de Fourier Rapide (T.F.R)* plus connue sous le nom anglais de *F.F.T. (Fast Fourier Transform).* L'algorithme que nous avons utilisé pour le calcul rapide de la TFD est celui du *Radix-2* (Harris, 1976). Ce calcul possède la particularité de ne pouvoir effectuer des TFD que sur un nombre de points *N* égal à une puissance de deux :

$N = 2^n$ $N \in \{2,4,8,16,32,64,128,256,512,1024,2048...\}$

III.2.F. Bilan sur la Transformée de Fourier

Pour obtenir une estimation de la composition spectrale d'un signal, il faut, d'une part limiter la durée d'analyse à l'aide de fenêtres de longueur finie *L* et, d'autre part, se contenter d'un échantillonnage de *N* points de mesures spectrales. La formule de la Transformée de Fourier Discrète à Court Terme issue des relations (Eq.6), p.91 et (Eq.12), p.106, est la suivante :

(Eq.13) $X(n,k) = \sum_{m=-\infty}^{+\infty} w(n-m).x(m).e^{-2j\pi m\frac{k}{N}}$

où
- *n* et *m* sont des indices temporels
- *k* est l'indice fréquentiel
- *1/N* est le pas d'échantillonnage entre deux valeurs discrètes de fréquences
- *W* est la fenêtre temporelle qui détermine la portion du signal d'entrée *x(m)* qui doit être considérée au temps *n*.

Il faut remarquer que la valeur $\Delta\nu$ de l'équation (Eq.11) représente la résolution de la mesure spectrale de la TFD, à ne pas confondre avec la résolution spectrale définie au § « Longueur de fenêtre et résolution spectrale », p.95. Généralement, par soucis d'optimisation, les analyseurs de Fourier utilisent l'égalité $N^* = L^{**}$. Cette relation n'est pas toujours possible dans le cas d'un calcul par FFT qui nécessite une valeur $N = 2^n$. N est alors choisie comme la puissance de deux supérieure à L. La solution consiste finalement à rajouter des zéros à la suite du signal tronqué. Par exemple, si L = 100 échantillons, la puissance de deux supérieure à 100 est 128 = 2^7. La FFT se calcule sur un signal $\underline{x}_{FFT}$ à 128 points défini de la façon suivante :

$$x_{FFT}(n) = \begin{cases} x(n) & \text{pour } n \in [n_0 - 50\ ;\ n_0 + 50\ ] \\ 0 & \text{pour } n \in [n_0 + 50\ ;\ n_0 + 78\ ] \end{cases}$$

Il est possible de généraliser ce procédé et prendre $N >> L$. Ce subterfuge permet d'augmenter la résolution de la mesure. En effet, l'égalité (Eq.11), p.106 montre que si N augmente, $\Delta\nu$ diminue. Mais il ne faut pas perdre de vue qu'un tel artifice ne permet pas d'augmenter la résolution spectrale réelle qui ne dépend que de L (cf. § « Longueur de fenêtre et résolution spectrale », p.95).

Il reste une dernière remarque à faire. Du fait de la symétrie de la TF d'un signal réel, ce qui est le cas de la parole (cf. § « Quelques propriétés », p.89), le spectre est généralement représenté uniquement sur les fréquences positives. La représentation spectrale ne laisse finalement apparaître que $N/2$ points de mesure.

III.2.G. Transformée de Fourier Discrète vs Ondelettes

III.2.G.a. Présentation générale des ondelettes

La transformée en ondelettes est une technique qui permet de décomposer un signal en un ensemble de fonctions élémentaires appelées « ondelettes ». Toute les ondelettes ont la même forme et se déduisent d'une fonction mère à partir de translations et d'homothéties (contraction ou dilatation). L'équation d'une ondelette est la suivante (Grossmann et al., 1987):

$$\phi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\phi\left(\frac{t-\tau}{a}\right)$$

où ϕ est l'ondelette mère
τ est un paramètre qui déplace l'ondelette sur l'axe des temps (effet de translation)
a est un facteur de changement d'échelle (effet de dilatation ou contraction).

L'ondelette $\phi_{a,\tau}$ est responsable de l'analyse du signal $\underline{x}$ en un point spatio-temporel (a,τ) (Figure 51). Le paramètre a s'apparente à une caractéristique spectrale. La transformée en ondelettes du signal $x(t)$ s'obtient en effectuant le produit scalaire du signal avec les ondelettes analysantes:

$$x(t) \xrightarrow{Ondelette(a,\tau)} X(a,\tau) = \left(x, \phi_{a,\tau}\right) = \int x(t).\overline{\phi_{a,\tau}}(t).dt = \frac{1}{\sqrt{a}}\int x(t).\overline{\phi}\left(\frac{t-\tau}{a}\right).dt$$

où $\overline{\phi}$ représente le complexe conjugué de ϕ
$X(a,\tau)$ sont les coefficients de décomposition du signal $\underline{x}$ dans la base des ondelettes

* nombre de points pour la DFT
** longueur de la fenêtre d'analyse

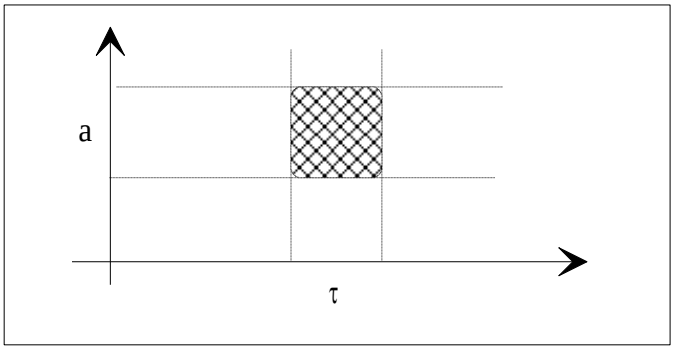


Figure 51: L'analyse spatio-temporelle de l'ondelette autour d'un point (a, τ)

III.2.G.b. Le cas des ondelettes de Morlet

Dans le cas de l'ondelette de Morlet, l'équation de l'ondelette mère est la suivante (Kronland-Martinet et al., 1987; Gérard & Baudry, 1994): $\phi(t) = e^{-\frac{t^2}{2}}.e^{j\omega_0 t}$, ce qui représente une gaussienne modulée en fréquence.

Posons $w(t) = e^{-\frac{t^2}{2}}$, l'ondelette de Morlet s'écrit alors comme $\phi(t) = w(t).e^{j\omega_0 t} = w(t).e^{2j\pi f_0 t}$, ce qui peut s'interpréter comme une sinusoïde complexe modulée en amplitude. L'analyse en ondelettes s'écrit alors:

$$X(a,t) = \frac{1}{\sqrt{a}}\int x(t).\left(w\left(\frac{t-\tau}{a}\right).e^{-2j\pi f_0\left(\frac{t-\tau}{a}\right)}\right).dt = \frac{1}{\sqrt{a}}\int\left(x(t).w\left(\frac{t-\tau}{a}\right)\right).e^{-2j\pi f_0\left(\frac{t-\tau}{a}\right)}.dt$$

Posons $y(t) = x(t).w\left(\frac{t-\tau}{a}\right)$ qui apparaît comme le fenêtrage du signal $\underline{x}$ par une gaussienne centrée en $t = \tau$. On obtient ainsi:

$$X(a,\tau) = \frac{1}{\sqrt{a}}\int y(t).e^{-2j\pi f_0\left(\frac{t-\tau}{a}\right)}.dt = \frac{1}{\sqrt{a}}.e^{2j\pi\frac{f_0}{a}\tau}.\int y(t).e^{-2j\pi\frac{f_0}{a}t}.dt$$

Plaçons-nous à $t = \tau$; posons $f = \frac{f_0}{a}$ et $A_\tau = \frac{1}{\sqrt{a}}.e^{2j\pi\frac{f_0}{a}\tau}$, la relation donnant X devient

$$X\ (f,\tau) = A_\tau.\int y(t).e^{-2j\pi ft}.dt$$

qui n'est rien d'autre, à une constante complexe près, que la transformée de Fourier du signal $\underline{y}$. Autrement dit, la transformée en ondelettes de Morlet d'un signal $\underline{x}$ en un point spatio-temporel *(f, τ)* revient à calculer la transformée de Fourier du signal $\underline{x}$ fenêtré par une gaussienne centrée en un point $t=\tau$. On rejoint ici la notion de Transformée de Fourier à Court Terme.

III.2.G.c. Ondelettes / Transformée de Fourier

Dans le cadre d'un signal numérique, les conclusions restent les mêmes quant à la forte ressemblance entre l'analyse par ondelettes de Morlet et l'analyse par TFCT. On discrétise ainsi l'espace spatio-temporel en prenant $a = a_0^m$ et $\tau = n\tau_0 a_0^m$ où $m, n \in Z$ (Daubechies, 1992). L'analyse s'effectue alors à largeur de bande relative $\frac{\Delta f}{f}$ constante et fournit une grille de valeurs non uniformément réparties (Figure 52).

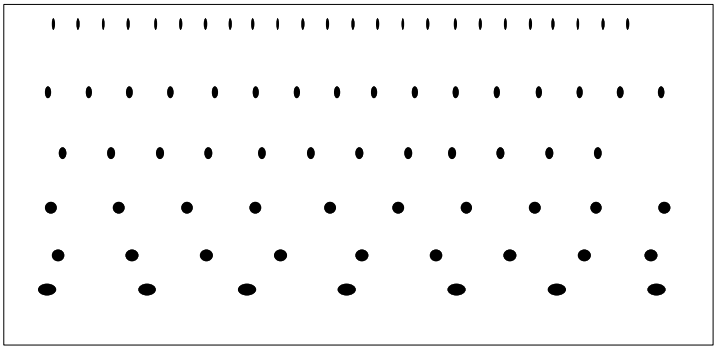


Figure 52: La répartition non-uniforme de l'analyse par ondelettes

Si la FFT, où Δf est constant entre chaque raie, ne permet pas un tel type d'analyse, la Transformée de Fourier à Court Terme (TFCT) définie par l'équation (Eq.6), p.91, autorise une prise de mesure libre. L'analyse par ondelettes de Morlet s'apparente complètement au calcul d'une TFCT en prenant une répartition des points d'analyse non uniforme. C'est d'ailleurs la technique qu'emploie d'Alessandro d'un point de vue informatique (d'Alessandro & Beautemps, 1991).

III.2.G.d. Discussion

Si les ondelettes comportent de nombreuses justifications mathématiques appréciables dans une étude théorique ou dans le cadre d'un dispositif d'analyse-synthèse (Montrésor, 1991), elles restent encore mal adaptées à la parole comme simple outil d'observation, à l'image d'autres transformées comme celle de Wigner-Ville (Flandrin, 1987). La seule ondelette interprétable physiquement est celle de Morlet mais elle s'apparente alors à une TFCT. Aussi avons-nous utilisé la transformée de Fourier classique pour mettre au point notre outil d'analyse spectrale: un vocodeur en bandes critiques.

III.3. « CRITIVOC » : un vocodeur en bandes critiques

Les caractéristiques d'un vocodeur sont importantes: choix sur le nombre de bandes, leur répartition, leur bande passante, leur forme, ainsi que sur la courbe de sensibilité de l'ensemble des bandes. Dans notre cas, les caractéristiques sont fondées sur le modèle auditif présenté au début de ce chapitre.

III.3.A. Les étapes de la réalisation du vocodeur à bandes critiques

L'analyse par notre vocodeur s'effectue en plusieurs étapes (Figure 53) :

- fenêtrage du signal par trame (de l'ordre de la centiseconde)
- transformée de Fourier puis obtention du module du spectre
- pondération auditive
- intégration en bandes critiques
- conversion en dB éventuellement
- décalage de la trame

Un exemple de ce traitement est fourni en (Figure 57, p.113 & Figure 58, p.114).

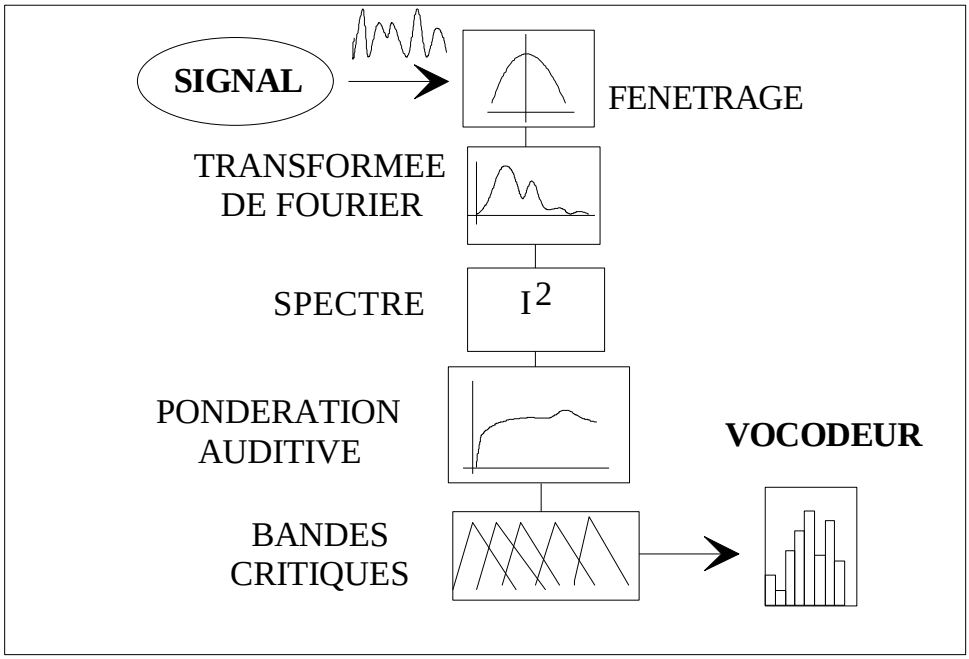


Figure 53: Diagramme fonctionnel du vocodeur à bandes critiques

III.3.B. La sortie de « CritiVoc »

Les valeurs obtenues dans les canaux du vocodeur auditif représentent les énergies dans les bandes critiques après correction du spectre par pondération sonique (cf. § « Un modèle auditif », p.76). La formalisation de ce calcul est:

$$E_k = \sum_{i=0}^{N} C_k(i).P(i).E(i)$$

où

Ek est l'énergie dans la bande *K*
Ck est la fonction de transfert de la bande *K*
P est la courbe de pondération sonique
E est le spectre de Fourier numérique
N est le nombre de points dans le spectre

Ce calcul est renouvelé régulièrement dans le temps en faisant glisser les trames d'analyse extraites avec des fenêtres de type Rectangulaire, Hamming ou Kaiser..., de longueur et de décalage paramétrables. La sortie du vocodeur est une représentation compacte temps-fréquence du signal analogue à une description de type spectrogramme.

III.3.C. L'utilisation de « CritiVoc »

« CritiVoc » a été expertisé par divers phonéticiens et considéré comme un outil d'analyse spectrale efficace pour effectuer une discrimination entre les sons de parole. De plus, il limite suffisamment le débit d'information (21 composantes toutes les 10ms) pour atténuer les phénomènes de variabilité.

Nous exploiterons notre vocodeur de deux façons:

- « CritiVoc » peut être exploité par l'expert phonéticien pour générer des règles de décodage acoustico-phonétique. Nous utilisons essentiellement la sortie du vocodeur sous forme quantitative en analysant directement les valeurs des énergies calculées dans les bandes (Figure 54) ou encore de façon graphique (Figure 55).
- « CritiVoc » peut servir de base pour l'extraction de paramètres acoustiques dans l'optique d'une reconnaissance globale. Dans cette optique a été développé la technique PLP que nous allons décrire.

	t	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19	K20	K21	SUM
	0	0	1	1	0	0	2	2	1	2	5	10	4	1	1	1	0	1	4	1	0	0	38
	1	14	66	150	207	483	1373	1712	598	269	501	1063	86	34	76	42	16	95	209	61	15	1	7071
	2	100	530	1481	1162	2345	10335	11689	1376	572	1799	3296	469	74	151	89	44	325	548	163	66	1	36617
	3	139	736	2719	2212	4746	22626	17652	1264	928	1667	3573	430	142	457	172	43	222	507	140	34	1	60409
a	4	206	1059	3906	3265	8397	18088	13208	511	758	1322	3335	544	201	399	87	78	403	1184	173	80	3	57206
	5	315	1643	5663	4195	16972	19572	3484	400	615	785	1946	1358	198	142	82	79	301	1090	121	88	1	59051
	6	365	1652	5774	3433	11051	8733	861	899	308	362	1235	821	137	66	127	67	164	843	123	62	2	37086
	7	450	1913	5119	2857	4504	2463	751	458	156	230	933	950	195	57	109	85	196	962	99	62	1	22549
	8	339	1089	1727	1349	1219	399	145	241	134	246	730	1203	78	86	222	116	182	497	74	114	3	10192
	9	138	275	126	216	168	64	74	24	14	31	83	208	40	70	272	848	996	315	387	163	4	4517
	10	42	80	20	43	79	22	24	16	11	13	19	36	75	254	889	1136	1713	1833	2007	707	12	9032
	11	2	3	2	6	12	3	5	6	4	3	10	21	37	273	1628	2178	2888	2661	4672	1633	14	16061
	12	0	1	2	7	7	8	8	7	5	14	37	26	107	670	2765	6486	2677	3016	5124	1614	13	22593
ʃ	13	0	1	2	2	3	3	1	2	4	10	11	24	81	449	1840	3946	6019	4948	5676	682	10	23714
	14	0	0	0	1	1	2	2	8	7	13	10	40	64	386	1319	3550	5024	3790	4052	874	11	19156
	15	0	1	0	1	1	1	0	1	5	6	14	31	74	424	2147	8527	6585	6477	6708	1209	8	32221
	16	0	0	0	1	2	1	0	1	1	4	9	11	25	158	787	3366	2271	2508	1651	632	7	11433
	17	0	0	0	0	0	0	0	0	0	1	2	6	9	56	302	1403	1045	933	951	218	3	4931
	18	0	0	0	0	1	0	0	0	0	0	1	2	2	12	76	476	545	159	167	44	1	1487
	19	0	0	0	0	0	1	0	0	0	0	0	0	0	1	7	45	75	14	11	6	0	161
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	2	3	2	0	0	17
	21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	2	2	2	1	0	0	10
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	5
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	5
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	4
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3
	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	3	1	0	0	13
t	27	0	1	3	2	5	1	2	7	21	30	25	68	15	37	114	161	287	294	31	3	0	1108
	28	1	2	3	4	9	5	4	3	5	3	11	19	3	11	65	63	88	147	46	6	0	500
	29	3	10	18	19	13	4	4	2	1	2	7	8	2	4	23	13	43	54	8	4	0	244
	30	176	711	1065	1475	1923	217	30	25	15	43	82	9	6	12	29	15	51	27	7	3	0	5923
	31	193	1251	1387	2972	5070	485	43	32	23	97	196	18	4	14	35	10	41	18	5	2	0	11896
	32	189	1163	1047	1510	2218	536	25	20	38	81	111	12	3	14	38	9	25	6	2	0	0	7046
	33	171	1031	679	865	1360	418	28	47	52	118	65	6	3	17	42	13	25	6	3	2	0	4950
	34	147	874	478	798	1698	525	77	95	144	127	25	4	2	7	20	11	17	4	4	1	0	5059
	35	120	717	333	444	1061	824	110	104	228	191	10	3	1	3	8	9	12	9	6	1	0	4195
ã	36	93	560	223	215	464	547	91	98	210	64	3	1	1	1	4	6	6	2	2	0	0	2592
	37	69	424	164	125	273	294	62	171	258	10	2	1	1	1	2	5	4	1	1	0	0	1866
	38	54	331	157	92	184	111	50	174	54	3	1	0	0	1	1	2	1	1	1	0	0	1216
	39	43	251	148	86	132	81	81	182	14	2	1	0	1	1	1	1	1	0	0	0	0	1025
	40	35	194	150	68	85	40	59	57	7	2	1	1	1	1	2	1	0	0	0	0	0	705
	41	24	119	92	35	48	27	49	29	3	1	1	0	0	0	0	0	0	0	0	0	0	429
	42	18	75	62	25	30	19	25	13	3	1	0	0	0	0	1	1	0	0	0	0	0	274
	43	14	60	46	20	17	11	6	3	2	1	0	0	0	0	1	1	0	0	0	0	0	184
	44	7	31	25	4	3	1	2	1	0	0	0	0	0	0	1	1	0	0	0	0	0	76
	45	3	14	25	4	3	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	55

Figure 54: Résultats de « CritiVoc » sur le mot "achetant". L'axe du temps est vertical et celui des fréquences, correspondant aux canaux, est horizontal. La colonne 't' indique le n° de trame, 'Ki' sont les canaux, 'SUM' indique la somme des canaux. Les conditions d'analyse sont les suivantes:

- *16 kHz de fréquence d'échantillonnage: bande passante de 8 kHz => 21 bandes critiques*
- *fenêtre de Hamming de 20 ms de long avec un recouvrement de 50%.*
- *pondération sonique de type X.*

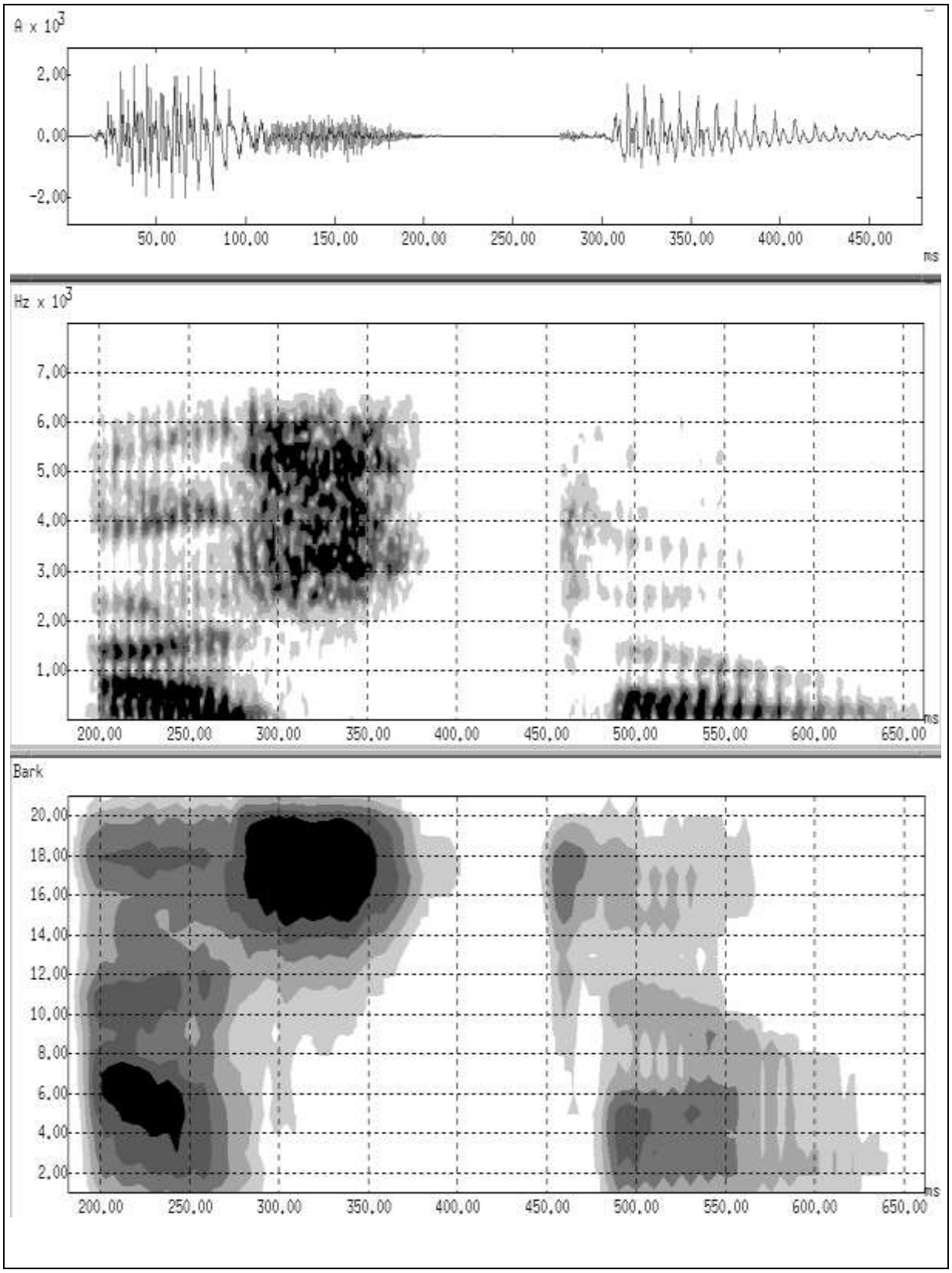


Figure 55: a) Signal de parole sur le mot "achetant" prononcé [a ʃ t ã],
b) Spectrogramme classique en bandes larges (temps en ms, fréquences en Hz),
c) « Auditogramme » (temps en ms, fréquences en BARKS) avec les mêmes conditions d'analyse que précédemment

III.4. La méthode de prédiction linéaire fondée sur un modèle auditif

La technique de Prédiction Linéaire fondée sur un spectre Perceptif (Perceptually-based Linear Prediction) est une approche qui permet d'extraire des informations à partir d'un spectre auditif modélisé par un filtre « à pôles ». Le résultat est une série de coefficients dits « PLP » qui peuvent servir d'éléments d'information en reconnaissance automatique de la parole (Hermansky, 1987, 1990 ; Yong & Mason, 1987; Junqua, 1990; Ghio, 1992).

III.4.A. Les étapes de l'extraction de coefficients P.L.P.

Le traitement PLP s'effectue en deux étapes: d'une part l'obtention du spectre auditif, d'autre part la recherche, par la technique de prédiction linéaire, d'une fonction de transfert liée à ce spectre. Le diagramme de la Figure 56 détaille les différents traitements. Un exemple de ce traitement est fourni en (Figure 58).

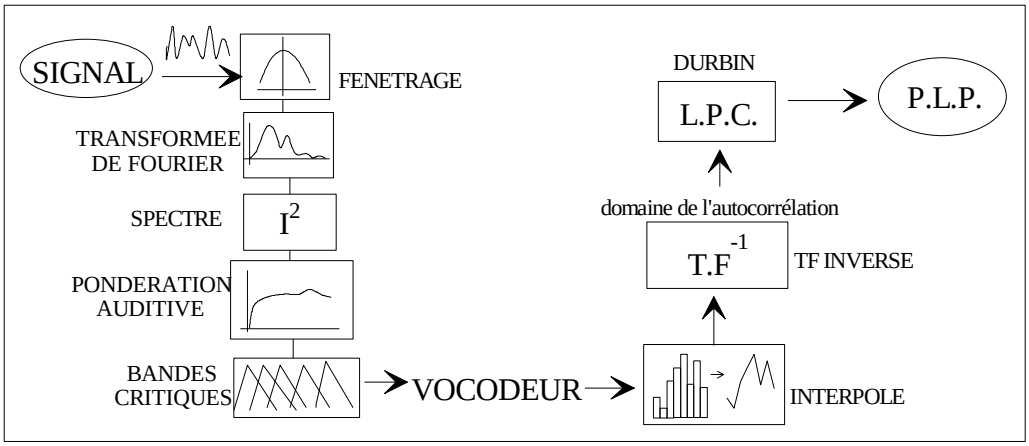


Figure 56: Diagramme fonctionnel de la technique d'analyse PLP

III.4.B. Le spectre auditif

Le calcul des énergies contenues dans les bandes critiques est celui du vocodeur décrit dans le paragraphe précédent. Suite à cette estimation, les bandes sont interpolées linéairement de façon à obtenir un spectre auditif (Figure 58). L'interpolation dans le domaine des barks revient à donner la même importance spectrale à chaque bande, ce qui signifie que la finesse spectrale en basses fréquences est plus importante qu'en hautes fréquences. Autrement dit, la localisation des pics d'énergie est précise en BF, ce qui permet une caractérisation fine des premiers formants des voyelles.

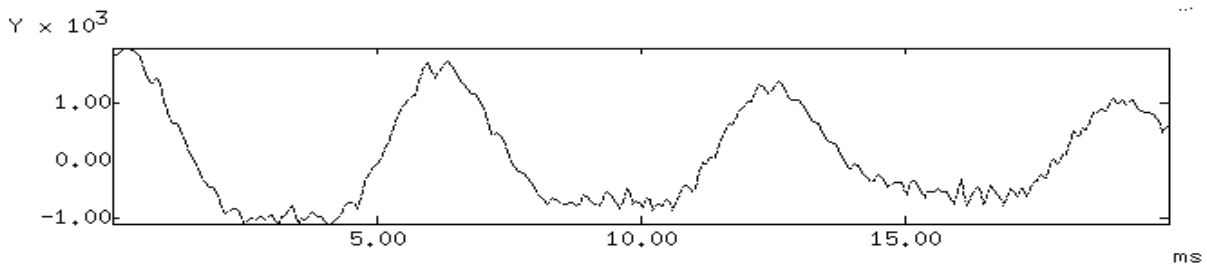


Figure 57: Signal de parole du 1er /i/ du mot " bicyclette"

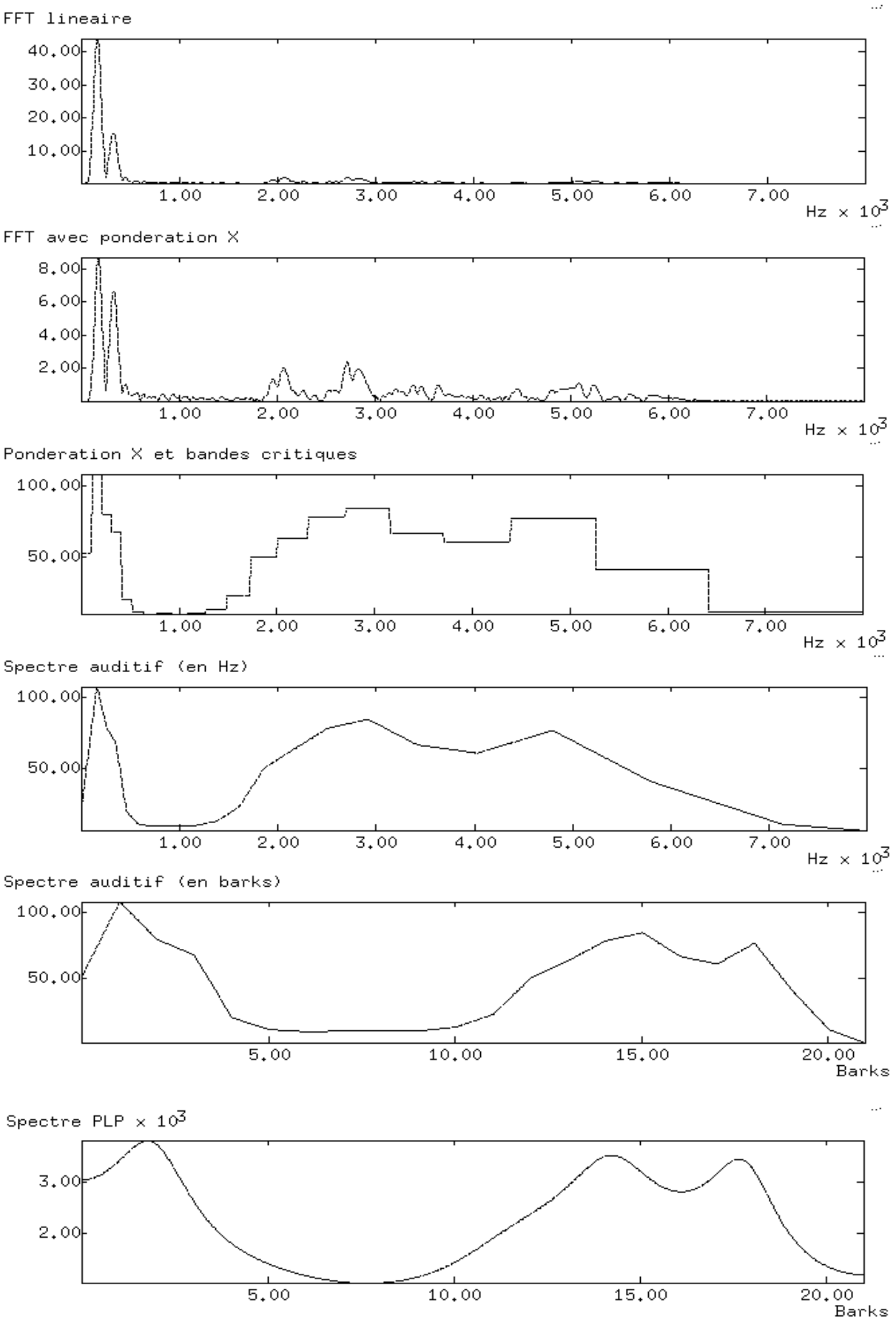


Figure 58: Traitement par « CritiVoc » et modélisation par prédiction linéaire sur le 1er /i/ du mot " bicyclette" (pondération X, 21 bandes critiques, ordre de PL = 10)

III.4.C. Le modèle à pôles

Le spectre auditif peut être considéré comme la réponse harmonique d'un filtre à pôles, ce qui a l'avantage de mettre en relief les pics du spectre et de modéliser de façon compacte l'information spectrale. Pour cela, nous avons utilisé la méthode de prédiction linéaire.

III.4.C.a. Le principe de la prédiction linéaire (Linear Predictive Coding ou L.P.C.)

La méthode de prédiction linéaire est une technique utilisée depuis longtemps dans le traitement de la parole (Makhoul,1975 ; Atal & Schroeder, 1978).Il s'agit d'un procédé qui a été mise au point originellement par les géologues dans l'analyse sismique des couches terrestres. Une onde puissante est envoyée dans le sol. Celle-ci, filtrée par les différents terrains, est réfléchie. L'analyse par prédiction linéaire de l'onde recueillie permet de connaître certains paramètres révélateurs de la nature et l'épaisseur des roches.

Plus généralement, la méthode de prédiction linéaire part du principe que le signal analysé est issu d'une source qui est filtrée. Dans le cas de la parole, il s'agit du filtrage de la source vocale par le tractus vocal (cf. § « L'aspect acoustique de la parole », p.53). **Le but de la prédiction linéaire est d'obtenir les paramètres de ce filtre**, celui-ci étant porteur d'information (Figure 27b, p.54). En effet, les caractéristiques du conduit vocal conditionnent le timbre sonore et apparaissent finalement comme des données pertinentes pour le décodage. Les paramètres du filtre modélisé par une prédiction linéaire sont appelés « coefficients LPC » . Ils s'obtiennent par un algorithme que nous allons décrire.

Considérons un signal numérique d'excitation *e(n)*, un filtre linéaire dont la réponse impulsionnelle est *h(n)* et le signal en sortie du filtre *s(n)*.

Par définition, $s(n) = h(n) * e(n)$ où * est le signe de convolution

ce qui, dans le domaine spectral donne la relation multiplicative :

$S(z) = H(z) \,.\, E(z)$ où les majuscules représentent les transformées en *Z* des signaux écrits en minuscules précédemment

Dans le cas où le filtre ne possède que des pôles (modèle à pôles), on peut expliciter sa fonction de transfert de la façon suivante:

$$H(z) = \frac{G}{1 + \sum_{i=1}^{p} a_i \,.\, z^{-i}}$$

où p est le nombre de pôles
les a_i sont les pôles
G est le gain du filtre (que nous ignorerons par la suite)

On obtient finalement: $$S(z) = \frac{1}{1 + \sum_i a_i \,.\, z^{-i}} \,.\, E(z) \Leftrightarrow S(z) + \sum_i \left(S(z).a_i \,.\, z^{-i} \right) = E(z)$$

Le retour dans le domaine temporel opéré par transformée en Z inverse de cette dernière équation donne:

$$s(n) + \sum_i a_i \,.\, s(n-i) = e(n) \;\Leftrightarrow\; s(n) = e(n) - \sum_{i=1}^{p} a_i \,.\, s(n-i)$$

Posons $s'(n) = -\sum_{i=1}^{p} a_i . s(n-i)$, on a alors $s(n) = e(n) + s'(n)$

s'(n) , qui est fonction de *s(n-1), s(n-2)...* peut être considéré comme la prédiction du $n^{ème}$ échantillon à partir des *p* précédents. Sur une fenêtre temporelle où le filtre est considéré comme stationnaire (environ 15 ms pour la parole), on peut alors choisir les coefficients a_i comme constants. Ceux-ci doivent être bien sûr renouvelés pour chaque fenêtre d'analyse. Les coefficients a_i étant fixés, la valeur prédite *s'(n)* peut être calculée à partir des *p* échantillons précédents et des a_i mais sa valeur ne sera pas exactement égale à la valeur exacte de l'échantillon *s(n).* On introduit alors une erreur de prédiction qui est égale à
s(n) - s'(n) , ce qui correspond à l'excitation *e(n).* Explicitons cette erreur.

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

Dans cette égalité, on remarque la relation de filtrage qui existe entre l'excitation *e(n)*, le filtre caractérisé par les coefficients a_i et le signal de sortie *s(n).* En général, on ne dispose que du signal en sortie et on cherche à accéder aux coefficients a_i. Dans le cas de modèle autorégressif où les échantillons sont fortement corrélés, l'erreur de prédiction est faible et possède une dynamique réduite. Cela se traduit par le fait que *les détails fréquentiels observés sur le signal de sortie peuvent être attribués au filtre.* Ceci est vérifié dans le cas de la parole où le modèle LPC distingue les contributions de la source et du filtre (Atal & Hananer, 1971). La source est modélisée par un bruit blanc ou un train d'impulsions selon que le segment modélisé est voisé ou non. Dans ce cas, l'erreur de prédiction est liée au bruit. Les paramètres a_i sont déterminés pour minimiser cette erreur. Telle est l'hypothèse clé de la prédiction linéaire.

III.4.C.b. La technique de prédiction linéaire par autocorrélation

Pour réaliser la minimisation évoquée précédemment, on fait alors intervenir ce que l'on appelle l'erreur quadratique moyenne notée M qui est définie par:

$$M = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left\{ \sum_{k=0}^{p} a_k . s(n-k) \right\}^2$$ où *N* est la longueur de la fenêtre d'analyse

Pour minimiser l'erreur de prédiction, on choisit les coefficients a_i dits de «prédiction linéaire» de telle façon que *M* soit minimale. Avant de réaliser cette opération, explicitons d'abord *M*.

$$M = \sum_{n=0}^{N-1} \left\{ \left[\sum_{i=0}^{p} a_i . s(n-i) \right] . \left[\sum_{j=0}^{p} a_j . s(n-j) \right] \right\} = \sum_{n=0}^{N-1} \left\{ \sum_{i=0}^{p} \sum_{j=0}^{p} a_i . s(n-i) . a_j . s(n-j) \right\}$$

$$M = \sum_{i=0}^{p} \sum_{j=0}^{p} \left\{ \sum_{n=0}^{N-1} s(n-i) . s(n-j) \right\} . a_i . a_j$$

On pose $d_{ij} = \sum_{n=0}^{N-1} s(n-i) . s(n-j) = \sum_{n=-i}^{N-1-i} s(n) . s(n+i-j)$

On remarque que $d_{ij} = d_{ji}$. Dans tout ce qui suit, on prendra par exemple $j < i$ sans pour autant que ce choix ne perturbe le raisonnement. On peut simplifier d_{ij} en tenant compte du fait que:

$s(n) = 0$ si $n \notin [\, 0 \,;\, N - 1 \,]$

$s(n+i-j) = 0$ si $n + i - j \notin [\, 0 \,;\, N - 1 \,] \iff n \notin [\, j - i \,;\, N + i - j - 1]$

Globalement, $s(n).s(n+i-j) = 0$ si $n \notin [\, 0 \,;\, N + j - i - 1 \,]$

Finalement, $d_{ij} = \sum_{n=0}^{N-1-|i-j|} s(n).s(n+i-j)$ qui représente la fonction d'autocorrélation R_{i-j} .

Aussi, on peut écrire l'erreur quadratique comme:

$$M = \sum_{i=0}^{p} \sum_{j=0}^{p} a_i . a_j . R_{|i-j|}$$

La dérivée partielle de M qui est une fonction de a_1, a_2, ... a_k,... a_p est:

$$\frac{\partial M(a_1, a_2, \ldots, a_p)}{\partial a_k} = 2. \sum_{i=0}^{p} a_i . R_{|i-k|}$$

La minimisation est réalisée en annulant cette dérivée partielle. On obtient alors la série d'équations suivante: $\sum_{i=1}^{p} a_i . R_{|i-k|} = -R_k$ qui est réalisée pour $k = 1, 2, \ldots p$

La forme matricielle de ces équations est: $[\mathcal{R}] \otimes \boldsymbol{a} = \boldsymbol{r}$

$$\begin{vmatrix} R_0 & R_1 & R_2 & \ldots & R_{p-1} \\ R_1 & R_0 & R_1 & \ldots & R_{p-2} \\ R_2 & R_1 & R_0 & \ldots & R_{p-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ R_{p-2} & R_{p-3} & R_{p-4} & \ldots & R_1 \\ R_{p-1} & R_{p-2} & R_{p-3} & \ldots & R_0 \end{vmatrix} \otimes \begin{vmatrix} a_1 \\ a_2 \\ a_3 \\ \ldots \\ a_{p-1} \\ a_p \end{vmatrix} = - \begin{vmatrix} R_1 \\ R_2 \\ R_3 \\ \ldots \\ R_{p-1} \\ R_p \end{vmatrix}$$

bilan

Finalement, la technique de Prédiction Linéaire, qui consiste à obtenir des coefficients a_i , se résume au calcul des coefficients d'autocorrélation R_{i-j} , à l'inversion de la matrice $[\mathcal{R}]$ et à la multiplication de la matrice résultante par le vecteur $[\boldsymbol{r}]$.

III.4.C.c. Le calcul des coefficients

En fait, les redondances de $[\mathcal{R}]$ permettent un calcul rapide des coefficients LP sans inverser explicitement $[\mathcal{R}]$. En effet, $[\mathcal{R}]$ est une matrice symétrique de Toeplitz (tous les coefficients appartenant à une diagonale sont égaux). Nous avons donc utilisé un algorithme récursif développé par *Durbin et Levinson* dont les détails sont donnés dans (Boite & Kunt, 1987). Si l'on reprend le bilan établi plus haut, on se rend compte que les éléments d'entrée de l'algorithme de prédiction linéaire par la technique de l'autocorrélation sont les *coefficients*

d'autocorrélation. Le signal de prédiction *s'(n),* le signal d'erreur *e(n)* ou l'erreur quadratique moyenne ne sont que des paramètres non explicités liés au modèle. Il existe deux façons de calculer les coefficients d'autocorrélation φ_X:

1/ *Par calcul direct sur le signal*

On utilise la définition: $\varphi_x(k) = \sum_{i=-\infty}^{+\infty} x(i).x(i+k)$

2/ *Par inversion du spectre de puissance*

Sachant que la transformée de Fourier Φ_X de la fonction d'autocorrélation φ_X est le spectre de puissance du signal, ce qui se traduit par l'équation TF[φ_X] = $\Phi_X(f) = |X(f)|^2$ où *X(f)* est la Transformée de Fourier (TF) de *x(n),* les coefficients d'autocorrélation peuvent alors s'obtenir par la formule $\varphi_X = TF^{-1}\{|X(f)|^2\}$ (Kunt, 1980). Il suffit pour cela de calculer la Transformée de Fourier *X(f)* du signal, de prendre le spectre de puissance $|X(f)|^2$ et d'effectuer une transformée de Fourier inverse. La TF inverse nous fait passer du domaine spectral de puissance au domaine temporel de l'autocorrélation.

résumé

La technique de prédiction linéaire part du principe que le signal analysé est issu d'une source qui est filtrée. Dans le cas de la parole, il s'agit du filtrage de la source vocale par le tractus vocal. Le but de la prédiction linéaire est d'obtenir les paramètres de ce filtre, celui-ci étant porteur d'information. Pour cela, il suffit de calculer les coefficients d'autocorrélation soit directement à partir du signal de parole, soit par la transformée de Fourier inverse du spectre de puissance de la parole. A partir des coefficients d'autocorrélation, l'algorithme de Durbin-Levinson fournit en sortie l'ensemble des coefficients LPC, caractérisant les pôles du filtre modélisé.

remarque

Il existe une autre méthode pour obtenir les coefficients a_i (Makhoul, 1975). Il s'agit de la technique des covariances, plus précise que celle de l'autocorrélation pour estimer les paramètres du filtre, mais ne garantissant pas la plausibilité physique du filtre (impulsion non amortie). Pour le repérage des formants, la méthode des covariances semble plus indiquée. Dans notre cas, elle apparaît impossible à mettre en place car notre prédiction linéaire porte sur un spectre auditif sur lequel il est facile d'obtenir des paramètres d'autocorrélation, contrairement au calcul de la matrice de covariance.

III.4.C.d. Le modèle auditif à pôles

Effectuer l'analyse par prédiction linéaire d'un spectre auditif revient à modéliser la « perception » des configurations du conduit vocal lors de la production de parole, ce qui semble particulièrement pertinent par rapport aux objectifs que nous nous étions fixés au § sur « Les questions préalables », p.9. Pour cela, il suffit d'extraire les coefficients d'autocorrélation à partir du spectre auditif de puissance et d'injecter ces données à l'algorithme de Durbin-Levinson qui fournit en sortie un ensemble de coefficients appelés non

plus LPC, mais PLP pour Prédiction Linéaire sur un spectre Perceptif. Le bilan de la modélisation est finalement le suivant (Figure 56, p.113):

- Transformée de Fourier inverse du spectre auditif de puissance
- saisie des p premières valeurs de la TF inverse (qui représentent les *p* premiers coefficients d'autocorrélation)
- algorithme de Durbin qui nous donne *p* coefficient a_i à partir des *p* coefficients d'autocorrélation

Ces *p* coefficients a_i représentent l'information que l'on garde pour caractériser le signal. Ils sont liés au filtre auditif modélisé. Pour une utilisation graphique, il est possible de visualiser la réponse harmonique du filtre à pôles obtenu. Pour cela, revenons à la fonction de transfert du filtre (cf. § III.III.4.III.4.C.III.4.C.a, p.115):

$$H(z) = \frac{G}{1 + \sum_{i=1}^{p} a_i . z^{-i}} = \frac{G}{\sum_{i=0}^{p} a_i . z^{-i}} \qquad \text{en prenant } a_0 = 1$$

Le module de la réponse harmonique est alors donné par:

$$\left|H(e^{j\omega})\right|^2 = \frac{G^2}{\left|\sum_{k=0}^{p} a_k . e^{-j\omega k}\right|^2} = \frac{G^2}{\left|TF\left[a(n)\right]\right|^2}$$

où $a(n) = a_n$ peut être considéré comme un signal (avec $a(n) = 0$ pour $n < 0$ ou $n > p$)

En conclusion, pour obtenir la réponse harmonique $H(e^{jw})$ du modèle à pôles, il suffit de calculer l'inverse du spectre de puissance du signal $a(n) = a_n$.

La Figure 58, p.114 permet de comparer graphiquement un spectre auditif et son modèle PLP, extrait avec la méthode décrite précédemment. L'interpolation dans le domaine des barks nous semble plus pertinente que dans celui des hertz (cf. § « Le spectre auditif », p.113). En effet, la localisation des amas d'énergie est précise en BF, ce qui autorise une caractérisation fine des premiers formants des voyelles, et grossière en HF, ce qui permet une réduction des phénomènes de variabilité.

A titre comparatif, nous présentons un traitement LPC traditionnel sur le même signal que la Figure 57, p.113, qui a servi d'exemple pour illustrer le traitement PLP. La modélisation LPC standard commence par un pré-traitement qui consiste à filtrer le signal en pré-accentuant les hautes fréquences. La modélisation agit ensuite sur ce signal filtré. Il est possible ensuite de visualiser le résultat sous forme de spectre LPC (Figure 59, p.120). On met en évidence les limites de la représentation linéaire en Hertz : la modélisation LPC se focalise sur des résonances peu pertinentes autour de 5 et 6 kHz. Elle assimile F2 et F3 entre 2 et 3 kHz. Elle ne tient aucunement compte de F1 autour de 300 Hz. Ceci est à comparer à la représentation PLP de la Figure 58, p.114, qui n'a nullement occulté F1, qui a repéré l'amas

énergétique autour de 2 et 3 kHz et qui a tenu compte de la présence d'énergies en hautes fréquences (5 à 6 kHz).

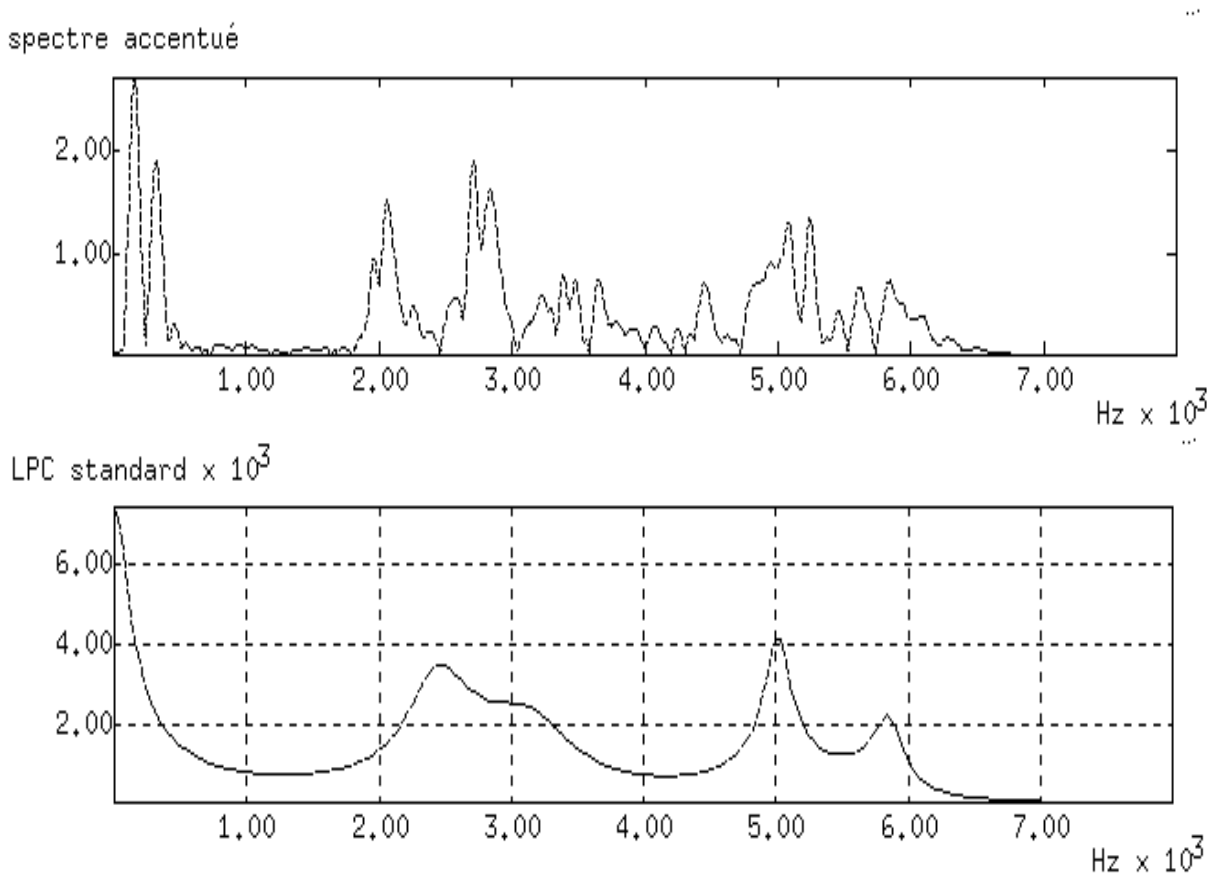


Figure 59: Répartition spectrale du 1er /i/ du mot " bicyclette" après pré-accentuation des aigus et représentation LPC avec un ordre = 10

La Figure 60 donne un exemple graphique de réponses harmoniques de modèles PLP issus des voyelles orales du français. Pour cela, un seul locuteur a été choisi et un unique extrait a été sélectionné. On se rend compte que le degré de séparabilité entre les voyelles apparaît intéressant. Il existe cependant des zones de recouvrement notamment entre /i/, /e/ et /y/.

Enfin, il reste à signaler que selon l'ordre de la prédiction linéaire, c'est à dire le nombre de coefficients retenus, la modélisation est plus ou moins proche du spectre (Figure 61). Il faut souligner qu'un ordre élevé (supérieur à 15) n'apporte pas forcément plus d'information pertinente pour le décodage acoustico-phonétique. En effet, une modélisation trop fine s'encombre de détails relatifs au locuteur, ce qui risque de dégrader les performances du système.

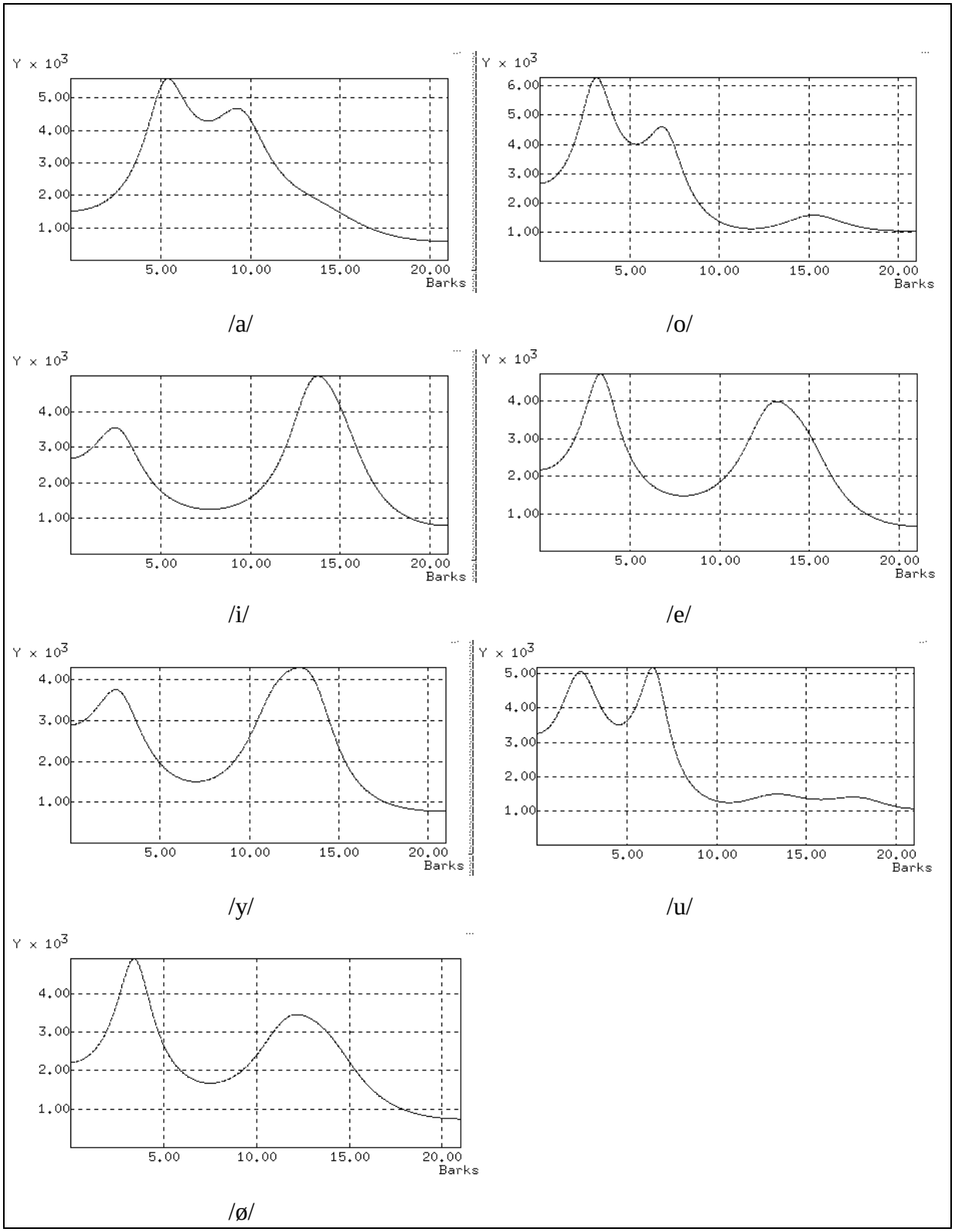


Figure 60: Réponses harmoniques de modèles PLP pour différentes voyelles (ordre de PL=8,abscisses en Barks)

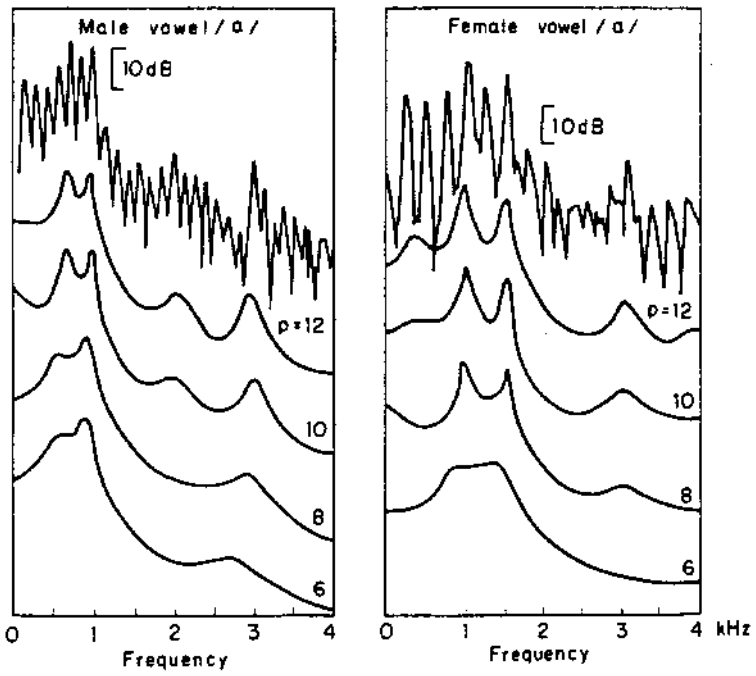


Figure 61: Comparaison de modèles LP en fonction de différents ordre de prédiction 'p' (Source: Furui, 1989)

IV. LE SYSTEME « ACHILE »

« Si les spécialités savent dépasser les disputes de chapelle, assumer la destruction des totems, résister à la tentation de l'ostracisme, alors, ensemble, elles accéderont à la lumière.»

Claude Allègre

Extrait de l'article « la Cognitique » paru dans le Point le 23 décembre 1994 (n°1162, p.87)

Plan du chapitre

RESUME

Dans ce chapitre, nous présentons ACHILE, un système de décodage ACoustico-pHonétique et d'Identification Lexicale. Ce dispositif est une alternative aux méthodes stochastiques. Il s'inspire, d'un point de vue fonctionnel, du traitement cognitif humain : le signal de parole alimente non pas un, mais plusieurs modules de décodage, qui varient à la fois par le traitement qu'ils opèrent sur le signal et par la conception de l'analyse. Le principe est de faire fonctionner en parallèle les différents modules, chacun d'entre eux analysant le signal de parole de façon indépendante. Ces différents modules sont les suivants: segmentation analytique couplée à une macro-classification, décodage analytique par règles, reconnaissance globale métrique. Le résultat de ces analyses montantes est fourni à un module superviseur qui, après une phase d'accès lexical, se charge de prendre une décision. La réussite de l'identification dépend d'une part de la bonne synchronisation des informations qui circulent, d'autre part, de l'étape fondamentale du décodage acoustico-phonétique, opération qui nécessite une étude en parallèle de l'axe syntagmatique et paradigmatique.

IV.1. La philosophie du système

IV.1.A. Ingénierie et connaissances

Les contraintes économiques ont conduit la plupart des ingénieurs et chercheurs en reconnaissance automatique de la parole à utiliser massivement les méthodes stochastiques, celles-ci donnant à court terme de bien meilleures performances. Le succès de ces techniques a même amené certains spécialistes - et non des moindres - à dénier toute efficacité, voire toute validité, aux systèmes à base de connaissances. Nous ne rappellerons pas le jugement fameux d'un non moins fameux spécialiste de la question (cf. note de bas de page, p.29). Cependant, de récentes études indiquent que les performances des méthodes stochastiques commencent à plafonner malgré les progrès informatiques (Bourlard, 1996). Tout porte à croire que la RAP ne pourra pas fonctionner uniquement sur la base de méthodes stochastiques.

Le phonéticien, et plus généralement le linguiste, a pour tâche d'apporter des connaissances sur le fonctionnement du langage et de tester ces connaissances. Les technologies vocales, en particulier la reconnaissance automatique, sont un bon banc d'essai pour cette validation. Le travail que nous présentons s'inscrit dans un courant scientifique destiné à faire progresser la recherche fondamentale pour résoudre des problèmes de RAP apparemment insolubles dans un cadre purement stochastique. Nous souhaitons examiner jusqu'à quel point un modèle à base de connaissances est capable de décoder de façon automatique la structure phonique de la parole sans recourir aux modèles stochastiques. Nous avons conscience, et l'histoire l'a montré, que les systèmes conçus à base de connaissances comme Hearsay dans le projet ARPA (Klatt, 1977) obtiennent bien souvent des performances moindres que des dispositifs ad hoc comme Harpy. Nous souhaitons malgré tout persister dans cet effort. Le mérite de tels systèmes repose sur la modélisation et donc la contrôlabilité des mécanismes. On peut donc espérer faire progresser leurs performances parallèlement aux connaissances. De plus, ils peuvent servir d'outil en recherche fondamentale pour comparer différentes théories.

IV.1.B. Connaissances ou probabilisme ?

Le langage et la parole sont structurés de façon complexe, sur plusieurs niveaux, de la conception à l'émission du message vocal. Ils le sont également lors de la réception. Pour reconnaître, l'auditeur doit décoder cette structure complexe. Dans cette tâche, les connaissances jouent un rôle fondamental. Ce savoir, qui n'est ni automatique, ni immédiat, nécessite plusieurs années d'acquisition à l'enfant en bas âge ou à l'adulte apprenant. Dans la communication orale homme-machine, le dispositif de décodage de la parole doit se substituer à l'auditeur. Dans ce cas-là, à nouveau, nous pensons que les connaissances sont essentielles. Contrairement à une onde sismique ou un bruit sous-marin, la parole est émise avec pour intention de communiquer du sens. Elle est organisée en conséquence et doit être traitée en conséquence. La question essentielle est de savoir quelles connaissances doivent être intégrées dans un système de R.A.P. et comment les organiser. (Rossi, 1995) fournit de bonnes pistes à ce sujet.

Contrairement aux idées reçues, les méthodes stochastiques telles que les modèles de Markov et les réseaux de neurones sont, a priori, celles qui se rapprochent le plus du comportement humain. D'une part du fait de leur aspect probabiliste, qui intervient aussi chez l'humain dans son aptitude à former des jugements de probabilités en situation d'incertitude. D'autre part, parce que ces techniques sont fondées sur l'apprentissage et l'accumulation de connaissances à partir des situations qui leur sont présentées. Enfin, pour leur possibilité de généralisation, en ce qui concerne les R.N.M. Cependant, même si ces dispositifs sont aptes à intégrer les variabilités de la parole lors de la phase d'apprentissage à partir de grandes bases de données, ils agissent ensuite de façon déterministe en phase de reconnaissance proprement dite. Ils restent, par construction, inaptes à une adaptation dynamique en cours de décodage, comme, par exemple, l'interaction des niveaux d'ordre supérieur venant réorienter l'analyse de façon descendante. La plupart des procédures d'adaptation proposées dans la littérature restent fondées sur un réapprentissage, ce qui s'éloigne d'un aspect dynamique propre au décodage humain. Même si les modèles connexionnistes copient efficacement le fonctionnement du cerveau, ils n'en copient que ce qui donne lieu à ce que Edelman appelle la conscience primaire. A un autre niveau, dans la conscience d'ordre supérieur, la catégorisation symbolique est intimement liée aux mémoires conceptuelle et symbolique, donc aux connaissances. L'information y est traitée par corrélation, à plusieurs niveaux d'information différents. S'en tenir, dans ces conditions, aux méthodes stochastiques équivaudrait à confondre la mémoire avec les mécanismes nécessaires à son établissement (Edelman, 1992). Le cerveau est un manipulateur de symboles (Millikan, 1984) ; la parole est un échange de symboles dont les valeurs sont fondées sur les connaissances acquises par apprentissage et sur une perpétuelle recatégorisation des stimuli sensoriels variables à l'aide des connaissances. Le fonctionnement du cerveau et du langage fonde ainsi la légitimité des systèmes à base de connaissances.

IV.1.C. Vers une imitation du traitement cognitif: la parallélisation et la modularité

Le cortex cérébral est organisé sur un ensemble de « cartes » qui constituent des structures locales stratifiées fortement interconnectées, dont les connexions sont massivement réentrantes (Edelman, 1992). Un même stimulus arrive parallèlement sur plusieurs cartes. La catégorisation provient des connexions réentrantes d'une part entre ces cartes, d'autre part entre ces cartes et les mémoires. La catégorisation est donc le résultat de corrélations entre les informations traitées en parallèle. Il semble à présent admis que les tâches de reconnaissance au niveau cortical soient réalisées par un ensemble de processus réalisant une analyse sur différents niveaux (Fodor, 1983 ; Minsky, 1985). Inspirés par la théorie du *Pandémonium*, Lindsay et Norman ajoutent que ces différents modules sont eux-mêmes composés d'agents spécialisés appelés *démons* (Lindsay & Norman, 1977). Chaque agent est responsable d'une tâche spécifique. Les auteurs distinguent alors les agents de bas niveaux, les analyseurs de propriétés, les démons cognitifs et enfin le module de décision (Figure 62). Tous ces processus peuvent communiquer entre eux par l'intermédiaire d'un tableau noir qui représente une zone de dialogue commune sur laquelle chaque démon lit et écrit des résultats. Un module *superviseur* gère les informations transitant par cette mémoire.

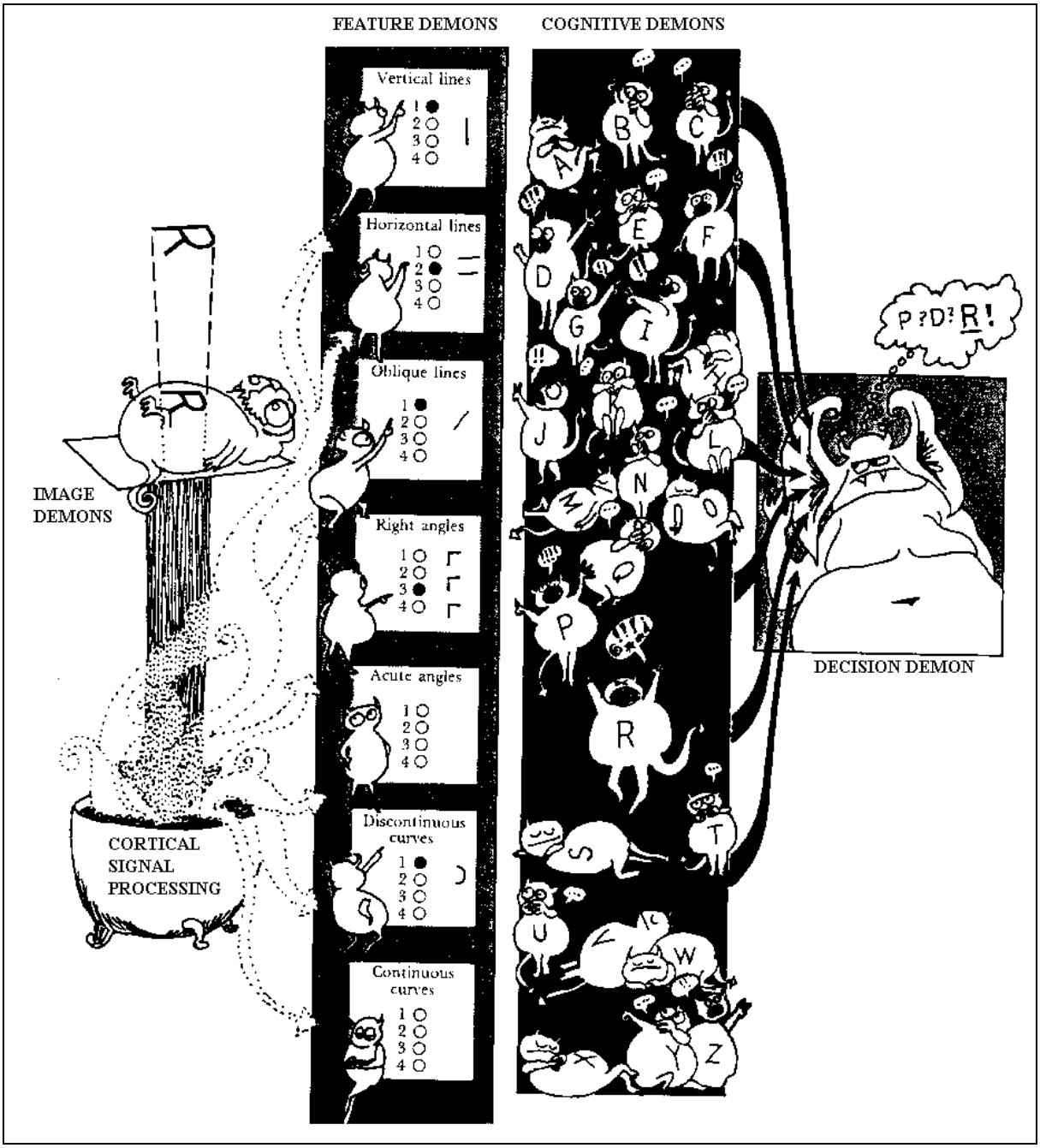


Figure 62: Parallélisme et modularité des processus de reconnaissance dans le cas de l'image (Source: Lindsay & Norman, 1977, p.266)

Ces théories sont sujettes à controverse, notamment sur l'encapsulation des processus. Nous ne rentrerons pas dans ce débat de spécialistes. Nous retiendrons, par contre, deux points qui semblent importants:

- le fait que le processus de reconnaissance soit fragmenté en sous tâches. Nous appelons cela la modularité.
- le fait que ces processus agissent en parallèle

Vouloir imiter ce mode de fonctionnement conduit à la mise en oeuvre de systèmes multi-analyses et multi-experts contrôlés par un moteur d'inférences. Ce type de fonctionnement est déjà connu (Giraud, 1991; Caelen et al., 1994). Nous nous sommes orientés dans la même direction.

IV.2. Présentation du système « ACHILE »

IV.2.A. Présentation générale

ACHILE est un système de décodage ACoustico-pHonétique et d'Identification LExicale (Ghio & Rossi, 1995b, 1995c). Il est fondé sur une approche modulaire et paralléliste s'inspirant, d'un point de vue fonctionnel, des observations faites précédemment sur les traitements cognitifs humains. Nous prétendons non pas imiter de tels processus, ceux-ci étant forts complexes et encore mal connus, mais simplement nous en inspirer. Nous pensons que la réussite d'un système de reconnaissance de la parole reste dépendante de la bonne collaboration des différentes sources d'informations disponibles dans le cadre de la communication parlée. Le projet MICRO (Caelen et al., 1994) nous semble séduisant car il intègre toutes sortes de connaissances à la fois acoustiques, phonétiques, phonologiques, lexicales, syntaxiques, sémantiques et aussi prosodiques. De tels systèmes sont très difficiles à mettre en oeuvre, ce qui explique le manque d'enthousiasme pour développer de tels dispositifs. De plus, cela nécessite la collaboration d'un grand nombre de compétences qu'il est bien souvent difficile de réunir dans un projet fédérateur.

IV.2.B. L'architecture du système

Les problèmes d'architecture logicielle restent mineurs. Toutefois, les considérer s'avère judicieux pour ne pas handicaper le développement progressif du dispositif. Bien souvent, la méthode impose l'architecture, mais par la suite, l'architecture « fige » la méthodologie. Ainsi, actuellement, la plupart des systèmes de communication homme-machine sont conçus autour de deux modules: un processus de Reconnaissance Automatique de la Parole (R.A.P.), un autre de Traitement Automatique des Langues Naturelles (T.A.L.N.). Ces deux processus sont quasiment indépendants, le premier alimentant le second de façon unidirectionnelle. Toute volonté d'établir des interactions entre les niveaux d'analyse (la syntaxe pouvant par exemple lever des ambiguïtés phoniques, cf. « L'interaction des sources d'information », p.46) se heurte à l'architecture adoptée.

L'architecture générale du dispositif « ACHILE » est exposée en Figure 63, p.129. La tâche de reconnaissance est effectuée par répartition du travail et interaction d'une société d'experts. Le signal de parole alimente tout d'abord les démons de bas niveaux d'analyse, c'est à dire les agents responsables de l'extraction de paramètres acoustiques (énergies, taux de passage par zéro, voisement, analyse spectrale...). Ces calculs sont suivis d'une évaluation de propriétés, d'indices et de traits phonétiques qui affinent l'analyse et permettent de s'affranchir d'une partie de la variabilité. Les informations sont ensuite transmises à plusieurs modules de décodage fonctionnant en parallèle.

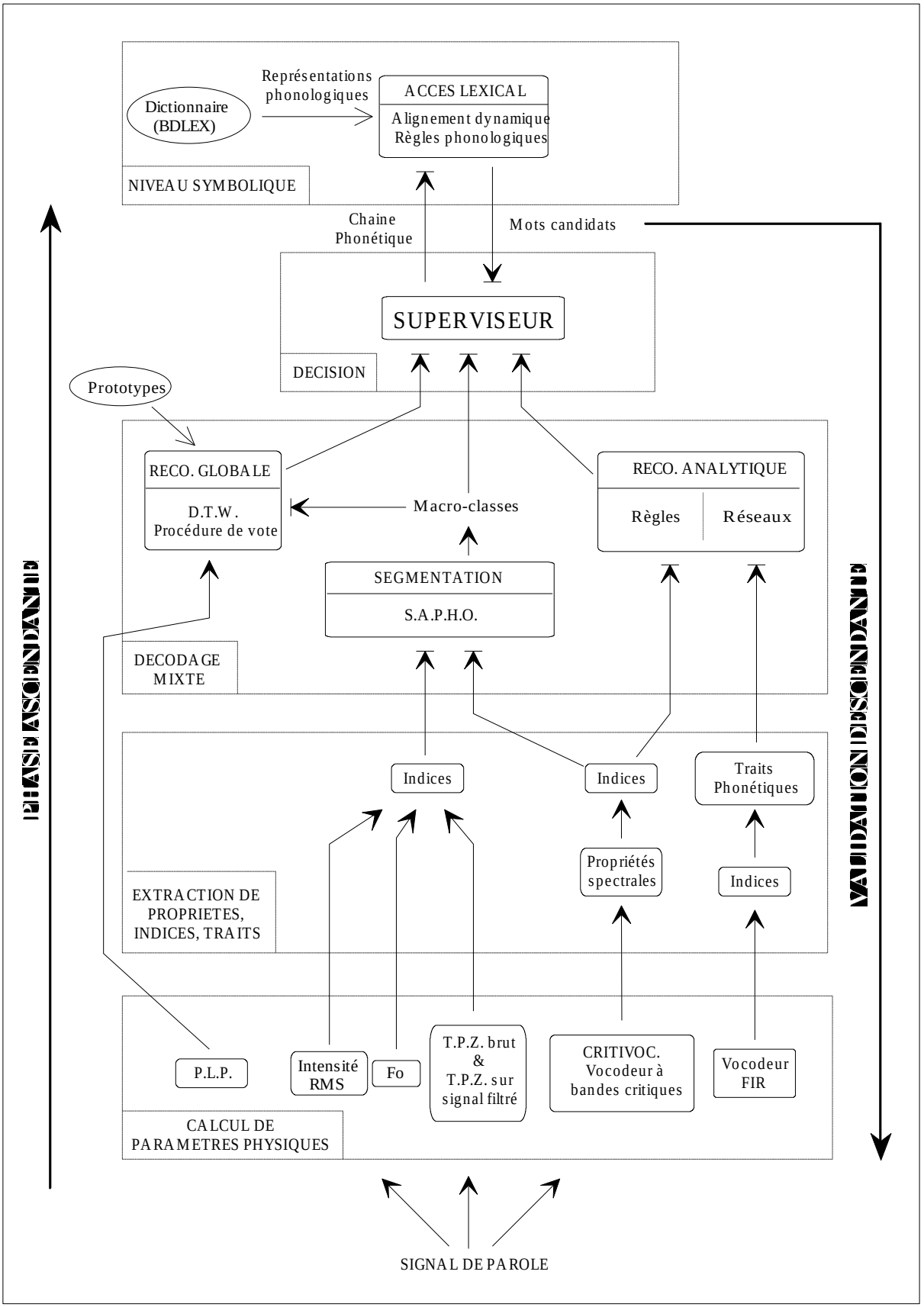


Figure 63: Schéma fonctionnel d'ACHILE

Ces modules sont les suivants :

- la *segmentation automatique* avec l'algorithme SAPHO, qui fournit un premier étiquetage des segments en macro-classes.
- la *reconnaissance globale* qui utilise la technique P.L.P. (Prédiction Linéaire sur un spectre Perceptif) pour extraire l'information spectrale puis un algorithme d'alignement dynamique permettant de comparer le stimulus à des références.
- un *premier module de reconnaissance analytique* fondé sur un ensemble de règles.
- un *second module de reconnaissance analytique* constitué d'un ensemble de graphes orientés à transitions d'état construits pour modéliser tous les allophones d'une voyelle. Cette partie ne sera pas développée dans cette étude mais peut être consultée dans (Ghio & Rossi, 1994)

Dans une approche multimodale, on pourrait éventuellement associer en juxtaposition avec le canal oral, un décodage du canal visuel. De par leur conception diverse, ces processus fournissent des informations différentes. Les résultats sont ensuite transmis aux démons cognitifs qui agissent en utilisant des connaissances de type symbolique (représentations phonologiques, accès lexical...) Un moteur d'inférences se charge tout d'abord de construire une ou plusieurs chaînes phonétiques « prétendantes » en fusionnant les données de l'analyse montante. L'accès à un dictionnaire lui permet ensuite de proposer une liste de mots-candidats. Une phase de vérification descendante est prévue dans le but de valider, infirmer ou réinterpréter le classement des candidats. Chacun des éléments du dispositif sera détaillé dans les chapitres ultérieurs. Il faut signaler qu'une telle architecture permet d'exploiter les redondances de la communication parlée, chaque module de décodage fournissant un résultat conforté ou infirmé par ses voisins (notion d'agents compétitifs). De plus, elle autorise la levée d'ambiguïtés et la résolution de problèmes parfois non résolus dans un fonctionnement linéaire. Ainsi, le repérage des frontières phoniques s'effectue par décision émergente, des retours arrière venant modifier progressivement la segmentation en fonction du contexte.

Un tel fonctionnement apparaît intéressant mais engendre un certain nombre de problèmes, notamment au niveau de la gestion simultanée de toutes les informations. Une bonne synchronisation devient vite nécessaire.

IV.3. La synchronisation de l'information

Le décodage ascendant repose sur l'extraction de paramètres physiques (Figure 63, p.129). Dans notre système, cette extraction s'effectue par trames d'analyse. Par exemple, l'intensité RMS est évaluée en intégrant 10 ms de signal toutes les 10 ms; le vocodeur à bandes critiques réalise une analyse sur des fenêtres de 20 ms décalées de 10 ms (recouvrement de 50%); le voisement peut être estimé sur des tranches de signal de 25, 30, voire 40 ms... L'utilisation combinée de ces paramètres pose alors un problème de synchronisation des mesures que ce soit au niveau de la fréquence des mesures, aussi bien que de la phase.

Le premier problème est relatif au pas de trames, c'est à dire au décalage entre deux mesures successives. En théorie, pour certaines transformées, il existe une relation stricte

entre longueur de la fenêtre d'analyse et fréquence d'échantillonnage des mesures (cf. § « Le pas de décalage de la fenêtre d'analyse », p.103). Nous avons suivi ces considérations.

Pour régler le problème du pas de trames, il existe deux solutions:

- évaluer chaque paramètre sans contrainte sur la fréquence de mesure puis interpoler pour être capable de fournir une valeur à des instants communs à tous les paramètres. Cette méthode est souvent adoptée dans les dispositifs de visualisation et d'édition de signal comme MES (Hirst et al., 1995; Espesser, 1996) ou Phonédit* .
- évaluer chaque paramètre avec une fréquence de mesure commune.

Nous avons choisi la deuxième solution pour sa simplicité. Le pas de trame choisi est de 10 ms (modèle centiseconde).

Le second problème de synchronisation est relatif au déphasage des mesures. La première valeur estimée d'un paramètre est locale au centre de la première fenêtre d'analyse de ce paramètre (Figure 64). Ainsi, la première valeur de l'intensité RMS est mesurée sur une section de signal centrée au temps t=5 ms (si la fenêtre est de 10 ms) ; le vocodeur à bandes critiques, qui réalise une analyse sur des fenêtres de 20 ms, fournit la première répartition spectrale sur une fenêtre centrée au temps t=10 ms; la détection du voisement, estimé sur des tranches 25 ms, propose un premier résultat valable pour le temps t=12.5ms... Les paramètres se retrouvent finalement en déphasage. A ces décalages propres à l'évaluation par trames s'ajoutent des écarts introduits par certaines techniques de filtrage. Ainsi, pour des raisons de causalité (Bellanger, 1980), un signal issu d'un filtrage de type FIR est décalé par rapport au signal d'entrée, ce qui peut, à nouveau, donner lieu à des incohérences sur les paramètres calculés sur ce signal filtré. Pour assurer une synchronisation des différents paramètres physiques, la solution consiste à centrer la première fenêtre d'analyse de chaque paramètre sur le temps t=0 du signal analysé. Une solution simple pour obtenir cette condition est d'ajouter au début de signal une amorce de longueur égale à une demie fenêtre d'analyse, composée d'un faible bruit blanc (Figure 65).

Une gestion rigoureuse de l'information temporelle est assurée ainsi au niveau du décodage acoustico-phonétique, partie cruciale du dispositif.

* logiciel développé par la société S.Q.Lab

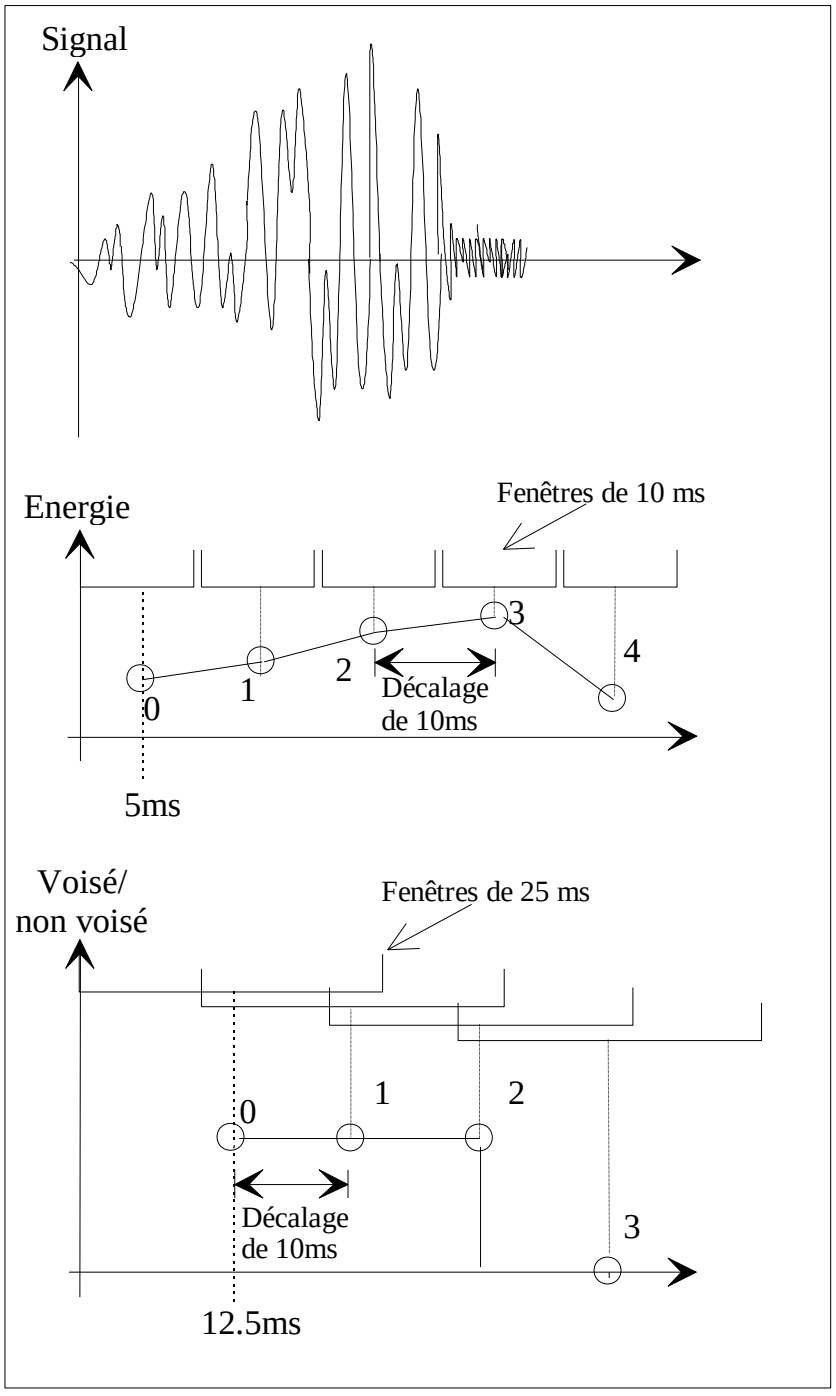


Figure 64: Les problèmes de synchronisation des paramètres physiques

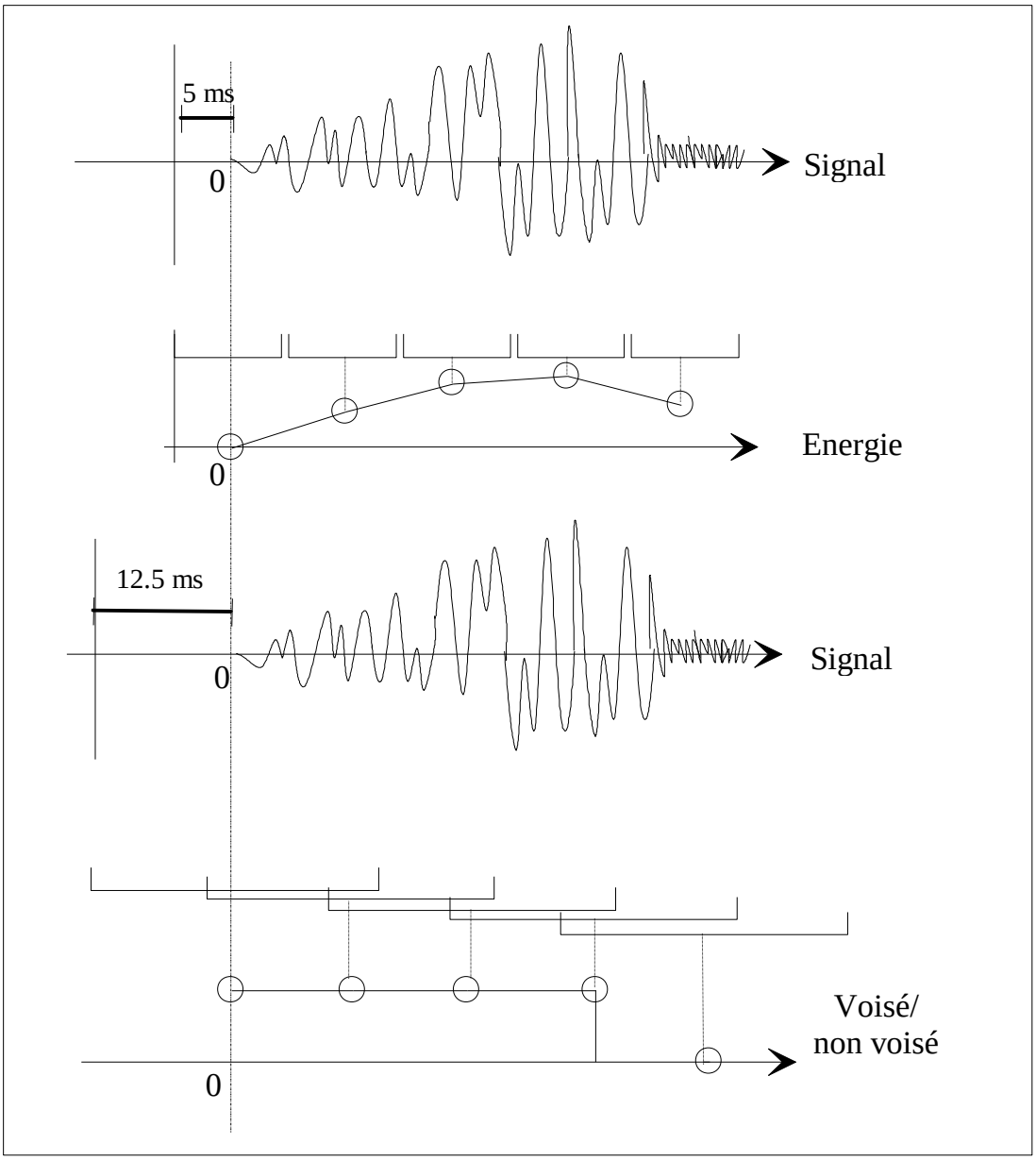


Figure 65: Mise en phase de l'extraction des paramètres

IV.4. Un système de reconnaissance fondé sur un Décodage Acoustico-Phonétique

IV.4.A. La nécessité d'un Décodage Acoustico-Phonétique

A partir des paramètres physiques calculés par les démons de bas niveaux et synchronisés, le système « ACHILE » effectue un décodage qui consiste à transformer l'information acoustique de la parole en une suite de phonèmes. Il s'agit de la phase de Décodage Acoustico-Phonétique (DAP). Ce passage du niveau physique au niveau symbolique nous semble absolument nécessaire car seule la manipulation d'unités discrètes en

nombre limité tels que le sont les phonèmes, reste envisageable pour pouvoir traiter des situations qui aspirent à :

- l'indépendance du locuteur
- la parole continue
- un grand vocabulaire
- les variantes stylistiques, sociolectales, dialectales...

Nous ne prétendons pas aborder de front tous ces problèmes, qui restent non résolus. Nous tenons seulement à signaler l'importance du DAP. Nous voulons aussi mettre en place une architecture qui d'emblée n'exclut pas la prise en compte de ces conditions.

IV.4.B. La difficulté du Décodage Acoustico-Phonétique

D'un point de vue quantitatif, nous avons vu au § II.II.1.II.1.C.II.1.C.d., p.45, que le décodage acoustico-phonétique est une opération qui consiste à compresser le débit d'information d'un rapport 4000 (de $D_{physique}$ = 200000 bits/s à $D_{phonétique}$ = 50 bits/s). Une telle tâche représente un immense défi. D'un point de vue qualitatif, cette transformation apparaît aussi très délicate. En effet, il s'agit de réaliser un transcodage entre le monde physique continu et le domaine abstrait des représentations mentales de type symbolique. Diverses études (Klatt, 1977; Liénard, 1977) mettent en avant l'importance du DAP dans un système de reconnaissance de la parole. Il semble qu'un dispositif dont le taux de décodage phonétique serait inférieur à 85 % ne pourrait espérer accéder à la compréhension du message oral malgré l'apport des niveaux supérieurs (lexique, syntaxe, sémantique). Ces niveaux sont souvent organisés pour apporter des contraintes descendantes.

IV.4.C. Les solutions

IV.4.C.a. Une gestion étroite entre axes syntagmatique et paradigmatique

Le DAP englobe deux tâches différentes qui peuvent être menées en parallèle ou en séquence (Haton, 1989):

- une tâche de segmentation du signal vocal en unités élémentaires (axe syntagmatique)
- une tâche d'étiquetage phonétique de ces segments (axe paradigmatique)

Nous pensons que les deux opérations de segmentation et d'étiquetage doivent être menées de front. En effet, l'identification est toujours meilleure quand on compare des voyelles à des voyelles et des consonnes à des consonnes, ce qui nécessite donc un repérage temporel de ces unités dans la chaîne phonique. Or, la segmentation apparaît plus efficace quand elle se fonde sur la nature des segments. Cela conduit à une situation paradoxale. La solution pour sortir de ce cercle vicieux réside dans le fait que certains indices, plus résistants au contexte, permettent de construire un contexte partiel (Rossi, 1980).

IV.4.C.b. La notion de candidats multiples: une stratégie à décision différée

Compte tenu du champ de dispersion des réalisations phonémiques, il apparaît illusoire d'espérer catégoriser de façon unique un segment phonétique. Il nous semble préférable d'obtenir un résultat imprécis mais juste plutôt que de proposer une catégorisation fine mais peu robuste. Cela conduit les différents modules de décodage à proposer très souvent plusieurs candidats et à différer la décision dans une étape ultérieure. Cela rejoint la théorie de la « sélection clonale » décrite dans (Rossi, 1995). Cette démarche reste en accord avec le principe de GRICE en vertu duquel il est préférable de ne pas induire un agent en erreur, ceci en maintenant un ensemble d'hypothèses ouvertes mais en nombre limité. Il s'agit de trouver un compromis entre fiabilité et économie. Une solution consiste, par exemple, à catégoriser une unité phonétique par une macro-classe d'appartenance (ex: fricative sourde) plutôt que par un phonème précis dans le cas où l'information montante est insuffisante.

Détaillons à présent chaque module.

V. LE MODULE DE SEGMENTATION ET DE MACRO-CLASSIFICATION « S.A.P.H.O. »

« Il est un âge où l'on enseigne ce que l'on sait; mais il vient ensuite un autre où l'on enseigne ce que l'on ne sait pas: cela s'appelle chercher.»

Roland Barthes.

Plan du chapitre

RESUME

L'étape de segmentation consiste à découper le continuum acoustique de la chaîne parlée en unités discrètes. Ce découpage étant délicat, il paraît plus juste de parler de « repérage temporel ». Cette analyse syntagmatique ne peut se faire a priori sans un minimum d'information paradigmatique. L'idée est de dégager progressivement les formes phonétiques en affinant le repérage temporel par étude du contexte, ce qui nécessite une catégorisation grossière des segments phoniques.

Notre algorithme SAPHO, de Segmentation Automatique à partir de connaissances PHOnétiques, s'apparente à un système expert à règles de production. Il est fondé sur le calcul de paramètres acoustiques, la déduction de propriétés et l'application d'un ensemble de règles simples du type SI [tel phénomène] AVEC [tel contexte] ET [telle contrainte] ALORS [c'est ça]. Le résultat est un étiquetage des trames d'analyse en macro-classes (voyelle, occlusive, fricative, consonne vocalique...) La mise au point du module SAPHO a été longue et a nécessité de longues séances de travail pour formaliser le savoir d'experts phonéticiens. Nous avons tenté d'éviter de tomber dans le piège de la mise au point « ad hoc » par l'adoption de paramètres ou règles sans fondement physique ou phonétique. A chaque calcul ou transformation correspond une réalité physique ou phonétique. L'obtention des différents seuils intervenant dans le module SAPHO a été effectuée sur un corpus de 40 mots comprenant 5 locuteurs. Il s'agit de mots isolés empruntés au travail de (Giraud, 1991). Ce corpus a ensuite été écarté des évaluations.

Réflexion préliminaire

En théorie, l'opération dite de « segmentation » consiste à découper le continuum acoustique de la chaîne parlée en unités discrètes (cf. § « La segmentation », p.19). Cette étape, délicate, s'avère nécessaire dans le cas d'un décodage de grand vocabulaire où il paraît judicieux de décomposer les mots en sous-unités plus petites dont le nombre est réduit. C'est, en plus, un pas en avant pour le traitement de la parole continue.

V.1. Le problème épineux de la segmentation

Comme le souligne Malmberg (Malmberg, 1979, p.1), « si nous avions à notre disposition un nombre illimité d'unités infiniment variables, aucune communication organisée ne serait possible ». Cette constatation plaide en faveur de l'existence d'unités fonctionnelles en nombre limité, qui, associées en séquences, vont permettre de générer une combinatoire suffisante à l'établissement d'un vaste vocabulaire. Le décodage de la parole passe donc par le repérage et l'identification de ces éléments de base.

On sait qu'il n'existe pas de correspondance biunivoque entre le plan de description acoustique et les niveaux linguistiques supérieurs. En effet, là où la linguistique met en évidence l'existence d'unités abstraites et discrètes telles que les mots ou les phonèmes, l'observation du signal acoustique ne laisse apparaître qu'un continuum, résultat de la mise en forme de la chaîne phonétique (cf. § « La segmentation », p.19). Aussi, le « terme (segmentation) est dangereux car il laisse croire qu'on peut trouver dans le signal les limites entre les phonèmes... Par la segmentation, on repère en réalité les zones de stabilité et d'instabilité du spectre... Les segments obtenus ne sont ni des représentants des phonèmes, ni même des unités phonétiques: ce sont des moments privilégiés pour extraire des indices et des traits.» (Rossi, 1980, p.104).

Le problème majeur de la segmentation est le suivant. La localisation des séparations éventuelles entre les unités à reconnaître nécessite un certain nombre d'indices. Or, ces indices varient en fonction de l'environnement. Pour pouvoir les obtenir, il est donc nécessaire d'identifier le contexte. Mais pour identifier le contexte, il faut d'abord segmenter le signal... L'une des solutions pour sortir de ce cercle vicieux consiste à effectuer une segmentation en plusieurs passes en s'appuyant sur des points d'ancrage appelés îlots de confiance repérés à partir d'indices plus résistants aux effets contextuels (Rossi, 1980). L'idée est de dégager progressivement les formes phonétiques en partant des plus évidentes pour arriver aux plus subtiles. Comme le propose Haton dans (Haton, 1989, p.428), nous pensons qu'il est nécessaire de mettre en place « une interaction aussi étroite que possible entre segmentation et étiquetage phonétique, une stratégie de décision retardée consistant à repousser des décisions définitives de segmentation le plus tard possible de façon à rassembler le maximum d'éléments décisifs ». Telle est notre démarche.

V.2. Présentation générale de S.A.P.H.O.

S.A.P.H.O. est un algorithme de Segmentation Automatique à partir de connaissances PHOnétiques. Il est fondé sur le calcul de paramètres acoustiques, la déduction d'indices et l'application d'un ensemble de règles organisées en système expert.

V.2.A. Un algorithme à base de connaissances

Afin d'identifier certaines discontinuités évidentes comme le passage d'un consonne occlusive à une voyelle, nous devons connaître la structure phonétique des occlusives et des voyelles ainsi que les règles de transformation du niveau phonétique aux niveaux articulatoire et acoustique. Par conséquent, l'accès aux unités abstraites de haut niveau est loin d'être direct et exige une connaissance approfondie du codage de la parole. Cette exigence est d'ailleurs admise par beaucoup. Ainsi, Zue affirme (Zue, 1982) que « la formalisation de la connaissance acoustique et phonétique est l'obstacle majeur pour le développement de systèmes de reconnaissance de la parole élaborés susceptibles d'approcher les performances humaines ».

L'algorithme de Segmentation Automatique à partir de connaissances PHOnétiques fournit un ensemble hiérarchisé de propriétés et de segments acoustiques et phonétiques congruents avec les unités phonétiques et leur structure interne (Rossi, 1990). Il ne s'agit pas d'une méthode aveugle de segmentation a priori fondée sur une fonction d'instabilité comme le proposent, par exemple (Barras et al., 1994; Wesfreid & Wickerhauser, 1994). Dans SAPHO, l'énergie, le taux de passage par zéro et certaines caractéristiques spectrales permettent l'identification sommaire des segments phonémiques. Seule la nature de ces éléments autorise a posteriori la localisation potentielle des frontières en tenant compte des effets de contexte. Le résultat est un étiquetage des trames d'analyse en macro-classes (voyelle, occlusive, constrictive, consonne vocalique, silence...).

V.2.B. Architecture de l'algorithme

Une première version de SAPHO a été présentée dans (Rossi, 1990). Depuis, de nombreuses modifications ont été apportées, notamment dans la rigueur et la validité physique des paramètres calculés et règles proposées. De plus, bien que non spécialiste d'Intelligence Artificielle, nous avons orienté l'architecture vers une structure en système expert comme le proposent (Le Beux & Fontaine, 1986). Nous avons adopté la représentation par règles de production, d'une part parce qu'elles sont aptes à exprimer des informations variées et à traiter des situations complexes, d'autre part parce que leur formalisme est séduisant et d'une relative simplicité. Tout fragment de savoir est exprimé sous la forme:

Si P1 & P2 & ... Pn Alors A

où les facteurs Pi sont les prémisses à une situation, A étant déclenchée lorsque toutes les prémisses sont satisfaites

La conjonction de prémisses est autorisée par l'utilisation de « ou »:

Si P1 ou P2 Alors A

Le déclenchement d'une règle modifie généralement la base de faits lors d'une consultation. L'action d'une règle est fréquemment reprise comme prémisse d'une autre règle. L'option adoptée est de chaîner les règles entre elles. Ce principe de chaînage s'étend à l'ensemble des règles de la base. Une règle qui n'admet pas de prédécesseur est dite « initiale ». Ce choix de représentation appelle la précision suivante: l'architecture est construite a priori, dès la

formalisation des règles et non en phase d'utilisation, comme cela peut être le cas dans certains systèmes experts, où le moteur d'inférences crée lui-même les liens entre les règles.

V.3. Les différentes étapes de la segmentation par S.A.P.H.O.

Il est vrai que la possibilité de réplication d'un travail est fondamentale pour le valider. Un organigramme complet de SAPHO serait le bienvenu. Toutefois, nous n'avons pas voulu tomber dans le piège de la rédaction d'une notice technique. C'est pourquoi, dans ce chapitre, nous ne rentrerons pas dans les détails de l'algorithme. Par soucis de concision, nous n'aborderons que les lignes directrices du processus de segmentation et macro-classification automatique (Figure 66).

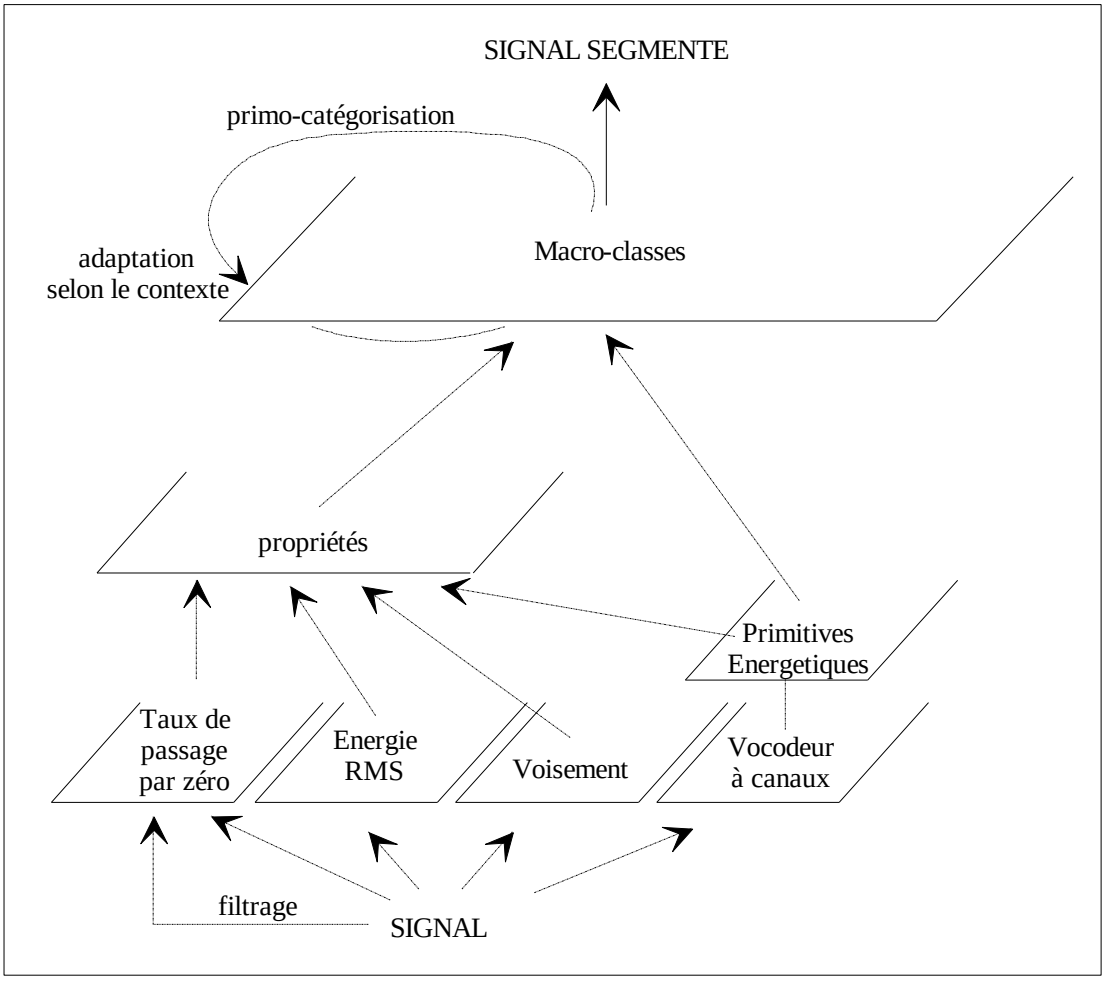


Figure 66: Schéma fonctionnel de S.A.P.H.O.

V.3.A. Le calcul des paramètres acoustiques

La première étape de la segmentation consiste à extraire un ensemble de données acoustiques qui sont les suivantes:

- l'intensité sonore
- le taux de passage par zéro
- une répartition spectrale par bandes sur lesquelles sont déduits de nombreux paramètres

V.3.A.a. L'intensité sonore

L'utilisation de l'intensité sonore est une opération délicate en traitement numérique des signaux. En effet, autant la mesure de l'intensité peut être fiable lorsqu'elle est prise par des dispositifs électroniques calibrés comme dans le système EVA (Teston & Galindo, 1995), autant elle s'avère inexacte si elle est évaluée après numérisation sur un calculateur. En effet, à moins de contrôler parfaitement la chaîne d'acquisition (cf. § « L'acquisition », p.17), les valeurs calculées a posteriori sur le signal de parole numérisé dépendront des niveaux d'enregistrement successifs. Les valeurs brutes de l'énergie n'étant pas fiables, il est nécessaire d'effectuer une normalisation de ce paramètre après calcul.

Le calcul de l'énergie RMS (Root Mean Square) consiste à sommer le carré de l'amplitude des échantillons du signal de parole sur une durée d'intégration de l'ordre de la centiseconde, puis à prendre la racine carrée.

$$E_i = \sqrt{\sum_{n=n_0}^{n_0+N} s(n)^2}$$ où

s(n) est le signal de parole
n_0 est l'indice du premier échantillon de la i^{ème} trame
N est la durée d'intégration en échantillons

L'opération est renouvelée à chaque trame d'analyse. Des tests préliminaires nous ont montré qu'il est préférable d'exprimer l'intensité en linéaire plutôt qu'en dB, cette mesure logarithmique réduisant trop fortement les contrastes sonores.

Pour effectuer l'opération de normalisation, l'algorithme commence par éliminer du calcul, 5% des trames les plus énergétiques pour exclure les éventuels cas singuliers. Une recherche de l'énergie maximale parmi les trames non exclues par ce tri préliminaire permet de fixer le niveau de référence. Chaque énergie est ensuite exprimée en pourcentage par rapport à ce niveau de référence. Nous noterons dorénavant ce paramètre par *[RMS]*.

La dérivée de ce paramètre normalisé peut être utile pour repérer les pics, les creux ou les sauts d'énergies. Nous le notons *[dRMS]*

Le paramètre [RMS] est intéressant pour effectuer une discrimination entre le signal et le « silence », par effet de contraste. Le silence apparaît non seulement avant et après l'énoncé, mais aussi lors de la tenue d'occlusives sourdes (cf. /t/, Figure 67). L'énergie normalisée permet aussi de discriminer les segments vocaliques des segments consonantiques, ces derniers étant souvent, par contraste, moins énergétiques (cf. /l/ et /ʒ/ par rapport à /e/, /i/ /ã/, Figure 67).

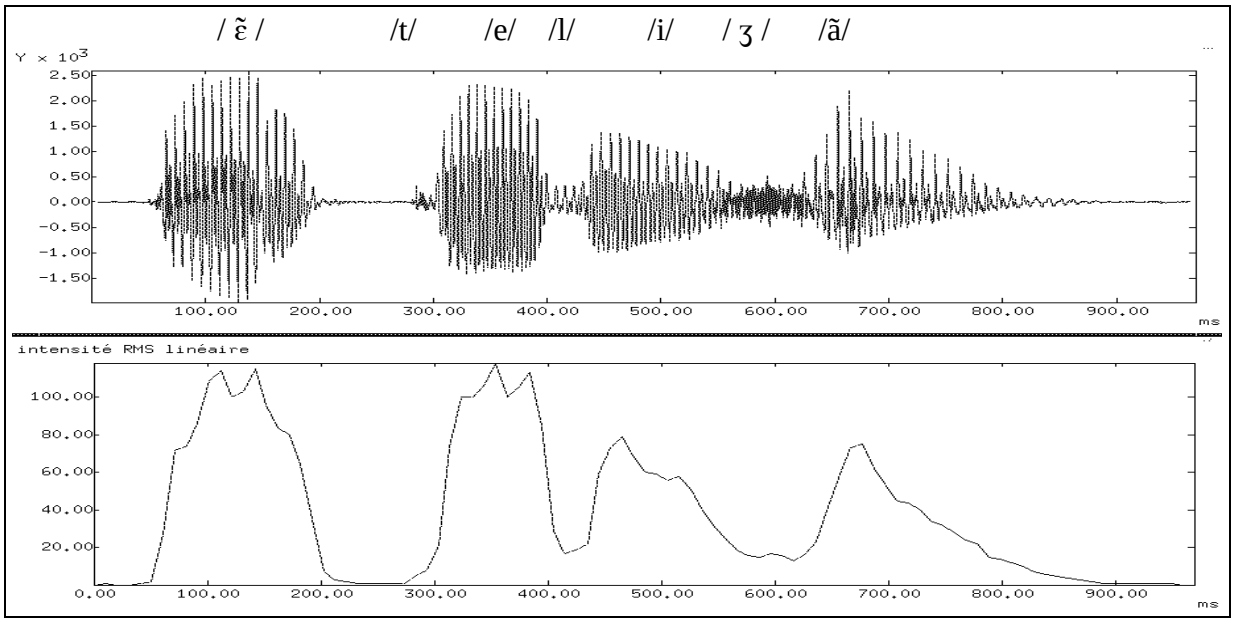


Figure 67: Signal et intensité RMS linéaire normalisée sur l'énoncé "intelligent", locuteur masculin.

V.3.A.b. Le taux de passage par zéro (T.P.Z.)

Le calcul du taux de passage par zéro consiste à compter le nombre de fois, par trame d'analyse, où le signal coupe l'axe des abscisses. Cette donnée, apparemment fort simpliste, est excellente pour repérer les phénomènes de friction, qui sont des bruits extrêmement variables. Ceux-ci peuvent intervenir dans la production des constrictives (cf. /s/, Figure 68) ou à l'explosion d'occlusives sourdes au contact d'une voyelle aiguë comme /i/, /y/ ou /e/ (cf. /ti/, Figure 68). Une importante erreur d'estimation du TPZ peut être introduite si le signal possède une malencontreuse composante continue souvent liée à une imperfection des convertisseurs analogique-numérique (Ghio, 1992). A ce moment-là, le signal d'un bruit de friction se trouve en dessous ou en dessus de l'axe des abscisses et ne le coupe plus, ce qui entraîne un faible taux de passage par zéro erroné. Pour remédier à ce problème, il est nécessaire d'effectuer un filtrage passe-haut qui neutralise alors la composante continue et rétablit un TPZ normal.

Dans le même ordre d'idée, il faut remarquer que le TPZ ne met en valeur que les phénomènes de friction associés à des segments phoniques non voisées. Sur des segments voisés, le bruit fricatif se superpose au signal pseudo-périodique de voisement, ce qui entraîne un faible passage par zéro. Pour obtenir un indice de friction y compris sur des segments voisés, il est judicieux de filtrer préalablement le signal de façon à réduire l'onde de voisement et permettre ainsi le rehaussement du bruit de friction (Ghio, 1992). Une façon simple d'effectuer ce traitement consiste à dériver le signal, ce qui revient à le filtrer passe-haut. On parle aussi de pré-accentuation des aigus. La transformée en Z d'un tel filtre est: $h(z) = 1 - z^{-1}$
Le calcul du TPZ sur le signal filtré permet alors de repérer les phénomènes de friction y compris sur des segments voisés. Sur la Figure 69, on peut remarquer l'émergence de la courbe de « TPZ sur le signal filtré » au niveau de / ʒ / et à un degré moindre sur /v/, ces deux phonèmes étant considérés comme fricatifs. Par contre, la courbe « TPZ sur signal brut » reste peu discriminante du fait du voisement continu.

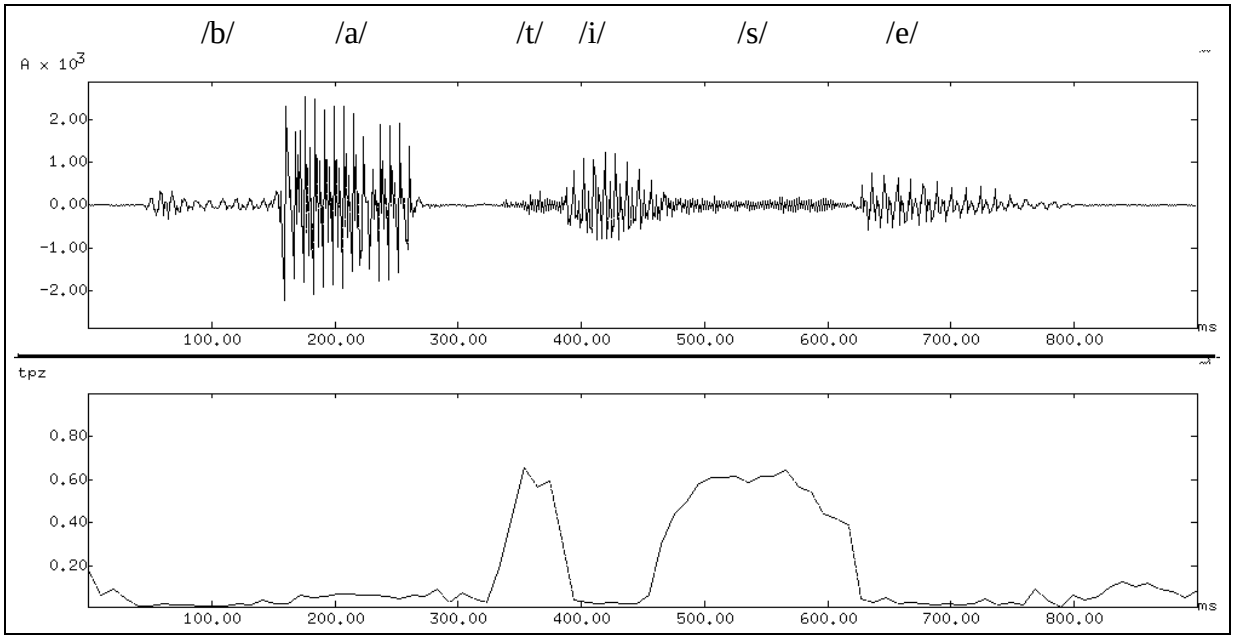


Figure 68: Signal et TPZ sur l'énoncé "bâtissez", locuteur masculin.

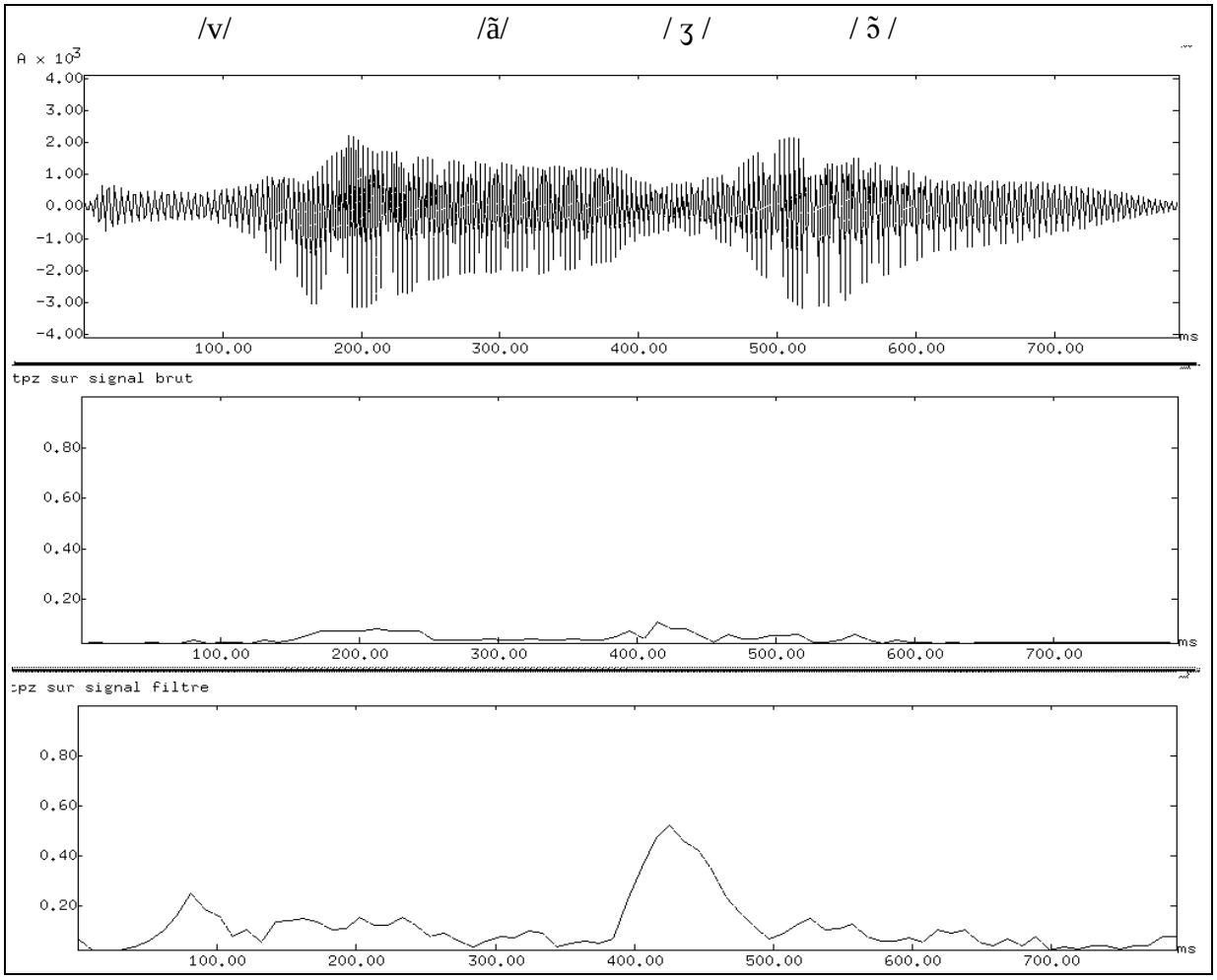


Figure 69: Signal, TPZ sur signal brut, TPZ sur signal filtré sur l'énoncé "vengeons", locuteur féminin.

V.3.A.c. Le vocodeur à canaux

Le rôle du vocodeur à canaux est de construire une représentation temps-fréquence du signal de parole. Dans le système ACHILE, cette estimation de la structure spectrale s'effectue à l'aide du vocodeur en bandes critiques décrit au § III.III.3, p.109. Celui-ci permet de déduire un certain nombre de paramètres spectraux qui sont les suivants.

V.3.A.c.i. L'énergie subjective

L'énergie subjective se calcule, pour chaque trame d'analyse, en sommant les énergies à amplitude linéaire présentes dans les canaux du vocodeur. Une normalisation temporelle du même type que présentée ci-avant pour l'énergie *RMS* permet de s'affranchir, en partie, des phénomènes de variabilité. L'énergie subjective normalisée est notée *[NRJ]*, sa dérivée dans le temps *[dNRJ]*.

Cette composante énergétique apparaît plus intéressante que l'intensité *RMS*. En effet, la pondération sonique opérée par l'analyseur « CritiVoc » permet de réduire l'influence de l'énergie de la fréquence fondamentale (cf. § « La pondération sonique », p.76). Cela réduit la corrélation entre énergie et voisement. Ainsi, une occlusive voisée, qui peut avoir une intensité RMS moyenne n'aura jamais un paramètre *[NRJ]* élevée. Un filtrage par la médiane permet de lisser temporellement ce paramètre *[NRJ]*, ce qui autorise un repérage des extrema d'énergie. La Figure 70 fournit un exemple de courbe d'énergie subjective brute et d'énergie subjective normalisée filtrée par la médiane. On remarque que les maxima correspondent de façon relativement fiable aux voyelles, propriété qui peut être exploitée dans la segmentation. A titre de comparaison, l'intensité RMS est moins fiable, entre autre pour séparer consonnes vocaliques et voyelles.

L'opération de filtrage par la médiane est expliquée ci-après:

> *Soit [s] un signal numérique et [s'] le signal résultant du filtrage par la médiane. Ce traitement s'effectue en classant les 2N+1 valeurs (nombre impair) centrées autour de s(n). La valeur s'(n) est la valeur médiane issue du classement.*
> *Exemple: [s] = {1,3,6,5,10,11,15,12,16,14,12,10,11,8,5}. Prenons un filtrage à 3 valeurs.*
> *s'(1)=médiane de {0,1,3} = 1, s'(2)=médiane de {1,3,6} = 3, s'(3)=médiane de {3,6,5} = 5, s'(4)=médiane de {6,5,10} = 6, s'(5)=médiane de {5,10,11} = 10, s'(6)=médiane de {10,11,15} = 11...*
> *On obtient le signal filtré [s'] = {1,3,5,6,10,11,12,15,14,14,12,11,10,8,5}*

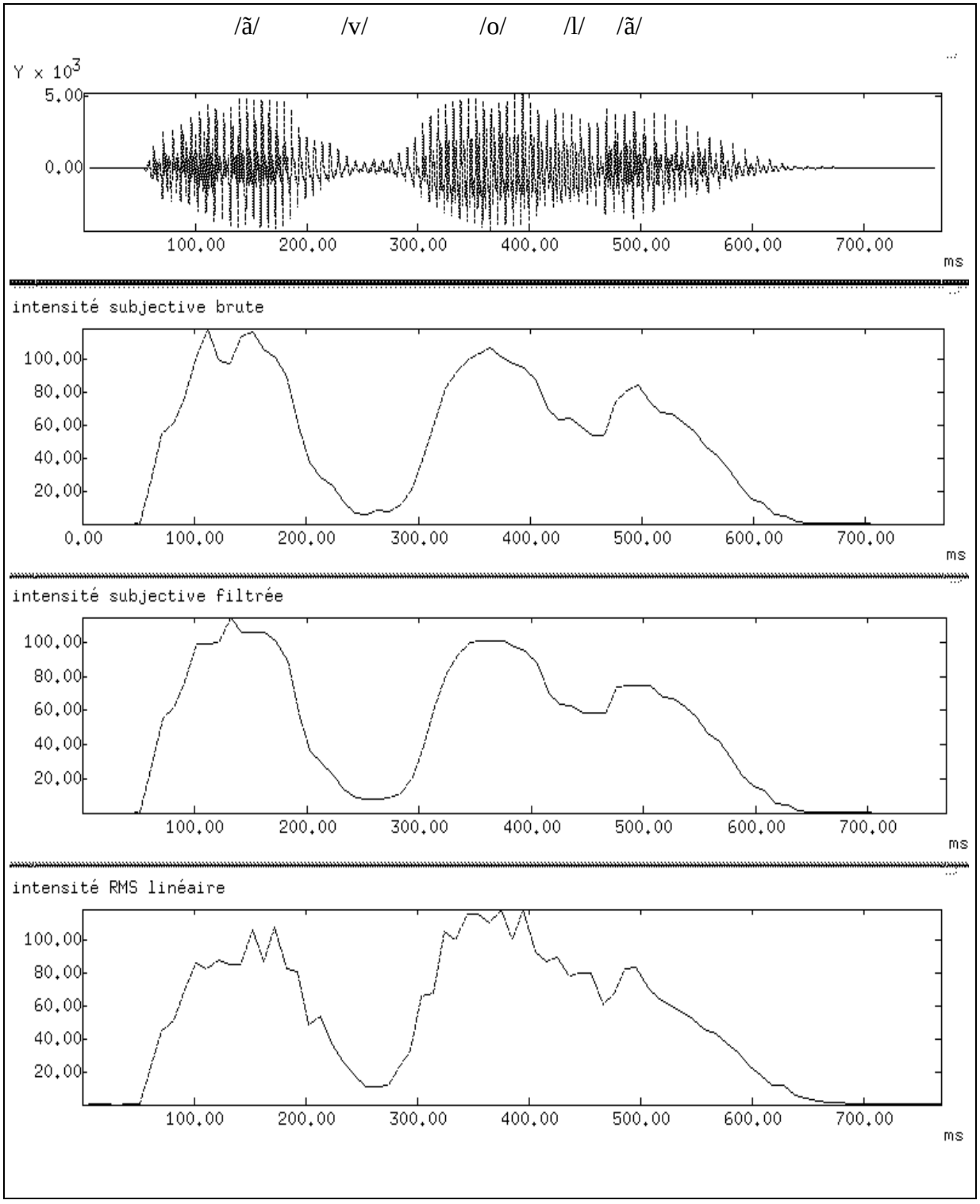


Figure 70: Signal, énergie subjective brute normalisée, énergie subjective normalisée filtrée par la médiane, intensité RMS normalisée sur l'énoncé "envolant", locuteur masculin.

V.3.A.c.ii. L'énergie de voisement

L'énergie de voisement se calcule, pour chaque trame d'analyse, en tenant compte de l'énergie présente dans les premiers canaux du vocodeur. Une normalisation du même type que présentée ci-avant permet de créer une distribution bimodale permettant une catégorisation voisé/non voisé, les segments voisés étant ceux dont l'énergie de voisement est non nulle. La Figure 71 fournit un exemple de courbe d'énergie de voisement ainsi qu'une courbe de détection du voisement par des techniques de « clustering » (Espesser, 1981). On se rend compte que les différences entre les deux méthodes sont faibles.

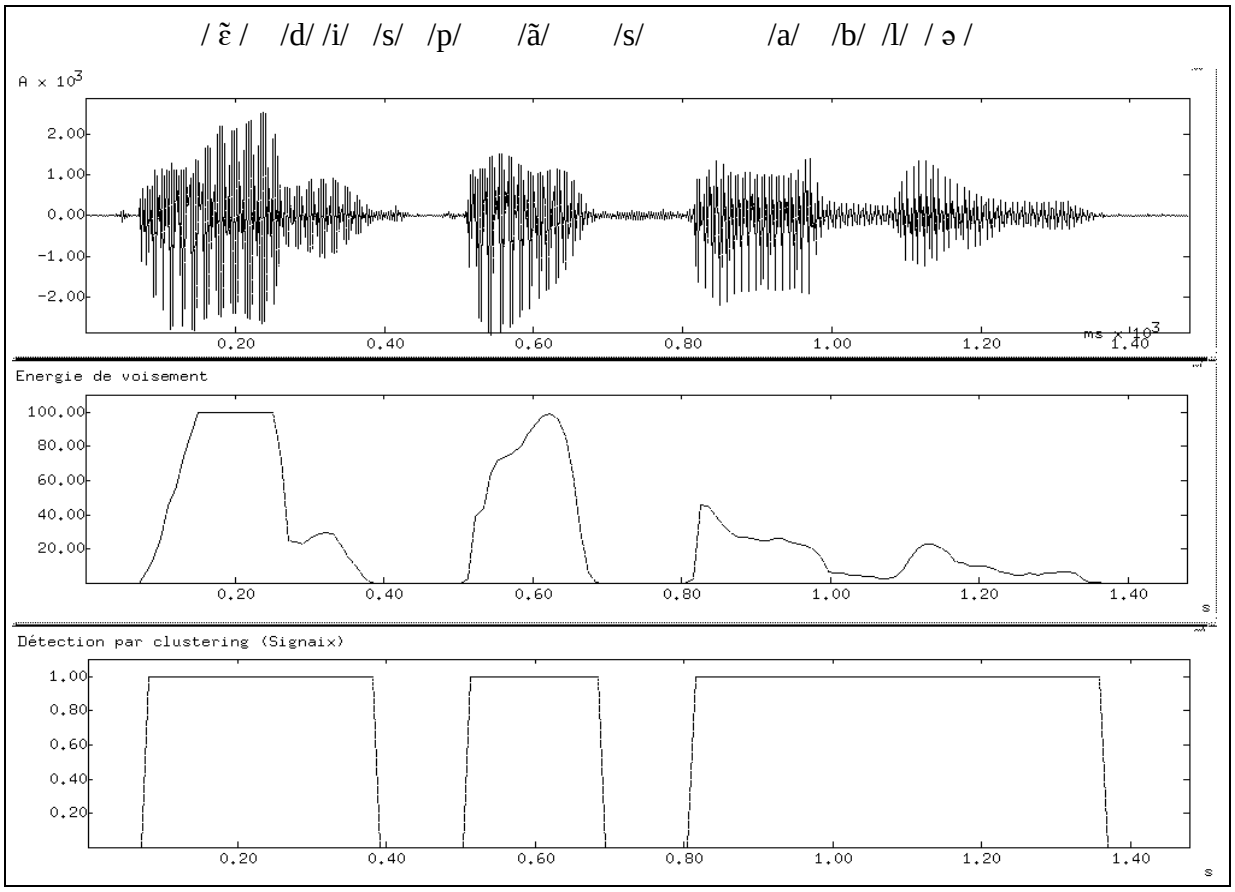


Figure 71: Signal, énergie de voisement normalisée et détection de voisement par «clustering» sur l'énoncé "indispensable", locuteur féminin.

V.3.A.c.iii. L'énergie formantique

L'énergie formantique se calcule, pour chaque trame d'analyse, en tenant compte de l'énergie présente dans les canaux médians du vocodeur, c'est à dire pour une plage de fréquences comprises entre 600 Hz et 3000 Hz. Une normalisation permet de mettre en évidence des zones à fortes ou faibles énergies formantiques. La Figure 72 fournit un exemple de courbe d'énergie formantique. On peut remarquer la très faible valeur de celle-ci au niveau des occlusives sonores.

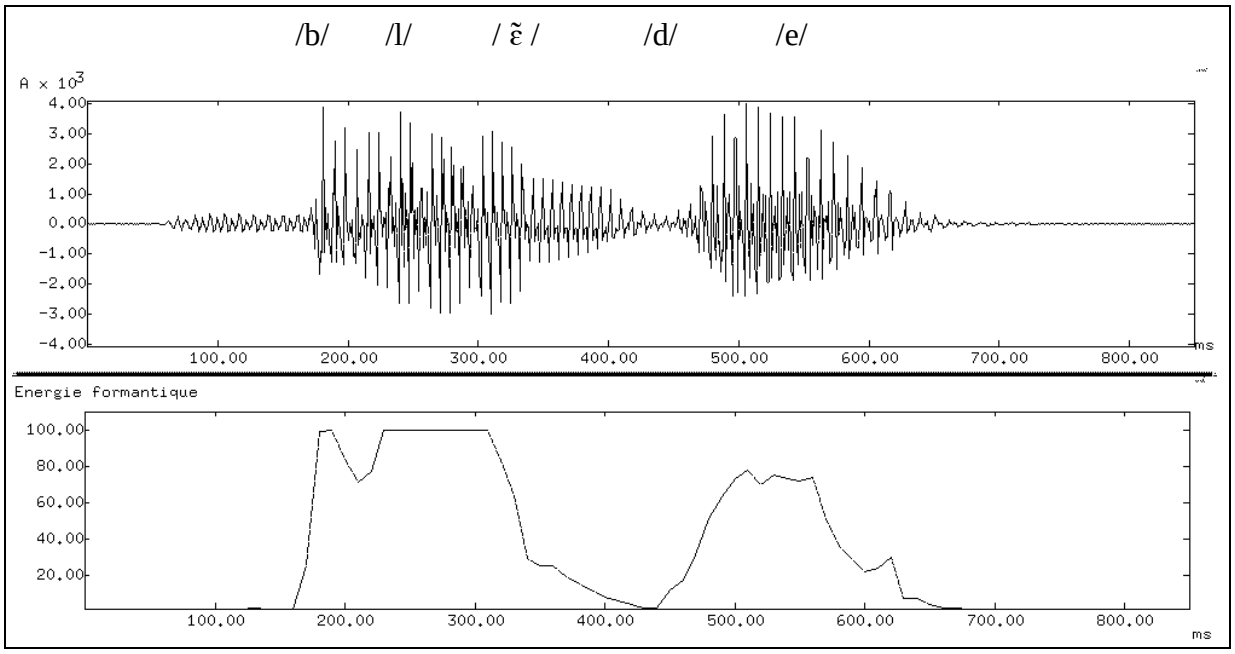


Figure 72: Signal et énergie formantique sur l'énoncé "blindé", locuteur masculin.

V.3.A.c.iv. L'indice de structure spectrale

L'indice de structure spectrale se calcule, pour chaque trame d'analyse, en évaluant la contribution des hautes fréquences [1500Hz;5000Hz] par rapport à la plage [600Hz;5000Hz] (Giraud, 1991). Le résultat est un pourcentage qui rend compte du degré d'acuité local à une trame d'analyse. La Figure 73 fournit un exemple de courbe d'indice d'acuité.

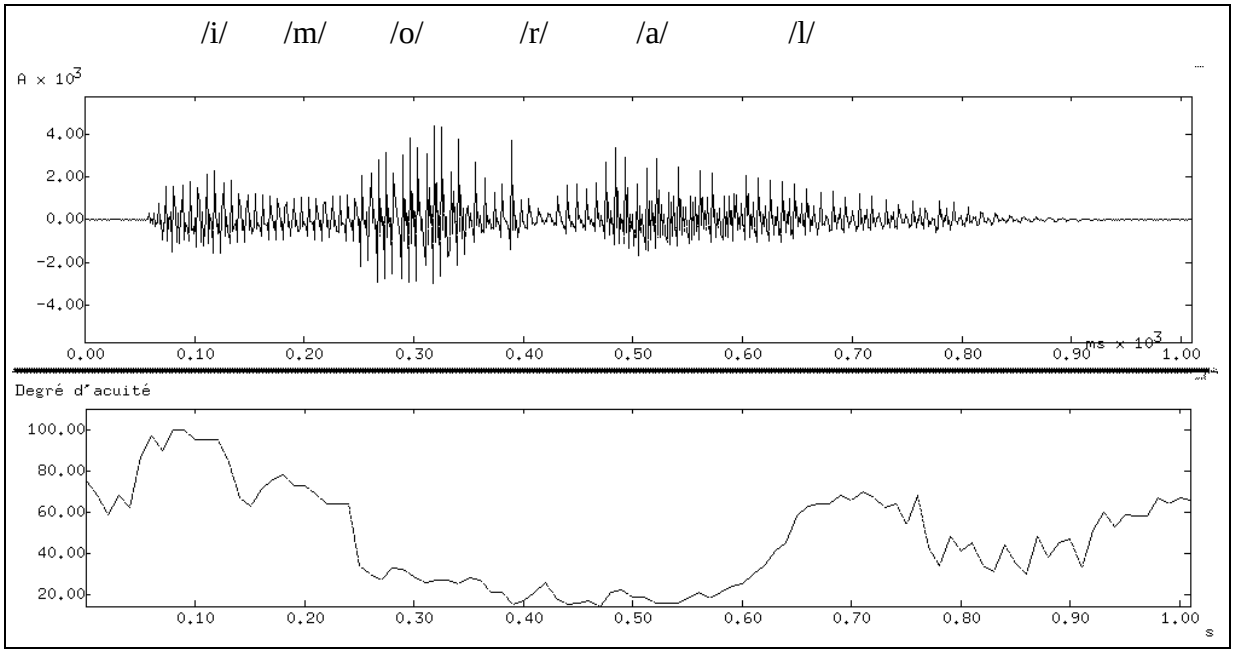


Figure 73: Signal et structure spectrale (degré d'acuité) sur l'énoncé "immoral", locuteur masculin.

V.3.A.c.v. Le degré de structure formantique

Le degré de structure formantique se calcule, tout d'abord, en considérant de façon isolée chaque trame du vocodeur. L'algorithme repère alors les pics émergeant de façon nette de la structure spectrale (Figure 74). Dans un second temps, un balayage temporel effectue un suivi des émergences et ne tient compte que de celles qui réalisent une certaine continuité. Une comptabilité précise permet alors de donner un degré de structure formantique relativement robuste. Sur la Figure 75, on remarque le fort degré de structure formantique au niveau des voyelles, un degré moindre pour /m/ et /r/, et l'absence totale pour /g/.

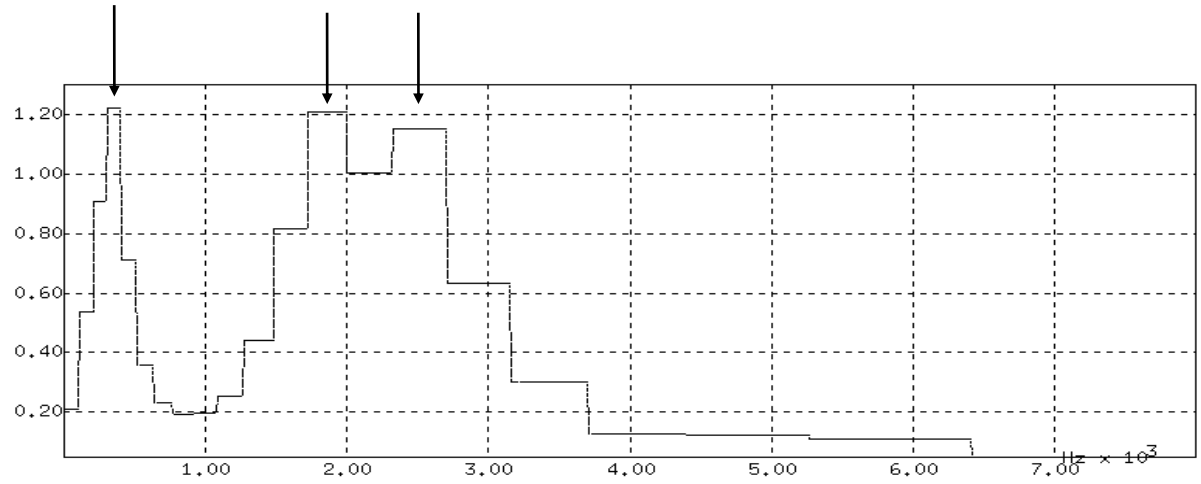


Figure 74: Repérage des pics sur le spectre en bandes critiques (X en kHz, Y en amplitude linéaire)

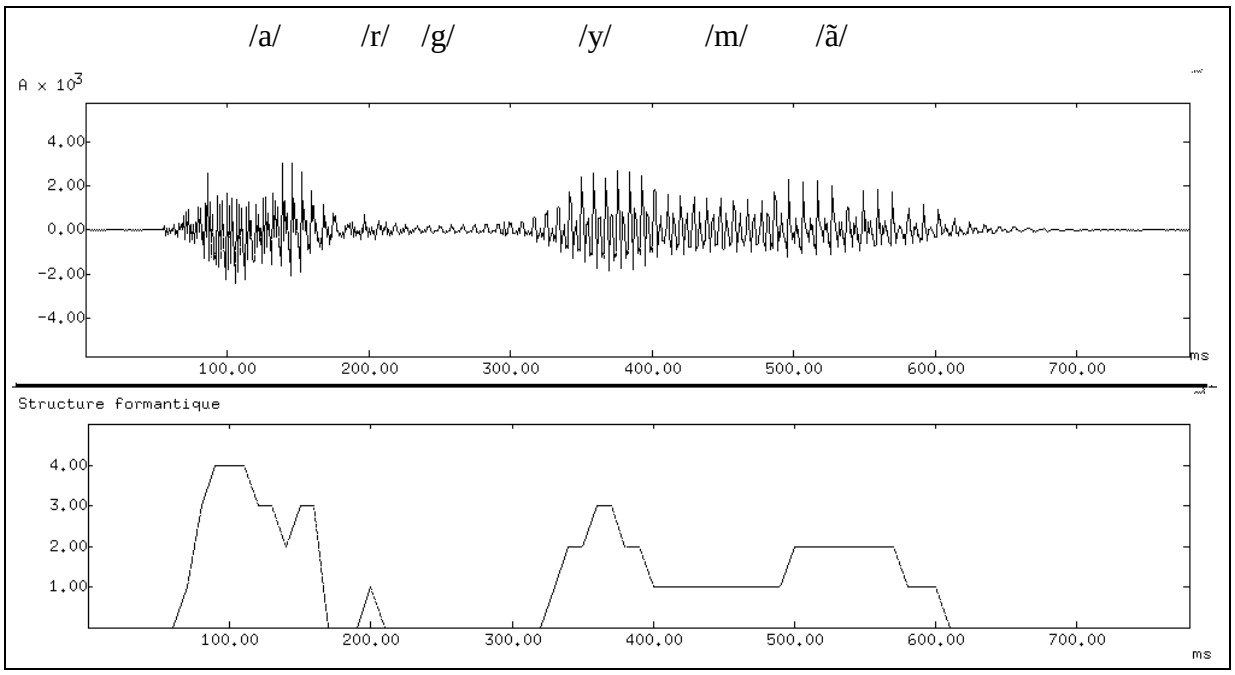


Figure 75: Signal et degré de structure formantique sur l'énoncé "argument", locuteur masculin.

V.3.A.c.vi. Le degré d'instabilité spectrale

Le degré d'instabilité spectrale se calcule à partir du vocodeur en bandes critiques. Le pas d'analyse étant de l'ordre de la centiseconde, cette mesure ne réagit que pour des variations rapides d'une dizaine de millisecondes. Le calcul est effectué de la façon suivante:

- le signal est analysé en bandes critiques (Figure 76a)

	t	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19	K20	K21	SUM
	0	0	1	1	0	0	2	2	1	2	5	10	4	1	1	1	0	1	4	1	0	0	38
	1	14	66	150	207	483	1373	1712	598	269	501	1063	86	34	76	42	16	95	209	61	15	1	7071
	2	100	530	1481	1162	2345	10335	11689	1376	572	1799	3296	469	74	151	89	44	325	548	163	66	1	36617
	3	139	736	2719	2212	4746	22626	17652	1264	928	1667	3573	430	142	457	172	43	222	507	140	34	1	60409
a	4	206	1059	3906	3265	8397	18088	13208	511	758	1322	3335	544	201	399	87	78	403	1184	173	80	3	57206
	5	315	1643	5663	4195	16972	19572	3484	400	615	785	1946	1358	198	142	82	79	301	1090	121	88	1	59051
	6	365	1652	5774	3433	11051	8733	861	899	308	362	1235	821	137	66	127	67	164	843	123	62	2	37086
	7	450	1913	5119	2857	4504	2463	751	458	156	230	933	950	195	57	109	85	196	962	99	62	1	22549
	8	339	1089	1727	1349	1219	399	145	241	134	246	730	1203	78	86	222	116	182	497	74	114	3	10192
	9	138	275	126	216	168	64	74	24	14	31	83	208	40	70	272	848	996	315	387	163	4	4517
	10	42	80	20	43	79	22	24	16	11	13	19	36	75	254	889	1136	1713	1833	2007	707	12	9032
	11	2	3	2	6	12	3	5	6	4	3	10	21	37	273	1628	2178	2888	2661	4672	1633	14	16061
	12	0	1	2	7	7	8	8	7	5	14	37	26	107	670	2765	6486	2677	3016	5124	1614	13	22593
ʃ	13	0	1	2	2	3	3	1	2	4	10	11	24	81	449	1840	3946	6019	4948	5676	682	10	23714
	14	0	0	0	1	1	2	2	8	7	13	10	40	64	386	1319	3550	5024	3790	4052	874	11	19156
	15	0	1	0	1	1	1	0	1	5	6	14	31	74	424	2147	8527	6585	6477	6708	1209	8	32221
	16	0	0	0	1	2	1	0	1	1	4	9	11	25	158	787	3366	2271	2508	1651	632	7	11433
	17	0	0	0	0	0	0	0	0	0	1	2	6	9	56	302	1403	1045	933	951	218	3	4931
	18	0	0	0	0	1	0	0	0	0	0	1	2	2	12	76	476	545	159	167	44	1	1487
	19	0	0	0	0	0	1	0	0	0	0	0	0	0	1	7	45	75	14	11	6	0	161
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	2	3	2	0	0	17
	21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	2	2	2	1	0	0	10
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	5
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	5
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	4
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3
	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	3	1	0	0	13
t	27	0	1	3	2	5	1	2	7	21	30	25	68	15	37	114	161	287	294	31	3	0	1108
	28	1	2	3	4	9	5	4	3	5	3	11	19	3	11	65	63	88	147	46	6	0	500
	29	3	10	18	19	13	4	4	2	1	2	7	8	2	4	23	13	43	54	8	4	0	244
	30	176	711	1065	1475	1923	217	30	25	15	43	82	9	6	12	29	15	51	27	7	3	0	5923
	31	193	1251	1387	2972	5070	485	43	32	23	97	196	18	4	14	35	10	41	18	5	2	0	11896
	32	189	1163	1047	1510	2218	536	25	20	38	81	111	12	3	14	38	9	25	6	2	0	0	7046
	33	171	1031	679	865	1360	418	28	47	52	118	65	6	3	17	42	13	25	6	3	2	0	4950
	34	147	874	478	798	1698	525	77	95	144	127	25	4	2	7	20	11	17	4	4	1	0	5059
	35	120	717	333	444	1061	824	110	104	228	191	10	3	1	3	8	9	12	9	6	1	0	4195
ã	36	93	560	223	215	464	547	91	98	210	64	3	1	1	1	4	6	6	2	2	0	0	2592
	37	69	424	164	125	273	294	62	171	258	10	2	1	1	1	2	5	4	1	1	0	0	1866
	38	54	331	157	92	184	111	50	174	54	3	1	0	0	1	1	2	1	1	1	0	0	1216
	39	43	251	148	86	132	81	81	182	14	2	1	0	1	1	1	1	1	0	0	0	0	1025
	40	35	194	150	68	85	40	59	57	7	2	1	1	1	1	2	1	0	0	0	0	0	705
	41	24	119	92	35	48	27	49	29	3	1	1	0	0	0	0	0	0	0	0	0	0	429
	42	18	75	62	25	30	19	25	13	3	1	0	0	0	0	1	1	0	0	0	0	0	274
	43	14	60	46	20	17	11	6	3	2	1	0	0	0	0	1	1	0	0	0	0	0	184
	44	7	31	25	4	3	1	2	1	0	0	0	0	0	0	1	1	0	0	0	0	0	76
	45	3	14	25	4	3	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	55

Figure 76a: Analyse en bandes critiques brute sur le mot "achetant" prononcé [a ʃ t ã] (la colonne 't' indique le n° de trame, 'Ki' sont les canaux, 'SUM' indique la somme des canaux)

- à partir de ce résultat brut, l'algorithme calcule la contribution de chaque canal par rapport à l'ensemble des canaux de la même trame. Cette opération de normalisation permet d'obtenir une représentation spectrale structurelle indépendante de l'intensité globale. Elle autorise bien un calcul d'instabilité spectrale structurelle. Sur la Figure 76b, les résultats sont donnés en pourcentages. La valeur *20,0* signifie qu'à la trame d'analyse n°7, l'énergie présente dans la 5ème bande critique représente 20 % de l'énergie totale.

	t	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19	K20	K21	SUM
	0	0	2.7	2.7	0	0	5.4	5.4	2.7	5.4	13.5	27	10.8	2.7	2.7	2.7	0	2.7	10.8	2.7	0	0	100
	1	0.2	0.9	2.1	2.9	6.8	19.4	24.2	8.5	3.8	7.1	15	1.2	0.5	1.1	0.6	0.2	1.3	3	0.9	0.2	0	100
	2	0.3	1.4	4	3.2	6.4	28.2	31.9	3.8	1.6	4.9	9	1.3	0.2	0.4	0.2	0.1	0.9	1.5	0.4	0.2	0	100
	3	0.2	1.2	4.5	3.7	7.9	37.8	29.2	2.1	1.5	2.8	5.9	0.7	0.2	0.8	0.3	0.1	0.4	0.8	0.2	0.1	0	100
a	4	0.4	1.9	6.8	5.7	14.7	31.6	23.1	0.9	1.3	2.3	5.8	1	0.4	0.7	0.2	0.1	0.7	2.1	0.3	0.1	0	100
	5	0.5	2.8	9.6	7.1	28.7	33.1	5.9	0.7	1	1.3	3.3	2.3	0.3	0.2	0.1	0.1	0.5	1.8	0.2	0.1	0	100
	6	1	4.5	15.6	9.3	29.8	23.5	2.3	2.4	0.8	1	3.3	2.2	0.4	0.2	0.3	0.2	0.4	2.3	0.3	0.2	0	100
	7	2	8.5	22.7	12.7	*20*	10.9	3.3	2	0.7	1	4.1	4.2	0.9	0.3	0.5	0.4	0.9	4.3	0.4	0.3	0	100
	8	3.3	10.7	16.9	13.2	12	3.9	1.4	2.4	1.3	2.4	7.2	11.8	0.8	0.8	2.2	1.1	1.8	4.9	0.7	1.1	0	100
	9	3.1	6.1	2.8	4.8	3.7	1.4	1.6	0.5	0.3	0.7	1.8	4.6	0.9	1.6	6	18.8	22.1	7	8.6	3.6	0.1	100
	10	0.5	0.9	0.2	0.5	0.9	0.2	0.3	0.2	0.1	0.1	0.2	0.4	0.8	2.8	9.8	12.6	19	20.3	22.2	7.8	0.1	100
	11	0	0	0	0	0.1	0	0	0	0	0	0.1	0.1	0.2	1.7	10.1	13.6	18	16.6	29.1	10.2	0.1	100
	12	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.5	3	12.2	28.7	11.8	13.3	22.7	7.1	0.1	100
ʃ	13	0	0	0	0	0	0	0	0	0	0	0	0.1	0.3	1.9	7.8	16.6	25.4	20.9	23.9	2.9	0	100
	14	0	0	0	0	0	0	0	0	0	0.1	0.1	0.2	0.3	2	6.9	18.5	26.2	19.8	21.2	4.6	0.1	100
	15	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2	1.3	6.7	26.5	20.4	20.1	20.8	3.8	0	100
	16	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.2	1.4	6.9	29.4	19.9	21.9	14.4	5.5	0.1	100
	17	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2	1.1	6.1	28.5	21.2	18.9	19.3	4.4	0.1	100
	18	0	0	0	0	0.1	0	0	0	0	0	0.1	0.1	0.1	0.8	5.1	32	36.7	10.7	11.2	3	0.1	100
	19	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0.6	4.4	28.1	46.9	8.8	6.9	3.8	0	100
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.3	35.7	14.3	21.4	14.3	0	0	100
	21	0	0	0	0	0	0	0	0	0	0	11.1	0	0	0	11.1	22.2	22.2	22.2	11.1	0	0	100
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	25	25	25	0	0	100
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	20	20	20	20	0	0	100
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33.3	33.3	33.3	0	0	0	100
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	50	0	0	0	100
	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.3	16.7	41.7	25	8.3	0	0	100
t	27	0	0.1	0.3	0.2	0.5	0.1	0.2	0.6	1.9	2.7	2.3	6.1	1.4	3.3	10.3	14.5	25.9	26.6	2.8	0.3	0	100
	28	0.2	0.4	0.6	0.8	1.8	1	0.8	0.6	1	0.6	2.2	3.8	0.6	2.2	13.1	12.7	17.7	29.5	9.2	1.2	0	100
	29	1.2	4.1	7.4	7.9	5.4	1.7	1.7	0.8	0.4	0.8	2.9	3.3	0.8	1.7	9.5	5.4	17.8	22.3	3.3	1.7	0	100
	30	3	12	18	24.9	32.5	3.7	0.5	0.4	0.3	0.7	1.4	0.2	0.1	0.2	0.5	0.3	0.9	0.5	0.1	0.1	0	100
	31	1.6	10.5	11.7	25	42.6	4.1	0.4	0.3	0.2	0.8	1.6	0.2	0	0.1	0.3	0.1	0.3	0.2	0	0	0	100
	32	2.7	16.5	14.9	21.4	31.5	7.6	0.4	0.3	0.5	1.1	1.6	0.2	0	0.2	0.5	0.1	0.4	0.1	0	0	0	100
	33	3.5	20.8	13.7	17.5	27.5	8.4	0.6	0.9	1.1	2.4	1.3	0.1	0.1	0.3	0.8	0.3	0.5	0.1	0.1	0	0	100
	34	2.9	17.3	9.5	15.8	33.6	10.4	1.5	1.9	2.8	2.5	0.5	0.1	0	0.1	0.4	0.2	0.3	0.1	0.1	0	0	100
	35	2.9	17.1	7.9	10.6	25.3	19.6	2.6	2.5	5.4	4.6	0.2	0.1	0	0.1	0.2	0.2	0.3	0.2	0.1	0	0	100
ã	36	3.6	21.6	8.6	8.3	17.9	21.1	3.5	3.8	8.1	2.5	0.1	0	0	0	0.2	0.2	0.2	0.1	0.1	0	0	100
	37	3.7	22.7	8.8	6.7	14.6	15.7	3.3	9.2	14	0.5	0.1	0.1	0.1	0.1	0.1	0.3	0.2	0.1	0.1	0	0	100
	38	4.4	27.2	12.9	7.6	15.1	9.1	4.1	14.3	4.4	0.2	0.1	0	0	0.1	0.1	0.2	0.1	0.1	0.1	0	0	100
	39	4.2	24.5	14.4	8.4	12.9	7.9	7.9	17.7	1.4	0.2	0.1	0	0.1	0.1	0.1	0.1	0.1	0	0	0	0	100
	40	5	27.6	21.3	9.7	12.1	5.7	8.4	8.1	1	0.3	0.1	0.1	0.1	0.1	0.3	0.1	0	0	0	0	0	100
	41	5.6	27.8	21.5	8.2	11.2	6.3	11.4	6.8	0.7	0.2	0.2	0	0	0	0	0	0	0	0	0	0	100
	42	6.6	27.5	22.7	9.2	11	7	9.2	4.8	1.1	0.4	0	0	0	0	0.4	0.4	0	0	0	0	0	100
	43	7.7	33	25.3	11	9.3	6	3.3	1.6	1.1	0.5	0	0	0	0	0.5	0.5	0	0	0	0	0	100
	44	9.2	40.8	32.9	5.3	3.9	1.3	2.6	1.3	0	0	0	0	0	0	1.3	1.3	0	0	0	0	0	100
	45	5.6	25.9	46.3	7.4	5.6	3.7	3.7	1.9	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Figure 76b: Analyse en bandes critiques avec canaux normalisés sur le mot "achetant" prononcé [a ʃ t ã] (la colonne 't' indique le n° de trame, 'Ki' sont les contributions en %, 'SUM' indique la somme des contributions)

- la dernière étape du calcul consiste en une dérivation trame à trame du vocodeur normalisé. Pour cela, il suffit de faire, pour chaque canal, la différence entre la contribution de la bande d'une trame *t* et celle de la trame *t-1*. La valeur de l'instabilité spectrale au temps *t* est le cumul des valeurs absolues de ces différences. Elle s'écrit:

$$dk(t) = \sum_{i=1}^{21} \left| K_i(t) - K_i(t-1) \right|$$ où $K_i(t)$ est la valeur normalisée de la $i^{ème}$ bande à la trame *t*

Sur la Figure 76c, la valeur **9.8** signifie qu'entre la trame d'analyse n°6 et 7, la contribution de la bande 5 a subi une variation de 9.8 points, qui correspond a un transfert d'énergie vers d'autres canaux et donc une instabilité spectrale.

Au total, entre ces deux trames, une différence cumulée de 46 points a été notée.

	t	dK 1	dK2	dK3	dK4	dK5	dK6	dK7	dK8	dK9	dK10	dK11	dK12	dK13	dK14	dK15	dK16	dK17	dK18	dK19	dK20	dK21	SUM
	0																						0
	1	0.2	1.8	0.6	2.9	6.8	14	18.8	5.8	1.6	6.4	12	9.6	2.2	1.6	2.1	0.2	1.4	7.9	1.8	0.2	0	98
	2	0.1	0.5	1.9	0.2	0.4	8.8	7.7	4.7	2.2	2.2	6	0.1	0.3	0.7	0.4	0.1	0.5	1.5	0.4	0	0	39
	3	0	0.2	0.5	0.5	1.5	9.2	2.7	1.7	0	2.2	3.1	0.6	0	0.3	0	0	0.5	0.7	0.2	0.1	0	24
	4	0.1	0.6	2.3	2	6.8	5.8	6.1	1.2	0.2	0.4	0.1	0.2	0.1	0.1	0.1	0.1	0.3	1.2	0.1	0.1	0	28
a	5	0.2	0.9	2.8	1.4	14.7	1.5	17.2	0.2	0.3	1	2.5	1.3	0	0.5	0	0	0.2	0.2	0.1	0	0	44
	6	0.5	1.7	6	2.2	1.7	9.6	3.6	1.7	0.2	0.4	0	0.1	0	0.1	0.2	0	0.1	0.4	0.1	0	0	28
	7	1	4	7.1	3.4	9.8	12.6	1	0.4	0.1	0	0.8	2	0.5	0.1	0.1	0.2	0.4	2	0.1	0.1	0	46
	8	1.3	2.2	5.8	0.6	8	7	1.9	0.3	0.6	1.4	3	7.6	0.1	0.6	1.7	0.8	0.9	0.6	0.3	0.8	0	46
	9	0.3	4.6	14.2	8.5	8.2	2.5	0.2	1.8	1	1.7	5.3	7.2	0.1	0.7	3.8	17.6	20.3	2.1	7.8	2.5	0.1	111
	10	2.6	5.2	2.6	4.3	2.8	1.2	1.4	0.4	0.2	0.5	1.6	4.2	0.1	1.3	3.8	6.2	3.1	13.3	13.7	4.2	0	73
	11	0.5	0.9	0.2	0.4	0.8	0.2	0.2	0.1	0.1	0.1	0.1	0.3	0.6	1.1	0.3	1	1	3.7	6.9	2.3	0	21
	12	0	0	0	0	0	0	0	0	0	0	0.1	0	0.2	1.3	2.1	15.1	6.1	3.2	6.4	3	0	38
	13	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	1.1	4.5	12.1	13.5	7.5	1.3	4.3	0	45
ʃ	14	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.9	1.9	0.8	1.1	2.8	1.7	0	10
	15	0	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.7	0.2	7.9	5.8	0.3	0.3	0.8	0	16
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.2	3	0.6	1.8	6.4	1.8	0	14
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.8	1	1.3	3	4.9	1.1	0	12
	18	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0.3	1	3.6	15.5	8.2	8.1	1.5	0	38
	19	0	0	0	0	0.1	0.6	0	0	0	0	0.1	0.1	0.1	0.2	0.7	3.9	10.2	1.9	4.4	0.8	0.1	23
	20	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0.6	9.9	7.6	32.6	12.7	7.4	3.8	0	75
	21	0	0	0	0	0	0	0	0	0	0	11.1	0	0	0	3.2	13.5	7.9	0.8	3.2	0	0	40
	22	0	0	0	0	0	0	0	0	0	0	11.1	0	0	0	11.1	2.8	2.8	2.8	13.9	0	0	44
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	5	5	5	5	0	0	40
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	13.3	13.3	13.3	20	0	0	80
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33.3	16.7	16.7	0	0	0	67
	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.3	16.7	8.3	25	8.3	0	0	67
	27	0	0.1	0.3	0.2	0.5	0.1	0.2	0.6	1.9	2.7	2.3	6.1	1.4	3.3	2	2.1	15.7	1.6	5.5	0.3	0	47
t	28	0.2	0.3	0.3	0.6	1.4	0.9	0.6	0	0.9	2.1	0	2.3	0.8	1.1	2.8	1.9	8.3	3	6.4	0.9	0	35
	29	1	3.7	6.8	7	3.6	0.6	0.8	0.2	0.6	0.2	0.7	0.5	0.2	0.6	3.5	7.3	0.1	7.2	5.9	0.4	0	51
	30	1.7	7.9	10.5	17.1	27.1	2	1.1	0.4	0.2	0.1	1.5	3.2	0.7	1.5	9	5.1	16.9	21.9	3.2	1.6	0	133
	31	1.4	1.5	6.3	0.1	10.1	0.4	0.1	0.2	0.1	0.1	0.3	0	0.1	0.1	0.2	0.2	0.5	0.3	0.1	0	0	22
	32	1.1	6	3.2	3.6	11.1	3.5	0	0	0.3	0.3	0.1	0	0	0.1	0.2	0	0	0.1	0	0	0	30
	33	0.8	4.3	1.1	4	4	0.8	0.2	0.7	0.5	1.2	0.3	0	0	0.1	0.3	0.1	0.2	0	0	0	0	19
	34	0.5	3.5	4.3	1.7	6.1	1.9	1	0.9	1.8	0.1	0.8	0	0	0.2	0.5	0	0.2	0	0	0	0	24
	35	0	0.2	1.5	5.2	8.3	9.3	1.1	0.6	2.6	2	0.3	0	0	0.1	0.2	0	0	0.1	0.1	0	0	32
	36	0.7	4.5	0.7	2.3	7.4	1.5	0.9	1.3	2.7	2.1	0.1	0	0	0	0	0	0.1	0.1	0.1	0	0	25
ã	37	0.1	1.1	0.2	1.6	3.3	5.4	0.2	5.4	5.7	1.9	0	0	0	0	0	0	0	0	0	0	0	25
	38	0.7	4.5	4.1	0.9	0.5	6.6	0.8	5.1	9.4	0.3	0	0.1	0.1	0	0	0.1	0.1	0	0	0	0	33
	39	0.2	2.7	1.5	0.8	2.2	1.2	3.8	3.5	3.1	0.1	0	0	0.1	0	0	0.1	0	0.1	0.1	0	0	20
	40	0.8	3.1	6.9	1.3	0.8	2.2	0.5	9.6	0.4	0.1	0	0.1	0	0	0.2	0	0.1	0	0	0	0	26
	41	0.6	0.2	0.2	1.5	0.9	0.6	3.1	1.3	0.3	0.1	0.1	0.1	0.1	0.1	0.3	0.1	0	0	0	0	0	10
	42	1	0.3	1.2	1	0.2	0.7	2.3	2	0.4	0.1	0.2	0	0	0	0.4	0.4	0	0	0	0	0	10
	43	1.1	5.5	2.6	1.8	1.6	0.9	5.9	3.1	0	0.2	0	0	0	0	0.2	0.2	0	0	0	0	0	23
	44	1.5	7.8	7.6	5.7	5.4	4.7	0.7	0.3	1.1	0.5	0	0	0	0	0.8	0.8	0	0	0	0	0	37
	45	3.7	14.9	13.4	2.1	1.6	2.4	1.1	0.5	0	0	0	0	0	0	1.3	1.3	0	0	0	0	0	42

Figure 76c: Dérivée canal à canal de l'analyse en bandes critiques sur le mot "achetant" prononcé [a ʃ t ã] (la colonne 't' indique le n° de trame, 'dKi' sont les dérivées canal à canal, 'SUM' indique la somme cumulée des différences canal à canal, c'est à dire l'instabilité spectrale)

Dans l'exemple précédent (Figure 76), on peut remarquer l'émergence de pics sur la transition [a ʃ] (trame n°9), sur le passage entre la fricative [ʃ] et le silence de [t] (trame n°20), et enfin sur l'explosion de l'occlusive [t] (trame n°30). La Figure 77 fournit un exemple graphique du degré d'instabilité spectrale.

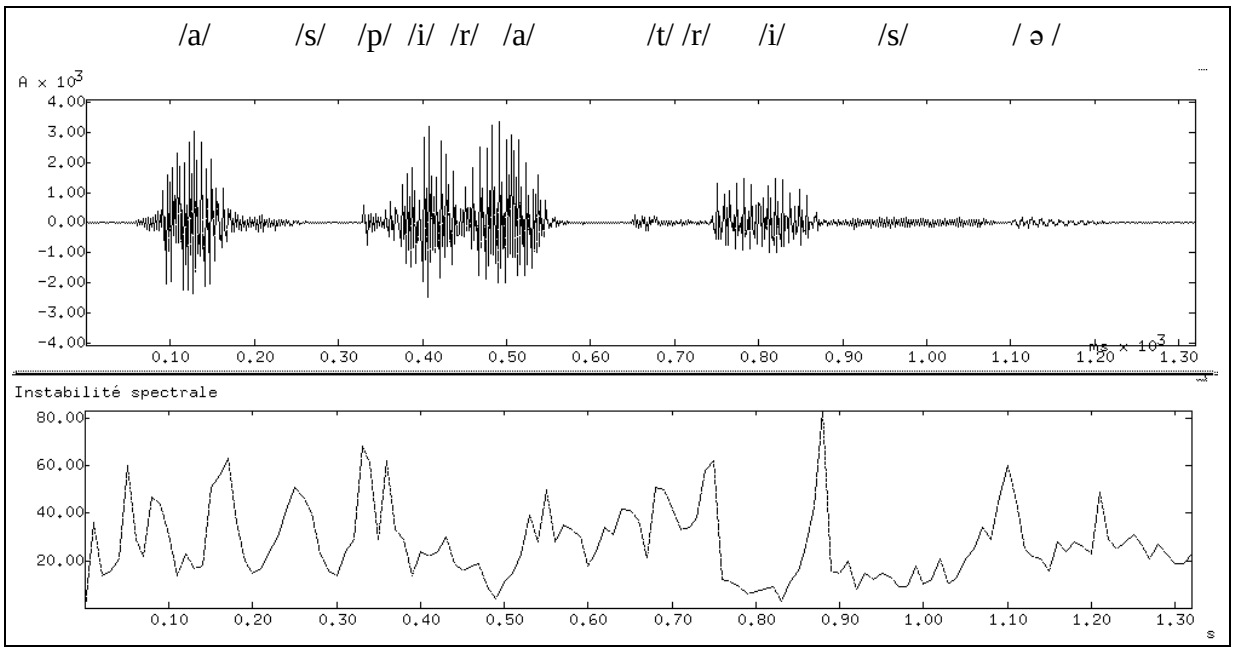


Figure 77: Signal et instabilité spectrale sur l'énoncé "aspiratrice", locuteur masculin.

V.3.B. Les propriétés de base

Le calcul de tous les paramètres décrits précédemment constitue une première étape. La suivante consiste à extraire des propriétés robustes et résistant aux problèmes de variabilité. Nous établissons, à la base, 4 propriétés majeures de type binaire calculées pour chaque trame d'analyse centiseconde:

- *la propriété de silence/signal* qui s'établit en comparant le niveau d'énergie *RMS* normalisée à un seuil critique.
- *la propriété de voisé/non voisé* qui s'obtient en validant comme sonore toute trame possédant une énergie de voisement normalisée suffisante.
- *la propriété de forte/faible énergie* qui se déduit de l'étude combinée des paramètres énergétiques normalisés *RMS* et *NRJ subjective.*
- *la propriété de présence/absence de friction* qui se calcule en comparant les taux de passage par zéro sur signal brut (*TPZ*) et sur signal filtré (*PZSD*) par rapport à deux seuils respectifs.

L'évaluation trame à trame de chaque propriété est complétée par une rapide analyse contextuelle à court terme qui filtre les singularités, c'est à dire la présence ou l'absence isolée à une trame d'une propriété.

Nous avons conscience que l'utilisation de seuils lors de la déclaration des propriétés reste délicate dans la mesure où ces choix peuvent paraître arbitraires. En fait, ce sont plutôt des décisions empiriques issues de l'analyse sur de vastes corpus. Ceci est rendu possible par la normalisation des paramètres qui réduit une partie des phénomènes de variabilité. Il est vrai que les techniques de classification automatique semblent plus satisfaisantes pour l'esprit; elles ne sont pas pour autant plus efficaces. L'utilisation de seuillage en décodage automatique peut se justifier par le fait que cette notion reste très présente dans le processus de décodage humain. Il suffit pour cela de se référer à toutes les études en perception de la parole (seuil de réception de la parole, seuil d'intensité, seuil de glissando...) Enfin, il faut signaler qu'il n'est pas impossible d'effectuer une adaptation des seuils en fonction du type de parole utilisée: nous pensons, par exemple, à la parole en milieu bruitée.

Les quatre propriétés de base permettent une première catégorisation des trames.

V.3.C. La primo-catégorisation

Nous avons déjà signalé que la segmentation proprement dite est dépendante d'une catégorisation grossière des segments phonémiques. Cette identification commence par un classement sommaire de chaque trame selon 3 catégories:

- SIL <=> étiquette de silence
- VOC <=> étiquette de vocalisme
- CTB <=> étiquette de consonantisme, transition, bruit

Pour réaliser cette opération, plusieurs passes sont nécessaires:

0) passe préliminaire de détection silence / signal
- présence de la propriété de silence => SIL

1) catégorisation des trames non silencieuses par étude de l'énergie
- absence de la propriété de forte énergie => CTB
- énergie formantique nulle => CTB

2) catégorisation des trames non silencieuses par étude du voisement
- absence de la propriété de voisement => CTB

3) catégorisation des trames non silencieuses par étude de la friction
- présence de la propriété de friction => CTB

4) sinon => VOC

Le vocalisme est finalement défini par un faisceau de propriétés avec forte énergie globale, énergie formantique non nulle, voisement et absence de friction.

Cette première classification se déroule trame à trame sans étude du contexte. L'étape suivante va affiner l'identification en effectuant un suivi temporel.

V.3.D. L'identification des macro-classes

Par macro-classe, nous entendons le regroupement d'unités phonologiques en catégories. Nous avons retenu 6 macro-classes principales: les voyelles, les consonnes vocaliques, les constrictives sonores, les constrictives sourdes, les occlusives sonores et enfin les occlusives sourdes (cf. Tableau 4, p.66). Il n'est pas toujours possible d'attribuer une macro-classe précise à un segment phonique, ceci par manque d'information permettant la distinction.

Par exemple, un segment identifié comme consonantique avec ambiguïté sur le voisement et le mode d'articulation sera étiqueté comme « consonne » et noté CS. Si les propriétés de voisement apparaissent, l'algorithme précisera s'il s'agit d'une consonne voisée (CSVX) ou non voisée (CSNV). Si le mode d'articulation est clair, fricatif par exemple, le résultat de la catégorisation fournira une classe FRIVO en présence de voisement, FRINV en absence de voisement ou FRICA en cas d'ambiguïté sur le voisement (Figure 78). Nous utilisons là une logique floue avec « décision retardée », terme emprunté à (Haton, 1989), ceci pour éviter une catégorisation hâtive peu robuste.

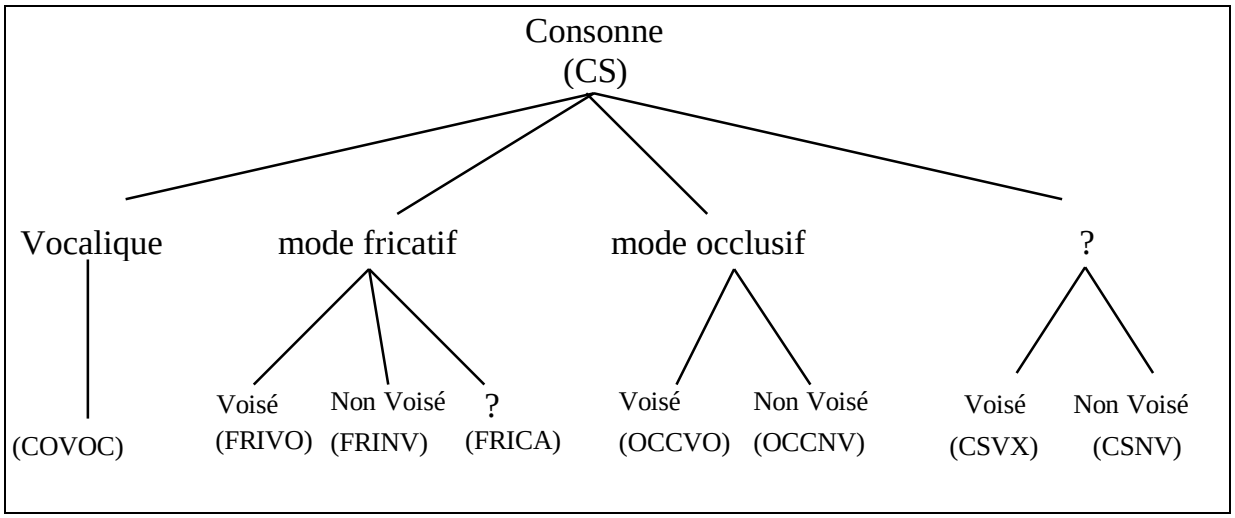


Figure 78: Catégorisation floue des segments consonantiques.

Une telle catégorisation donne lieu à des « super-macro-classes » regroupant diverses macro-classes (Tableau 12). Ainsi, par exemple, les consonnes sonores englobent les consonnes vocaliques, les constrictives et les occlusives sonores. Cela signifie, aussi, que /l/, qui est une consonne vocalique pourra être catégorisé, sauf erreur, comme consonne vocalique (COVOC), segment vocalique (VOC), consonne (CS), consonne voisée (CSVX) ou consonne avec barre voisée * (CSBVX) selon les propriétés qui auront été détectées. L'identification des macro-classes s'effectue en plusieurs passes.

* Par le terme de « consonne avec barre voisée », nous entendons des segments réduits à une barre de voisement. Il s'agit essentiellement de certaines formes d'occlusives sonores ou de consonnes vocaliques.

Tableau 12: Les macro-classes et leur codage dans S.A.P.H.O.

Macro-Classes de base		Super-Macro-classes					
		Vocalique	Consonne	Consonne sourde	Consonne sonore	Consonne avec barre voisée	Fricative
	codage	VOC	CS	CSNV	CSVX	CSBVX	FRICA
voyelle	VOY	✓					
consonne vocalique	COVOC	✓	✓		✓	✓	
constrictive sonore	FRIVO		✓		✓		✓
constrictive sourde	FRINV		✓	✓			✓
occlusive sonore	OCCVO		✓		✓	✓	
occlusive sourde	OCCNV		✓	✓			

V.3.D.a. Le repérage des constrictives

> Si une série suffisante de trames CTB successives sont repérées avec les propriétés:
> - présence de la propriété de friction
>
> ou - TPZ sur signal dérivé et degré d'acuité au dessus de seuils critiques
>
> alors le segment constitué par ces trames est identifié comme une constrictive (FRICA).

Le nombre de trames successives possédant ces propriétés doit être important pour éviter une confusion avec l'explosion longue d'une occlusive sourde. Une étude de la propriété de voisement sur l'ensemble du segment catégorisé comme constrictive (FRICA) permet d'affiner la catégorisation en constrictive sonore (FRIVO) si une majorité de trames sont voisées ou en constrictive sourde (FRINV) si une majorité de trames sont non voisées. Dans tous les autres cas, le segment garde l'étiquette FRICA car aucune décision n'est prise quant au voisement. Nous introduisons ici une notion de flou médian, où seule les situations franches sont complètement catégorisées. Les positions mitigées entraînent une absence de décision.

Cette catégorisation en plusieurs étapes est efficace pour gérer les problèmes de transitions. Ainsi, la recherche directe de segments de type constrictive sonore ou constrictive sourde entraîne souvent un phénomène de sursegmentation. En effet, le passage d'une voyelle à une constrictive sourde, par exemple, laisse apparaître un segment médian ayant les propriétés de la constrictive accompagné de voisement correspondant à la coarticulation avec la voyelle précédente. Si la segmentation s'effectue sans précaution de façon trop directe, elle fera apparaître une voyelle, une constrictive sonore suivie d'une constrictive sourde, d'où insertion d'une unité.

Une dernière passe s'attache enfin au repérage des segments CTB non voisés avec présence de friction. Si le nombre de trames successives possédant ces propriétés dépasse un seuil critique, le segment est directement catégorisé comme une constrictive sourde FRINV. Sinon, on peut soupçonner la présence d'un explosion d'occlusive sourde. L'algorithme ne prend donc pas de décision précise et catégorise le segment comme consonne sourde CSNV.

V.3.D.b. Le repérage des occlusives sourdes

La macro-classe des occlusives sourdes est à la fois la plus facile et la plus difficile des catégories à repérer. L'identification est relativement aisée lorsque ces unités apparaissent à l'intérieur d'un mot du fait de la présence de silence due à la tenue de l'occlusive. Seule quelques rares réalisations particulière du /r/ font apparaître aussi un tel phénomène. Le repérage de silence est donc une propriété forte pour localiser les occlusives sourdes. L'hypothèse peut être renforcée si, après la tenue, une explosion franche est détectée. Celle-ci peut se manifester soit par un bruit de friction, soit par une courte barre d'explosion, soit par les deux.

La détection de barre d'explosion reste une opération extrêmement délicate, ce qui explique les études spécifiques réalisées à ce sujet (Caelen et al., 1988; Malbos et al., 1994). La difficulté provient de l'extrême brièveté du phénomène, ainsi que de sa faible énergie. Parfois, l'explosion n'apparaît même pas sur le signal de parole (ex: avec la bilabiale /p/ en contexte /a/ comme dans "papa"), ce qui laisse penser à l'existence d'autres propriétés révélant la présence de l'occlusive: nous pensons évidemment aux transitions formantiques. Cependant, nous n'avons pas voulu intégrer cette information dans l'algorithme de segmentation car ce type d'observation est une opération impossible à réaliser de façon isolée. Cela pourrait par contre être effectué dans une interaction haut-bas .

La présence de court bruit de friction reste une propriété robuste pour la détection des explosions d'occlusives sourdes. Il existe toutefois une possibilité de confusion avec les constrictives sourdes, spécialement dans certaines variétés de français méridional où les explosions sont particulièrement affriquées comme dans le mot "voiture" quasiment prononcé /vwatʃyrə/ ou en français québécois, par exemple dans le pronom "tu" réalisé phonétiquement en /tsy/. Dans ces cas-là, ou dans le cas de groupes consonantiques (ex: "accès", "triste"...), la détection apparaît comme extrêmement difficile.

Enfin, le problème des occlusives sourdes à l'initiale reste entier du fait de l'ambiguïté entre le silence initial de l'énoncé et la tenue de l'occlusive. Sauf présence très marquée d'une explosion, la détection apparaît comme impossible sans faire appel à d'autres sources d'information. La difficulté est d'autant plus accrue que, bien souvent, du fait de l'inertie des organes articulateurs, apparaissent des phénomènes impulsionnels au démarrage de l'élocution, ce qui pourrait donc laisser présager la présence d'une explosion, y compris dans les cas où l'énoncé commence, par exemple, par une voyelle.

V.3.D.c. Le repérage des consonnes avec barre voisée

Par le terme de « consonne avec barre voisée », nous entendons des segments réduits à une barre de voisement. Il s'agit essentiellement de certaines formes d'occlusives sonores (Figure 79) ou de consonnes vocaliques.

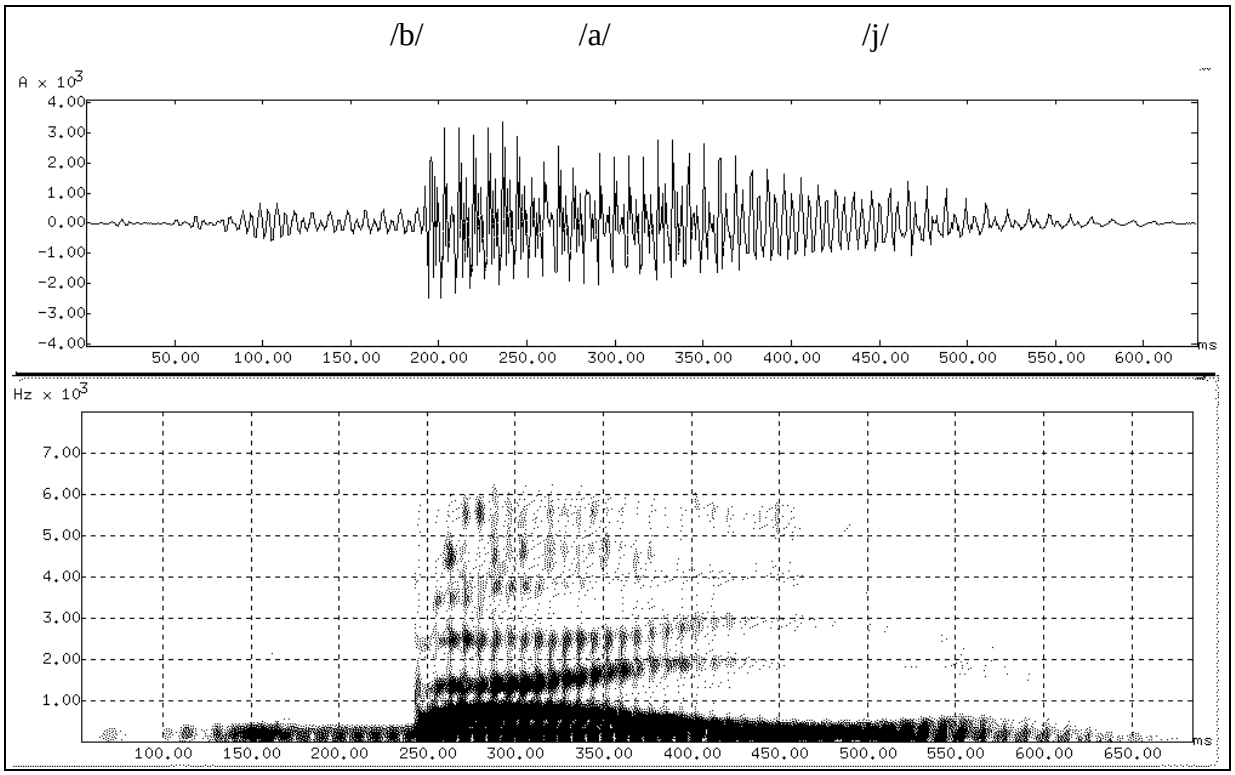


Figure 79: Signal et spectrogramme sur l'énoncé "bail", locuteur masculin. Le segment /b/ se réduit à une barre de voisement.

La détection de ce type de segment est issue de l'étude de plusieurs facteurs.

Sont déclarées barres consonantiques voisées les trames possédant les propriétés :
- présence de la propriété de voisement

et - faible énergie formantique

et - faible taux de passage par zéro calculé sur le signal filtré

L'évaluation trame à trame de ces propriétés est complétée par une rapide analyse contextuelle à court terme qui consiste à filtrer les singularités, c'est à dire la présence ou l'absence isolée à une trame des caractéristiques décrites précédemment. Dans une troisième passe, si une série suffisante de trames CTB successives ont été étiquetées comme barres consonantiques voisées, le segment est interprété soit comme une occlusive sonore, soit comme une consonne vocalique. Le nombre de trames doit être suffisant pour éviter une confusion avec des phénomènes d'amortissement.

La distinction entre occlusive sonore ou consonne vocalique s'effectue grâce à l'évaluation locale de propriétés favorisant soit l'hypothèse de l'occlusive (présence d'une explosion, faible degré de structure formantique), soit celle de la consonne vocalique (absence d'explosion, degré de structure formantique moyen). Si les propriétés occlusives sont fortes, le segment sera étiqueté OCCVO. Si les propriétés vocaliques sont fortes, le segment sera étiqueté COVOC. Si le cas est mitigé, le segment sera codé comme consonne à barre voisée CSBVX. A nouveau est utilisée la notion de flou médian.

A ce niveau d'analyse, la plupart des étiquettes CTB faciles à décoder ont été interprétées. L'algorithme poursuit alors son analyse sur les trames VOC.

V.3.E. La segmentation des continuums vocaliques

Par continuum vocalique, nous entendons une suite suffisamment longue de trames successives étiquetées VOC lors de la primo-catégorisation. Une telle suite peut laisser supposer l'existence de plusieurs segments vocaliques contigus, que l'algorithme de segmentation va s'efforcer de repérer en mettant en évidence des cassures de type énergétique et/ou spectrale. Pour cela, différentes étapes sont nécessaires.

Le repérage des fractures énergétiques s'effectue entre deux extrema de la courbe d'énergie subjective filtrée par la médiane (cf. § « L'énergie subjective », p.144). S'il existe un contraste net entre des zones de forte et faible énergie sur le continuum vocalique, comme cela est visible sur le segment /anali/ à la Figure 80, l'hypothèse de l'existence de plusieurs unités vocaliques est sûre: une frontière provisoire est placée entre les extrema au niveau de la trame qui possède la plus forte dérivée d'énergie subjective, si celle-ci est significative. Le repérage des fractures spectrales s'effectue globalement sur le continuum vocalique. Si la courbe d'instabilité spectrale dépasse un seuil critique sur l'une des trames du continuum, une frontière provisoire est placée à ce niveau-là.

Une étude contextuelle de l'emplacement des frontières placées à l'intérieur du continuum permet d'éliminer la sursegmentation éventuellement entraînée par ces repérages. La Figure 80 fournit un exemple de segmentation d'un continuum vocalique. Il s'agit de l'énoncé "analyse" où la primo-catégorisation a étiqueté VOC l'ensemble des 5 segments /anali/. L'étude des extrema de la courbe d'énergie subjective (Figure 80, 2ème courbe) permet de mettre en évidence un contraste sur le continuum et de localiser 4 unités qui correspondent à:

- /a/ (forte énergie)
- /n/ (faible énergie)
- /a/ (forte énergie)
- /li/ (faible énergie)

Les 3 frontières entre ces 4 segments sont posées au niveau des trames où la dérivée d'énergie subjective est extrémale (Figure 80, 3ème courbe). Le segment /li/, de longueur importante est ensuite analysé lui-même en tant que continuum vocalique. Aucune cassure énergétique significative n'est notée. Par contre, une forte instabilité spectrale au sein du segment autorise le marquage d'une frontière à ce niveau-là. /l/ et /i/ sont ainsi repérés distinctement. Finalement, le continuum est segmenté correctement en 5 segments vocaliques.

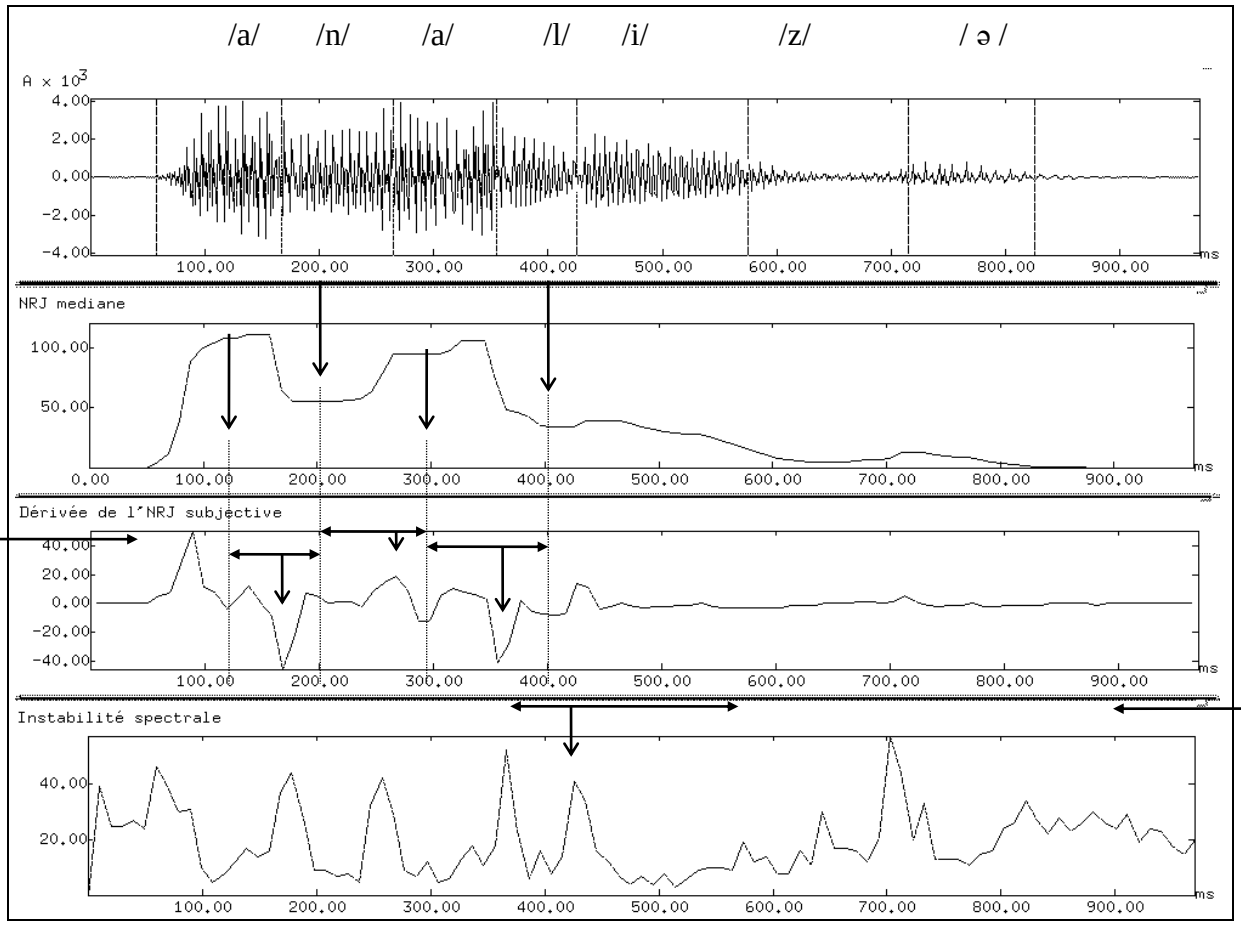


Figure 80: Signal et paramètres utilisés pour la segmentation du continuum vocalique /anali/ de l'énoncé "analyse", locuteur masculin.. Les traits verticaux sur le signal sont les frontières repérées par SAPHO.

Il faut noter que, fréquemment, dans le cas d'un continuum vocalique constitué de voyelles (ex: "aérienne"), la segmentation ne laisse apparaître qu'un segment par manque de contraste énergétique et du fait de la variation lente de la structure spectrale, ce qui passe inaperçu dans la courbe d'instabilité. Seule une étude plus fine permet de distinguer deux unités. Cette analyse interviendra ultérieurement.

A cette étape de traitement, l'émergence de segments commence à apparaître. Les trames étiquetées VOC ont été rassemblées pour former des unités cohérentes. Une partie des étiquettes CTB ont été catégorisées (cf. § « L'identification des macro-classes », p.154);

d'autres ne le sont pas, soit par manque d'information permettant d'affiner l'identification, soit parce qu'elles correspondent à des phases de transitions. La suite des opérations, dans un premier temps, va consister à adapter les frontières, c'est à dire à étudier plus précisément les régions frontalières des premiers segments repérés et ceci en tenant compte du contexte.

V.3.F. L'adaptation des frontières

Il existe différents types d'adaptation aux frontières :

- l'observation des débuts et fins d'énoncés
- l'adaptation des transitions
- l'adaptation des frontières de segments consonantiques

V.3.F.a. L'observation des débuts et fins d'énoncés

- l'observation de la phase initiale

L'analyse de la phase initiale de l'énoncé s'avère extrêmement délicate. Dans la plupart des cas, les premières trames non silencieuses sont étiquetées CTB. Si la suite est rapidement étiqueté VOC, le problème consiste à savoir si nous sommes en présence d'une ou deux unités. En effet, les premières trames CTB peuvent rendre compte soit de la présence d'une consonne non vocalique à l'initiale, soit de l'amorce d'un segment vocalique. L'étude du contexte et de la durée peut permettre de prendre une décision:

- si le segment précède un segment vocalique, s'il est court et si l'énergie est monotone croissante, il s'agit très certainement de l'amorce du segment vocalique: la frontière CTB/VOC est supprimée et les trames CTB sont intégrées au segment vocalique. Cette initiative généralement efficace peut par contre effacer la présence d'occlusives sourdes à l'initiale.
- sinon, la présence de consonne est possible. Aucune modification de frontière n'est effectuée.

- l'observation de la phase finale

De même, au niveau de la phase finale de l'énoncé, apparaissent très souvent des trames étiquetées CTB qui peuvent rendre compte de trois phénomènes :

⇒ l'existence d'une consonne finale
⇒ la présence d'une unité affaiblie de type schwa (ex: aspiratrice)
⇒ l'amortissement d'un segment vocalique.

L'étude du contexte et de la durée peut permettre de prendre une décision:

- si le segment n'est pas trop long et est précédé par un segment vocalique, il s'agit d'un amortissement. La frontière VOC/CTB est supprimée et les trames CTB sont intégrées au segment vocalique.
- sinon, la présence de consonne est possible. Aucune modification de frontière n'est effectuée.

V.3.F.b. L'étude des relâchements vocaliques

Dans un premier temps, l'observation va porter sur les trames CTB encore non catégorisées, qui peuvent rendre compte d'une phase de transition. L'objectif, alors, consiste à attribuer les trames de cette partie indéfinie aux unités adjacentes. Très souvent, l'opération de primo-catégorisation laisse apparaître, à la suite d'un segment vocalique, une série de trames étiquetées CTB, qui peuvent rendre compte du relâchement de l'unité vocalique. Il s'agit alors de détecter un tel phénomène et d'intégrer les trames CTB de transition au segment vocalique. L'étude du contexte droit ainsi qu'un critère de durée peut permettre de prendre une décision:

- si le segment est long, la présence de consonne est possible. Aucune modification de frontière n'est effectuée.
- sinon, le segment est court

  ⇒ si le segment est placé entre deux segments vocaliques, il rend compte d'une discontinuité certaine et mérite de rester en évidence. Aucune modification de frontière n'est effectuée.

  ⇒ si le segment suit un segment vocalique et précède un silence, il s'agit d'un relâchement. Ce phénomène est probablement lié à une phase d'occlusion. La frontière VOC/CTB est supprimée et les trames CTB sont intégrées au segment vocalique.

  ⇒ si le segment est placé entre un segment vocalique et un segment consonantique et s'il ne laisse apparaître aucune cassure énergétique significative à gauche, l'algorithme considère que cette phase de transition appartient au segment vocalique. La frontière VOC/CTB est supprimée et les trames CTB sont intégrées au segment vocalique.

V.3.F.c. L'adaptation des frontières des segments consonantiques

A cette étape du traitement, l'observation va porter sur les segments consonantiques clairement catégorisés, c'est à dire les constrictives sourdes ou sonores, les occlusives sourdes ou sonores et les consonnes vocaliques. L'objectif est de préciser les frontières avec les unités adjacentes. Généralement, les segments ont été identifiés de façon réduite. Par exemple, les constrictives sourdes correspondent au regroupement des trames non voisées extrêmement affriquées. Les trames périphériques, qui ne sont pas définies aussi clairement du fait des phénomènes de coarticulation, ont provisoirement été exclues. L'algorithme de segmentation va alors se fonder sur le noyau robuste et étendre les frontières en tâches d'huile en tenant compte, de façon moins restrictive, des paramètres caractéristiques du segment. Toujours dans l'exemple des constrictives sourdes, l'algorithme va intégrer au segment certaines trames périphériques moins bruitées et éventuellement légèrement voisées, comme cela peut arriver au contact d'une voyelle (Figure 13, p.21). C'est dans ce type d'adaptation que nous sommes convaincu que l'efficacité du repérage temporel est fortement dépendant de la pré-identification du segment. En effet, les frontières seront adaptées différemment selon les propriétés fortes de chaque macro-classe.

Dans un deuxième temps, l'algorithme va s'attacher à adapter les frontières de type CTB/VOC et VOC/CTB. Pour cela, il repère, tout d'abord, les extrema d'énergie subjective sur chacune des unités adjacentes puis recherche la cassure la plus importante de point de vue énergétique et spectral entre les deux extrema respectifs. Si cette cassure est significative, la frontière est déplacée.

V.3.G. La catégorisation des segments vocaliques

A la suite de l'adaptation des frontières, les segments VOC sont étudiés à nouveau et particulièrement les frontières posées à l'intérieur d'un continuum vocalique. Certaines peuvent être supprimées selon leur position. A ce niveau-là, les segments VOC correspondent soit à des consonnes vocaliques, soit à des voyelles.

Pour catégoriser ces deux classes, l'étude du contraste énergétique semble le moyen le plus pertinent. Il semble que les voyelles (notées VOY) se distinguent par la présence d'un unique maximum sur la courbe d'énergie subjective filtrée par la médiane; l'existence d'un minimum révèle plutôt la présence de consonne vocalique (notées COVOC). En cas d'ambiguïtés, aucune décision n'est prise. Nous avons conscience que cette catégorisation est incomplète. Nous pensons que la segmentation d'un continuum vocalique est plus efficace en analysant des ruptures de propriétés et traits. Ainsi, dans la séquence /ae/ de l'énoncé "aérienne", une identification de traits d'ouverture/fermeture peut mettre en évidence une cassure entre /a/ et /e/ (Figure 92, p.195). Il s'agit là d'une segmentation a posteriori qui nécessite déjà des informations de plus haut niveau.

V.3.H. L'étude des groupes consonantiques

Les groupes consonantiques constituent une difficulté importante lors de la segmentation. Plus les constituants du groupe sont différents d'un point de vue phonétique, plus le repérage est facile. Ainsi, les combinaisons du type [obstruante* sourde + consonne vocalique et vice versa] ne posent pas trop de problème, les composantes étant distinguées nettement (ex: bicyclette, cultivateur). Par contre, l'association de deux consonnes vocaliques commence à poser des problèmes (ex: format), surtout dans le cas de /r/ dont les multiples formes empêchent un décodage efficace. La difficulté majeure provient du groupement de deux consonnes sourdes (ex: aspiratrice). Dans ce cas-là, la fusion des unités est très fréquente. L'apport de connaissances sur ces phénomènes, tel que le travail de (Meunier, 1994) s'avérerait certainement très fructueux.

* une obstruante est une occlusive ou une constrictive (fricative)

V.4. Bilan

V.4.A. Récapitulatif

Le module de segmentation et de macro-classification SAPHO fait émerger progressivement les formes phonétiques d'un signal d'entrée en réalisant les étapes suivantes :

- calcul de paramètres acoustiques :RMS, TPZ, TPZSD, NRJ, EnF, dK (Figure 81, p.164)
- obtention des propriétés de base : sil, vx, nrj, fric (Figure 82, p.165)
- primo-catégorisation en SIL, VOC ou CTB (Figure 82, p.165)
- identification des macro-classes et réadaptation progressive des frontières en plusieurs passes. Commentons cette opération avec la Figure 82, p.165.

  ⇒ La primo-catégorisation (colonne [Primo.Frict.]) a identifié un certain nombre de segments qui vont être analysés

  ⇒ La passe [Détect.Fricati.] a repéré un segment fricatif (l'explosion de /k/, trames 26 à 30) mais ne l'a pas déclaré comme une fricative pour des raisons de durée. Elle l'a donc qualifié de ConSonne Non voisée (CSN)

  ⇒ La passe [Détect.Ocn.] a repéré une zone de silence (trames 31 à 38) ainsi qu'une explosion (trame 39). Deux indices très forts qui permettent de catégoriser le segment comme une occlusive sourde (OCN) qui correspond à /t/

  ⇒ La passe [Détect.Bvx.] est chargé de repérer des consonnes occlusives ou vocaliques formées essentiellement d'une barre voisée (bvx). Le repérage d'une explosion à la trame 12 ainsi que d'autres indices (faible NRJ, faible structure formantique, Figure 81, p.164) permettent le repérage d'une occlusive voisée, qui correspond à /d/. D'autres segments (trames [21 ;22], [57 ;59]) sont étiquetés barre voisée (bvx) sans plus amples informations.

  ⇒ La passe [Adapt.Bnd] adapte les frontières (boundaries). Les explosions d'occlusives sont intégrées aux tenues pour ne former qu'un segment (trame 12 pour /d/, trame 39 pour /t/). La fin de l'énoncé est interprété comme un relâchement vocalique et intégré au segment vocalique qui précède (trames 57 à 60).

  ⇒ La passe [Analys.Vocali] analyse les deux segments vocaliques (voc). Aucune coupure franche n'est notée à l'intérieur de chacun d'eux. Leur position phonotactique et la présence de forte énergie les identifient comme des voyelles (VOY).

  ⇒ les dernières passes adaptent légèrement les frontières

- Le résultat est un étiquetage des trames d'analyse en macro-classes: voyelle, occlusive, constrictive, consonne vocalique, silence... (Figure 82, p.165, colonne « output »). Ces résultats seront utilisés ultérieurement par le module superviseur du système.

TEMPS	PARAMETRES ACOUSTIQUES											
trame	RMS	NRJ	med	EnF	H/M	O/F	sf	dK	TPZ	TPZsd	Vx	F0
1	0	0	0	1	62%	O	0	24	9	41	0	0
2	1	0	0	1	65%	O	0	24	1	46	0	0
3	1	0	0	1	65%	F	0	36	12	38	0	0
4	3	2	2	1	64%	F	0	64	2	13	0	0
5	13	7	7	1	56%	F	0	40	2	5	5	197
6	16	7	7	1	64%	F	1	26	2	2	8	197
7	16	7	7	1	64%	F	1	2	2	3	9	196
8	18	8	8	2	60%	F	1	4	2	3	11	196
9	19	9	9	2	67%	F	1	6	2	1	12	198
10	20	10	10	2	67%	F	1	1	2	2	13	200
11	20	10	10	2	60%	F	1	8	2	3	13	203
12	21	19	19	27	76%	F	1	109	5	28	13	219
13	35	23	23	16	69%	F	2	65	3	39	29	239
14	52	35	35	21	82%	F	2	23	3	33	52	254
15	76	58	58	35	82%	F	2	21	3	26	91	275
16	102	81	81	63	85%	F	3	19	3	23	113	292
17	116	100	93	78	86%	F	3	13	3	30	126	301
18	122	107	93	92	88%	F	2	8	4	26	130	306
19	114	93	93	73	86%	F	3	17	3	20	110	306
20	51	37	37	25	82%	F	1	22	3	15	29	289
21	20	14	14	3	64%	F	1	39	3	5	5	287
22	7	5	5	2	68%	F	0	19	3	13	1	279
23	2	1	1	1	79%	F	0	50	6	36	0	0
24	1	1	1	1	78%	F	0	45	10	43	0	0
25	1	1	1	1	80%	F	0	15	11	43	0	0
26	5	9	9	23	89%	O	0	74	15	40	0	0
27	7	13	9	32	79%	O	1	41	27	39	0	0
28	6	11	9	23	90%	O	0	40	28	49	0	0
29	6	6	6	10	92%	O	0	32	7	58	0	0
30	2	1	1	1	81%	F	0	60	2	44	0	0
31	0	0	0	1	73%	F	0	29	15	47	0	0
32	1	0	0	1	69%	O	0	35	1	48	0	0
33	0	0	0	1	73%	O	0	25	10	45	0	0
34	1	0	0	1	72%	F	0	36	0	48	0	0
35	0	0	0	1	72%	O	0	34	3	43	0	0
36	0	0	0	1	74%	O	0	15	3	42	0	0
37	1	0	0	1	74%	O	0	17	0	46	0	0
38	1	2	2	3	73%	O	0	38	2	47	0	0
39	8	12	12	27	67%	O	0	34	16	40	0	0
40	10	14	14	19	70%	F	0	51	14	44	2	283
41	84	74	74	88	77%	F	1	31	3	15	80	283
42	100	88	88	100	80%	F	5	22	6	28	112	278
43	94	98	98	124	82%	F	5	14	11	36	107	274
44	93	114	114	170	87%	O	5	15	13	33	100	271
45	107	129	129	208	88%	O	4	7	10	36	95	268
46	102	136	129	221	88%	O	4	3	10	34	93	268
47	100	134	129	213	88%	O	4	3	10	34	92	268
48	108	127	127	190	87%	O	5	6	11	36	90	269
49	99	112	112	146	84%	O	5	15	6	26	91	270
50	94	101	101	100	79%	O	4	16	8	25	92	273
51	95	96	96	76	72%	O	3	14	4	31	95	275
52	106	94	94	59	66%	O	3	20	3	23	115	275
53	89	72	72	36	63%	F	1	20	3	11	110	273
54	65	47	47	14	67%	F	1	30	3	6	73	266
55	47	30	30	5	64%	F	1	25	3	5	41	258
56	38	25	25	4	67%	F	1	19	3	5	27	258
57	24	17	17	3	73%	F	1	4	3	3	10	264
58	13	9	9	3	72%	F	1	17	3	5	3	266
59	7	5	5	2	65%	F	0	28	3	7	1	266
60	2	2	2	2	66%	F	0	37	3	30	0	0
61	1	1	1	2	67%	O	0	43	3	36	0	0
62	1	1	1	1	65%	O	0	26	6	33	0	0
63	1	1	1	1	68%	O	0	19	9	38	0	0
64	1	1	1	1	66%	O	0	19	8	35	0	0

Figure 81: Calcul des paramètres acoustiques sur le mot " dictée". Le temps se déroule de bas en haut. Chaque trame représente un pas de 10ms. Chaque colonne représente un paramètre acoustique : RMS = intensité RMS, NRJ ='intensité subjective, med = NRJ filtrée par la médiane, EnF = énergie formantique, H/M = degré d'acuité, sf = degré de structure formantique, dK =fonction d'instabilité spectrale, TPZ = taux de passage par zéro, TPZsd = TPZ sur signal filtré, Vx =énergie de voisement, F0 = fondamentale

Tps	PROPRIETES				Primo-Catégorisation			ETAPES DE LA SEGMENTATION								
trame	Sil	Vx	Fric	Nrj	Primo-Energie	Primo-Voise.	Primo-Frict.	Detect. Fricati.	Detect. Ocn	Detect. Bvx	Adapt. Bnd	Analys. Vocali.	Adapt. Bnd	Analys Cons.	Output	
1	si	-	fr	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
2	si	-	fr	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
3	si	-	-	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
4	-	-	-	-	CTB	CTB	CTB	CTB	CTB	CTB	CTB	CTB	.?.	.?.	.?.	
5	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
6	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
7	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
8	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
9	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	**/d/**
10	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
11	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	OCV	OCV	OCV	OCV	OCV	OCV	
12	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	!!!	OCV	OCV	OCV	OCV	OCV	
13	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
14	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
15	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
16	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	**/i/**
17	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
18	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
19	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
20	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
21	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	bvx	bvx	bvx	bvx	CSV	VOY	
22	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	bvx	bvx	bvx	bvx	CSV	VOY	
23	-	-	-	-	CTB	CTB	CTB	CTB	CTB	CTB	CTB	CTB	.?.	.?.	.?.	
24	si	-	fr	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	CSN	
25	si	-	fr	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	CSN	
26	-	-	fr	-	CTB	CTB	CTB	CSN	CSN	CSN	CSN	CSN	CSN	CSN	CSN	**/k/**
27	-	-	fr	-	CTB	CTB	CTB	CSN	CSN	CSN	CSN	CSN	CSN	CSN	CSN	
28	-	-	fr	-	CTB	CTB	CTB	CSN	CSN	CSN	CSN	CSN	CSN	CSN	CSN	
29	-	-	fr	-	CTB	CTB	CTB	CSN	CSN	CSN	CSN	CSN	CSN	CSN	CSN	
30	-	-	fr	-	CTB	CTB	CTB	CSN	CSN	CSN	OCN	OCN	OCN	OCN	OCN	
31	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
32	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
33	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
34	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
35	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	**/t/**
36	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
37	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
38	si	-	fr	-	SIL	SIL	SIL	SIL	OCN	OCN	OCN	OCN	OCN	OCN	OCN	
39	-	-	fr	-	CTB	CTB	CTB	CTB	!!!	!!!	OCN	OCN	OCN	OCN	OCN	
40	-	vx	fr	-	CTB	CTB	CTB	CTB	CTB	CTB	OCN	OCN	OCN	OCN	OCN	
41	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
42	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
43	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
44	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
45	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
46	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
47	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
48	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
49	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	**/e/**
50	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
51	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
52	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
53	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
54	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
55	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
56	-	vx	-	+n	voc	voc	voc	voc	voc	voc	voc	VOY	VOY	VOY	VOY	
57	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	bvx	voc	VOY	VOY	VOY	VOY	
58	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	bvx	voc	VOY	VOY	VOY	VOY	
59	-	vx	-	-	CTB	CTB	CTB	CTB	CTB	bvx	voc	VOY	VOY	VOY	VOY	
60	-	-	-	-	CTB	CTB	CTB	CTB	CTB	CTB	voc	VOY	VOY	VOY	VOY	
61	si	-	-	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
62	si	-	-	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
63	si	-	-	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	
64	si	-	-	-	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	

Figure 82: Propriétés et étapes de segmentation sur le mot " dictée". Le temps est synchronisé avec la Figure 81. Les 4 colonnes propriétés se rapportent aux propriétés de base (cf p.152). Les colonnes 5 à 7 sont relatives à la primo-catégorisation (cf p.153). Les suivantes illustrent les différentes passes de la segmentation et la macro-classification. La colonne Output est le résultat final de l'algorithme.

V.4.B. Mise au point

La mise au point du module SAPHO a été longue et a nécessité de longues séances de travail pour formaliser le savoir d'experts phonéticiens. Le coeur du dispositif est un ensemble de règles simples du type SI [tel phénomène] AVEC [tel contexte] ET [telle contrainte] ALORS [c'est ça]. Nous avons tenté d'éviter de tomber dans le piège de la mise au point « ad hoc » par l'adoption de paramètres ou règles sans fondement physique ou phonétique. Aucune propriété n'est obtenue en multipliant la valeur du canal 15 du vocodeur par le taux de passage par zéro divisé par la fréquence fondamentale. A chaque calcul ou transformation correspond une réalité physique ou phonétique.

La mise au point des différents seuils intervenant dans le module SAPHO a été effectuée sur un corpus de 40 mots comprenant 5 locuteurs. Il s'agit de mots isolés empruntés au travail de (Giraud, 1991). Ce corpus a ensuite été écarté des évaluations.

V.4.C. L'évaluation

L'évaluation d'un algorithme de segmentation est une opération délicate. En principe, elle nécessite l'utilisation d'un corpus étiqueté manuellement par un expert phonéticien, ceci afin de comparer les frontières marquées à la main à celles fournies par le processus automatique. Un score de divergence entre les deux étiquetages fournit un résultat chiffré plus ou moins exploitable. Pour éviter les problèmes de stratégie de segmentation propre à chaque segmentateur, il est préférable d'utiliser plusieurs références manuelles et d'écarter les propositions atypiques. De plus, pour que les résultats soient représentatifs, le corpus de test doit être suffisamment important pour contenir divers locuteurs, divers contextes... La difficulté d'une telle évaluation réside dans le temps nécessaire à la constitution et l'étiquetage d'un tel corpus. Nous n'avons pas suivi cette voie.

Le but de l'algorithme SAPHO est de fournir la distribution syntagmatique des unités fonctionnelles d'un signal à décoder. La recherche de frontières n'est qu'une étape pour obtenir cette distribution. Une évaluation très précise des frontières ne nous apparaît pas nécessaire. Par contre, l'évaluation de la qualité de la catégorisation en macro-classes est loin d'être sans intérêt pour connaître les points forts et faibles de SAPHO. L'algorithme a donc été évalué comme pré-traitement à l'accès lexical. Ses performances seront présentées au chapitre exposant les performances globales du système (cf. § « Evaluation du système », p.216).

V.4.D. Quelques réflexions

L'algorithme de segmentation et de macro-classification revêt pour nous une importance cruciale. Il permet l'analyse de l'axe syntagmatique et autorise le repérage des unités fonctionnelles. L'identification qui en découle est le résultat de la mise en compétition de ces unités. Ainsi, par exemple, pour les occlusives voisées, le but de la reconnaissance sera de fournir le lieu d'articulation (compétition entre /b/, /d/ et /g/). Cette tâche est facilitée par le fait que l'on compare des « poires à des poires » et « des pommes à des pommes », autrement dit que les unités mises en compétition sont homogènes.

VI. LA RECONNAISSANCE ANALYTIQUE

« Dans le domaine de la science, jamais coeur timide n'a gagné belle dame. »
Hocart A.M.

« Je ne connais pas la clé du succès, mais la clé de l'échec est d'essayer de plaire à tout le monde. »
Bill Cosby.

Plan du chapitre

RESUME

Le module de reconnaissance analytique est fondé sur l'utilisation de règles phonétiques simples (Ghio & Rossi, 1995). Seules les voyelles sont décodées, les consonnes l'étant déjà partiellement dans le module de macro-classification SAPHO. Pour l'identification des voyelles, des paramètres acoustiques sont extraits de l'analyse en bandes critiques par trames d'analyse de l'ordre de la centiseconde. De cette information sont déduits deux traits fondamentaux (un trait d'ouverture et un trait d'acuité), ce qui permet une catégorisation en 4 classes. A l'intérieur de chaque catégorie, un repérage des maxima d'énergie dans la représentation spectrale en bandes critiques permet de distinguer en partie les voyelles entre elles. Un suivi temporel à court terme permet d'éliminer toute apparition instable de ces propriétés. Le module a été évalué de façon indépendante puis intégré au dispositif complet de décodage. Nous rappelons qu'il fonctionne indépendamment et parallèlement aux autres modules ascendant.

VI.1. Le principe de la reconnaissance analytique par règles

Dans ce chapitre, nous présentons le module de reconnaissance analytique des voyelles, fondé sur un ensemble de règles simples. Dans l'architecture du système (cf. Figure 63, p.129), il existe un second module de reconnaissance analytique constitué d'un ensemble de graphes orientés à transitions d'état construits pour modéliser tous les allophones d'une voyelle. Cette partie peut être consultée dans (Ghio & Rossi, 1994). Ce travail, exclusivement développé par Rossi, ne sera pas développée dans notre étude.

VI.1.A. La reconnaissance analytique des consonnes

La reconnaissance analytique des consonnes est réalisée partiellement par le module de macro-classification SAPHO, qui fournit un étiquetage des segments en macro-classes (ex: fricative sourde, occlusive sonore, consonne vocalique). Une étape supplémentaire pour affiner l'analyse consiste à la recherche du lieu d'articulation. Ainsi, pour les occlusives sourdes, il s'agit de distinguer la labiale /p/, de la dentale /t/ et la vélaire /k/. Dans cet exemple, la solution réside, d'une part, dans l'identification fréquentielle du bruit d'explosion et, d'autre part, dans les transitions formantiques avec les voyelles adjacentes. Cette analyse nécessite une robuste étude de l'environnement phonémique du fait de la très grande variabilité contextuelle (Figure 7, p.15). Bien que certains travaux existent comme ceux de (Bonneau et al., 1996), nous avons provisoirement renoncé à cette tâche pour accentuer l'étude sur l'identification des voyelles.

VI.1.B. La reconnaissance analytique des voyelles

La notion de propriété, indice et trait a été présentée au § II.II.4.II.4.D., p.65. Nous savons qu'il existe un classement des phonèmes par traits à la fois articulatoires et à la fois acoustiques. Aucun système ne nous semble a priori meilleur. Il faut garder à l'esprit que la notion de trait est du domaine de l'abstraction et que, dans sa conception phonologique, il est le moyen de distinguer deux phonèmes. L'erreur serait de croire que la recherche de traits est immédiate et que ceux-ci sont directement visibles dans le signal de parole. En fait, nous pensons que l'identification par traits est un moyen structuré de rechercher des informations discriminantes pour séparer deux catégories abstraites différentes.

VI.1.B.a. Le cas de la nasalité

Bien qu'existantes, comme le prouvent les travaux de (Chafcouloff, 1994), les études acoustiques sur la nasalité dans la parole ne permettent pas vraiment de tirer des informations robustes susceptibles d'être exploitées dans les systèmes de décodage automatique. Certaines anti-résonances acoustiques dues à la cavité nasale sont parfois visibles sur le signal mais ce phénomène est loin d'être constant et reste très délicat à détecter. Il semble que les meilleurs indices de nasalité soient des indices temporels de rupture spectrale liés à l'ouverture et la fermeture du voile du palais. Les voyelles nasales sont généralement plus longues et sont très souvent formées de deux parties: une première phase non nasale, qui la rend similaire à son équivalent oral, et une deuxième partie nasalisée (ex: /ã/ = [a]+[~]). Certaines expériences réalisées au laboratoire « Parole et Langage » avec la station de travail EVA (Teston & Galindo, 1995), qui permet de mesurer le débit d'air nasal et donc de détecter réellement les

phénomènes de nasalité, semblent conforter l'idée que le voile du palais, responsable de la nasalisation, n'est pas nécessairement synchrone à la phonation. Autrement dit, lors de la production de /ã/, l'émission débute sans nasalisation, ce qui donne un [a] puis le voile s'abaisse avec un certain retard pour ouvrir alors la cavité nasale.

Malgré ces résultats intéressants, aucun indice robuste n'a pu être parfaitement formalisé pour notre application en décodage automatique. L'information discriminante pour séparer voyelles nasales et orales étant absente, le trait de nasalité est mis de côté, ce qui entraîne la non distinction des voyelles /a/ et /ã/, /ɔ/ et /ɔ̃/, /ɛ/ et /ɛ̃/, /œ/ et /œ̃/. Dans la suite, nous ne mentionnerons plus que les formes orales de ces voyelles.

VI.1.B.b. Une première classification des voyelles en quatre classes

La reconnaissance analytique par règles s'appuie sur les informations fournies par « *CritiVoc* », décrit au § III.III.3, p.109. Ce vocodeur utilise une modélisation psycho-acoustique (pondération sonique, intégration par bandes critiques) et effectue une représentation compacte temps - fréquence du signal de parole. Un ensemble de propriétés estimées par l'analyse des canaux du vocodeur situés entre 200 Hz et 800 Hz, autorise la distinction entre deux groupes de voyelles qui se différencient phonologiquement par une opposition d'ouverture/fermeture:

- voyelles ouvertes: /œ/, /ɛ/, /ɔ/, /a/
- voyelles fermées: /y/, /i/, /e/, /u/, /o/, /ø/

Parmi les voyelles ouvertes, une propriété comparant les énergies de moyennes fréquences par rapport aux hautes fréquences permet de mettre en évidence:

- /ɔ/ voyelle grave
- /ɛ/ voyelle aiguë

Cette propriété ne permet pas d'estimer l'aspect grave/aigu de /a/ qui est classé grave en contexte vélaire et aigu ailleurs. De même, la position centrale de la voyelle /œ/ laisse une ambiguïté sur son caractère grave/aigu.

Parmi les voyelles fermées, la même propriété met en évidence la distinction entre:

- /u/, /o/ voyelles graves
- /y/, /i/, /e/, /ø/ voyelles aiguës

Cette analyse permet d'effectuer une catégorisation en 4 macro-classes:

- voyelles ouvertes/aiguës: /œ/[1] , /ɛ/, /a/[2]
- voyelles ouvertes/graves: /œ/[1], /ɔ/, /a/[3]
- voyelles fermées/aiguës: /y/, /i/, /e/, /ø/
- voyelles fermées/graves: /u/, /o/

[1] nous avons choisi de placer ce phonème dans deux macro-classes du fait de l'ambiguïté grave/aigu
[2] en contexte vélaire
[3] hors contexte vélaire

La présence d'un phonème dans deux classes n'est pas gênante car pour chaque macro-classe de voyelles, les maxima d'énergie permettent l'identification de la voyelle comme nous allons le voir dans le paragraphe suivant. De plus, le trait grave/aigu de /a/ apporte finalement une information contextuelle.

VI.1.B.c. Une analyse des maxima d'énergie dans le spectre en bandes critiques

La catégorisation en macro-classes obtenue, l'algorithme va s'attacher à détecter les propriétés autorisant la discrimination à l'intérieur d'une catégorie. Pour cela, il effectue une recherche des maxima d'énergie pertinents dans la distribution spectrale en bandes critiques. Cette opération, qui consiste à repérer les pics d'énergie dans la représentation spectrale, correspond à une analyse formantique. L'avantage de l'analyse en bandes critiques réside dans la fusion des composantes harmoniques en différentes masses énergétiques pertinentes (Figure 83).

Des problèmes peuvent apparaître pour les voix particulièrement aiguës. En effet, dans ce cas-là, la pondération sonique atténue peu la fréquence fondamentale et celle-ci émerge alors nettement dans l'organisation spectrale du modèle auditif. De ce fait, les bandes critiques les plus basses se calquent sur la fondamentale et ses harmoniques (Figure 83), ce qui introduit des pics non pertinents dans l'analyse en bandes critiques. La détection de ce type de configuration est important et peut être réalisée très simplement. Par exemple, si les bandes K2 ($\Leftrightarrow$ 100;200Hz) et K4 ($\Leftrightarrow$ 300;400Hz) sont émergentes, il s'agit manifestement de la fondamentale et son harmonique car aucune voyelle n'a de configuration formantique de la sorte. Il suffit alors de réévaluer les pics d'énergie en tenant compte de ce phénomène. Nous opérons donc à une correction de décision.

VI.1.B.d. L'identification de la voyelle

Une première phase a déjà séparé les voyelles en 4 macro-classes. A l'intérieur de chaque catégorie, une analyse des emplacements des maxima d'énergie permet d'identifier un ou plusieurs candidats à chaque trame d'analyse. Pour effectuer cette opération, l'algorithme consulte une table de données fournissant cette information (Tableau 13). Ces données ont été dans une premier temps proposées a priori par l'expertise de phonéticiens. Dans un second temps, nous avons affiné les résultats grâce à l'étude d'un corpus de 40 mots comprenant 5 locuteurs. Il s'agit de mots isolés empruntés au travail de (Giraud, 1991). Ce corpus a ensuite été écarté des évaluations.

Les phénomènes de recouvrement de zones sont voulus et tiennent compte de la dispersion bien connue des emplacements formantiques des voyelles (Figure 28, p.55). Nous rappelons que cette analyse est effectuée trame à trame de façon automatique et indépendante. Il est clair que les résultats ne prennent de sens que s'il existe une continuité temporelle des caractéristiques identifiées. Aussi, un suivi trame à trame permet d'éliminer les apparitions isolées de traits ou pics d'énergies. La Figure 84 fournit un exemple du décodage analytique des voyelles.

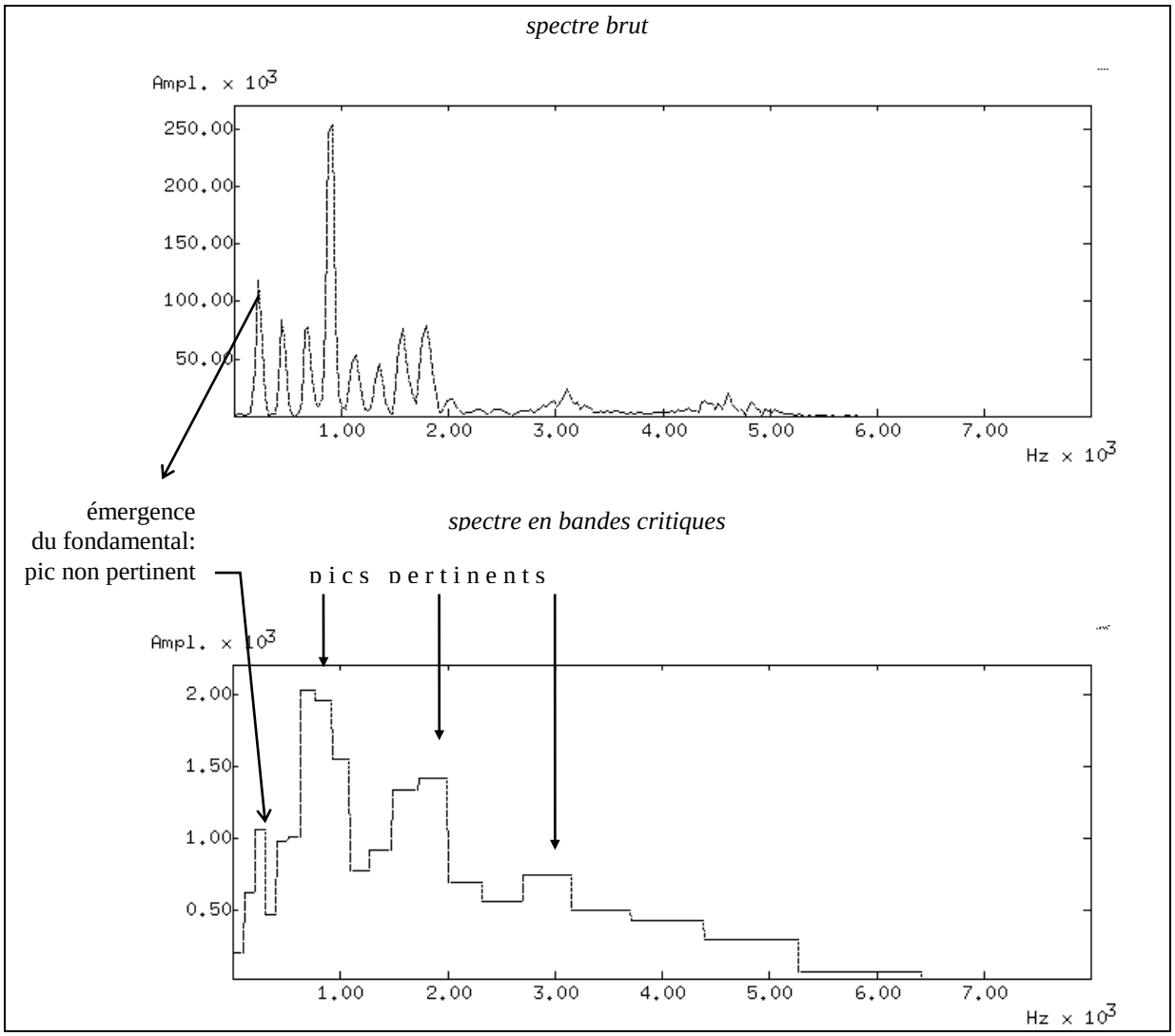


Figure 83: Recherche des pics sur le spectre en bandes critiques du /e/ de " heurté".

Tableau 13: Table des maxima d'énergie dans l'analyse en bandes critiques des voyelles (chaque colonne 'N' correspond à la N^{ème} bande critique)

N→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
i	.	x	x	x	.	.	.	.	.	.	.	.	x	x	x	x	x	.	.	.	.
e	.	.	x	x	x	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.
ɛ	.	.	.	x	x	x	x	.	.	.	x	x	x	x	.	.	.	.	.	.	.
a	.	.	.	.	x	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.
ɔ	.	.	.	x	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.
o	.	.	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
u	.	x	x	x	.	.	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
y	.	x	x	x	.	.	.	.	.	.	x	x	x	x	.	.	.	.	.	.	.
ø	.	.	x	x	x	x	.	.	x	x	x	x	x	.	.	.	.	.	.	.	.
œ	.	.	.	x	x	x	x	.	.	x	x	x	x	.	.	.	.	.	.	.	.

n° de la trame centiseconde — *Analyse CritiVoc* — *taux d'acuité* — *trait fermeture* — *résultats*

t	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K10	K11	K12	K13	K14	K15	K16	K17	K18	NRJ	Acu	Fer	RES	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	M	.	
1	0	0<	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	F	.	*transcription manuelle*
2	0	0	0<	0	0	0	0	0	0	0#	0	0	0	0	0	0	0	0	0	0%	F	.	
3	0	0	0<	0	0	0	0	0	0	0	0#	0	0	0	0	0	0	0	0	0%	O	o	
4	0	0*	0	0	0	0	0<	0	0	0	0#	0	0	0	0	0	0	0	0	10%	O	o&	
5	0	1	1	2*	2	3<	2	1	0	1	2#	1	0	0.	0	0	0	0	2	11%	O	o&	
6	0	2	4	2	2<	5	5	2	1	1	3#	1	0	0.	0	0	1*	0	3	14%	O	o&	
7	2	9	21*	9	11	34	80<	21	5	10	25#	7	2	4.	4	1	1	0	25	11%	O	o&	
8	5	21	51#	21	29	82	183<	29	12	17	33*	12	3	11.	9	2	3	0	52	12%	O	.	
9	9	36	87#	48	58	175	325<	89	16	13	25*	5	1	3	5.	4	2	0	90	3%	O	.	
10	12	48	126#	63	55	150	200<	83	6	6	24.	11	6	26*	14	7	2	0	84	16%	O	.	
11	17	70	199	107	135<	361	331	111	23	35	101#	87	19	52*	30	28	4	0	171	25%	O	&	/a/
12	23	114	321	183	325	1176<	809	179	47	73	301	547#	57	114*	73	54	7	0	440	36%	O	a&	
13	27	121	326	183	343	1204<	659	62	36	63	282	774#	144	128	122	176*	23	1	467	54%	O	aE&	
14	30	129	403	275	828	2161<	1128	60	34	48	216	947#	312	235	194	174	13	1	719	54%	O	aE&	
15	34	151	469	362	926<	952	484	57	26	37	115	577#	495	240	196	156	32	2	531	69%	O	aE&	
16	39	174	560	471	1490<	1567	230	69	18	21	56	309	646#	262	191	147	34	1	628	79%	O	aE&	
17	43	186	646	685	1802	2007<	101	63	26	19	40	194	591#	197	114	82	19	1	682	81%	O	aE&	
18	46	198	613	595<	772	741	46	28	17	15	25	120	513#	154	151	283*	139	4	446	90%	O	.	
19	50	218	742	752	877<	418	27	17	11	10	13	60	303#	110	121	294*	195	6	423	92%	F	e	
20	53	243	874	1190	1256<	364	28	19	9	10	15	82	478#	259	216	348*	184	3	564	94%	F	e	
21	55	254	973	1669	1821<	225	29	13	7	7	14	77	488*	293	294	513#	388	6	713	95%	F	e	
22	55	265	1108	2116	2793<	164	26	12	8	10	15	61	382*	212	197	397	455#	8	829	95%	F	e	
23	54	270	1128	2585	3428<	142	24	9	5	6	11	44	270#	157	128	205	265*	4	873	94%	F	e	
24	48	232	986	2547	3493<	207	29	9	5	6	10	35	182*	130	138	169	244#	6	848	92%	F	e	/e/
25	42	201	777	2043	2522<	412	34	14	8	10	18	75	304*	168	180	205	332#	10	736	92%	F	e	
26	37	170	612	1189	1307<	670	30	26	16	15	34	162	472*	191	276	291	482#	15	600	92%	F	e	
27	32	147	524	807	1085<	904	45	49.	21	18	56	272*	249	141	188	206	346#	7	510	87%	M	e	
28	29	137	424	609	1012<	748	81	56	16	20	86	319*	90	132	155	201	356#	4	448	81%	M	e	
29	17	81	223	364	535<	262	79	23	8	9	42	85*	26	76.	61	126	240#	3	226	78%	F	e	
30	8	36	106	213<	144	61	21	6	5	8	46	52*	11	24.	10	25	54#	1	83	66%	F	e	
31	4	20	63	136<	57	15	7	4	3	3	17	30#	4	12*	3	3	7.	0	39	63%	F	y@	
32	3	22	74	120<	51	12	3	3	4	6	12	26#	3	11*	4	3	6.	1	36	65%	F	y@	/ɾ/
33	5	25	82	181<	54	9	5	3	1	2	8	12#	3	7.	5	3	8*	1	41	66%	F	y@	
34	14	62	232	467<	202	28	6	4	2	3	18	54#	17	23	24*	8	14.	1	118	80%	F	y@	
35	30	145	503	825<	405	50	14	5	3	3	11	50.	43	64	94#	52	85*	2	239	90%	F	iey	
36	44	233	854	1393<	581	50	22	6	3	3	8	33	88	105	203#	82	137*	3	385	93%	F	iey	
37	49	263	1029	1610<	495	28	16	2	2	1	3	15	76	114	169#	62	83*	2	402	94%	F	iey	
38	50	281	951	1261<	233	13	9	1	1	1	1	5	26	67	71	74	86#	5	313	95%	F	ie	
39	47	255	676<	555	82	8	3	1	1	1	1	2	12	44	47	102#	60	3	190	96%	F	ie	
40	42	223	530<	129	23	5	2	0	0	0	0	1	8	38	46	187#	158	8	140	98%	F	ie	
41	37	187	528<	63	9	2	1	1	0	0	0	1	6	27*	22	111	117#	7	112	98%	F	ie	
42	37	186	526<	54	8	2	1	0	0	0	0	1	5	21*	19	111	153#	7	113	98%	F	ie	/j/
43	42	226	634<	74	8	2	1	0	0	0	0	1	5	22*	20	119	132#	6	129	98%	F	ie	
44	45	268	851<	126	12	2	1	0	0	0	0	1	8	30	32	164	174#	8	173	98%	F	ie	
45	56	379	1271<	201	33	6	3	1	1	1	1	2	17	53	60	259	281#	11	264	98%	F	ie	
46	74	556	1604<	323	78	12	7	2	1	1	1	5	36	90*	69	200#	197	6	326	97%	F	ie	
47	75	561	1646<	875	156	13	7	2	2	1	2	10	72	143*	113	224#	212	6	412	97%	F	ie	
48	67	483	1659	2202<	407	24	10	2	2	2	3	15	113	166#	65	108	140*	3	547	96%	F	ie	
49	58	389	1447	3935<	893	73	14	4	4	3	7	34	192#	132	90	132	185*	6	760	95%	F	e@	
50	47	285	836	1980<	794	140	14	7	6	5	11	60	278#	121	102	142	227*	7	506	94%	F	e@	
51	47	281	710	1115	1499<	224	27	10	8	8	22	130	487#	76	74	106	159*	5	499	92%	F	e@	
52	42	244	639	1354	3589<	510	43	17	10	11	33	197	541#	95	69	127	190*	6	772	90%	F	e@	
53	37	218	596	1542	6040<	989	88	20	10	13	46	272	433#	128	78	136	202*	7	1086	86%	F	e@	
54	36	207	556	1383	6269<	1184	145	18	15	19	73	397	484#	167	91	157	217*	9	1143	83%	F	e@	
55	38	223	543	1096	4544<	1239	193	20	22	27	90	485#	473	211	120	184*	184	6	970	81%	M	a@E&	
56	39	224	517	950	3211<	1261	145	22	23	30	113	614#	548	249	122	138	145*	7	836	83%	M	a@E&	/e/
57	37	213	487	639	1344<	929	117	21	21	25	96	574#	562	237	163	212*	204	8	589	86%	M	a@E&	
58	34	189	444	500	529	944<	84	25	19	19	65	392	409#	112	101	131*	123	6	413	84%	O	aE&	
59	33	185	442	527#	315	862<	81	26	17	15	49	277*	273	50	80.	70	58	3	336	80%	O	.	
60	36	201	496	616#	201	687<	74	28	14	11	27	144*	127	22	54.	37	26	2	280	71%	O	.	
61	38	221	567	736<	174	483#	61	27	13	10	23	117*	86	13	37.	27	21	2	266	68%	F	.	
62	38	230	613	796<	154	243#	38	19	10	10	20	114*	60	9	28.	16	15	2	241	70%	F	.	
63	35	223	637	794<	100	69	14	8	5	5	13	71#	23	4	11*	3	4.	1	202	71%	F	y	
64	26	170	490	511<	34	9	2	1	2	3	11#	3	1	1.	1	2*	0	1	127	65%	F	y	
65	24	182	322<	69	15	7	1	0	0	0	1	2#	1	1	2*	0	0	0.|	63	68%	F	y	
66	30	240	413<	66	23	11	1	0	0	0	1*	0	0	1	3#	0	0	0.|	79	70%	F	iey	
67	33	259	394<	75	29	16	2	0	0	0	1*	1	0	1	3#	0	0	0.|	82	60%	F	iey	
68	34	266	396<	84	34	25	4	0	0	1	1*	1	1	3	6#	1	0	0.|	86	64%	F	iey	
69	34	264	418<	78	30	26	7	0	0	0	1	1*	1	2	8#	1	0	0	87	61%	F	iey	/n/
70	31	250	405<	78	25	24	6	0	0	0	1	1*	1	2	8#	1	0	0	84	63%	F	iey	
71	29	230	360<	82	24	22	5	0	0	0	1	1	1	3	8#	2	1	0	77	68%	F	iey	
72	27	212	305<	90	26	18	4	0	0	0	1	1	1	3	7#	1	1	1	70	72%	F	iey	
73	26	198	261<	98	29	18	3	0	0	0	2*	1	1	3	6#	1	1	0	65	68%	F	iey	
74	26	193	323<	131	20	8	5	2	2	4	9#	4	3	3	2	1	2*	1	74	41%	F	y@	
75	36	237	567	976<	136	12	18	12	5	6	20	32#	4	3	3	2	4*	1	207	44%	F	@	
76	35	223	562	751<	134	16	31	17	6	6	20	33#	5	3	6*	3	5.	1	186	40%	F	@	
77	32	203	446	528<	183	14	37	22	5	6	21	31#	4	2	5*	3	5.	1	155	35%	F	@	/ø/
78	27	172	336	393<	188	13	36	20	4	7	26	39#	4	2	4*	2	2.	0	128	35%	F	@	
79	23	146	267	345<	180	11	32	19	4	7	33	43#	3	1	3*	1	1	0	112	34%	F	@	
80	19	122	198	233<	141	7	22	15	4	6	22	27#	2	1	1*	1	1.	0	82	31%	F	@	
81	15	100	136	139<	108	5	15	10	4	4	13#	13	1	1	1	1	2*	0	57	28%	F	o@	
82	12	74	91<	82	66	2	4	7#	1	1	5*	4	0	0	0	0	1.	0	35	24%	F	o@	
83	8	42	56<	37	21	2	3	3*	1	1	4#	2	0	0	0	0	0.	0	18	21%	F	uo	
84	6	26	54	74<	6	1	2	2#	1	0	1*	1	0	0	0	0	0.	0	17	18%	F	uo	
85	4	17	33	44<	4	1	2	2#	0	0	1*	0	0	0	0	0.	0	0	11	8%	F	uo	
86	3	10	16	17<	2	1	1	4#	0	0	1*	0	0	0	0	0	0.	0	6	5%	F	uo	

Figure 84: Répartition spectrale sur le mot "aérienne", locuteur masc. La colonne RES propose les résultats (E=/ ɛ /, @=/ ø /, &=/ œ /, o= /o/ ou / ɔ/)

VI.2. L'évaluation du module de reconnaissance analytique

VI.2.A. Les conditions de l'évaluation

Afin d'évaluer sa pertinence, le module de reconnaissance analytique a été testé indépendamment du dispositif complet. Les tests ont porté sur les sept voyelles orales [a,i,u,e,y,o,ø] issues du corpus acoustique SYL de BD-SON (Cervantès et al, 1986). Les énoncés apparaissent sous la forme de logatomes du type CVCV (ex: titi, bubu...). Nous avons utilisé un contexte consonantique multiple (six occlusives, six constrictives, deux nasales, deux liquides). Dix locuteurs (cinq femmes, cinq hommes) ont été choisis, soit un total de 7 voyelles * 2 réalisations * 16 contextes * 10 locuteurs = 2240 énoncés.

VI.2.B. Les résultats bruts

Le Tableau 14 présente les résultats de façon globale. La signification des colonnes est la suivante:

- [Total] représente le nombre de tests effectués
- [Correct] représente le nombre de cas où la voyelle est correctement reconnue
- [Taux] représente le taux de reconnaissance
- [Ncnd] représente le nombre de candidats total fournis par le système
- [Mcnd] représente le nombre moyen de candidats par test
- [Effic] représente un coefficient d'efficacité qui donne le pourcentage de réponse correcte par rapport au nombre total de candidats fournis par le système

Tableau 14: Résultat global de la reconnaissance analytique

Total	Correct	Taux	Ncnd	Mcnd	Effic
2237*	2068	92,4 %	5264	2,35	39,3 %

L'analyse succincte des résultats est la suivante: en moyenne, le dispositif de reconnaissance analytique par règles propose 2,35 candidats parmi les 10 voyelles [a, i, u, o, e, y, ø, œ, ɔ, ɛ] ; dans 92,4 % des cas, la bonne réponse est parmi les candidats. Ces résultats sont à nuancer en tenant compte de chaque voyelle (Tableau15).

Tableau15: Résultats par voyelle

Voyelle	Total	Correct	Taux	Ncnd	Mcnd	Effic
a	320	299	93%	688	2,2	43%
i	320	303	95%	648	2	47%
u	320	292	91%	703	2,2	42%
e	319	307	96%	819	2,6	37%
o	318	256	81%	723	2,3	35%
y	320	306	96%	900	2,8	34%
ø	320	305	95%	783	2,4	39%

* et non 2240 car le corpus contient parfois des erreurs (un locuteur a dit « maman » au lieu de « mama »...)

Le Tableau 16 illustre le classement des voyelles en fonction du taux de reconnaissance, du nombre moyen de candidats et de l'efficacité du décodage relatifs à chacune d'elle.

Tableau 16: Classement des voyelles en fonction
(a) du taux de reconnaissance, (b) du nombre moyen de candidats, (c) de l'efficacité

Voyelle	Taux (a)
e	96%
y	96%
ø	95%
i	95%
a	93%
u	91%
o	81%

Voyelle	Mcnd (b)
i	2
a	2,2
u	2,2
o	2,3
ø	2,4
e	2,6
y	2,8

Voyelle	Effic (c)
i	47%
a	43%
u	42%
ø	39%
e	37%
o	35%
y	34%

La difficulté d'identification des voyelles graves peut s'expliquer par la faiblesse de F_2 et F_3, ce qui entraîne un contraste très faible dans l'analyse en bandes critiques et donc un manque certain d'information. Les bonnes performances dans le décodage de /y/ et /e/ sont à modérer en tenant compte de l'efficacité. Il semble que ces deux entités aient tendance à faire déclencher un lot important de règles, entraînant un nombre plus élevé de candidats et donc l'apparition de la bonne réponse.

VI.2.C. Les influences

VI.2.C.a. Les effets du contexte consonantique

Le Tableau 17 présente les résultats de la reconnaissance des voyelles en fonction du contexte consonantique, qui est du type CVC. Les colonnes « Correct » et « Taux » ne sont donc pas relatives au décodage des consonnes, mais bien des voyelles.

Tableau 17: Résultats de la reconnaissance analytique des voyelles en fonction du contexte consonantique

Contexte	Total	Correct	Taux	Ncnd	Mcnd	Effic
p	140	129	92,1%	327	2,34	39,4%
t	140	132	94,3%	355	2,54	37,2%
k	140	133	95%	346	2,47	38,4%
b	140	135	96,4%	318	2,27	42,5%
d	140	137	97,9%	333	2,38	41,1%
g	140	134	95,7%	348	2,49	38,5%
f	140	131	93,6%	319	2,28	41,1%
s	140	128	91,4%	315	2,25	40,6%
ʃ	139	127	91,4%	311	2,24	40,8%
v	140	135	96,4%	329	2,35	41%
z	140	130	92,9%	353	2,52	36,8%
ʒ	140	132	94,3%	354	2,53	37,3%
m	140	127	90,7%	295	2,11	43,1%
n	140	127	90,7%	327	2,34	38,8%
l	140	134	95,7%	326	2,33	41,1%
r	138	97	70,3%	308	2,23	31,5%

De façon très nette, le contexte /r/ se distingue par le nombre d'erreurs qui interviennent dans le décodage des voyelles qui lui sont adjacentes. Sa faculté à déformer son entourage est incontestable. La plupart des erreurs proviennent d'un abaissement des maxima spectraux vers les basses fréquences, ce qui est un phénomène connu. Il conviendra d'en tenir compte à l'avenir. Le Tableau 18 illustre le regroupement des contextes consonantiques par macro-classes. La nasalité du contexte consonantique entraînent quelques difficultés dans le décodage des voyelles adjacentes.

Tableau 18: Résultats de la reconnaissance analytique des voyelles en fonction des classes du contexte consonantique (a) par contexte de macro-classes, (b) dans le contexte obstruant seulement

Contexte	Total	Correct	Taux
occlusif sourd	420	394	93,8%
occlusif sonore	420	406	96,7%
fricatif sourd	419	386	92,1%
fricatif sonore	420	397	94,5%
nasal	280	254	90,7%
l	140	134	95,7%
r	138	97	70,3%

Contexte	Total	Correct	Taux
occlusif	840	800	95,2%
fricatif	839	783	93,3%
sonore	840	803	95,6%
sourd	839	780	93%

VI.2.C.b. Les effets « locuteurs »

Le Tableau 19 présente les résultats de la reconnaissance des voyelles en fonction des locuteurs. La variabilité due au locuteur entraîne incontestablement une différence de performances du système. Les résultats médiocres du locuteur «jo» peuvent s'expliquer par le fait que son débit de parole est beaucoup plus rapide que celui des autres locuteurs (20 % au dessus de la moyenne). Les cibles vocaliques sont donc plus difficilement atteintes, ce qui entraîne des erreurs dans l'analyse. Cet exemple permet aussi de prendre conscience des difficultés à résoudre si le décodage agissait sur de la parole spontanée à débit rapide.

Tableau 19: Résultats de la reconnaissance analytique des voyelles en fonction du locuteur

Locuteur	Total	Correct	Taux
lt (f)	224	217	96,9%
bp (m)	224	214	95,5%
nc (f)	222	210	94,6%
rs (f)	224	209	93,3%
jb (m)	224	206	92%
sl (m)	223	204	91,5%
lc (m)	224	204	91,1%
md (f)	224	204	91,1%
po (m)	224	200	89,3%
jo (f)	224	200	89,3%

Le Tableau 20 présente les résultats de la reconnaissance des voyelles en fonction du sexe du locuteur. Il semble se dessiner une tendance légèrement favorable aux voix de femmes, ce qui va à l'encontre de l'idée généralement admise à propos de la difficulté de décodage des voix aiguës.

Tableau 20: Résultats de la reconnaissance analytique des voyelles en fonction du sexe du locuteur

Sexe	Total	Correct	Taux
f	1118	1040	93 %
m	1119	1028	91,9 %

VI.2.D. Les zones de recouvrement

Dans le Tableau 21 sont indiqués le nombre et le pourcentage d'apparitions des voyelles candidates en fonction de la nature du stimulus. Il ne s'agit pas d'une matrice de confusion. Nous faisons référence au phénomène de candidature multiple. L'information que l'on peut extraire de ce tableau permet d'évaluer les zones de recouvrement des règles. Dans l'ensemble, 5264 candidats (cf.*1) ont été proposés pour 2237 stimuli (cf.*2). Sur 5264 candidats, /a/ est apparu 364 fois (cf.*3). Sur 364 candidatures, 299 (cf.*4) correspondaient à un stimulus /a/, 34 (cf.*5) correspondaient à un stimulus /o/, 23 (cf.*6) correspondaient à un stimulus / ø /...

Tableau 21a: Nombre d'apparition des voyelles en tant que candidat en fonction des stimuli présentés

	Stimulus	a	i	u	e	o	y	ø	Total
	Nb	320	320	320	319	318	320	320	2237*2
Candidat									
a		299*4	0	2	6	34*5	0	23*6	364*3
i		1	303	19	148	1	147	17	636
u		0	8	292	1	181	5	27	514
e		17	249	8	307	2	237	86	906
o		10	5	249	0	256	15	118	653
y		5	63	15	205	1	306	141	736
ø		28	9	64	103	91	185	305	785
ɔ		74	0	0	0	74	0	1	149
ɛ		10	0	0	15	0	0	20	45
œ		188	0	0	5	2	0	24	219
xxx		56	11	54	29	81	5	21	257
									5264*1

Le tableau ci-dessous présente les mêmes résultats sous forme de pourcentages. On peut mettre en évidence plusieurs faits:

- les classes de voyelles sont correctement identifiées:
  - voyelles aiguës: /i/, /e/, /y/
  - voyelles graves: /a/, /o/

- au sein d'une classe, la confusion entre les voyelles reste importante:

- les règles pour /e/ fonctionnent dans 27% des cas avec /i/ comme stimulus et 26% avec /y/.
- les règles pour /u/ fonctionnent dans 35% des cas avec /o/ comme stimulus et vice versa

Tableau 21b: Taux d'apparition des voyelles en tant que candidat en fonction des stimuli présentés

	Stimulus	a	i	u	e	o	y	ø	Tot
Candidat									
a		***82%***	0%	0,5%	1,6%	9,3%	0	6,3%	100%
i		0,2%	***48%***	3%	23%	0,2%	23%	2,7%	100%
u		0	1,6%	***57%***	0,2%	35%	1%	5,3%	100%
e		1,9%	27%	0,9%	***34%***	0,2%	26%	9,5%	100%
o		1,5%	0,8%	38%	0%	***39%***	2,3%	18%	100%
y		0,7%	8,6%	2%	28%	0,1%	***42%***	19%	100%
ø		3,6%	1,1%	8,2%	13%	12%	24%	***39%***	100%
ɔ		50%	0	0	0%	50%	0	0,7%	100%
ɛ		22%	0	0	33%	0	0	44%	100%
œ		86%	0	0	2,3%	0,9%	0	11%	100%
xxx		22%	4,3%	21%	11%	32%	1,9	8,2%	100%

VI.3. Bilan

Malgré ses imperfections, le module de reconnaissance analytique semble donner de bonnes informations. De nombreuses améliorations sont possibles comme l'introduction de règles contextuelles qui tiennent compte de la variabilité structurelle. Le cas du contexte /r/ met en évidence les perturbations occasionnées à la voyelle par les unités contiguës. Il faut garder à l'esprit qu'un système qui décode le signal de parole du mot "vocaliser" en une séquence [vokɛlize], ou le mot "casserole" en [kasorol] n'est pas un système qui se trompe. En effet, [ø] peut être une forme allophonique de /a/ s'il précède la voyelle /i/, et [o] peut être une variante de /ø/ s'il précède la consonne /r/.Un tel décodage n'est pas incorrect. Un accès lexical et une vérification descendante permettent de lever des ambiguïtés non résolues de façon montante.

Une autre façon d'aborder les choses est d'introduire un suivi dynamique plus sophistiqué. En effet, pour le moment, l'identification est effectuée trame à trame de façon statique, quasi-indépendante et porte son analyse sur des parties « stables » de voyelles. Seul un rapide suivi à court terme élimine les identifications localisées à une ou deux trames. Il se pourrait très bien que l'introduction de notions tels que « l'overshoot perceptif » (Lindblom & Studdent-Kennedy, 1967) permette d'identifier comme /a/ le 4ème phonème de "vocaliser" même si la partie « stable » ressemble plus à un [ɛ]. Nous pensons, entre autre, que ce genre de technique rendrait l'identification plus robuste notamment dans le cas de parole continue, voire spontanée, où les cibles vocaliques sont rarement atteintes, ce qui nécessite une analyse cinématique perfectionnée.

VII. LA RECONNAISSANCE GLOBALE

« L'imagination est plus importante que la connaissance. »

« Il est plus facile de détruire un atome que de détruire un préjugé. »

Albert Einstein

Plan du chapitre

RESUME

Dans le cadre d'une reconnaissance de grand vocabulaire, il est nécessaire de faire porter le décodage sur des unités de type syllabique, en quantité limitée par rapport au nombre important d'éléments d'un grand lexique. C'est pourquoi le module de reconnaissance globale a pour rôle de catégoriser des segments de type consonne/voyelle (CV). Nous avons choisi les 10 « voyelles » /a, i, u, o+ɔ, e+ɛ, y, œ+ø, ɔ̃, ã, ɛ̃+ œ̃/ et les 16 consonnes /p t k b d g f s ʃ v z ʒ m n l R/ , ce qui donne 160 combinaisons différentes. A partir d'un signal d'entrée, le module de reconnaissance globale donne finalement un résultat sous la forme /pi/, /to/, /fa/... Pour chaque couple /CV/, le dispositif possède 10 exemplaires extraits de 10 locuteurs différents (5 hommes, 5 femmes) pour tenir compte des phénomènes de variabilité, ce qui fait un total de 1600 exemplaires.

La première étape du décodage global consiste à extraire une information du signal. Celle-ci est fournie par le calcul de coefficients P.L.P. L'identification proprement dite consiste à mesurer une distance entre les coefficients du signal d'entrée et ceux des 1600 exemplaires stockés. Les distances les plus courtes correspondent aux plus sérieux candidats. La prise de décision s'effectue indirectement par analyse des premiers prétendants.

Le module a été testé et optimisé de façon indépendante sur le corpus Bd-Sons. Il a ensuite été intégré dans le dispositif ACHILE, où il est piloté par le résultat de la macro-classification et segmentation. Il apporte une information paradigmatique sur les séquences CV repérées par SAPHO.

VII.1. Principe général du module de reconnaissance globale

VII.1.A. Une méthode de type métrique avec des exemplaires

Dans les techniques de classification et de catégorisation, il convient de distinguer la notion de prototype et d'exemplaire. Un prototype est un élément « idéalisé » représentant sa classe d'appartenance. Un exemplaire est plus simplement l'exemple d'élément d'une classe. Une classe est donc représentée soit par un unique prototype, soit par une collection d'exemplaires.

La méthode que nous avons développée est de type métrique et utilise des exemplaires. Elle effectue une comparaison spectrale entre un stimulus et des références. Dans une phase préliminaire, il s'agit de saisir des caractéristiques acoustiques du signal de parole pour les stocker sous forme d'archives. Dans la phase de décodage proprement dite, le dispositif cherche ensuite à mettre en relation les paramètres du stimulus avec ceux qui sont conservés dans le dictionnaire des archives. Les exemplaires sont issus d'énoncés de différents locuteurs hommes et femmes pour tenir compte des phénomènes de variabilité. Le décodage est donc conçu pour être indépendant du locuteur.

VII.1.B. Une unité de décodage du type consonne/voyelle

La reconnaissance globale considère l'élément à décoder comme une forme insécable qu'il s'agit d'identifier en lui attribuant une classe d'appartenance. On parle de catégorisation. Dans notre cas, le système cherche à reconnaître globalement une unité de type consonne/voyelle (abréviation CV). Par exemple, il attribuera à un segment l'étiquette /la/ ou /si/ ou /ty/... Nous ne prétendons pas croire à l'invariance d'une telle unité. Nous l'avons choisie car elle apparaît comme un bon compromis. Les couples CV semblent en quantité suffisamment restreinte pour permettre un décodage de grands vocabulaires par association d'unités, contrairement à une reconnaissance globale de mots qui nécessite au moins autant de prototypes que d'éléments lexicaux. De plus, sachant que « d'une manière générale, le français est caractérisé par une anticipation vocalique » (Landercy & Renard, 1977, p.94), les couples CV apparaissent comme une bonne unité pour inclure une grande partie des phénomènes de coarticulation contrairement aux phonèmes, trop petits et trop dépendant du contexte. Ceci est conforté par la théorie du pivot (Dogil & Braun, 1988) qui met en évidence le rôle porteur d'information des transitions consonne/voyelle.

Le décodage global s'effectue sur un ensemble d'éléments qui se compose des combinaisons entre consonnes et voyelles. Pour le français, nous avons choisi les 10 « voyelles » /a, i, u, o+ɔ, e+ɛ, y, œ+ø, ɔ̃, ã, ɛ̃+ œ̃/ et les 16 consonnes /p t k b d g f s ʃ v z ʒ m n l R/ ce qui donne 160 combinaisons différentes. Pour chaque couple Consonne/Voyelle, le système possède 10 exemplaires extraits de 10 locuteurs différents (5 hommes, 5 femmes) pour tenir compte des phénomènes de variabilité, ce qui fait un total de 1600 exemplaires.

Nous sommes conscients qu'il serait intéressant de pouvoir traiter les groupes consonantiques en introduisant des unités de type CCV. Nous avons toutefois limité notre étude aux groupes de type CV.

VII.1.C. Un module dépendant de la segmentation automatique

La mise au point du module de reconnaissance globale s'est effectuée sur un corpus composé de logatomes étiquetés manuellement afin de repérer les unités de type CV. Les données sonores appartiennent au corpus SYL de BD-SONS (Cervantès et al., 1986). Les phrases prononcés sont par exemple « papa n'est pas parti à Paris »... Nous avons extrait le 2éme /pa/ de papa placé sur la syllabe accentuée.

Dans le fonctionnement définitif du système, se pose le problème du repérage des couples de type CV, le module de décodage global ne pouvant fonctionner que sur de telles unités.

Le repérage temporel est fournit par l'algorithme de segmentation automatique SAPHO décrit dans le chapitre V. A la suite de l'étude syntagmatique des unités phoniques, le superviseur du système (Figure 63, p.129) repère les pivots entre Consonnes et Voyelles puis oriente le décodage global sur ces zones adaptées.

Dans la suite du chapitre, nous distinguerons, tout d'abord, la mise au point du décodage global indépendamment de la segmentation, puis le fonctionnement définitif du module.

VII.2. Les étapes de la reconnaissance globale

La première étape du décodage consiste à extraire du stimulus une information pertinente qui servira à le caractériser. La catégorisation de l'unité est dépendante ensuite du calcul d'un degré de ressemblance entre celle-ci et les éléments de référence, cette évaluation étant dans notre cas de type métrique (cf. « La notion d'espace métrique », page 26). La prise de décision est l'étape finale qui construit la solution.

VII.2.A. L'extraction de l'information acoustique

« *CritiVoc* » est un vocodeur qui effectue une représentation compacte temps-fréquence du signal de parole (cf. § « CRITIVOC » : un vocodeur en bandes critiques, p.109). Il serait possible de garder comme paramètres acoustiques les valeurs énergétiques présentes dans les canaux du vocodeur. L'inconvénient d'une telle représentation réside dans le fait qu'elle ne nous permet pas d'avoir une vue d'ensemble de la répartition spectrale. Nous avons donc choisi un autre type de paramètres acoustiques: les coefficients PLP (cf. § « La méthode de prédiction linéaire fondée sur un modèle auditif », p.113). La prédiction linéaire a l'avantage de modéliser le spectre auditif dans son ensemble en fournissant un codage par pôles (Boite & Kunt, 1987).

L'analyse du signal s'effectue par trames d'échantillons desquelles sont extraites, pour chacune d'elle, un ensemble de coefficients PLP. Si N est le nombre de trames d'analyse et p l'ordre de la prédiction linéaire, l'extraction PLP nous fournit alors N séries de p paramètres acoustiques caractérisant l'ensemble du signal (Figure 85).

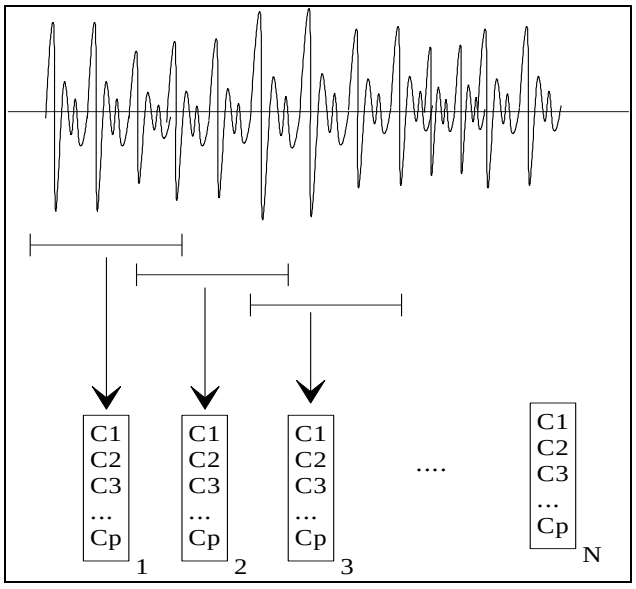


Figure 85: L'extraction des paramètres acoustiques

VII.2.B. La mesure d'un degré de similitude

VII.2.B.a. La comparaison entre trames

Les coefficients PLP sont sources d'information acoustique. Au nombre de *p*, *p* étant l'ordre de la prédiction linéaire, ils caractérisent pleinement une trame de signal. A une trame est donc associé un vecteur de dimension *p* (Figure 14, p.26). L'estimation d'un degré de similitude entre deux trames de signaux se ramène à l'évaluation d'un degré de proximité entre les vecteurs associées à chaque trame, ceci par le calcul d'une distance. Nous avons testé plusieurs distances (distance euclidienne simple, pondérée, Mahalanobis...). Les résultats ne montrent pas de différences majeures. Celle que nous avons retenue est la distance euclidienne classique du fait de sa simplicité. Elle s'écrit sous la forme :

$d = \sqrt{\sum_{i=1}^{p} x_i - a_i ^2}$ où x_i et a_i sont les coefficients PLP extraits de deux trames

VII.2.B.b. La comparaison entre un stimulus et un exemplaire

Un signal de parole est constitué d'une série de *N* trames, *N* dépendant de la longueur du signal et du pas de trames. A chaque trame est associé un vecteur à *p* composantes (coefficients PLP). Le signal est donc caractérisé par une matrice de dimensions [$N \otimes p$]. L'évaluation du degré de similitude entre un exemplaire et le stimulus consiste finalement à calculer une distance cumulée entre une matrice [$N \otimes p$] et [$M \otimes p$], *N* n'étant pas nécessairement égal à *M*.

La difficulté provient justement du fait qu'il n'y ait que rarement une correspondance temporelle entre les deux réalisations à comparer, la différence étant non linéaire. Ainsi, le début d'un des signaux sera, par exemple, plus rapide que l'autre au départ, puis plus lent dans la seconde partie. Aussi, il apparaît nécessaire d'aligner les deux représentations spectrales

pour que, lors de la mesure du degré de proximité, les vecteurs correspondant aux mêmes sonorités dans les deux réalisations soient alignés.

VII.2.B.c. La technique d'alignement dynamique temporel

La technique d'alignement dynamique temporel (Data Time Warping ou encore D.T.W.) est une méthode répandue depuis longtemps (Vintsyuk, 1968 ; Haton, 1974). Bien qu'ancienne, elle reste très efficace pour de petites unités phonétiques de quelques syllabes. Elle a pour but d'ajuster les échelles temporelles de deux éléments à comparer. Cette transformation non linéaire permet ainsi de synchroniser des segments acoustiques de même nature entre deux signaux. Ainsi, dans le cas de diphones de type Consonne/Voyelle, l'algorithme effectue une mesure globale de l'ensemble mais la partie consonantique du stimulus est mise en correspondance avec la partie consonantique de l'exemplaire, de même pour les parties vocaliques. La conséquence est l'alignement des transitions Consonne/Voyelle (Figure 86). La technique DTW établit le chemin optimal entre les points (1,1) et (I,J) (tracé sur la Figure 86) en calculant les distances locales entre les trames des deux éléments. Elle fournit ensuite la distance globale cumulée après repérage du parcours optimal qui correspond à l'alignement des signaux.

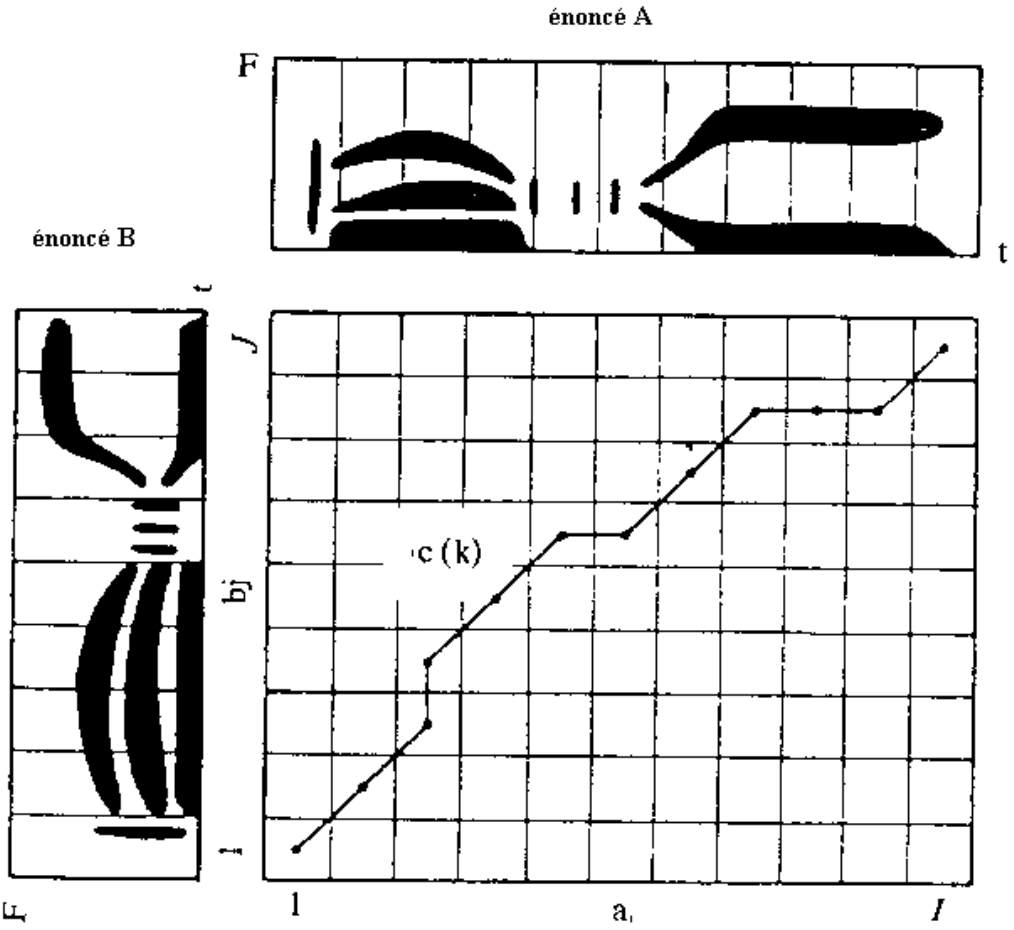


Figure 86: Comparaison dynamique de deux représentations spectrales. La grille centrale représente le chemin permettant un alignement optimal des signaux (Source: Calliope, 1989, p.517)

VII.2.B.d. La liste des prétendants

Le décodage global est dépendant de la mesure de distance entre le stimulus et chacun des 1600 exemplaires archivés. Ces évaluations permettent de constituer une liste de distances globales entre le diphone stimulus et les diphones de référence. Chacune de ces valeurs nous donne ainsi un degré de similitude entre deux réalisations acoustiques. Ces résultats nous permettent de classer les diphones prétendants par ordre croissant, la distance minimale correspondant au prétendant le plus sérieux (Tableau 22).

Tableau 22: Classement des exemplaires. Exemple de résultats sur le /de/ de "défi"

Ne sont présentés que les 30 exemplaires les plus proches sur les 1600 constituant la base de références.

Abréviations: Ê = /ɛ/ ou /e/, @ = / œ / ou /ø/

rs, md, po, jo, bp... sont les initiales des locuteurs

Position	Exemplaire	Distance
1	dÊ.rs	4.126955
2	dy.md	4.184851
3	bÊ.po	4.351903
4	gÊ.md	4.352420
5	bÊ.rs	4.529477
6	bÊ.md	4.735341
7	d@.rs	4.938497
8	gÊ.rs	4.948412
9	bÊ.nc	5.028987
10	dÊ.po	5.036116
11	g@.rs	5.137128
12	dÊ.bp	5.146370
13	gy.md	5.213777
14	bÊ.sl	5.219402
15	dÊ.lc	5.262511
16	bi.md	5.292888
17	bÊ.jb	5.354694
18	di.rs	5.358647
19	dÊ.jo	5.395143
20	d@.md	5.410806
21	gÊ.jo	5.413367
22	dÊ.sl	5.420863
23	by.po	5.426408
24	di.bp	5.450330
25	li.po	5.454205
26	d@.po	5.455389
27	dÊ.nc	5.463726
28	bi.bp	5.520593
29	gÊ.lc	5.526266
30	bi.nc	5.556241

VII.2.C. La prise de décision

Une fois le classement établi, la prise de décision dépend de l'analyse des meilleurs prétendants, c'est à dire des exemplaires les plus proches du stimulus (Tableau 22).

VII.2.C.a. La catégorisation de la voyelle

Le processus de catégorisation de la voyelle suppose, dans un premier temps, la prise en compte du nombre d'apparitions de chaque voyelle parmi les N premiers prétendants. Les voyelles retenues sont celles qui apparaissent de façon significative. Dans l'exemple du Tableau 22, si le système ne retient que les voyelles qui apparaissent au moins 5 fois parmi les trente premiers exemplaires, il ne reste que /Ê/ et /i/. On obtient ainsi un nombre variable de voyelles candidates (de une à trois, deux étant le chiffre le plus fréquent).

Dans un second temps, le système prend en considération les valeurs des distances associées à ces voyelles retenues. La voyelle déclarée la plus probable sera celle qui aura les distances les plus faibles... D'après l'exemple Tableau 22, la distance moyenne associée aux 5 meilleurs exemplaires contenant /Ê/ est d_{moy}= 4,4192192, et pour /i/, d_{moy}= 5,4153326. /Ê/ est donc premier candidat, /i/ est second candidat.

VII.2.C.b. Catégorisation de la consonne

Une prise de décision telle que celle utilisée pour identifier les voyelles ne nous permet pas une catégorisation efficace des consonnes. Cette impossibilité provient essentiellement de la grande sensibilité des consonnes au contexte. Nous avons conscience que des unités de type CV sont empreintes de phénomènes de coarticulation externes à l'unité. De plus, la quantité d'information acoustique extraite sur la consonne peut être extrêmement réduite dans le cas des occlusives sourdes entraînant une difficulté de catégorisation.

La solution envisagée pour la catégorisation des consonnes dans le module de décodage global est fondée sur une méthode par vote. Le résultat n'est pas immédiat et est issu d'une construction. Nous nous inspirons ici d'expériences perceptives où un sujet humain soumis à une écoute dichotique de deux consonnes différentes (ex: /b/ dans l'oreille gauche, /g/ dans l'oreille droite) perçoit une troisième consonne qui correspond à la fusion des traits des deux stimuli (/d/ dans notre exemple). Ce phénomène est connu sous le nom d'effet Mac Gurk (Mc Gurk & Mc Donald, 1976). Dans le cas de notre système, l'algorithme de comparaison décrit précédemment fournit une liste de candidats CV (Tableau 22). A partir de ces prétendants, le dispositif scrute chaque couple CV et évalue les traits les plus fréquents relatifs à la consonne. Ainsi, dans l'exemple du Tableau 22, parmi les 10 premiers candidats, 10 consonnes sont voisées, 4 sont aiguës, 2 sont compactes, 0 sont continues, 0 sont nasales, 0 sont vocaliques. Par une technique de vote, le module propose alors /b/ et /d/ comme solutions car ces phonèmes sont voisés, non compacts, non continues, non nasaux et non vocaliques. La matrice de traits utilisée est celle proposée par Rossi (Tableau 7b, *p.*70). Il faut remarquer que cette technique de vote permet de placer différents candidats à une place ex aequo, propriété particulièrement adaptée à un flou de l'information.

VII.2.D. Bilan

Si le module de reconnaissance globale effectue un décodage sur des unités de type CV, il fournit finalement une ou plusieurs solutions pour la consonne et pour la voyelle. Un degré de confiance peut être attaché au décodage de la voyelle en tenant compte de la valeur de la distance associée, ainsi qu'au décodage de la consonne en fonction du caractère marqué ou non des traits.

VII.3. Le réglage des paramètres

La performance du système repose sur le choix judicieux des caractéristiques d'analyse. Une analyse trop fine sur le signal a pour conséquence de saisir trop d'éléments relatifs au locuteur. En revanche, l'analyse doit être suffisamment précise pour conserver un pouvoir discriminant et séparer les réalisations de phonèmes différents aux images acoustiques proches. L'étape d'extraction des caractéristiques acoustiques du signal nous semble essentielle. Aussi, visons nous à rechercher la pertinence maximale des caractéristiques retenues. Il convient à cet effet de choisir les conditions d'analyse qui donneront les meilleurs résultats. Plusieurs paramètres entrent en jeu pour l'extraction des coefficients PLP*:*

- les paramètres d'analyse spectrale

  $\Rightarrow$ le type de fenêtre d'analyse (rectangulaire, Hamming, Kaiser...)

  $\Rightarrow$ la longueur de la trame d'analyse (de 5 à 30 ms)

  $\Rightarrow$ le pas de décalage des trames (en % de la longueur).

- le filtrage

  $\Rightarrow$ la pré-emphase à 6 dB/oct des aigus

  $\Rightarrow$ la pondération sonique

- la représentation spectrale

  $\Rightarrow$ représentation sur une échelle de Barks ou de Hertz

- la modélisation LP

  $\Rightarrow$ la modélisation sur le spectre de puissance ou d'amplitude

  $\Rightarrow$ la modélisation sur le spectre avec amplitude linéaire ou en dB

  $\Rightarrow$ l'ordre de la modélisation (nombre de pôles)

Ces groupes de paramètres agissent différemment sur les performances du système de reconnaissance. Ces actions sont les suivantes*:*

- action sur la finesse spectrale d'analyse
- action sur la finesse en amplitude (dynamique du spectre)
- action sur la pertinence de l'analyse (pré-emphase, pondération sonique, échelle Barks)
- action sur la finesse de modélisation (ordre de la prédiction linéaire)

La recherche de la meilleure configuration se traduit par un calcul de taux d'erreurs de reconnaissance en fonction de diverses valeurs de paramètres. Une première estimation du choix des paramètres a été effectuée grâce à quelques tests préliminaires. La démarche qui a suivi a consisté à étudier systématiquement l'évolution des performances du système en faisant varier les paramètres, tout en maintenant constant le fonctionnement global de la reconnaissance. Nous avons conscience d'une part que compte tenu du nombre de paramètres à régler, l'optimum obtenu ne restera qu'un optimum local de la fonction de coût. D'autre part, nous pensons aussi que cette optimisation des paramètres est subordonnée au type de décodage utilisé. Toutefois, certaines études (Hamada et al., 1989) laissent penser que la pertinence de paramètres est relativement indépendante de la méthode de décodage.

L'optimisation des performances est subordonnée à une recherche de maximum dans un espace multidimensionnel, les dimensions étant les différents paramètres de l'extraction PLP. Une telle étude est délicate car nous avons conscience que ces variables ne sont pas complètement indépendantes et que la quête du maximum global ne se restreint pas à une recherche des positions des maxima sur chacun des axes indépendamment.

Nous présentons, dans les sections suivantes, l'évolution des performances en fonction des variations de paramètres. Nous rappelons que la base des exemplaires est constituée de la combinaison des 10 « voyelles » /a, i, u, o+ɔ, e+ɛ, y, œ+ø, ɔ̃, ã, ɛ̃+ œ̃/ et les 16 consonnes /p t k b d g f s ʃ v z ʒ m n l R/ , ce qui donne 160 combinaisons différentes. Nous disposons de 10 locuteurs (5 hommes, 5 femmes). Les données sonores appartiennent au corpus SYL de BD-SONS (Cervantès et al., 1986). Les phrases prononcées sont par exemple « papa n'est pas parti à Paris »... Nous avons extrait le 2ème /pa/ de papa placé sur la syllabe accentuée. Pour effectuer nos réglages, nous avons utilisé tour à tour chaque locuteur pour fournir les stimuli et les 9 autres comme base de données des exemplaires. Bien que l'unité CV soit décodée globalement, nous avons séparé les résultats de l'identification de la consonne et de la voyelle. Le décodage de la consonne est considéré comme correct si, parmi les candidats placés ex aequo en première position se trouve la consonne stimulus. Le terme « efficacité » désigne le taux de décodage divisé par le nombre de candidats placés ex aequo en première position. Le décodage de la voyelle est considéré comme correct si, parmi les candidats proposés par le l'algorithme de décision (cf. § « La catégorisation de la voyelle », p.186) se trouve la voyelle stimulus, indépendamment de la distance associée. Le terme « efficacité » désigne le taux de décodage divisé par le nombre de candidats proposés. Nous présentons à chaque fois les résultats relatifs à la consonne et à la voyelle.

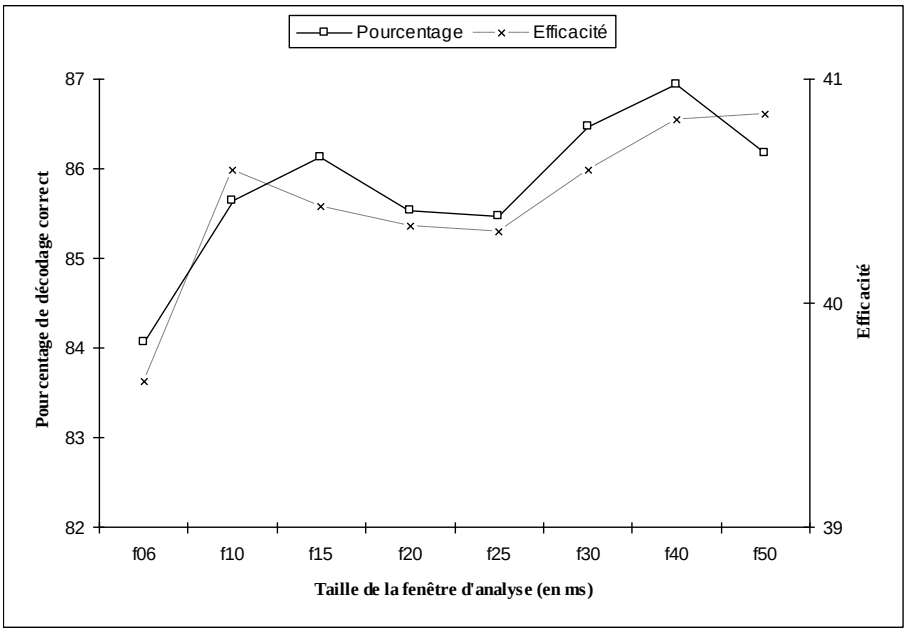
VII.3.A. Optimisation de la fenêtre d'analyse

Le rôle et les propriétés des fenêtres d'analyse sont décrits au § III.III.2.III.2.C.III.2.C.e., p.98. Mise à part la fenêtre rectangulaire qui accentue les distorsions spectrales et qui par conséquence fournit de moins bonnes performances, la nature de la fenêtre (Kaiser, Hamming...) ne semble pas affecter de façon notable la pertinence de l'analyse. Comme nous l'avons indiqué (cf. § « Les fenêtres classiques », p.99), nous avons choisi la fenêtre de Hamming du fait de son bon compromis entre la hauteur du premier lobe

secondaire et la demi-largeur du lobe principal. De plus, la forme de cette fenêtre reste simple à calculer. Si, la forme de la fenêtre importe relativement peu, en revanche, le résultat reste sensible à la taille de la fenêtre d'analyse (Figure 87). La valeur optimale du paramètre est différente selon qu'il s'agit de voyelles ou de consonnes. Le compromis se situe aux environs de 20 à 25ms.

Il peut paraître surprenant que les résultats de l'identification de la consonne soient meilleur que ceux de la voyelle. Il faut relativiser ce résultat en tenant compte du nombre de candidats proposés par l'algorithme de décision : en moyenne 2 pour les voyelles, 3 à 4 pour les consonnes. Cette rectification apparaît clairement dans l'observation du facteur d'efficacité.

Voyelles



Consonnes

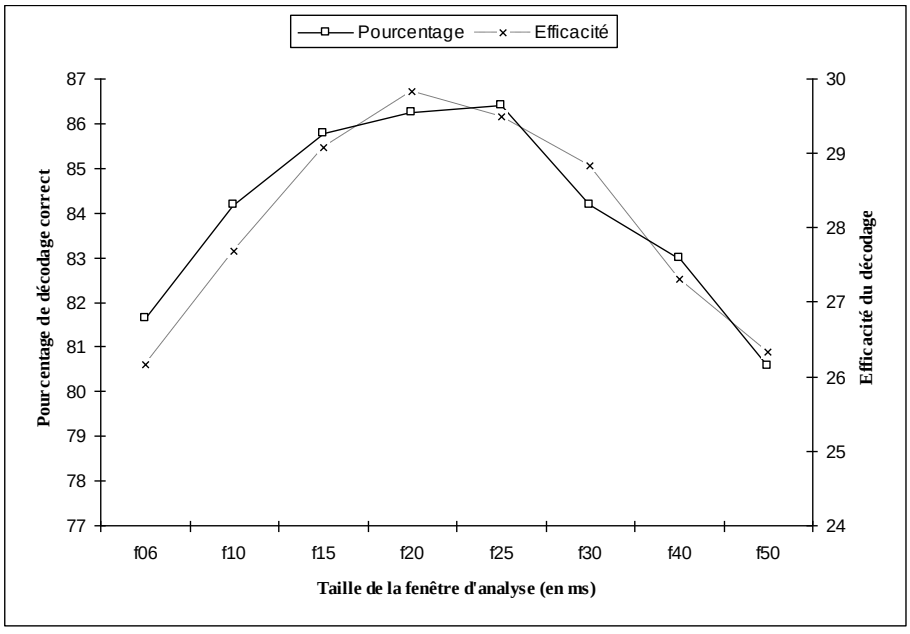


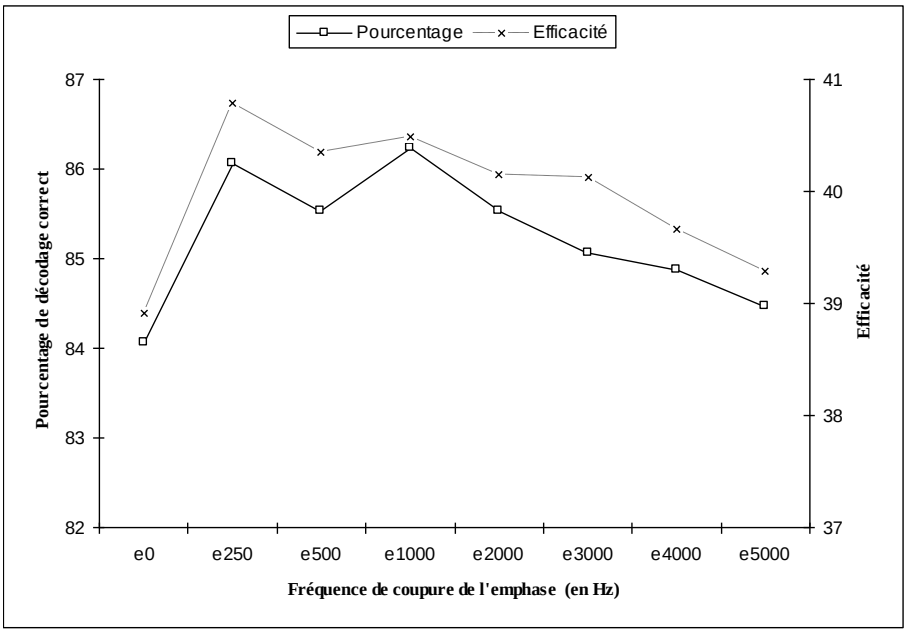
Figure 87: effets de la taille de la fenêtre d'analyse sur le décodage des voyelles et des consonnes

VII.3.B. Optimisation des paramètres de filtrage

VII.3.B.a. La pré-emphase des aigus

La pré-emphase des aigus consiste à rééquilibrer la répartition spectrale en rehaussant les fréquences hautes. Pour cela, est appliquée en post-traitement sur le signal de parole un filtrage du type $h(z)=1-\alpha z^{-1}$ (Figure 9, *p.*17). La modification d'alpha permet de régler la fréquence de coupure. La valeur optimale du paramètre se situe aux environs de 500 Hz (Figure 88).

Voyelles



Consonnes

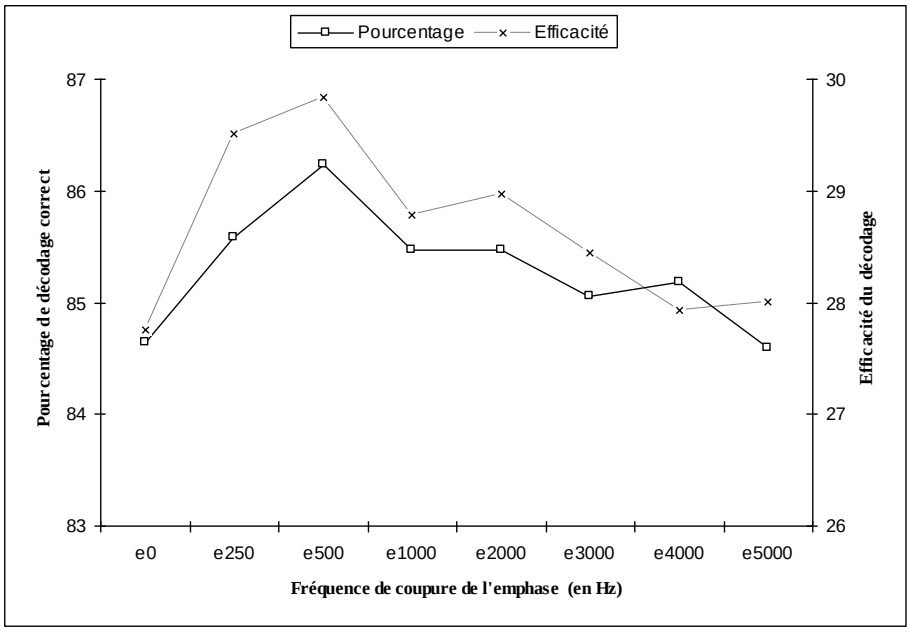


Figure 88: effets de la place de la pré-emphase sur le décodage des voyelles et des consonnes

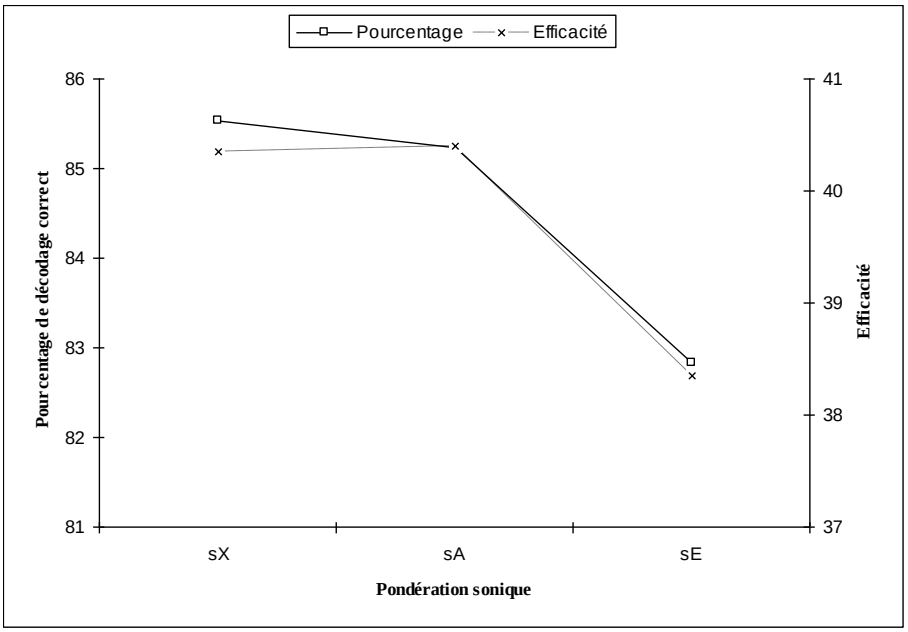
VII.3.B.b. La pondération sonique

La pondération sonique est une correction spectrale imitant le comportement auditif humain (cf. §.III.III.1.III.1.B.III.1.B.b. *p.*78). Nous avons testé trois types de pondérations :

- la pondération A, qui correspond à la courbe isosonique de 40 phone
- la pondération E, proposé par Stevens
- la pondération X, un amalgame des deux précédentes, présentée au § « La modélisation des pondérations soniques », p.79.

La meilleure est la pondération de type X (Figure 89).

Voyelles



Consonnes

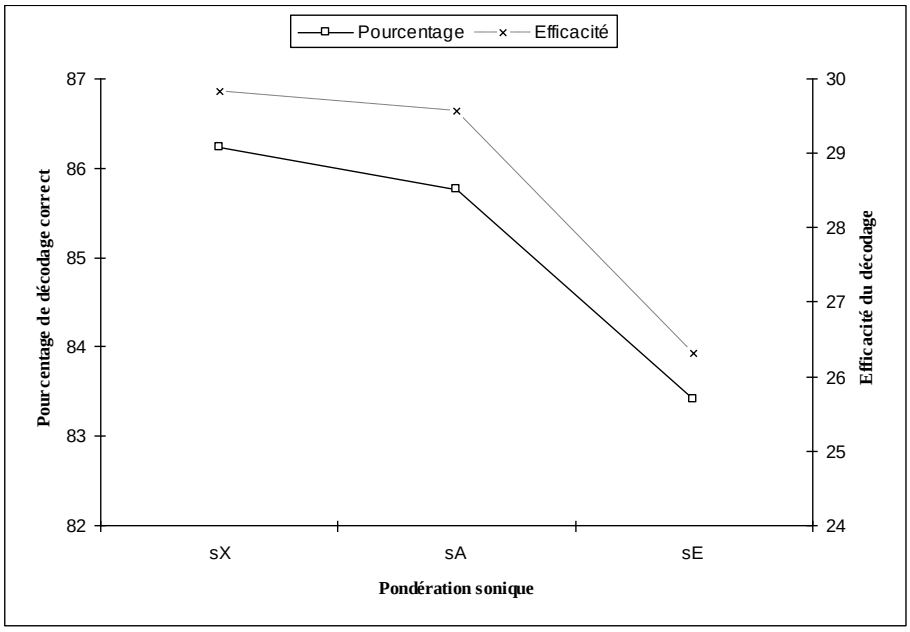


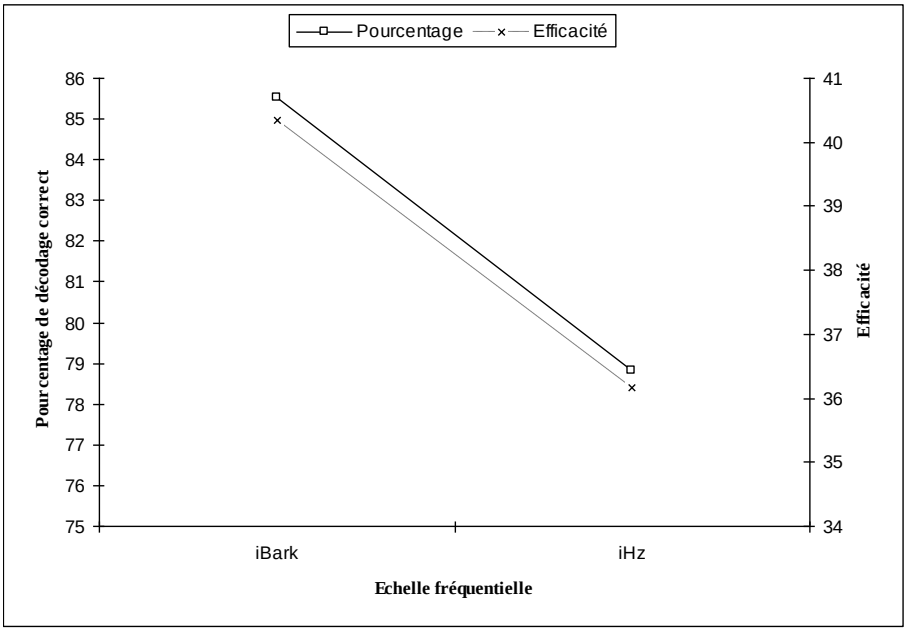
Figure 89: effets de la pondération sonique sur le décodage des voyelles et des consonnes

VII.3.C. Optimisation de la représentation spectrale

VII.3.C.a. Modélisation en barks/hertz

Il est possible de modéliser le spectre auditif selon deux échelles fréquentielles (Figure 57, p.113)*:* l'une linéaire (en Hz), l'autre logarithmique (en barks). La meilleure est la modélisation en barks (Figure 90).

Voyelles



Consonnes

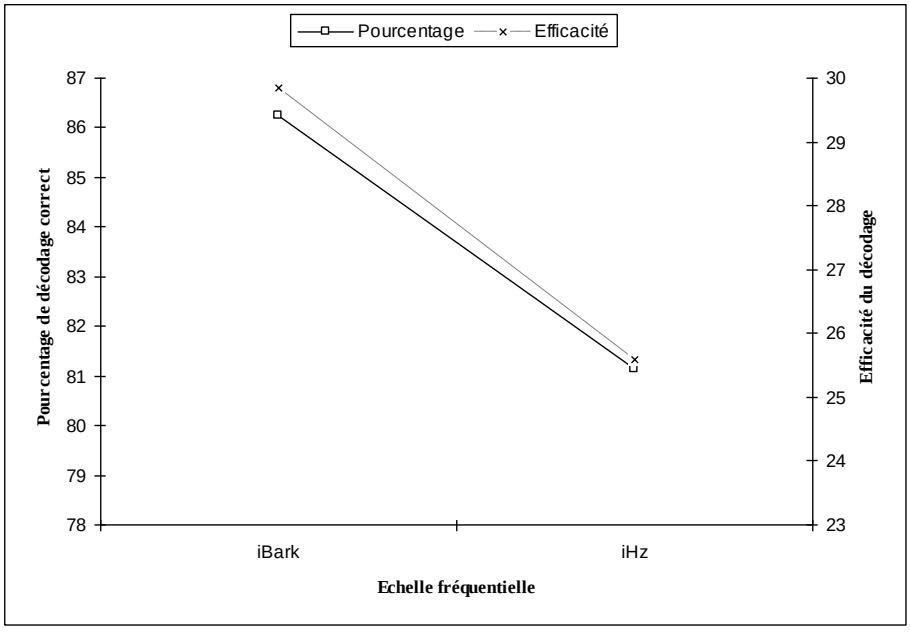
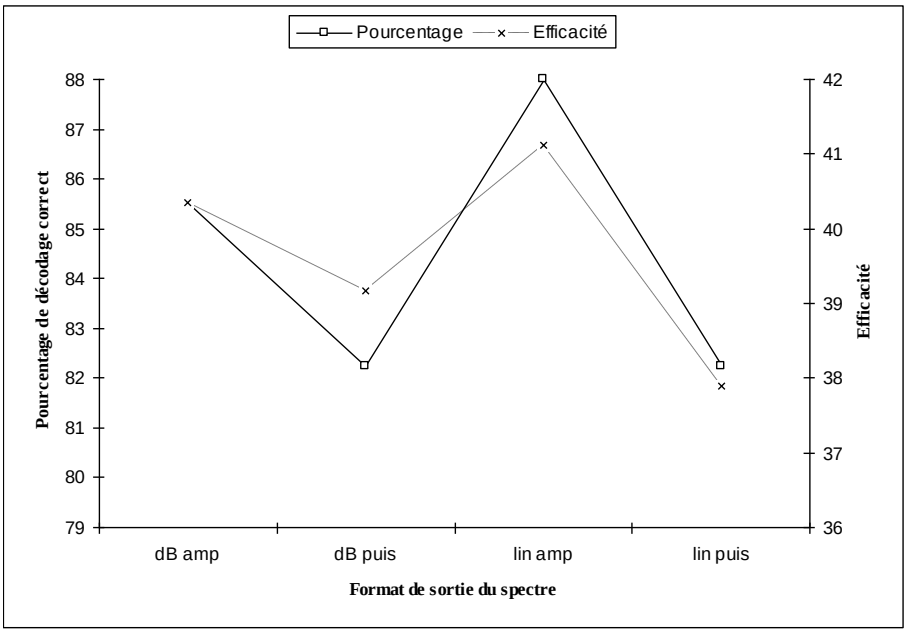


Figure 90: effets de la modélisation Hz/barks sur le décodage des voyelles et des consonnes

VII.3.C.b. Dynamique du spectre

Un spectre peut être représenté en amplitude ou en puissance selon deux modes: l'un linéaire, l'autre logarithmique (en dB). Ces représentations influencent la dynamique du spectre sur laquelle la modélisation par prédiction linéaire agit par la suite. Un spectre trop dynamique, comme celui exprimé en puissance linéaire, aura du mal à être modélisé correctement par la PL. La meilleure représentation, surtout pour les consonnes, est un spectre d'amplitude exprimé en dB (Figure 91).

Voyelles



Consonnes

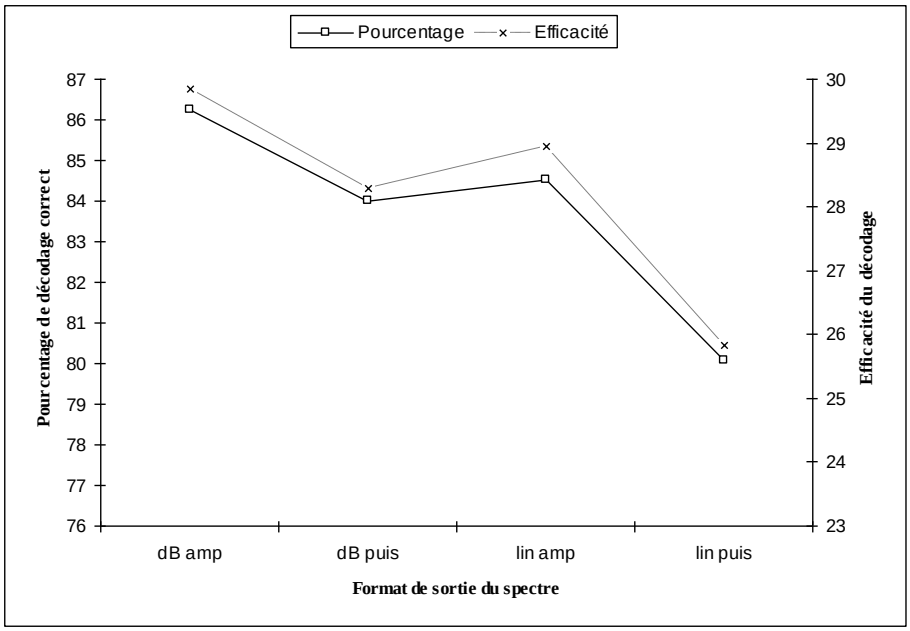
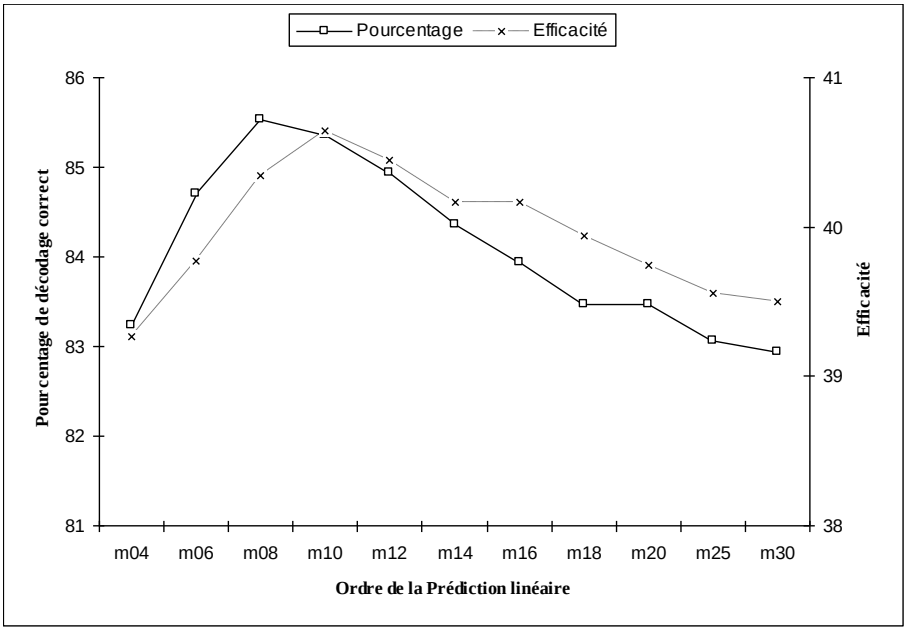


Figure 91: effets de la dynamique du spectre sur le décodage des voyelles et des consonnes

VII.3.D. Optimisation des paramètres de modélisation LP

L'ordre de la prédiction linéaire correspond à la finesse de modélisation (Figure 61, p.122). La valeur 8 semble être un bon compromis (Figure 92) entre trop grande grossièreté d'analyse (faible valeur) et empreinte du locuteur ou autre cause de variabilité (forte valeur).

Voyelles



Consonnes

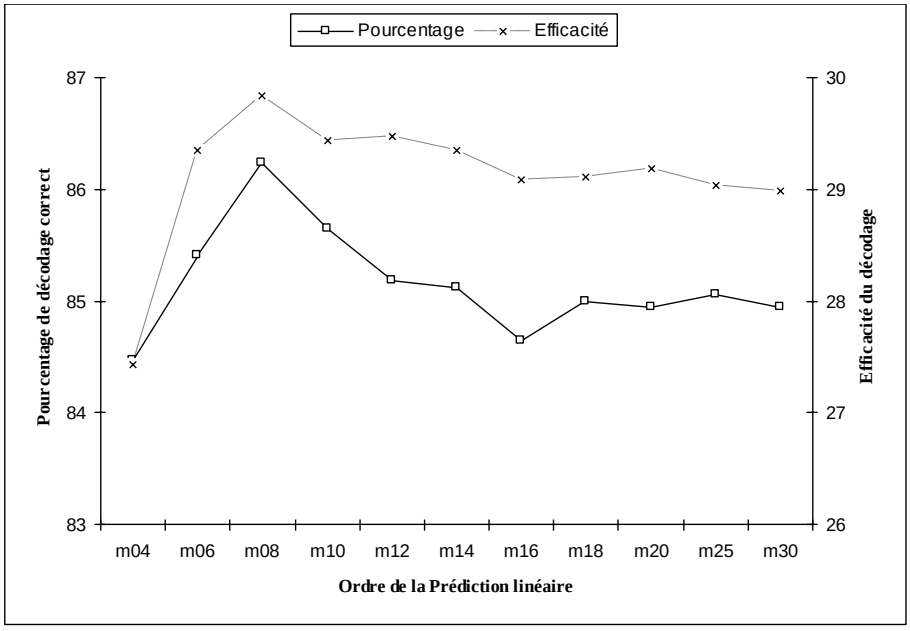


Figure 92: effets de l'ordre de la prédiction linéaire sur le décodage des voyelles et des consonnes

VII.3.E. Bilan

Après avoir obtenu un certain nombre d'optima locaux, nous avons réitéré le processus en se plaçant dans la position de ces optima. La position n'a pas changé, ce qui laisse penser que nous avons obtenu minimum local en terme de coût. Cette configuration optimale locale des paramètres d'extraction PLP est la suivante:

- fenêtre de Hamming de longueur égale à 20ms avec un recouvrement de 50%
- pré-emphase à partir de 500Hz et pondération sonique de type X
- échelle fréquentielle en Barks
- spectre d'amplitude en dB
- ordre de prédiction égal à 8

On peut constater que la finesse dans l'analyse n'est pas forcément le meilleur choix. Ainsi, la configuration « spectre d'amplitude en dB » est celle qui a le moins de dynamique, l'ordre LP à 8 est une valeur moyenne... Ceci s'explique par le fait qu'une analyse grossière s'affranchit d'une partie de la variabilité de la parole. Une trop grande précision entache l'analyse de l'empreinte du locuteur, ce qui est à bannir dans le cadre d'un décodage indépendant du locuteur.

VII.4. L'évaluation

La phase d'évaluation consiste à analyser les résultats de façon fine. Nous rappelons que les unités à reconnaître sont des séquences consonne/voyelle. La combinatoire retenue laisse apparaître 160 unités. 10 locuteurs (5 hommes, 5 femmes) extraits du corpus SYL de BD-SON (Cervantes et al., 1986) permettent d'établir la base des 1600 exemplaires. Un étiquetage manuel nous autorise à ne faire porter l'évaluation que sur le module de reconnaissance globale sans introduire d'erreur de segmentation.

En toute rigueur, l'évaluation du module de reconnaissance globale devrait utiliser un corpus spécifique. Or, la constitution d'un ensemble de données sonores organisées incluant un nombre conséquent de locuteurs est un travail long et fastidieux, surtout s'il nécessite un étiquetage phonémique. Par souci d'économie de temps, nous avons donc utilisé le corpus SYL de BD-SON de la façon suivante. Sur les 10 locuteurs, nous en avons exclu tour à tour un, celui-ci servant de test tandis que les 9 autres constituaient la base des exemplaires. Pour chaque locuteur, 160 tests ont été réalisés. L'opération étant renouvelée pour les 10 locuteurs, cela représente un volume de 1600 tests dont nous présentons les résultats déjà exposés dans (Ghio & Rossi, 1994).

VII.4.A. Reconnaissance de la voyelle

Il s'agit de distinguer trois types de résultats (Tableau 23):

- le cas où la voyelle a été reconnue correctement en première position (*1er candidat)*
- le cas où la voyelle a été reconnue correctement en 2, 3, 4ème position (*2nd rang*)
- le cas où la voyelle n'a pas été reconnue (*Omission*)

Tableau 23: Résultats de la reconnaissance globale des voyelles

	1er candidat	*2nd rang*	*Omission*
/a/	96%	3%	1%
/i/	89%	9%	2%
/u/	82%	15%	3%
/o, ɔ/	81%	16%	3%
/e, ɛ /	75%	21%	4%
/y/	80%	16%	4%
/ ø, œ /	88%	10%	2%
Moyenne	84%	13%	3%

Nous rappelons que les voyelles nasales sont exclues. Les raisons sont expliquées au § « Le cas de la nasalité », p.169.

Le Tableau 24 représente la matrice de confusion des voyelles. On remarque que, bien souvent, dans les cas de mauvaise reconnaissance, le système se reporte sur un voisin acoustique du stimulus: /o/ est confondu avec /u/, tous deux graves ; /y/ est confondu avec /e/, tous deux aigus. La confusion générale avec /*f, ʃ*/ s'explique par la position centrale de ces phonèmes dans l'espace acoustique. La réalisation en [*f, ʃ*] est souvent le résultat d'une centralisation d'autres phonèmes due à des effets de contexte. La confusion est donc fréquente.

Nous retrouvons ici ce que nous avions entrevu dès la construction de nos outils d'analyse, c'est à dire les limites de discrimination de l'analyse PLP (Figure 60, p.121).

Tableau 24: Matrice de confusion des voyelles pour la reconnaissance globale. Résultats exprimés en %. Le signe « . » correspond à une absence de confusion.

réponse / stimulus	/a/	/i/	/u/	/o, ɔ/	/e, ɛ /	/y/	/ ø, œ /
/a/ =>	**96**	.	.	1	.	.	3
/i/ =>	.	**89**	1	.	5	4	1
/u/ =>	.	1	**82**	13	1	1	2
/o, ɔ/=>	3	.	14	**81**	1	.	1
/e, ɛ /=>	.	4	.	.	**75**	8	13
/y/ =>	.	2	.	.	8	**77**	13
/ ø, œ /=>	1	.	1	1	8	1	**88**

VII.4.B. Reconnaissance de la consonne

Dans le cas de la reconnaissance des consonnes, le système fournit plusieurs prétendants dont aucun n'est privilégié a priori. Il n'existe pas de classement parmi les candidats. On distingue alors deux types de résultats (Tableau 25): le cas où la consonne stimulus se trouve parmi les candidats (*Correct),* le cas où la consonne stimulus ne se trouve pas parmi les candidats (*Omission).* La colonne *Nb moyen de candidats* indique le nombre moyen de candidats fournis par le système dans le décodage de chacune des consonnes. Ce paramètre est important car intuitivement, on comprend que plus ce nombre sera grand, moins bonne sera la discrimination mais meilleurs seront les résultats bruts.

Tableau 25: Résultats de la reconnaissance globale des consonnes

	Correct	*Omission*	*Nb moyen de candidats*	*Efficacité*
/p/	93%	7%	6,04	15,4%
/t/	97%	3%	5,16	18,8%
/k/	90%	10%	4,69	19,2%
/b/	99%	1%	4,64	21,3%
/d/	97%	3%	4,2	23,1%
/g/	79%	21%	4,81	16,4%
/f/	91%	9%	2,44	37,3%
/s/	93%	7%	2,16	43,1%
/j/	91%	9%	2,3	39,6%
/v/	87%	13%	4,43	19,6%
/z/	97%	3%	3,11	31,2%
/ɋ/	93%	7%	3,26	28,5%
/m/	90%	10%	3,16	28,5%
/n/	94%	6%	3,7	25,4%
/l/	96%	4%	3,99	24,1%
/R/	97%	3%	8,73	11,1%
Moyenne	93%	7%	4,18	22,2%

L'analyse succincte des résultats est la suivante: en moyenne, le dispositif de reconnaissance globale propose 4 consonnes parmi les 16 possibles ; dans 93% des cas, la bonne réponse est parmi ces candidats. Ce chiffre de 4 n'est pas anodin: il traduit la quasi-impossibilité d'identifier le lieu d'articulation des consonnes. En effet, parmi les candidats proposés, sont presque toujours présents les éléments d'une même macro-classe comme par exemple {v z ʒ}, {p t k}, {m n}... Il faut remarquer que la difficulté de décodage ne se traduit pas forcément par un mauvais score mais par le nombre de candidats proposés. Ainsi, les fricatives sourdes, unités relativement faciles à détecter, sont correctement placées comme premier candidat dans plus de 90% des cas et ceci avec un seul autre élément ex aequo (le plus souvent, une autre fricative sourde). En revanche, /R/ est identifié comme tel dans 97% des cas mais dans un lot de 8 candidats potentiels. On retrouve là l'aspect polymorphe de cette unité.

VII.4.C. Comparaison avec d'autres types d'analyse

Afin d'estimer la pertinence de l'analyse PLP, nous avons effectué des tests comparatifs en utilisant d'autres types de paramètres acoustiques tels que ceux issus de la technique LPC, du Cepstre, du Mel Cepstre (Davis & Mermelstein, 1980 ; Brancaccio et al., 1992). Nous avons conservé le même module de comparaison DTW ainsi que celui de la décision. Tout comme nous l'avons fait pour la technique PLP, nous avons optimisé les jeux de paramètres de ces méthodes, opération relativement rapide du fait du faible nombre de paramètres impliqués dans ces analyses (taille et pas de fenêtre, nombre de coefficients, pré-emphase éventuelle). Les résultats sont présentés dans le Tableau 26. Les différents types d'analyse sont les suivants:

- *LPC* est la technique de Prédiction Linéaire.
- *LPCC e*st la technique de coefficients cepstraux calculés à partir de la LPC.
- *MFCC* est la technique d'extraction de coefficients cepstraux sur une échelle MEL.
- *PLP* est relative à cette étude.

Tableau 26: Comparaison de différents types d'analyse pour la reconnaissance globale

	LPC	*LPCC*	*MFCC*	*PLP*
VOYELLE				
1er candidat	*45%*	*74%*	*83%*	*84%*
2nd rang	*25%*	*21%*	*10%*	*13%*
Omission	*30%*	*5%*	*7%*	*3%*
CONSONNE				
Candidat	*82%*	*90%*	*92%*	*93%*
Omission	*18%*	*10%*	*8%*	*7%*
Nb moy. cand	*5.01*	*4.56*	*4.23*	*4.18*
Efficacité	16,4%	19,7%	21,7%	22,2%

VII.4.D. Analyse des résultats

Il faut garder à l'esprit que les résultats présentés ici ne sont relatifs qu'à un décodage sur l'axe paradigmatique, le découpage temporel étant fourni par un étiquetage manuel. Tout risque d'omission ou d'insertion sont évités. Il semble, d'après le Tableau 26, que la technique PLP ait un degré de pertinence analogue à la méthode Mel Cepstre. Les erreurs de décodage en utilisant l'une ou l'autre des techniques ne sont pas les mêmes. La technique PLP est loin d'être figée et nous gardons l'espoir d'améliorer son pouvoir discriminant en tenant compte des causes responsables des mauvais résultats du décodage, entre autres de /e/ et /y/. L'exemple de /y/, pour laquelle F2 et F3 sont confondus dans un seul pôle autour de 1700-2000Hz, montre ici les limites de l'analyse PLP.

Il est intéressant de noter que les performances pour les consonnes sont relativement peu dépendants du type de coefficients. Il semble que seule une classification en macro-classe est réalisée. Cela signifie que la méthode est limitée ou que l'information n'est réellement pas présente dans le signal acoustique, en tous cas dans la façon où nous l'appréhendons.

VII.5. Bilan

Nous venons de décrire les 3 modules d'analyse montante :

- l'algorithme SAPHO, qui fournit une analyse de l'axe syntagmatique et autorise le repérage des unités fonctionnelles que sont les phonèmes. De plus, il propose une classification en macro-classes de ces éléments, notamment des consonnes.
- la reconnaissance analytique, qui apporte des informations paradigmatiques sur les voyelles.
- la reconnaissance globale, qui s'attache à catégoriser des unités de type Consonne/Voyelle. Ce module se greffe dans l'ensemble du dispositif pour apporter une information paradigmatique sur les éléments syntagmatiques préalablement identifiés par le module de macro-classification et de segmentation SAPHO. Il s'apparente en cela au module de reconnaissance analytique.

Ces trois modules constituent le décodage acoustico-phonétique, base du dispositif de reconnaissance. Il reste à présent à accéder au lexique. Cette opération est effectuée par les modules de haut niveau.

VIII. LES MODULES DE HAUT NIVEAU

« Le génie, c'est 10 % d'inspiration et 90 % de transpiration. »

Thomas Edison.

Plan du chapitre

RESUME

Les modules de haut niveau sont tous les traitements qui agissent sur la matière symbolique par opposition aux modules de décodage acoustico-phonétique. Ce chapitre est consacré à la description des processus d'accès lexical et de prise de décision par le module superviseur. Ce dernier pilote tous les processus de décodage ascendant décrits dans les chapitres précédents. A partir de l'information montante ainsi fournie, il procède à un accès lexical en consultant un dictionnaire. Il fournit finalement sa réponse en comparant les données du lexique aux données décodées. Le module d'accès lexical est fondé sur un algorithme de programmation dynamique qui s'appuie sur une matrice de coût. Celle-ci est établie à partir de la théorie des traits. Ce processus permet de comparer deux chaînes phonétiques en intégrant les opérations d'insertion, d'omission et de substitution de phonèmes. Dans le cadre de notre système ACHILE, le module d'accès lexical a pour rôle de fournir le mot correspondant à la chaîne phonétique identifiée par les modules de décodage ascendant. Pour cela, il s'appuie sur un lexique de 500 mots codés à la fois sous la forme orthographique et phonémique.

Notre système effectue un décodage de mots isolés de façon indépendante du locuteur, sans apprentissage, ni adaptation. L'évaluation du dispositif nous permet de quantifier la pertinence des connaissances, des analyses et des algorithmes employés. C'est aussi le moyen de pouvoir faire évoluer le dispositif en apportant des modifications. Pour cela, il nous semble nécessaire de garder un critère de contrôlabilité, c'est à dire la possibilité de prévoir, analyser et comprendre l'évolution des performances en fonction des modifications apportées. Nous avons exploité la modularité du dispositif pour prendre conscience des différentes sources d'informations qui autorisent un décodage robuste. Les résultats sont présentés dans cette optique-là.

VIII.1. Que sont les modules de haut niveau ?

Par modules de haut niveau, nous entendons tous les traitements qui agissent sur la matière symbolique par opposition aux modules de décodage acoustico-phonétique, dont la matière première est le signal de parole, phénomène physique continu. D'un point de vue linguistique, nous abordons ici les questions de manipulation du lexique, de la syntaxe, du sens et du sous-entendu. En se replaçant dans le contexte précis de notre étude, nous rappelons que notre système effectue un décodage de mots isolés. Il n'est donc question que de lexique. Dans le cadre de l'architecture du dispositif ACHILE (cf. Figure 63, p.129), nous allons décrire les processus d'accès lexical et de prise de décision par le module superviseur.

Comme nous le verrons plus en détails au § VIII.3 page 213, le module superviseur pilote tous les processus de décodage ascendant décrits dans les chapitres précédents. A partir de l'information montante ainsi fournie, le superviseur procède à un accès lexical en consultant un dictionnaire. Il fournit finalement sa réponse en comparant les données du lexique aux données décodées. Une phase de vérification descendante est prévue dans le but de valider, infirmer ou réinterpréter le résultat. Cette procédure s'inscrit dans le cadre des théories de « top-down processing » proposées, entre autre, dans (Ohala, 1986). Elle ne sera pas décrite dans notre travail. Par souci de clarté, nous présentons d'abord le procédé d'accès lexical puis, dans un second temps, le module superviseur.

VIII.2. L'accès lexical

VIII.2.A. Une mesure de distance

Certaines méthodes de correction orthographique automatique calculent une distance entre un mot-référence et un mot-test. Elles nécessitent une série d'opérations qui tiennent compte des erreurs d'insertion, d'omission et de substitution. Un tel processus est réalisable en utilisant un algorithme de programmation dynamique (Véronis, 1994). Notre module d'accès lexical s'inspire de cette méthode.

Dans le cas de la correction orthographique, la distance est évaluée entre graphèmes. Par contre, dans notre cas, la distance est calculée entre phonèmes, en particulier entre ceux de la chaîne phonétique décodée et ceux des représentations phonémiques des mots stockés dans le dictionnaire. L'algorithme de programmation dynamique confronte un à un les phonèmes des deux chaînes à comparer. Il établit une matrice qui rassemble alors les score de ressemblance entre phonèmes (Figure 93). Ces scores de ressemblance représentent une distance locale entre les unités élémentaires de chaque chaîne. Dans le cas de chaînes orthographiques, la distance locale est simple: 0 si les graphèmes sont identiques, 1 s'ils sont différents. En revanche, pour des chaînes phonétiques, il est nécessaire de faire intervenir une distance plus sophistiquée. En effet, la confusion entre /a/ et /o/ est moins grave que celle entre /a/ et /s/. C'est la raison pour laquelle a été introduite une matrice de « coût », qui indique le degré de différences entre phonèmes. De plus, la chaîne phonétique décodée peut comporter des unités de type macro-classe, c'est à dire non complètement définies. Par exemple, sur la Figure 93, la 5ème unité du stimulus a été décodée comme une consonne vocalique. Nous devons donc positionner chaque phonème par rapport aux éventuelles macro-

classes proposées par les modules de décodage ascendant. Une fois recueillis tous les scores de dissemblance, un processus automatique recherche ensuite le chemin qui fournira la distance cumulée la plus faible, qui correspond à l'alignement optimum des deux chaînes (Figure 93). Cet algorithme est intéressant car il intègre, de façon efficace et ceci dans un unique processus, tous les phénomènes d'insertion, d'omission et de substitution qui peuvent apparaître dans le décodage ascendant.

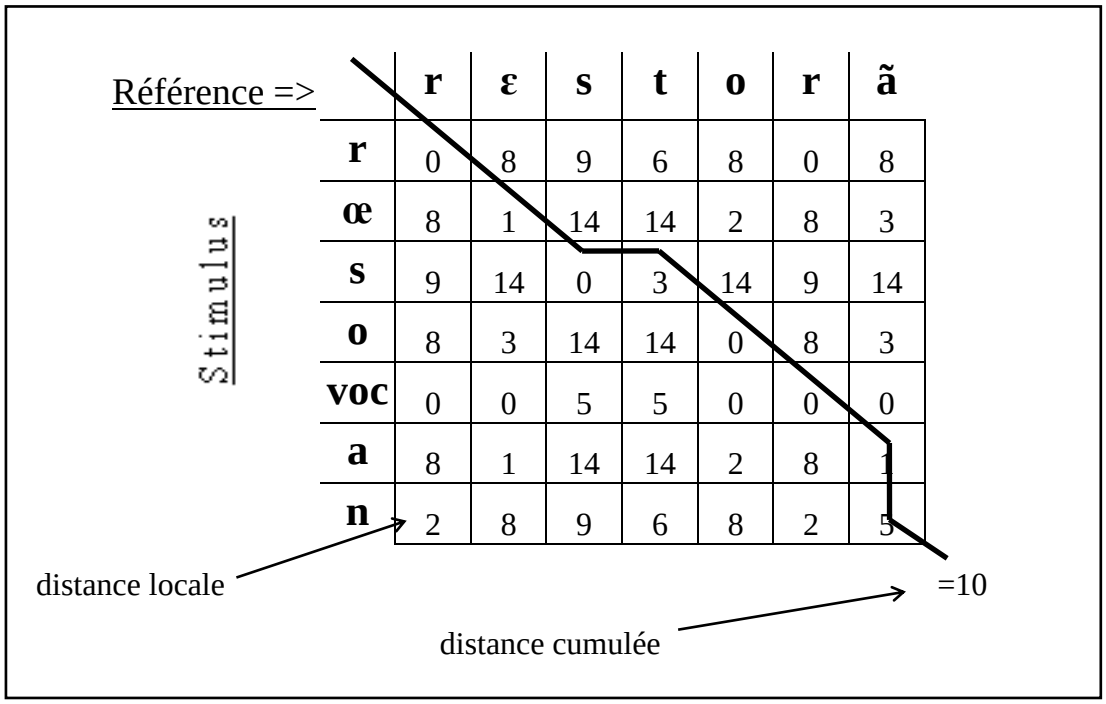


Figure 93: Calcul de distance entre deux chaînes phonétiques par la méthode de programmation dynamique

VIII.2.B. L'importance de la matrice de « coût »

VIII.2.B.a. Présentation de la matrice

La matrice de « coût » est un tableau qui contient le degré de dissimilitude entre phonèmes. Elle comporte horizontalement les 35 phonèmes / a i u o ɔ e ɛ y œ ø ɔ̃ ã ɛ̃ œ̃ p t k b d g f s ʃ v z ʒ m n l R j w ɥ ñ ŋ / auxquels s'ajoutent 4 archiphonèmes

/ Ê / qui représente /e/ ou /ɛ/,
/Ô/ qui représente /o/ ou /ɔ/,
/@/ qui représente /ø/ ou /œ/,
/ Ë / qui représente / ɛ̃ / ou / œ̃ /.

Elle inclut également les macro-classes définies au Tableau 12, p.155, c'est à dire:

1. voyelle
2. vocalique
3. consonne
4. consonne non voisée

5. consonne voisée
6. consonne occlusive voisée ou vocalique
7. consonne vocalique
8. fricative
9. fricative sonore
10. fricative sourde
11. occlusive sonore
12. occlusive sourde
13. schwa final ou consonne

VIII.2.B.b. Problèmes liés à la constitution de la matrice

Se pose le problème de la constitution de la matrice. Deux stratégies peuvent être adoptées:

- une mesure fondée sur les données. Dans ce cas-là, des procédures automatiques calculent statistiquement l'écart moyen entre phonèmes. Il s'agit alors de choisir un corpus représentatif ainsi qu'une méthode pertinente de comparaison.
- une mesure fondée sur les connaissances. Dans ce cas-là, la distance entre phonèmes est attribuée a priori à partir de données connues.

La première méthode est délicate car pouvant nous conduire dans un cercle vicieux. En effet, nous avons besoin de quantifier des différences entre phonèmes, scores à intégrer dans un système de décodage automatique. Or, pour que les mesures soient statistiquement pertinentes, nous devons utiliser un corpus important et donc des techniques automatiques d'analyse... Nous avons finalement choisi la seconde méthode, fondée sur les connaissances. Afin de réduire son aspect arbitraire, nous avons fondé la comparaison sur la théorie des traits développée au §. II.II.4.II.4.D., p.65. Nous savons que les phonèmes peuvent être décomposés en un ensemble de traits qui les distinguent. Il est facile de construire, à partir de cette décomposition, un espace multidimensionnel dans lequel chaque phonème est repéré géométriquement. L'établissement d'une distance euclidienne ou autre permet d'établir la matrice de coût.

Plusieurs options s'offrent à nous. Tout d'abord, le choix de la décomposition en traits. Nous savons qu'elle n'est pas unique et qu'elle peut être articulatoire, acoustique ou mixte. La question de la redondance de certains traits doit être clairement posée. Ensuite, le choix de la distance. Celle-ci peut être euclidienne $\left(d = \sqrt{\sum_i (x_i - y_i)^2}\right)$, de norme 1 $\left(d = \sum_i |x_i - y_i|\right)$ ou pondérée en donnant plus ou moins de « poids » à certains traits.

VIII.2.B.c. Le choix de la distance entre traits

Nous rappelons que la notion de traits, au sens orthodoxe du terme, impose une caractéristique binaire: trait présent ou absent. Par conséquent, les coordonnées des phonèmes dans l'espace multidimensionnel des traits ne prennent que les valeurs 0 ou 1. Cela diminue grandement l'importance du choix de la distance. En effet, dans ce cas-là, la valeur donnée par une distance euclidienne est la racine carrée de celle fournie par une norme 1. Il n'existe qu'un

effet de contraction que nous n'étudierons pas. Nous nous contenterons d'utiliser la distance de norme 1, qui consiste finalement à compter le nombre de traits différents entre deux phonèmes.

Par contre, il nous semble intéressant d'analyser la pertinence de la pondération des traits. En effet, il est bien connu que le traitement des traits par l'humain n'est pas uniforme, certains étant plus résistant que d'autres. La communication humaine tendant à l'optimisation, il est fort probable que ces traits plus robustes soient porteurs de plus d'information et, par conséquent, il nous semble alors judicieux de les amplifier. Ainsi, l'expérience de Chistovitch et al. présentée dans (Rossi, 1977, p.75) sur l'identification des consonnes par l'humain montre que « le sujet reconnaît en premier lieu des traits distinctifs. Ces traits sont reconnus séparément et toujours dans l'ordre:

a) trait de mode

b) trait de voisement

c) trait de lieu. »

Nous tiendrons compte de ces résultats pour hiérarchiser l'espace des traits en donnant plus d'importance aux informations portant sur le mode d'articulation, puis le voisement et enfin le lieu d'articulation.

VIII.2.B.d. Le choix de la base de décomposition en traits

La décomposition proposée par Rossi et présentée à la Figure 34, page 67, fournit une classification des voyelles en 5 traits articulatoires. En utilisant une distance de norme 1 non pondérée, elle permet de dresser la matrice de coût suivante:

Tableau 27: Matrice de coût pour les voyelles

.	a	i	u	o	e	y	ø	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃	Ê	Ô	@	Ë
a	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
i	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
u	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
o	2	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
e	2	1	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.
y	4	1	1	2	2	0	.	.	.	.	.	.	.	.	.	.	.	.
ø	3	2	2	1	1	1	0	.	.	.	.	.	.	.	.	.	.	.
ɛ	1	2	4	3	1	3	2	0	.	.	.	.	.	.	.	.	.	.
ɔ	1	4	2	1	3	3	2	2	0	.	.	.	.	.	.	.	.	.
œ	2	3	3	2	2	2	1	1	1	0	.	.	.	.	.	.	.	.
ã	1	4	4	3	3	5	4	2	2	3	0	.	.	.	.	.	.	.
ɛ̃	2	3	5	4	2	4	3	1	3	2	1	0	.	.	.	.	.	.
ɔ̃	2	5	3	2	4	4	3	3	1	2	1	2	0	.	.	.	.	.
œ̃	3	4	4	3	3	3	2	2	2	1	2	1	1	0	.	.	.	.
Ê	1	1	3	2	0	2	1	0	2	1	2	1	3	2	0	.	.	.
Ô	1	3	1	0	2	2	1	2	0	1	2	3	1	2	2	0	.	.
@	2	2	2	1	1	1	0	1	1	0	3	2	2	1	1	1	0	.
Ë	2	3	4	3	2	3	2	1	2	1	1	0	1	0	1	2	1	0

La décomposition des voyelles en traits acoustiques (cf. Tableau 7a, p.70) fournit une matrice quasiment analogue à part pour /a/ et /ã/ qui sont considérés comme acoustiquement [extrême] à l'image de /iuy/ mais [non haut] d'un point de vue articulatoire comme /oɔeɛøœ /. Cette nuance place donc ces deux voyelles légèrement différemment dans l'espace multidimensionnel des traits selon la décomposition choisie. Les conséquences restent toutefois minimes.

La décomposition des consonnes proposée par Rossi et présentée dans le Tableau 7b, p.70, fournit une classification en 6 traits. Une distance de norme 1 non pondérée permet de dresser la matrice de coût suivante:

Tableau 28: Matrice de coût pour les consonnes sans pondération des traits

.	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	r	j	w	ɥ	ñ
p	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	1	2	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	2	1	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	3	2	1	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	1	2	3	2	3	4	0	.	.	.	.	.	.	.	.	.	.	.	.	.
s	2	1	2	3	2	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.
ʃ	3	2	1	4	3	2	2	1	0	.	.	.	.	.	.	.	.	.	.	.
v	2	3	4	1	2	3	1	2	3	0	.	.	.	.	.	.	.	.	.	.
z	3	2	3	2	1	2	2	1	2	1	0	.	.	.	.	.	.	.	.	.
ʒ	4	3	2	3	2	1	3	2	1	2	1	0	.	.	.	.	.	.	.	.
m	3	4	5	2	3	4	4	5	6	3	4	5	0	.	.	.	.	.	.	.
n	4	3	4	3	2	3	5	4	5	4	3	4	1	0	.	.	.	.	.	.
l	3	2	3	2	1	2	4	3	4	3	2	3	2	1	0	.	.	.	.	.
r	4	3	2	3	2	1	5	4	3	4	3	2	3	2	1	0	.	.	.	.
j	4	3	4	3	2	3	3	2	3	2	1	2	3	2	1	2	0	.	.	.
w	5	4	3	4	3	2	4	3	2	3	2	1	4	3	2	1	1	0	.	.
ɥ	4	3	4	3	2	3	3	2	3	2	1	2	3	2	1	2	0	1	0	.
ñ	5	4	3	4	3	2	6	5	4	5	4	3	2	1	2	1	3	2	3	0

En tenant compte des observations faites ci-avant, notamment à propos de l'expérience de Chistovitch et al., il nous semble judicieux de donner plus de poids aux traits de mode puis de voisement et enfin de lieu. Pour cela, nous avons attribué une importance triple aux traits [vocalique] et [interrompu], une pondération double au trait de [voisement] et enfin un poids simple aux traits [nasal], [compact] et [aigu]. Cette distribution pondérée permet de dresser la matrice de coût suivante:

Tableau 29: Matrice de coût pour les consonnes avec pondération des traits

.	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	r	j	w	ɥ	ñ
p	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	2	3	4	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	3	2	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	4	3	2	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	3	4	5	5	6	7	0	.	.	.	.	.	.	.	.	.	.	.	.	.
s	4	3	4	6	5	6	1	0	.	.	.	.	.	.	.	.	.	.	.	.
ʃ	5	4	3	7	6	5	2	1	0	.	.	.	.	.	.	.	.	.	.	.
v	5	6	7	3	4	5	2	3	4	0	.	.	.	.	.	.	.	.	.	.
z	6	5	6	4	3	4	3	2	3	1	0	.	.	.	.	.	.	.	.	.
ʒ	7	6	5	5	4	3	4	3	2	2	1	0	.	.	.	.	.	.	.	.
m	6	7	8	4	5	6	9	10	11	7	8	9	0	.	.	.	.	.	.	.
n	7	6	7	5	4	5	10	9	10	8	7	8	1	0	.	.	.	.	.	.
l	6	5	6	4	3	4	9	8	9	7	6	7	2	1	0	.	.	.	.	.
r	7	6	5	5	4	3	10	9	8	8	7	6	3	2	1	0	.	.	.	.
j	9	8	9	7	6	7	6	5	6	4	3	4	5	4	3	4	0	.	.	.
w	10	9	8	8	7	6	7	6	5	5	4	3	6	5	4	3	1	0	.	.
ɥ	9	8	9	7	6	7	6	5	6	4	3	4	5	4	3	4	0	1	0	.
ñ	8	7	6	6	5	4	11	10	9	9	8	7	2	1	2	1	5	4	5	0

Il reste à régler deux problèmes :

- le positionnement des macro-classes.

  Cette question est relativement facile à régler compte tenu du fait que les macro-classes sont simplement des catégories présentant la neutralisation d'un ou plusieurs traits. Ainsi, les consonnes vocaliques sont définies par les traits [vocalique] et [voisé], les autres traits étant indéfinis. Cela place cet élément à distance nulle des consonnes /mnlrjw/ et non nulles des autres. Dans l'espace multidimensionnel des traits, les macro-classes constituent non pas des points comme le sont les phonèmes mais des objets multidimensionnels.

  L'analyse des résultats nous a révélé deux faits connus dans la communauté phonéticienne. Le premier phénomène est le polymorphisme de /r/ (Straka, 1965). Nous avons constaté que, fréquemment, pour certains locuteurs, ce phonème prend une forme gutturale [x], qui correspond à une constrictive. Nous avons donc tenu compte de cette forme gutturale en associant une faible distance entre les macro-classes fricatives et /r/. De plus, la labiodentale /v/ est, de toutes les constrictives, celle qui produit le moins de bruit de friction. Parfois même, pour certains locuteurs, ce bruit n'apparaît pas dans la structure spectrale. /v/ prend alors une forme quasi-périodique qui l'apparente à une consonne vocalique. Là encore, nous avons tenu compte de ce type de réalisation en associant une faible distance entre les macro-classes vocaliques et /v/.

- le positionnement respectif des consonnes et des voyelles.

  Les matrices de coût précédemment décrites sont établies à l'intérieur des classes de consonnes et de voyelles. Nous ne disposons pour le moment d'aucun moyen de quantifier les écarts entre une voyelle et une consonne. Pour établir ce lien, nous avons utilisé les macro-traits définis par Dell (cf. Tableau 4, p.66). Malgré son existence phonologique, le trait de syllabicité, permettant de distinguer voyelles et semi-voyelles (cf. § II.II.4.II.4.D.II.4.D.c., p.66), reste un concept contestable. Nous avons donc traité de façon indépendante le lien entre voyelles et semi-voyelles. Par contre, en ajoutant aux traits [consonantique] et [vocalique] le trait de voisement, nous définissons alors le système suivant :

Tableau 30: Macro-traits définissant les macro-classes

	consonantique	vocalique	voisé
voyelles et semi-voyelles	-	+	+
liquides et nasales	+	+	+
obstruantes sonores	+	-	+
obstruantes sourdes	+	-	-

La plus grande distance entre deux voyelles étant de 5 (ex: entre /u/ et / ɛ̃ /), nous avons fixé cette valeur comme offset entre voyelles et consonnes. En attribuant un poids triple à l'écart de macro-traits, auquel s'ajoutent l'offset, cela positionne les voyelles de la façon suivante :

Tableau 31: Position des voyelles par rapport aux macro-classes consonantiques

	liquides et nasales	obstruantes sonores	obstruantes sourdes
voyelles et semi-voyelles	5 + 1 * 3 = ***8***	5 + 2 * 3 = ***11***	5 + 3 * 3 = ***14***

offset

pondération

nombre de traits différents

Il faut remarquer qu'un tel positionnement place les consonnes vocaliques plus près des voyelles que des constrictives sourdes. En effet, la distance entre /m/ et /s/ est évaluée à 10 (cf. Tableau 29, p.208) alors que la distance voyelles ↔ /m/ (consonne nasale) est fixée à 8. Un tel fait est prémédité car il reflète une certaine réalité acoustico-phonétique.

Par contre, le positionnement proposé ne laisse pas apparaître la proximité qui peut exister entre voyelles nasales et consonnes nasales. Pourtant, en général et plus particulièrement dans certaines variétés de français du sud-ouest et du Languedoc, le trait de

nasalité sur la consonne peut se transmettre aux voyelles adjacentes comme dans le mot « année » qui se prononce non pas [ane] mais [ãne]. De même, nous avons vu que la production d'une voyelle nasale laisse souvent apparaître deux phases qui peuvent être décodées séparément par le système en une voyelle orale et une unité nasale (cf. § « Le cas de la nasalité », p.169). Afin de tenir compte de ces phénomènes, nous avons réduit la distance entre voyelles nasales et consonnes nasales.

Il reste, à présent, à positionner voyelles et semi-voyelles. La solution que nous avons adoptée consiste à assimiler semi-voyelles et voyelles dans une même classe. En considérant le système de traits des voyelles présenté à la Figure 34, page 67 , en assimilant la semi-voyelle à son homologue syllabique (ex : /j/ ↔ /i/, /w/ ↔ /u/) et en ajoutant un trait de syllabicité permettant la distinction voyelles/semi-voyelles nous obtenons la matrice suivante:

Tableau 32: Position des semi-voyelles par rapport aux voyelles

	a	i	u	o	e	y	ø	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃	Ê	Ô	@	Ë
j	4	1	3	4	2	2	3	3	5	4	5	4	6	5	2	4	3	4
w	4	3	1	2	4	2	3	5	3	4	5	6	4	5	4	2	3	5
ɥ	5	2	2	3	3	1	2	4	4	3	6	5	5	4	3	3	2	4

En tenant compte des différentes considérations précédentes, nous sommes en mesure de proposer une matrice de coût générale présentée au Tableau 33.

Le dispositif d'accès lexical utilise les colonnes comme éléments des représentations phonotypiques des mots du dictionnaire et les lignes pour les unités de la chaîne phonétique décodée (cf. § « Une mesure de distance », p.203). Ceci explique le fait que la matrice ne soit pas complètement symétrique par rapport à la diagonale. En effet, les macro-classes n'interviennent que dans de la chaîne phonétique décodée et jamais dans les représentations du dictionnaire.

VIII.2.C. L'évaluation du degré de difficulté de décodage d'un lexique

Nous savons que, selon la composition phonique des mots du lexique, les performances d'un système de reconnaissance automatique de la parole peuvent être grandement affectées (cf. § « Petit/grand lexique, vocabulaire facile/difficile », p.23). Des mots acoustiquement proches seront difficiles à reconnaître. La possibilité de chiffrer la dissemblance entre deux phonèmes peut être utile pour évaluer le degré de difficulté de décodage d'un lexique particulier. En effet, deux mots acoustiquement proches sont reliés par une faible distance du fait de la similarité probable des traits relatifs aux phonèmes de ces mots. Si le calcul des distances entre chaque mot du lexique et le reste du dictionnaire laisse apparaître un résultat globalement important, cela indique, au contraire, une tâche de reconnaissance plutôt facile.

A l'aide de l'algorithme de comparaison dynamique (cf. page 203) s'appuyant sur la matrice de coût, il est possible d'effectuer cette opération. Il suffit de comparer chaque mot du lexique à tous les autres éléments du dictionnaire. Le mot dont la distance relative est la plus faible est alors considéré comme le plus proche voisin.

Tableau 33: Matrice de coût

	a	i	u	o	e	y	ø	ɛ	ɔ	œ	ã	ɛ̃	ɔ̃	œ̃	Ê	Ô	@	Ë	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	r	j	w	ɥ	ñ
a	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
i	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
u	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
o	2	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
e	2	1	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
y	4	1	1	2	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ø	3	2	2	1	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɛ	1	2	4	3	1	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɔ	1	4	2	1	3	3	2	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
œ	2	3	3	2	2	2	1	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ã	1	4	4	3	3	5	4	2	2	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɛ̃	2	3	5	4	2	4	3	1	3	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ɔ̃	2	5	3	2	4	4	3	3	1	2	1	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
œ̃	3	4	4	3	3	3	2	2	2	1	2	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ê	1	1	3	2	0	2	1	0	2	1	2	1	3	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ô	1	3	1	0	2	2	1	2	0	1	2	3	1	2	2	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
@	2	2	2	1	1	1	0	1	1	0	3	2	2	1	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ë	2	3	4	3	2	3	2	1	2	1	1	0	1	0	1	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
p	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	2	3	4	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	3	2	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	4	3	2	2	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
f	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	3	4	5	5	6	7	0	.	.	.	.	.	.	.	.	.	.	.	.	.
s	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	4	3	4	6	5	6	1	0	.	.	.	.	.	.	.	.	.	.	.	.
ʃ	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	5	4	3	7	6	5	2	1	0	.	.	.	.	.	.	.	.	.	.	.
v	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	5	6	7	3	4	5	2	3	4	0	.	.	.	.	.	.	.	.	.	.
z	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	6	5	6	4	3	4	3	2	3	1	0	.	.	.	.	.	.	.	.	.
ʒ	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	7	6	5	5	4	3	4	3	2	2	1	0	.	.	.	.	.	.	.	.
m	8	8	8	8	8	8	8	8	8	8	5	5	5	5	8	8	8	5	6	7	8	4	5	6	9	10	11	7	8	9	0	.	.	.	.	.	.	.
n	8	8	8	8	8	8	8	8	8	8	5	5	5	5	8	8	8	5	7	6	7	5	4	5	10	9	10	8	7	8	1	0	.	.	.	.	.	.
l	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	6	5	6	4	3	4	9	8	9	7	6	7	2	1	0	.	.	.	.	.
r	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	7	6	5	5	4	3	10	9	8	8	7	6	3	2	1	0	.	.	.	.
j	4	1	3	4	2	2	3	3	5	4	5	4	6	5	2	4	3	4	9	8	9	7	6	7	6	5	6	4	3	4	5	4	3	4	0	.	.	.
w	4	3	1	2	4	2	3	5	3	4	5	6	4	5	4	2	3	5	10	9	8	8	7	6	7	6	5	5	4	3	6	5	4	3	1	0	.	.
ɥ	5	2	2	3	3	1	2	4	4	3	6	5	5	4	3	3	2	4	9	8	9	7	6	7	6	5	6	4	3	4	5	4	3	4	0	1	0	.
ñ	8	8	8	8	8	8	8	8	8	8	5	5	5	5	8	8	8	5	8	7	6	6	5	4	11	10	9	9	8	7	2	1	2	1	5	4	5	0
ŋ	8	8	8	8	8	8	8	8	8	8	5	5	5	5	8	8	8	5	7	6	7	5	4	5	10	9	10	8	7	8	1	0	1	2	4	5	4	1
VOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	5	3	3	3	5	5	5	0	3	3	0	0	0	0	0	0	0	0
VOY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	14	14	11	11	11	14	14	14	11	11	11	8	8	8	8	0	0	0	8
CS	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CSNV	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	0	0	0	2	2	2	0	0	0	2	2	2	5	5	5	3	14	14	14	5
CSVX	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	2	2	2	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0
BCVX	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	5	5	5	0	0	0	8	8	8	5	5	5	0	0	0	0	0	0	0	0
COVOC	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	3	3	3	5	5	5	0	3	3	0	0	0	0	0	0	0	0
FRICA	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	3	3	3	3	3	3	0	0	0	0	0	0	7	7	6	3	12	12	12	7
OCCVO	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	2	2	2	0	0	0	5	5	5	3	3	3	4	4	3	3	11	11	11	4
OCNV	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	0	0	0	2	2	2	3	3	3	5	5	5	6	6	5	5	14	14	14	6
FRVX	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	5	5	5	3	3	3	2	2	2	0	0	0	7	7	6	3	11	11	11	7
FRNV	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	3	3	3	5	5	5	0	0	0	2	2	2	9	9	8	4	14	14	14	9

Abréviations :

VOC = vocalique, VOY = voyelle,
CS = consonne, CSNV = consonne non voisée, CSVX = consonne voisée,
BCVX = occlusive voisée ou consonne vocalique, COVOC = consonne vocalique,
FRICA = fricative, OCCVO = occlusive voisée, OCNV = occlusive non voisée,
FRVX = fricative voisée, FRNV = fricative non voisée

Dans l'exemple du Tableau 34, le mot testé est « accès ». Son plus proche voisin dans le lexique de 500 mots que nous utilisons est « hausse ». La distance associée donne une valeur quantitative de cette proximité.

Tableau 34: Calcul des plus proches voisins du mot "accès" sur un lexique de 500 mots

Position	Mot	Transcription	Distance
1)	accès	aksÊ	0
2)	hausse	os@	7
3)	achetant	aStä	8
4)	accueilli	akæji	8
5)	pacte	pakt@	9
6)	abstiens	abstjë	10
7)	percer	pÊrsÊ	11
8)	penche	päS@	11
9)	heurte	ært@	11
10)	pêche	pÊS@	11

L'itération de cette opération pour tous les mots du lexique autorise le calcul d'une distance moyenne qui peut être vue comme le facteur moyen de facilité de décodage de ce lexique. Dans l'absolu, cette démarche n'apporte pas vraiment d'informations, si ce n'est de connaître a priori les mots qui peuvent poser problème. Par contre, elle apparaît intéressante dans le cadre d'études comparatives. Nous pensons, par exemple, à l'évaluation des systèmes de reconnaissance où le choix du lexique n'est pas sans importance.

VIII.2.D. Bilan

Le module d'accès lexical est fondé sur un algorithme de programmation dynamique qui s'appuie sur une matrice de coût. Celle-ci a été établie à partir de la théorie des traits. Nous avons conscience de l'aspect fort arbitraire de notre démarche. Cette recherche des poids de la matrice ressemble d'ailleurs un peu à celle de l'estimation des probabilités utilisées dans les chaînes de Markov ou dans les poids d'un réseau de neurones. Néanmoins, une différence importante sépare les deux approches : on sait où est l'information. En d'autres termes, si le dispositif sépare mal deux phonèmes, nous pouvons essayer de jouer avec la matrice. C'est d'ailleurs ainsi que nous avons fonctionné. La matrice a été améliorée au fur et à mesure que les résultats du décodage étaient dépouillés et analysés. Les nuances décrites précédemment se sont mises en place petit à petit. La matrice risque d'ailleurs d'évoluer à nouveau si nous prenons conscience qu'un phénomène est mal exprimé.

Le module d'accès lexical permet finalement de comparer deux chaînes phonétiques en intégrant les opérations d'insertion, d'omission et de substitution de phonèmes. Dans le cadre de notre système ACHILE, le module d'accès lexical a pour rôle de fournir le mot correspondant à la chaîne phonétique identifiée par les modules de décodage ascendants. Pour cela, il s'appuie sur un lexique de 500 mots codés à la fois sous la forme orthographique et phonémique. Nous décrirons ce dictionnaire ultérieurement.

VIII.3. Le superviseur

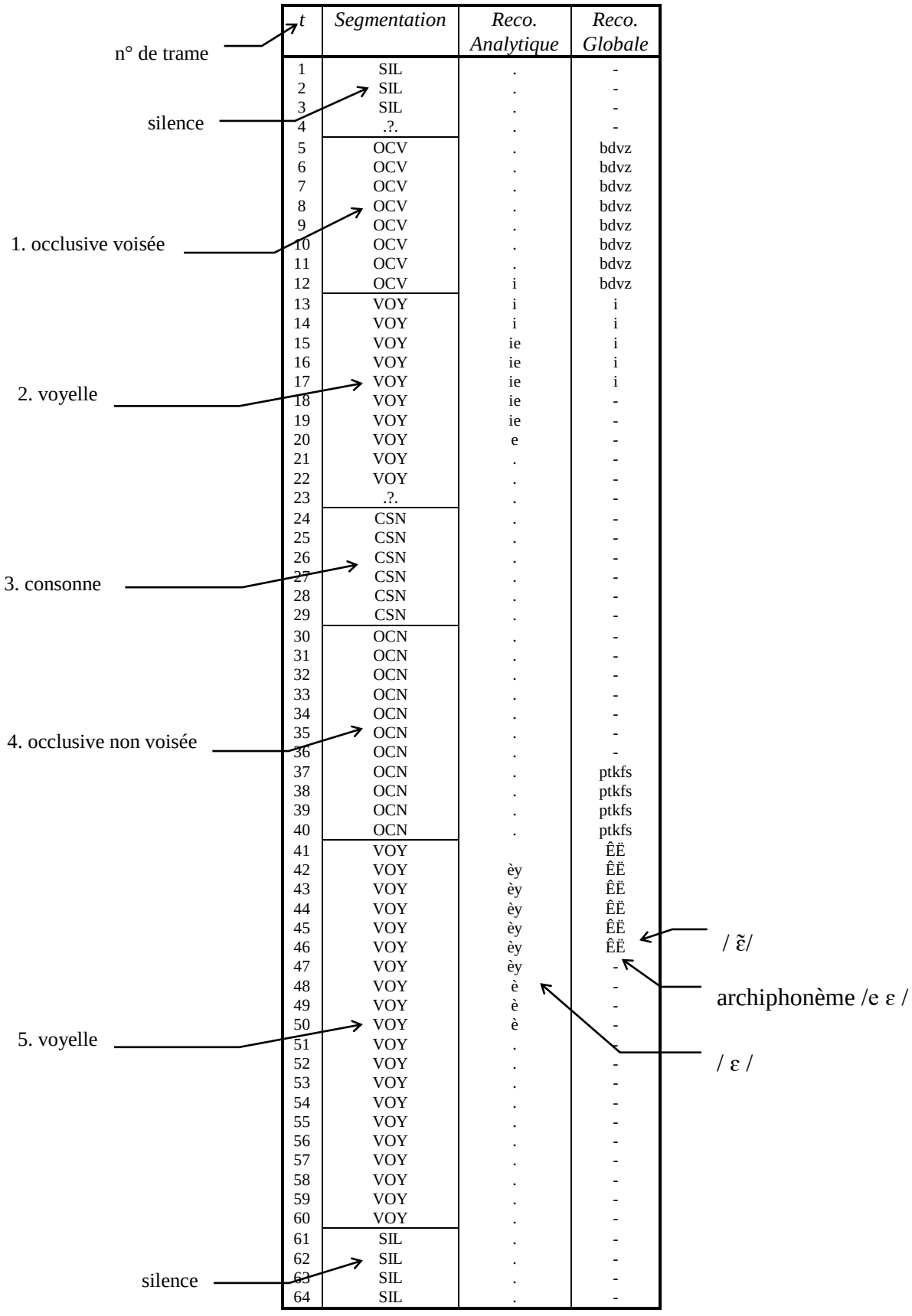
Le module superviseur est probablement la partie la plus délicate du dispositif et la plus difficile à mettre en oeuvre. Son rôle est de piloter les modules de décodage ascendant, de recueillir l'information ainsi fournie, de construire une ou plusieurs chaînes phonétiques prétendantes et de proposer une réponse à partir d'un accès lexical (Figure 63, p.129). Nous touchons ici aux stratégies de décision et à la fusion de l'information. Nous regrettons de n'avoir pas pu approfondir ce travail. Nous présentons donc un module superviseur plutôt sommaire, qui fonctionne de la façon suivante.

Dans un premier temps, le module superviseur collecte les différents résultats du décodage ascendant (Tableau 35). L'information syntagmatique, c'est à dire la distribution temporelle des phonèmes, est apportée par la segmentation automatique (Tableau 35, colonne [*Segmentation*]). Ce module fournit aussi les macro-classes consonantiques. L'identification exacte des voyelles est fournie par les modules de reconnaissance globale et analytique (Tableau 35, colonne [*Reco. Analytique*] et [*Reco. Globale*]). En fusionnant toutes ces informations, le superviseur construit différentes chaînes phonétiques candidates. Cette opération est, pour le moment, très rudimentaire. Le principe est de prendre tour à tour chaque candidat à une position syntagmatique précise, à commuter ces candidats et à générer ainsi toute la combinatoire. Dans l'exemple du Tableau 35, le décodage ascendant a identifié 5 unités. La 1ère est une occlusive voisée (OCV). La 2ème est une voyelle [i] ou [e]. La 3ème est une consonne sourde (CSN). La 4ème une occlusive sourde (OCN). La 5ème est [ɛ e], [ɛ̃] ou [y]. Ceci fournit finalement une combinatoire à 6 candidats (Tableau 36).

Nous avons conscience que cette fusion des informations est extrêmement simpliste et gagnerait à être améliorée, notamment dans le cas de conflit d'informations. Nous pensons, par exemple, aux techniques de raisonnement hypothétique (Coste-Marquis, 1994). Dans le cas d'un décodage de mots isolés, l'accès à un dictionnaire permet au superviseur de fournir une liste ordonnée de mots-candidats à partir des chaînes phonétiques candidates (Tableau 37). Cette opération est réalisée en comparant chaque chaîne phonétique décodée avec les entrées phonémiques du dictionnaire comme décrit au § « L'accès lexical », p.203.

Dans l'exemple du Tableau 37, le mot testé "dictée" a été placé en première position. Un processus de vérification descendante, qui reste à développer, pourrait éliminer des candidats comme "goûter" ou "bonté" car dans ces mots, la première voyelle est grave, ce qui se trouve en contradiction avec le mot testé où la première voyelle est clairement aiguë.

Tableau 35: Décodage ascendant sur le mot "dictée"



t	*Segmentation*	*Reco. Analytique*	*Reco. Globale*
1	SIL	.	-
2	SIL	.	-
3	SIL	.	-
4	.?.	.	-
5	OCV	.	bdvz
6	OCV	.	bdvz
7	OCV	.	bdvz
8	OCV	.	bdvz
9	OCV	.	bdvz
10	OCV	.	bdvz
11	OCV	.	bdvz
12	OCV	i	bdvz
13	VOY	i	i
14	VOY	i	i
15	VOY	ie	i
16	VOY	ie	i
17	VOY	ie	i
18	VOY	ie	-
19	VOY	ie	-
20	VOY	e	-
21	VOY	.	-
22	VOY	.	-
23	.?.	.	-
24	CSN	.	-
25	CSN	.	-
26	CSN	.	-
27	CSN	.	-
28	CSN	.	-
29	CSN	.	-
30	OCN	.	-
31	OCN	.	-
32	OCN	.	-
33	OCN	.	-
34	OCN	.	-
35	OCN	.	-
36	OCN	.	-
37	OCN	.	ptkfs
38	OCN	.	ptkfs
39	OCN	.	ptkfs
40	OCN	.	ptkfs
41	VOY		ÊË
42	VOY	èy	ÊË
43	VOY	èy	ÊË
44	VOY	èy	ÊË
45	VOY	èy	ÊË
46	VOY	èy	ÊË
47	VOY	èy	-
48	VOY	è	-
49	VOY	è	-
50	VOY	è	-
51	VOY	.	-
52	VOY	.	-
53	VOY	.	-
54	VOY	.	-
55	VOY	.	-
56	VOY	.	-
57	VOY	.	-
58	VOY	.	-
59	VOY	.	-
60	VOY	.	-
61	SIL	.	-
62	SIL	.	-
63	SIL	.	-
64	SIL	.	-

Tableau 36: Chaînes phonétiques candidates

[OCV][i][CSN][OCN][ɛ e]
[OCV][i][CSN][OCN][ɛ̃]
[OCV][i][CSN][OCN][y]
[OCV][e][CSN][OCN] [ɛ e]
[OCV][e][CSN][OCN] [ɛ̃]
[OCV][e][CSN][OCN][y]

Tableau 37: L'accès lexical

position	chaîne phonétique décodée	forme orthographique du mot dans le dictionnaire	forme phonétique du mot dans le dictionnaire	distance entre la chaîne phonétique décodée et la forme phonétique du mot dans le dictionnaire
1	[OCV]i[CSN][OCN]Ê	dictée	diktÊ	0
2	[OCV]i[CSN][OCN]y	discute	diskyt	2
3	[OCV]e[CSN][OCN]Ê	goûter	gutÊ	5
4	[OCV]i[CSN][OCN]Ê	quitter	kitÊ	5
5	[OCV]i[CSN][OCN]Ê	bonté	bõtÊ	5
6	[OCV]e[CSN][OCN]y	dessus	døsy	5
7	[OCV]Ê[CSN][OCN] ɛ̃	veston	vÊstõ	7
8	[OCV]i[CSN][OCN]Ê	discret	diskrÊ	8
9	[OCV]Ê[CSN][OCN] ɛ̃	latins	latɛ̃	8
10	[OCV]e[CSN][OCN]Ê	poster	postÊ	10

VIII.4. Evaluation du système

VIII.4.A. Le « pourquoi ? » de l'évaluation

Généralement, l'évaluation d'un dispositif de reconnaissance automatique permet de comparer les systèmes entre eux. Une telle démarche nécessite l'adoption de standards comme le choix d'un corpus d'apprentissage et de test, la sélection d'un lexique et l'adoption d'un format de sortie. Dans le cadre de vastes projets de recherche comme ARPA ou AUPELF, de grandes campagnes de tests sont entreprises. Cette course aux chiffres reste le moyen de positionner les équipes de recherche. Nous ne sommes pas entrés dans cette compétition. Dans notre cas, l'évaluation est très importante car elle permet de quantifier la pertinence des connaissances, des analyses et des algorithmes employés. C'est aussi le moyen de pouvoir faire évoluer le dispositif en apportant des modifications. Pour cela, il nous semble nécessaire de garder un critère de contrôlabilité (cf. § « Notre approche », *p.*33), c'est à dire la possibilité de prévoir, analyser et comprendre l'évolution des performances en fonction des modifications apportées. Nous pensons avoir atteint cet objectif. A titre d'exemple, la nécessité d'introduire dans la matrice de « coût » la forme gutturale de /r/ , la version « vocalique » de /v/ , les phénomènes dus à la nasalité... est issue de cette possibilité d'analyse des résultats. En effet, c'est la prise de conscience que 80% des plus mauvaises identifications étaient liés à la présence de /r/ qui nous a poussé à examiner les données. Nous nous sommes alors aperçu de la réalité du polymorphisme de /r/, phénomène pourtant connu mais qui n'apparaissait pas dans nos formalisations. Nous avons donc corrigé cette imperfection en introduisant une forme gutturale de ce phonème. Les performances s'en sont ressenties. C'est alors que les plus mauvaises identifications se sont principalement portées sur les mots comprenant un /v/. Nous avons donc introduit une forme vocalique...

VIII.4.B. Les données sonores

Le corpus d'évaluation de notre système a été réalisé par le laboratoire d'informatique d'Avignon et des Pays du Vaucluse* (Béchet, 1994). Il est composé de 500 mots du lexique français extraits de la base de données BD-Lex (Pérennou & De Calmès, 1986). Six locuteurs (5 hommes, 1 femme) ont participé à l'enregistrement des données sonores. Les données ont été numérisées sur 16 bits à une fréquence d'échantillonnage de 12 kHz. Le signal est peu bruité. Pour nos besoins locaux et garder ces données compatibles avec les nôtres, nous avons rééchantillonné ce corpus à une fréquence d'échantillonnage de 16 kHz.

Une procédure de détection automatique « début/fin de signal » permet d'extraire la parole du silence. L'analyse est alors focalisée sur cette partie utile.

* nous le remercions pour le soutien qu'il nous a apporté

VIII.4.C. Les données lexicales

VIII.4.C.a. Quelle taille de vocabulaire ?

La tendance actuelle en matière de Reconnaissance Automatique de la Parole consiste à augmenter systématiquement la taille du vocabulaire pour atteindre plusieurs dizaines de milliers de mots: 65 000 mots au LIMSI (Matrouf & Gauvain, 1996). Cela s'inscrit dans le cadre de la « machine à dicter » où l'univers de conversation peut être très grand. En réalité, dans le cadre d'une communication humaine ou homme-machine dans un contexte restreint, le nombre de mots potentiellement activables est inférieur à 5000. Cette valeur est approximative car l'on comprend très bien que cela dépend du type de conversation et du vocabulaire actif du locuteur. Toujours est-il que le choix d'un vocabulaire d'environ un millier de mots paraît suffisant. C'est le cas dans (Pousse et. al, 1996) avec 1515 mots, dans El Méliani & O'Shaughnessy, 1996) avec 1018 et 4842 mots.

VIII.4.C.b. Composition et structure du dictionnaire

Notre dictionnaire est composé des 500 mots du corpus d'Avignon. Ci-après est donnée la liste orthographique de ces mots. Le dictionnaire est formé par les formes orthographique et phonémique du mot. Celles-ci sont extraites de la base de données BD-Lex (Pérennou & De Calmès, 1986).

exemple:

argument	argymã
bourgeoise	burʒwazə

Nous avons conscience de l'écart qui existe entre les formes phonologiques des entrées du dictionnaire et les réalisations phonétiques (Pousse et al., 1996). Pour combler cet écart dû à la non-biunivocité des plans de description, plusieurs solutions sont envisageables:

- l'utilisation en direct de règles phonologiques permettant le passage des formes phonologiques vers les réalisations phonétiques. Le sens transformationnel de tels processus est utile en synthèse de la parole mais se heurte à l'aspect interprétatif de la reconnaissance automatique de la parole.
- la dérivation préalable des formes phonologiques en réalisations phonétiques.
  Cette méthode consiste à associer plusieurs des formes phonétiques à un élément du lexique. C'est la solution que nous avons adoptée.

Notre lexique comporte 500 graphies, ce qui donne lieu à 758 entrées lorsqu'on utilise plusieurs prononciations pour un même mot par dérivation de la forme phonologique du schwa. Les entrées phonétiques tiennent compte des élisions ou réalisations effectives du 'e muet' (ex: seconde => /səgɔ̃də/ => (1)[sgɔ̃d] ou (2)[søgɔ̃d] ou (3)[sgɔ̃dø] ou (4)[søgɔ̃dø]). Une telle dérivation nous semble d'ailleurs insuffisante: une forme phonétique du type [zgɔ̃d] nous semblerait préférable à la forme (1)[sgɔ̃d] compte tenu des phénomènes connus d'assimilation. Nous n'avons toutefois pas abordé ces phénomènes complexes et nous sommes contentés de dérivations simples. Nous avons conscience de ce point faible.

1. à
2. abandonnerai
3. aborder
4. abrégerons
5. abstiens
6. accéderons
7. accès
8. accompagnerons
9. accourons
10. accueilli
11. achetant
12. adapterai
13. adjectif
14. administrant
15. adoucirai
16. aérienne
17. afficherai
18. affronterai
19. agissez
20. agréant
21. aïeul
22. aligner
23. allée
24. allongerai
25. ambitieux
26. aménagerai
27. analyse
28. anéantissez
29. anniversaire
30. apaisons
31. aplatissez
32. appelle
33. appétit
34. apportant
35. approuverai
36. argument
37. arranger
38. arroser
39. aspiratrice
40. assassinerons
41. assiégerons
42. assurant
43. atteignez
44. atterrissez
45. attribut
46. attristons
47. autonomie
48. avancerai
49. aviateur
50. avoue
51. bail
52. baise
53. baptise
54. barbare
55. bâtissez
56. bébé
57. bénissons
58. bicyclette
59. blanchissant
60. blindé
61. bombarder
62. bonté
63. bornant
64. bouleverse
65. bourgeoise
66. braient
67. brillant
68. brillons
69. brouillerai
70. brunissent
71. brutalité
72. cage
73. calculerai
74. cane
75. capitulerons
76. carburant
77. casserole
78. causant
79. certain
80. chance
81. chargerai
82. chasseuse
83. chercher
84. cheval
85. chrétien
86. cirer
87. ciseaux
88. climat
89. clorai
90. coïnciderons
91. colle
92. combattrons
93. comité
94. commettre
95. compagnon
96. compléterai
97. comprise
98. conception
99. concernerons
100. condenser
101. conformer
102. confus
103. conquièrent
104. conserver
105. consomme
106. consonne
107. construisant
108. contesterai
109. contrebande
110. contredite
111. convenez
112. correct
113. correspondu
114. coulerons
115. couronnant
116. courte
117. craignent
118. crèverons
119. cri
120. croissons
121. cuisent
122. cultivateur
123. danserai
124. débarrasserons
125. débouche
126. débrouillarde
127. décède
128. déchausserons
129. décollerons
130. découragerons
131. dédirai
132. défendez
133. définition
134. déguiser
135. délibérer
136. déménage
137. démissionnerons
138. démocratique
139. démoralisons
140. dépassons
141. déplacer
142. déposerai
143. déroberai
144. désaccord
145. déshonorer
146. désobéis
147. dessus
148. détournant
149. devenons
150. dévoue
151. dictée
152. diminuerai
153. discret
154. discute
155. disposerons
156. dissoudrai
157. diviser
158. document
159. dormir
160. drap
161. dressant
162. éblouissons
163. échapper
164. éclairerons
165. éclatant
166. écoulant
167. écrouerai
168. effectue
169. effrayerons
170. égale
171. éléphant
172. élèverons
173. emballe
174. embrasse
175. émigrerons
176. empêche
177. employé
178. empresserai
179. enclouerai
180. enduire
181. enflerai
182. engloutissons
183. enlèverons
184. enrichirai
185. entendez
186. enterrement
187. entreprendre
188. envahir
189. envolant
190. épellerons
191. épidémie
192. épuiserons
193. espionne
194. essayerai
195. établissent
196. étendant
197. éternel
198. étranglant
199. évaluons
200. évidemment
201. éviter
202. examiner
203. exclure
204. exécution
205. exiger
206. expliquons
207. exposons
208. extrait
209. fabrique
210. familial
211. fauche
212. féminin
213. fendrai
214. figurons
215. flambe
216. fleur
217. foncerai
218. forger
219. format
220. fouillons
221. fraîche
222. franchis
223. frémirai
224. frottons
225. gâchons
226. gamme
227. garerai
228. geins
229. générale
230. gênerons
231. glaciale
232. goûter
233. grâce
234. gratter
235. grave
236. grince
237. grossissant
238. guère
239. habiller
240. hausse
241. heurte
242. hier
243. hôtesse
244. identique
245. ignorons
246. imiterai
247. immoral
248. imposant
249. imposerai
250. inconscient
251. indignerai
252. indispensable
253. infinitif
254. infliger
255. inondation
256. inscris
257. inspecterons
258. insulte
259. intelligent
260. interpellant
261. intervenu
262. intime
263. invoquons
264. israélite
265. jaunir
266. jeûne
267. jouerai
268. jumeau
269. justifierons
270. lancer
271. latins
272. léguer
273. lentement
274. lierai
275. lieu
276. lirons
277. locution
278. loyer
279. luis
280. magistrat
281. maigrirons
282. maladresse
283. manche
284. manierons
285. maquillant
286. marins
287. matériel
288. maudire
289. mécanisme
290. mélangeons
291. ménager
292. mériterons
293. message
294. microscope
295. minéral
296. mobile
297. modère
298. mondiale
299. montons
300. mouche
301. mouchons
302. mouvons
303. mûrissent
304. musicien
305. nation
306. nationalisons
307. négligence
308. nettoie
309. Noël
310. notons
311. nu
312. nuisons
313. obligerons
314. oblong
315. obtenons
316. odorat
317. offrir
318. opérant
319. opinion
320. ordonnance
321. organisant
322. originalité
323. oublier
324. pacte
325. paierai
326. palis
327. papier
328. parcourent
329. paresseuse
330. parlerons
331. partielle
332. parviennent
333. patience
334. pêche
335. penche
336. percer
337. périlleux
338. permettons
339. peser
340. peuple
341. pierre
342. pincerons
343. plafond
344. plaignant
345. pleurant
346. plombier
347. poétique
348. pomme
349. pondre
350. pose
351. poster
352. pouce
353. poursuivrons
354. pratiquerai
355. précieuse
356. précisant
357. préméditons
358. préparatifs
359. prescrire
360. présiderons
361. prêtons
362. prévision
363. prive
364. privilégié
365. produisez
366. profitable
367. promener
368. propageons
369. propos
370. prouvons
371. punir
372. qualificative
373. quitter
374. quoi
375. raccourcissent
376. raffolons
377. raisonnement
378. rajeunirons
379. ramenons
380. ranimerons
381. rapportant
382. rassurant
383. ravi
384. ravissons
385. réalisant
386. recevrons
387. réclamation
388. récolter
389. réconcilierai
390. reconnaîtrons
391. recrute
392. récupérer
393. redoublerai
394. refaire
395. réfléchi
396. reflétons
397. refusons
398. réglerai
399. rein
400. rejetterons
401. relâchant
402. remettons
403. remplissent
404. renferme
405. renseigne
406. renvoi
407. répandez
408. répartissez
409. repentent
410. répondons
411. repoussant
412. reprocherons
413. réserve
414. résignons
415. respectant
416. resserrons
417. résultant
418. retarderons
419. retourne
420. retrouvant
421. réveil
422. révèle
423. reverdirons
424. rincerons
425. risque
426. ronger
427. rouage
428. rougirons
429. ruinerai
430. ruse
431. salissons
432. salive
433. sauf
434. sauver
435. séchons
436. seconde
437. séduite
438. sens
439. sentant
440. serrure
441. siégeons
442. signifierons
443. sincérité
444. soin
445. sollicitant
446. sorte
447. soufflons
448. souillant
449. soumettrai
450. souscription
451. soutenant
452. spectateur
453. stupéfait
454. subissent
455. sucrerai
456. suite
457. suivons
458. supprime
459. surprendrai
460. sursis
461. survenir
462. suspends
463. synthétique
464. tapisserie
465. tardant
466. télégraphier
467. témoignage
468. tenons
469. terre
470. tigresse
471. timbrerons
472. tombant
473. torpillerons
474. torture
475. traduisez
476. tragédie
477. trairons
478. transformerai
479. transportant
480. travaux
481. tricher
482. trompe
483. tronc
484. tu
485. tuerons
486. une
487. utiliserons
488. vantons
489. vaste
490. vengeons
491. venue
492. verrons
493. veston
494. vexerai
495. vigilante
496. violet
497. visons
498. voici
499. voterai
500. vrai

VIII.4.C.c. Composition phonétique des éléments du dictionnaire

L'analyse du corpus en terme de nombre de phonèmes par mot laisse apparaître une répartition normale (Figure 94). Le mot le plus court est "à" (1 phonème). Le plus long est "qualificative" avec 12 phonèmes. La longueur moyenne est de 6,3 phonèmes par mot. 57 % des mots sont constitués de 5 à 7 phonèmes. 90 % des mots font entre 4 et 9 phonèmes. Le vocabulaire apparaît convenablement réparti.

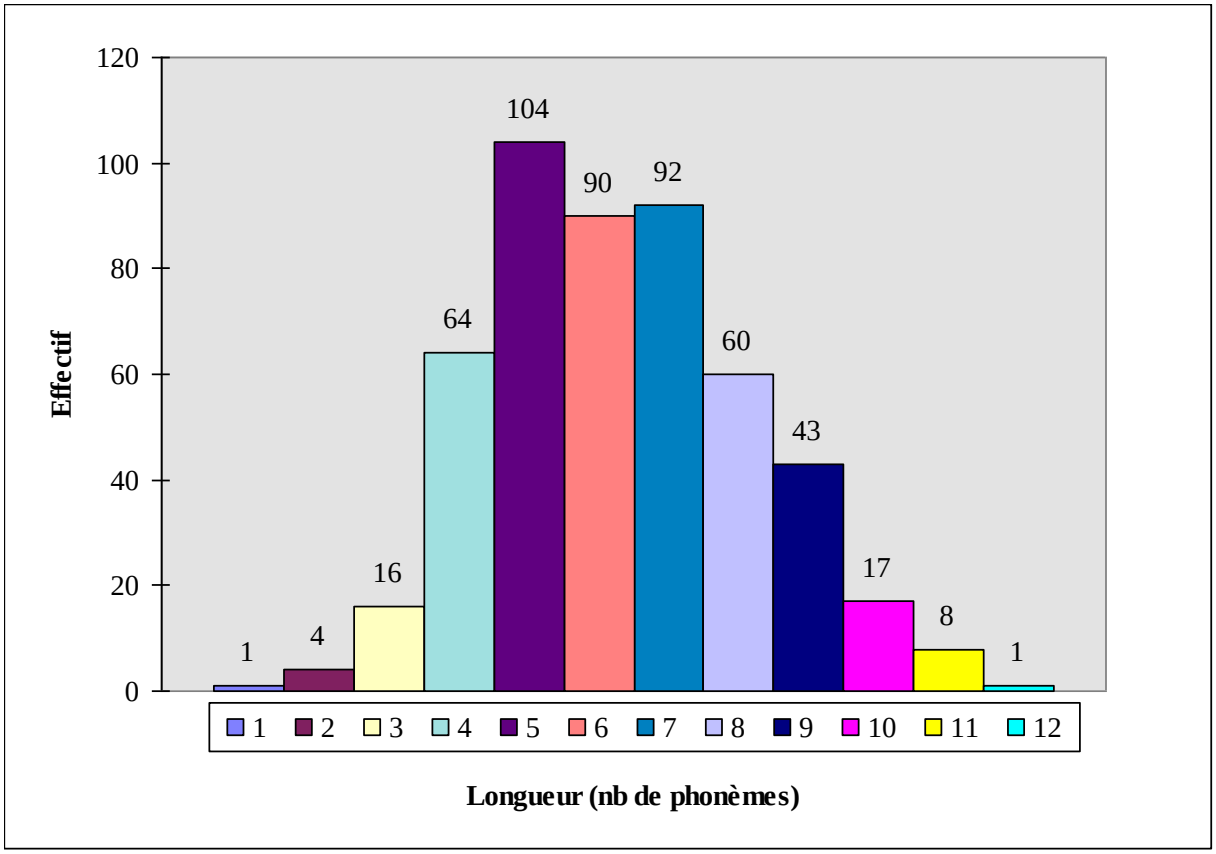


Figure 94: Composition des mots du lexique en terme de longueur phonémique

L'analyse en terme de combinatoire consonne/voyelle laisse apparaître 101 structures différentes. La plus fréquente est la forme CVCVCV qui représente 12% des mots du vocabulaire, puis se succèdent des structures CVCV avec 11%, CCVCV avec 6,8%... (Figure 95). Le vocabulaire ne laisse apparaître aucun biais dû à une importance nette d'une structure spécifique.

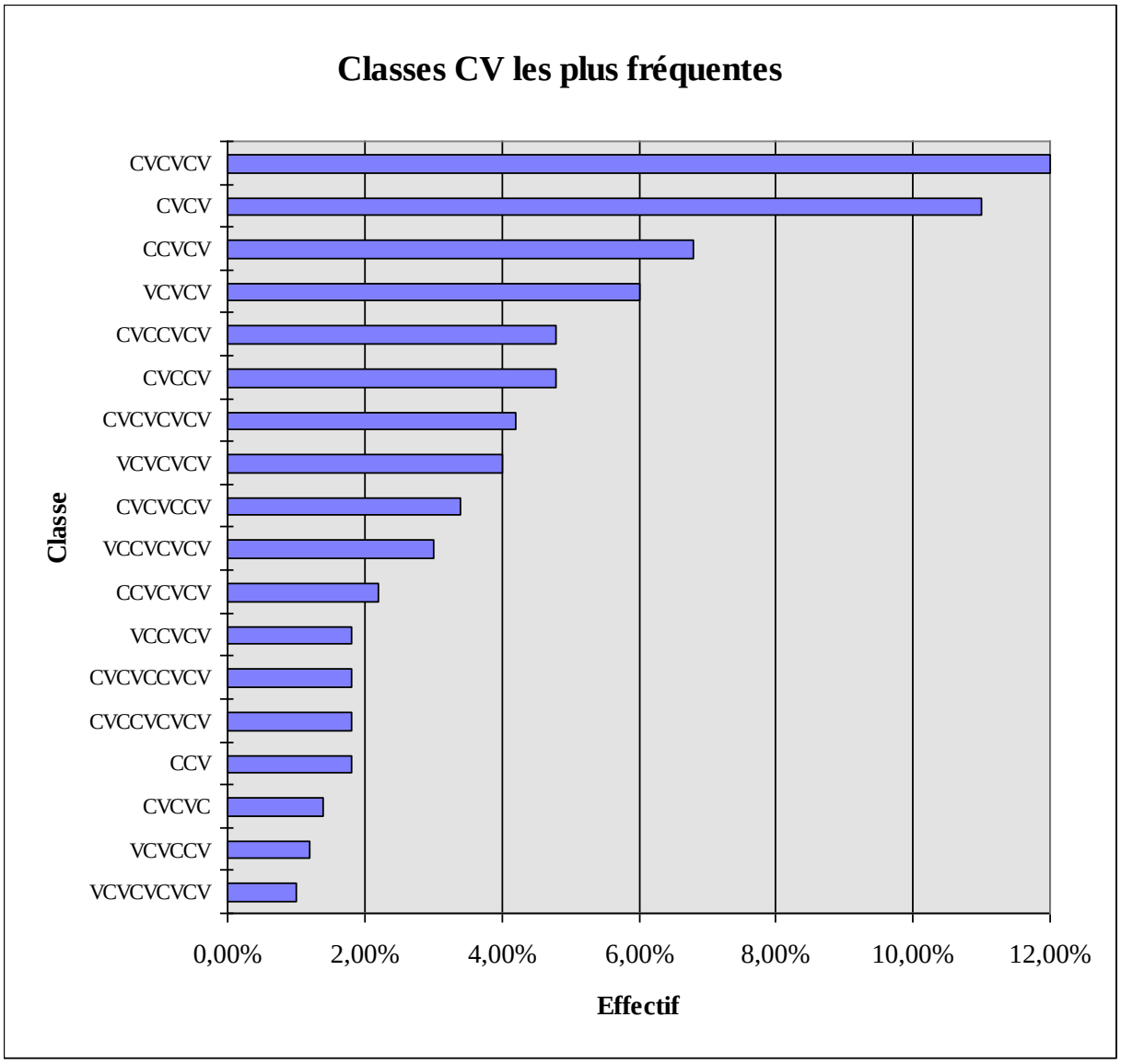


Figure 95: Composition phonotactique des mots du lexique

VIII.4.D. Résultats de l'évaluation

VIII.4.D.a. Les conditions d'évaluation

L'évaluation des modules de reconnaissance analytique et globale a été présentée dans les chapitres précédents. Il s'agissait alors de tests spécifiques aux modules avec un corpus particulier. Nous présentons maintenant l'évaluation globale du système ACHILE dans lequel coopèrent à la fois les modules de reconnaissance ascendante et ceux de haut-niveau. Un stimulus sonore est présenté en entrée du dispositif et celui-ci fournit en sortie le résultat de l'identification. Les données sonores sont produites par 6 locuteurs ayant prononcé les 500 mots du lexique, ce qui représente 3000 tests. Nous rappelons qu'il n'y a aucune phase d'apprentissage, ni d'adaptation. L'analyse des résultats consiste à repérer le classement du mot stimulus dans la liste des candidats proposés par le système. L'accès lexical laissant

apparaître plusieurs mots à égalité, il convient de repérer non seulement la position du mot testé mais aussi le nombre d'éléments ex aequo.

exemple: le décodage du mot "casserole" donne comme résultat

→	[mot]	[distance]
	essayerai	2
	casserole	2
	consomme	3
	consonne	3
	recevrons	5
	...	

le mot "casserole" est placé en première position ex aequo avec "essayerai".

Nous nous plaçons d'abord dans une évaluation tous mots et tous locuteurs confondus. Afin d'appréhender l'efficacité des différents modules de décodage ascendant, plusieurs évaluations sont effectuées pour lesquelles chaque module est tour à tour désactivé. Le superviseur et l'accès lexical sont eux toujours actifs.

Sur les Figures qui vont suivre sont indiqués les performances du système. En abscisse est représentée la position du mot testé dans la liste des candidats identifiés par le dispositif. En ordonné est mentionné le nombre de fois (ramené en %) où le mot testé a été identifié entre la 1ère et la position x du graphique. Autrement dit, la Figure 96 montre que dans la configuration testé du dispositif, 49.4% des mots testés étaient en première position, 69.6% des mots étaient entre la 1ère et la 20ème position, 80.7% étaient entre la 1ère et la 50ème .
Le symbole ——————┤
représente le phénomène de candidatures ex aequo. Plus le trait est allongé, plus grand est le nombre de mots placés à la même position.

VIII.4.D.b. L'évaluation de l'ensemble [segmentation + modules de haut-niveau]

.La première évaluation consiste à n'utiliser que l'information montante apportée par le module de segmentation et de macro-classification SAPHO.

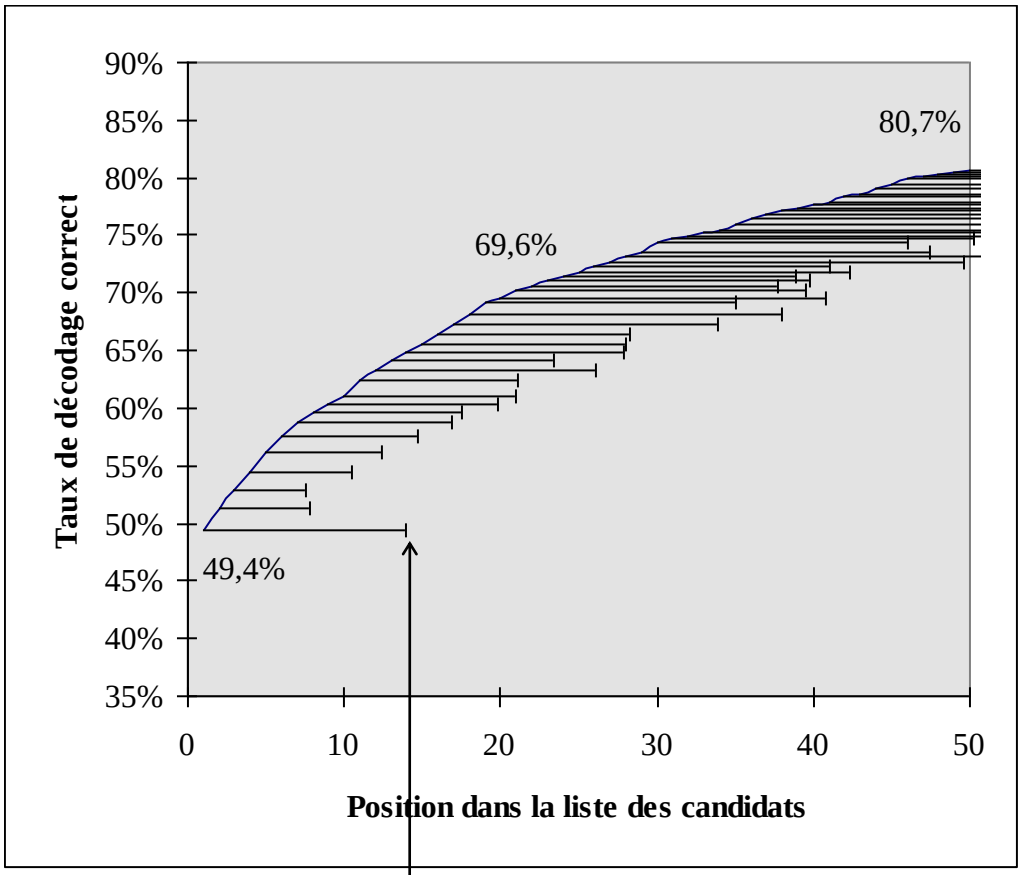


Figure 96: Résultats de la reconnaissance où seul le module de segmentation apporte une information montante

Les résultats de la Figure 96 montrent que dans 49.4% des cas, le mot testé est correctement identifié. En moyenne, une quinzaine d'autres mots sont placés ex aequo avec le mot-test. Dans 69.6 % des cas, le mot testé est parmi les 20 premiers candidats. Dans 80.7 % des cas, il se trouve parmi les 50 premiers prétendants. Ces résultats apparaissent encourageant pour un système indépendant du locuteur avec un lexique de 500 mots. Ils montrent aussi que l'identification correcte de la structure syntagmatique de la chaîne parlée avec quelques compléments paradigmatiques (macro-classes) permet déjà un bon décodage de mots isolés. Le nombre important de candidats ex aequo est compréhensible car de nombreux mots sont identiques en terme de macro-classes. Ainsi, "causant", "confus", "pêche", "penche", "peser", "pose", "pouce"... sont tous de la forme [occlusive sourde]+[voyelle]+[fricative]+[voyelle] et sont donc classés ex aequo si le décodage ascendant propose une telle chaîne phonétique. La discrimination pourrait être apportée par les modules de reconnaissance analytique et globale.

VIII.4.D.c. L'évaluation de l'ensemble [segmentation + reconnaissance analytique + modules de haut-niveau]

Nous présentons l'évaluation du dispositif où l'information montante est apportée par le module de segmentation SAPHO ainsi que par le module de reconnaissance analytique. Les résultats de la Figure 97 montrent que dans 39.8 % des cas, le mot testé est correctement identifié. Nous assistons à une dégradation des performances par rapport à l'architecture précédente qui n'incluait pas la reconnaissance analytique. Ce phénomène s'explique très bien par la réduction du nombre moyen de candidats, qui est divisé par 3. En fin de compte, le dispositif prend plus de risques en fournissant moins de candidats ex aequo grâce à l'information paradigmatique qu'il apporte. La conséquence est une baisse du taux de décodage mais une augmentation de l'efficacité globale. Cette hausse de performances se retrouve aussi dans le fait que dans 73.5 % des cas, le mot testé est parmi les 20 premiers candidats (vs 69.6 % avec la segmentation seulement, cf. Figure 96). De même, dans 84.4 % des cas, le mot testé se trouve parmi les 50 premiers prétendants (vs 80.7 % avec la segmentation seulement, cf. Figure 96).

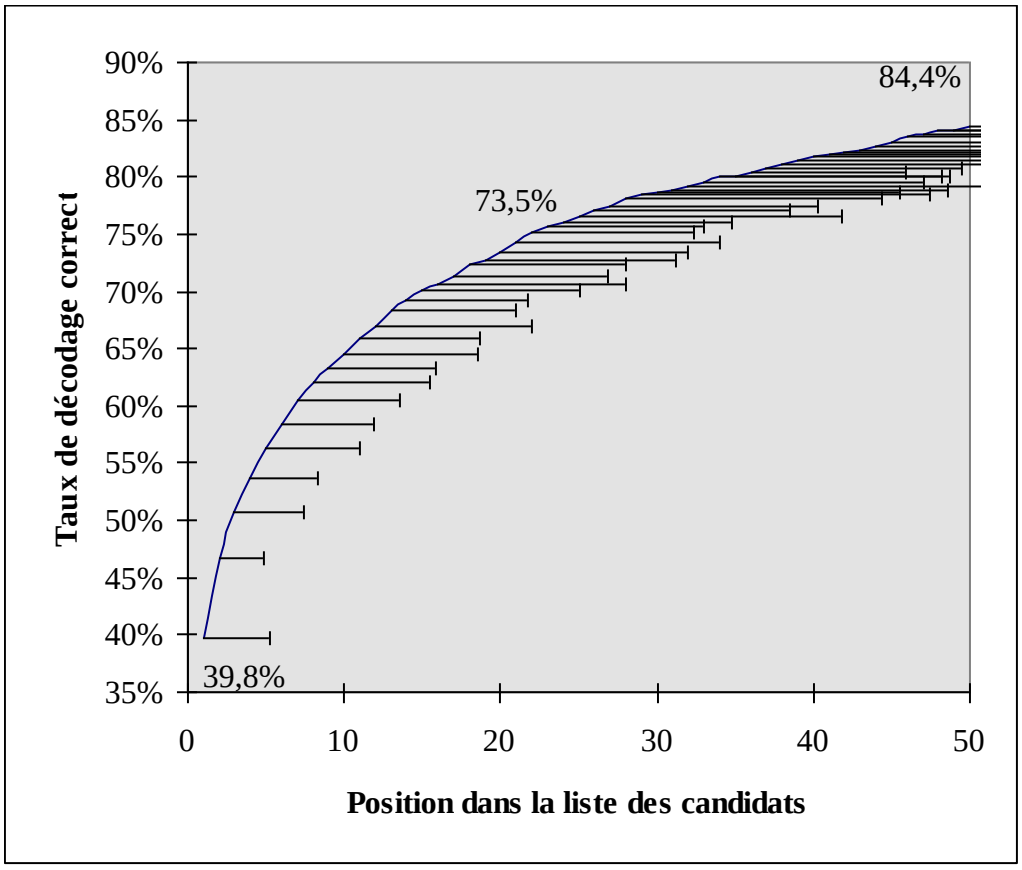


Figure 97: Résultats de la reconnaissance où seuls les modules de segmentation et de reconnaissance analytique apportent une information montante

VIII.4.D.d. L'évaluation de l'ensemble [segmentation + reconnaissance globale + modules de haut-niveau]

Nous présentons l'évaluation du dispositif où l'information montante est apportée par le module de segmentation SAPHO ainsi que par le module de reconnaissance globale. Les résultats de la Figure 98 montrent que dans 41.1 % des cas, le mot testé est correctement identifié. Nous assistons, là aussi, à une modification des performances par rapport à l'architecture où seul l'algorithme de segmentation apporte l'information montante. L'explication est la même que précédemment: le dispositif prend plus de risques en réduisant la cohorte de mots ex aequo. Le mot testé est plus souvent présent parmi les premiers candidats: dans 75.4 % des cas parmi les 20 premiers (vs 69.6%), dans 85.9 % des cas parmi les 50 premiers (vs 80.7 %). Par contre, il est plus rarement placé en 1ère position (41.1% vs 49.4%). Il faut aussi noter que l'information paradigmatique fournie par la reconnaissance globale est légérement supérieure à celle de la reconnaissance analytique (41.1% vs 39.8%). Cette tendance pourrait s'inverser si les améliorations potentielles de l'algorithme analytique étaient concrétisées.

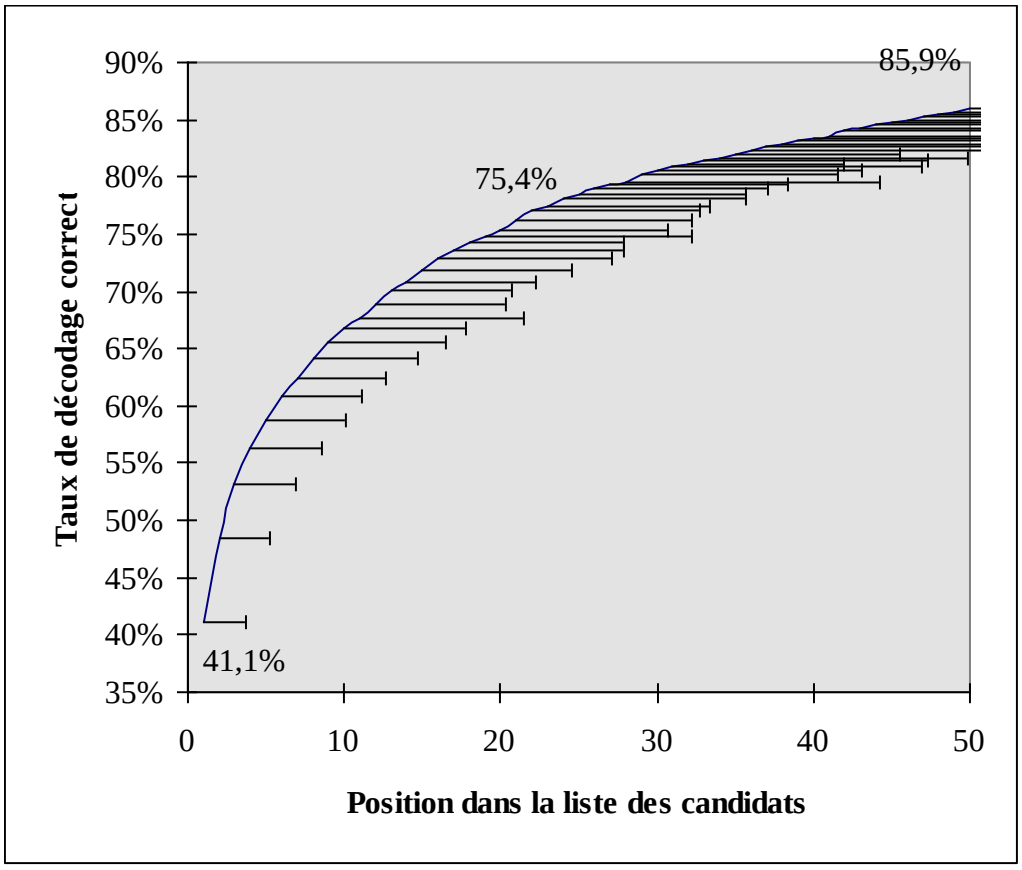


Figure 98: Résultats de la reconnaissance où seuls les modules de segmentation et de reconnaissance globale apportent une information montante

VIII.4.D.e. L'évaluation du dispositif complet

Nous présentons l'évaluation du dispositif où l'information montante est apportée par le module de segmentation SAPHO ainsi que par les modules de reconnaissance globale et analytique. Le dispositif est complet dans sa version actuelle. Les résultats de la Figure 99 montrent que dans 45 % des cas, le mot testé est correctement identifié. Dans 75 % des cas, le mot testé est parmi les 20 premiers candidats. Dans 86 % des cas, il se trouve parmi les 50 premiers prétendants. Nous assistons à une amélioration des performances par rapport aux architectures précédentes. En plus, le nombre de candidats est réduit, ce qui prouve la meilleure efficacité du dispositif grâce à l'apport de l'information paradigmatique fournie à la fois par la reconnaissance globale et analytique. Il faut noter que la combinaison des deux permet seulement une légère amélioration des performances, ce qui laisserait penser que les deux méthodes fournissent une information de même nature. Les perspectives pourraient être d'une part de faire porter la reconnaissance globale sur d'autres paramètres que les coefficients PLP, tirés du modèle auditif. Nous pensons, pourquoi pas, à des données articulatoires issues d'un processus d'inversion. Une deuxième solution serait d'approfondir la reconnaissance analytique en introduisant un aspect dynamique et en formalisant le décodage des lieux d'articulation des consonnes en fonction du contexte. Dans tous les cas, nous apprécions grandement l'aspect modulaire du dispositif pour mettre en place ces expérimentations.

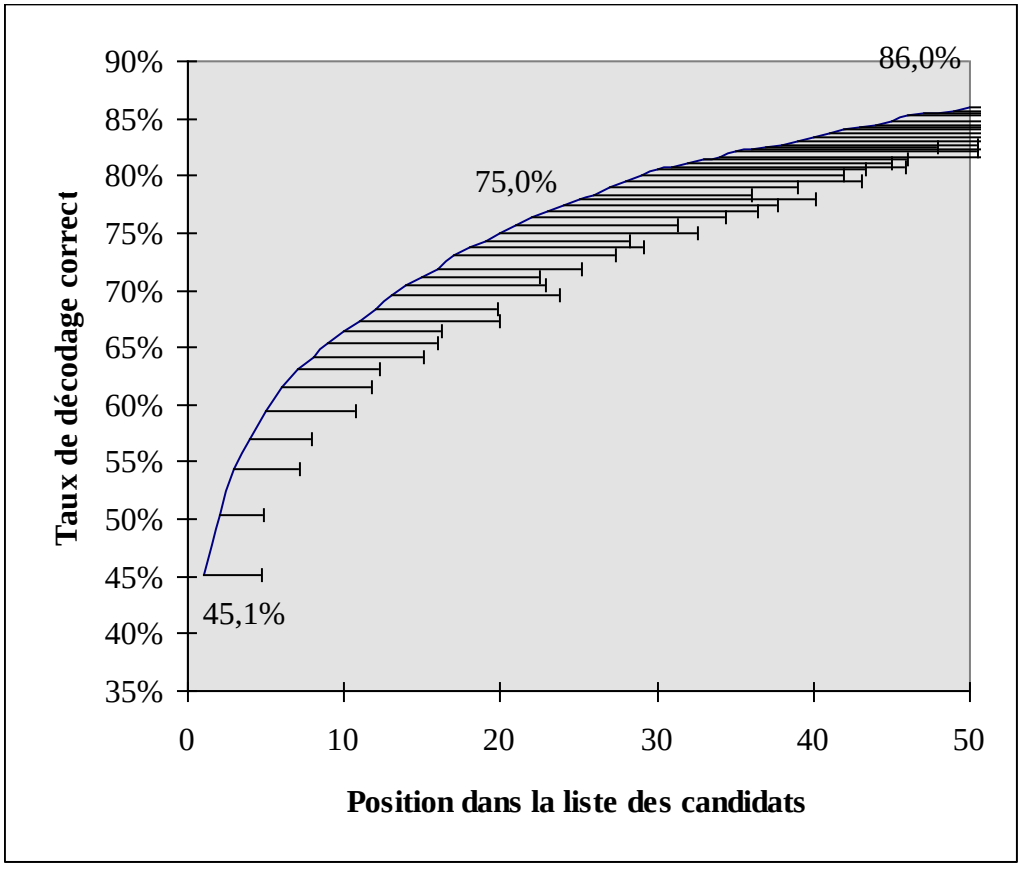


Figure 99: Résultats de la reconnaissance où tous les modules de décodage ascendant collaborent

VIII.4.D.f. L'effet « locuteurs »

Nous rappelons que les données sonores sont fournies par 6 locuteurs (5 hommes, 1 femme). Nous avons analysé les performances de décodage du dispositif complet en fonction de chaque individu (Figure 100). En prenant le classement des 500 mots testés comme paramètre et les locuteurs comme facteurs, nous avons effectué une analyse de la variance. Celle-ci laisse apparaître des différences peu significatives dans les résultats (p=0,018), ce qui laisse à penser que le dispositif est bien indépendant du locuteur. Les moins bonnes performances sont réalisées par le locuteur féminin (Figure 100, locuteur lc). La faiblesse d'effectifs des locuteurs féminins nous empêche d'analyser les différences entre sexes.

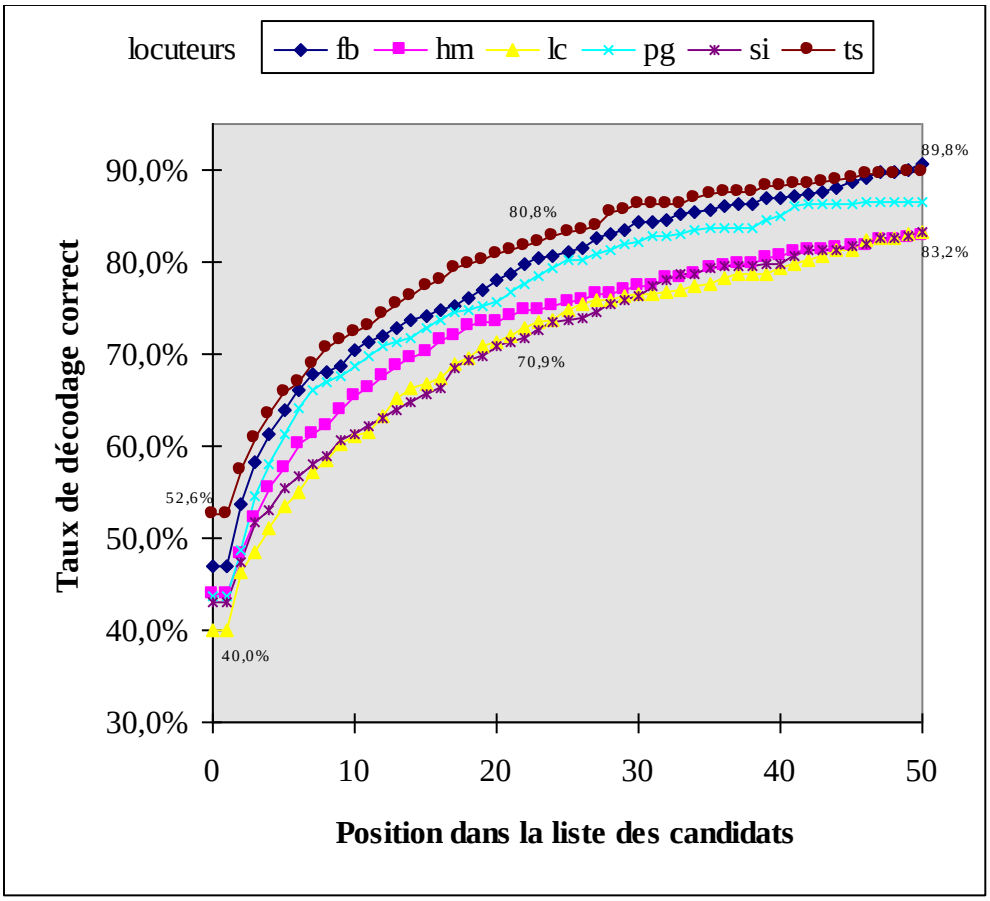


Figure 100: Résultats de la reconnaissance en fonction des locuteurs

VIII.4.D.g. L'effet « longueur de mots »

Il est généralement admis que les mots « longs » sont plus faciles à identifier que les « courts ». L'explication logique consiste à dire que la longueur contribue à la redondance de l'information et donc à la robustesse de l'identification. Nous avons vérifié cette hypothèse en calculant a priori la distance moyenne du plus proche voisin des éléments du lexique en fonction du nombre de phonèmes par mot. Pour cela, nous avons utilisé la méthode décrite dans le Tableau 34, page 212. Nous avons exclu les classes de longueurs de mots dont les

effectifs étaient insuffisants pour être statistiquement pertinents (au moins 5 éléments). Les résultats sont présentés à la Figure 101. Ils confirment l'hypothèse de départ: plus un mot contient de phonèmes, plus la distance avec son plus proche voisin a tendance à être importante. C'est en tous cas ce qui est visible dans notre lexique de 500 mots.

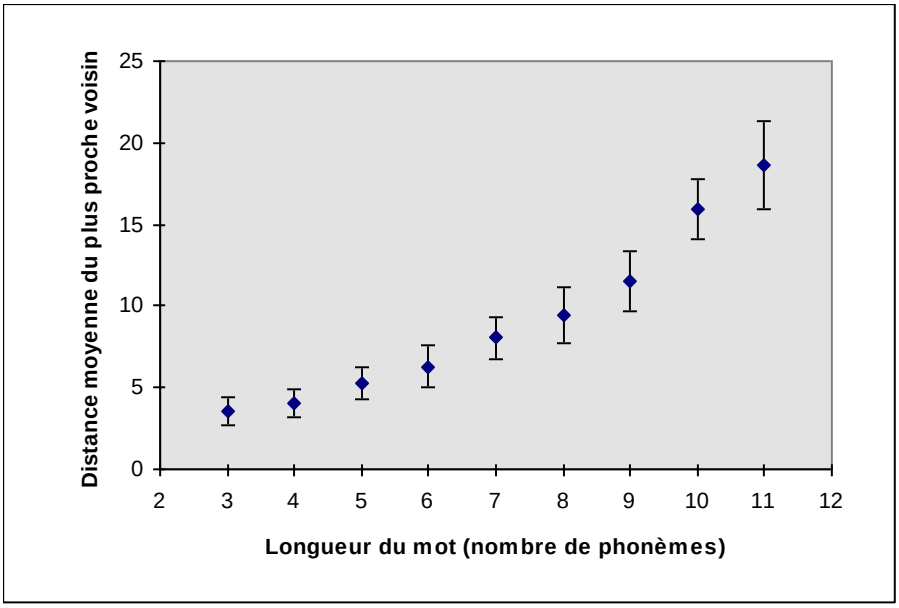
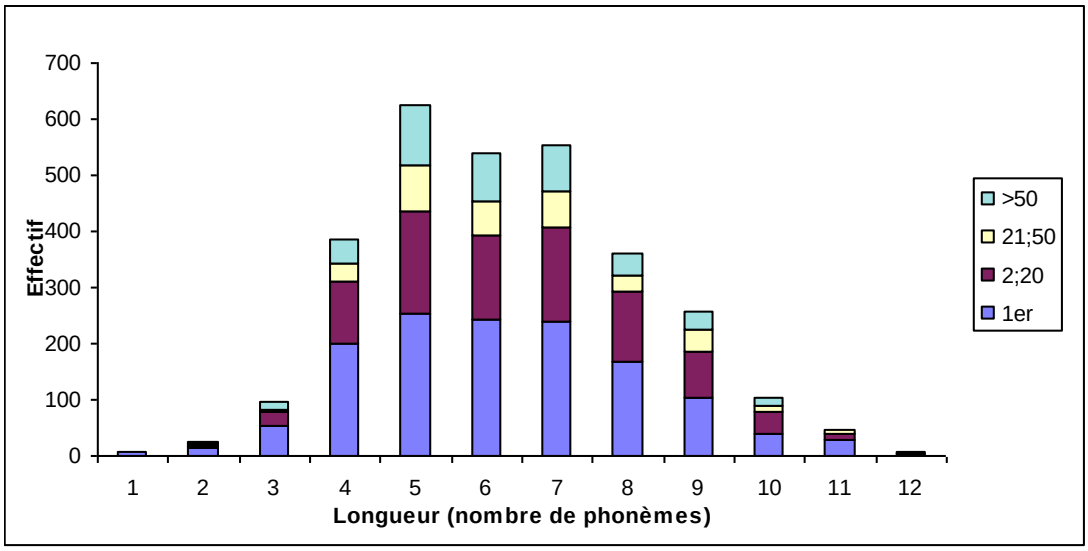


Figure 101: Distance moyenne du plus proche voisin d'un mot du lexique en fonction de sa longueur

Pour vérifier cette hypothèse en situation de reconnaissance, nous avons analysé les performances de décodage du dispositif complet en fonction de la taille des mots. En prenant le classement des 3000 mots testés comme paramètre et le nombre de phonèmes comme facteurs, nous avons effectué un calcul de moyenne et une analyse de la variance. Celle-ci laisse apparaître une valeur critique de 4 phonèmes en dessous desquels les résultats sont significativement meilleurs ($p=0{,}00019$). Cette tendance est visible sur la Figure 102b, qui regroupe les résultats par tranches pour des raisons de clarté de présentation.

Le fait que les mots de moins de 4 phonèmes soient plus facilement identifiés se trouve en contradiction avec notre hypothèse de départ. Pour tenter de donner une explication à ce phénomène, nous avons analysé les résultats du décodage pour chaque mot en fonction de la distance de son plus proche voisin. Aucune tendance nette ne se dessine, ce qui laisse supposer que les confusions ne se font pas nécessairement entre proches voisins. En fait, les mauvaises identifications proviennent plus d'erreurs grossières comme l'omission ou l'insertion d'un phonème, plutôt que de la substitution de deux unités acoustiquement proches. Or, dans ces cas là, les mots courts, qui sont essentiellement composés de séquences CVCV, sont faciles à identifier par rapport à des structures plus longues comportant des continuums vocaliques ou des groupes consonantiques variés.

(a)



(b)

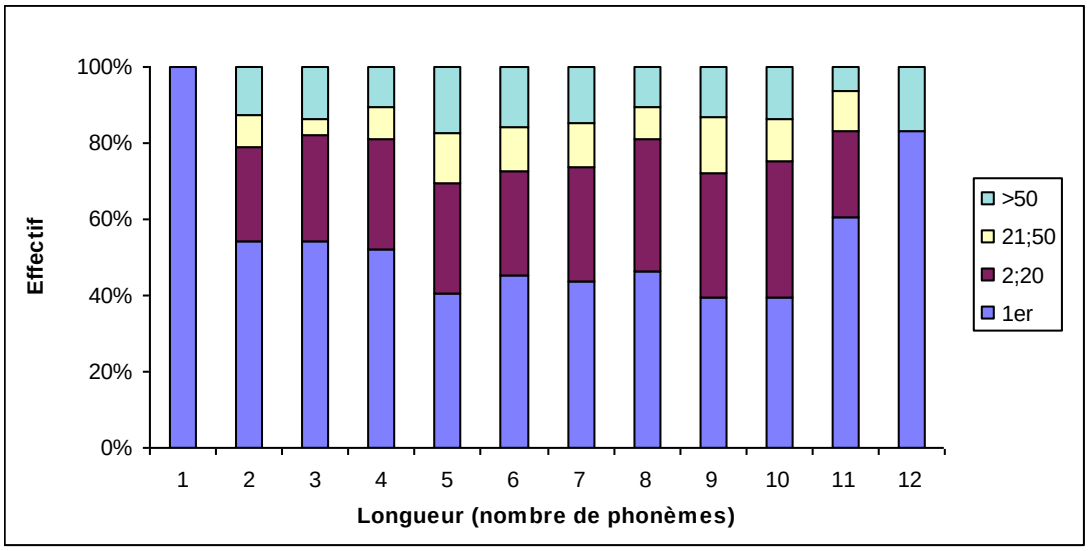


Figure 102: Résultats de la reconnaissance en fonction de la taille des mots. Résultats regroupés en catégories: classement en 1ère position, entre 2 et 20, entre 21 et 50, au dessus de 50. (a) effectifs bruts, (b) effectifs en pourcentage

VIII.4.D.h. L'apport d'une composante syntaxique

Nous savons que l'ajout d'un module d'analyse syntaxique apporterait des contraintes de haut niveau permettant d'améliorer les performances en réduisant la cohorte de mots candidats (cf. « L'interaction des sources d'information », p.46). A titre expérimental, nous avons testé le dispositif en simulant l'apport d'une composante syntaxique. Nous sommes partis de l'hypothèse théorique où un module d'analyse syntaxique fournirait dans 100 % des cas la catégorie grammaticale du mot testé. Les classes syntaxiques des mots du lexique sont extraites de la base de données BD-Lex (Pérennou & De Calmès, 1986). Les effectifs sont présentés au Tableau 38.

Tableau 38: Catégories grammaticales des mots du lexique

Codage	Catégorie grammaticale	Effectif dans le lexique
A	adverbe	5
c	conjonction	0
d	déterminant	2
F	adjectif ou nom féminin	0
G	adjectif ou nom (masculin ou féminin)	0
J	adjectif	55
i	interjection	0
M	adjectif ou nom masculin	0
N	nom	124
p	préposition	2
P	pronom	2
V	verbe	310
		500

De façon évidente, le lexique est mal équilibré en terme de classes grammaticales. Il n'a pas été conçu pour cela. Un fait notable est la surabondance de verbes qui sont dans la majorité des cas sous la forme de participe présent (ex: agréant, apportant, blanchissant...), d'indicatif présent à la 1ère personne du pluriel (ex: capitulerons, justifierons, ravissons...) ou de futur à la 1ère personne du singulier (ex: abandonnerai, déposerai, imposerai...). Cela limite un peu l'effet du filtrage syntaxique mais l'expérience reste malgré tout intéressante.

L'utilisation d'information syntaxique autorise une limitation de la cohorte de mots possibles, ce qui revient à réduire le lexique de façon dynamique. Les résultats de la reconnaissance avec l'ajout d'un tel procédé sont présentés à la Figure 103. Le gain de performances est notable d'une part en terme de meilleur décodage en première position (51.2 % vs 45.1 %) mais d'autre part sur le nombre fortement réduit de candidats ex aequo. Ce résultat n'est pas surprenant car le procédé de filtrage syntaxique apporte suffisamment de contraintes pour limiter les confusions acoustico-phonétiques. L'effet pourrait d'ailleurs être plus important si le lexique comportait moins de verbes. Par contre, il ne faut pas oublier que nous sommes partis de l'hypothèse théorique que le module d'analyse syntaxique fournissait dans 100 % des cas la bonne catégorie grammaticale du mot testé, ce qui reste évidemment faux en réalité. Les erreurs d'analyse syntaxique en situation réelle peuvent entraîner une mauvaise orientation de l'accès lexical. Tout cela reste à expérimenter avec des techniques de Traitement Automatique des Langues Naturelles sur de la parole continue. Le procédé pourrait d'ailleurs être analogue avec un module sémantique. Il faudrait à ce moment là repenser les composantes syntaxique et sémantique comme quelques choses de plus sophistiquées que de simples filtres lexicaux. Nous pensons entre autre à une approche descendante.

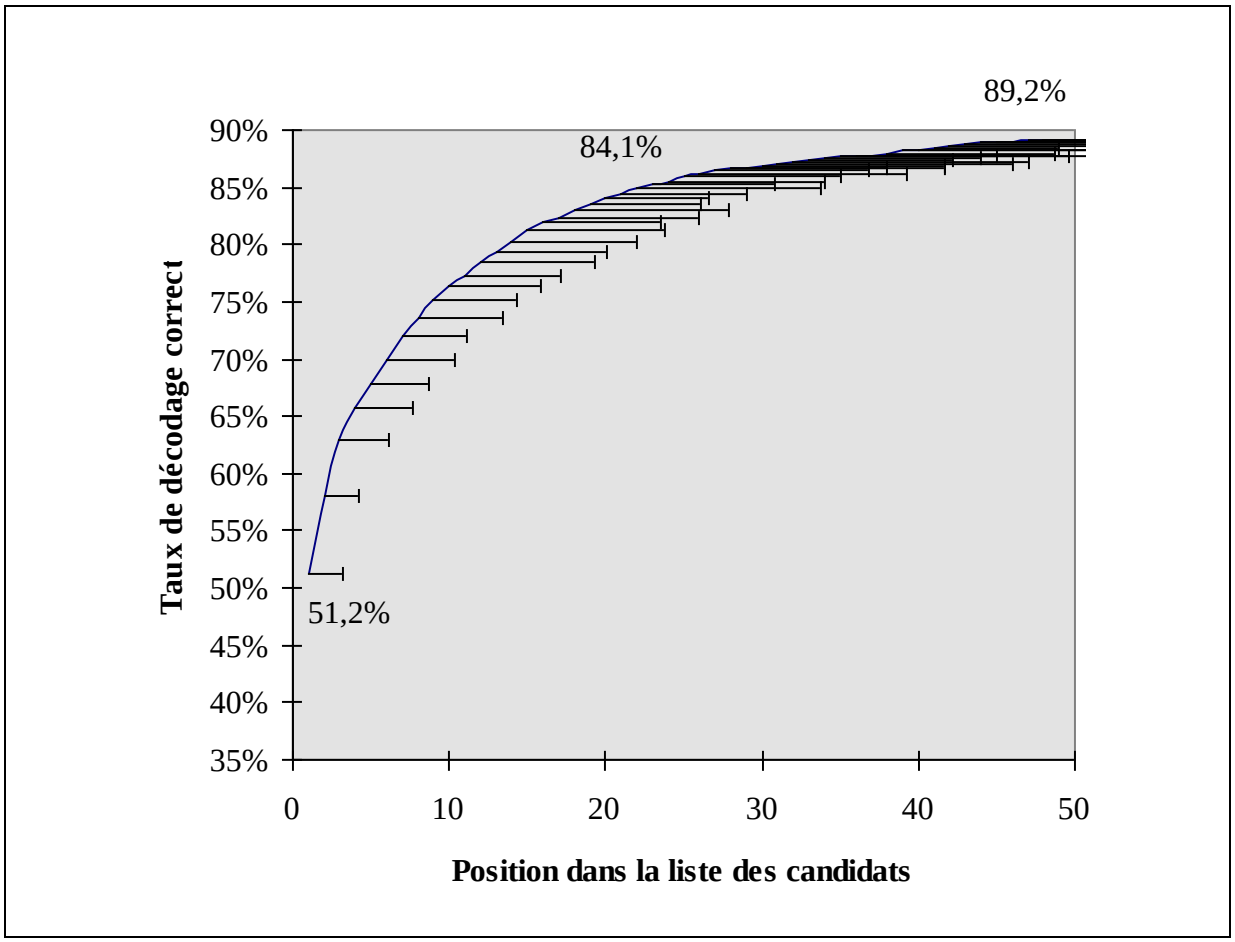


Figure 103: Résultats de la reconnaissance avec l'apport d'une composante syntaxique

Conclusion

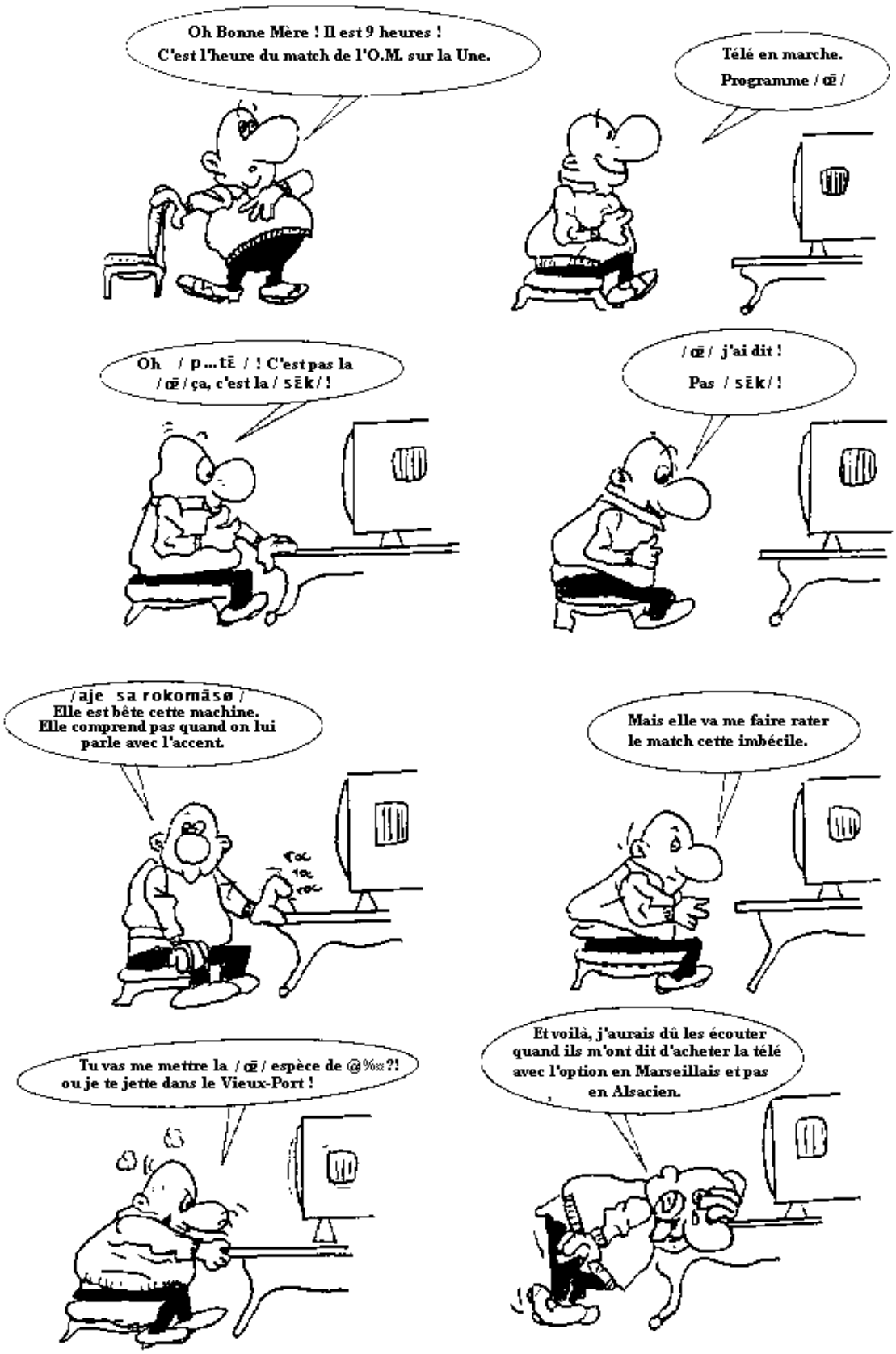
La reconnaissance automatique de la parole est un domaine qui est intéressant à plus d'un titre. Tout d'abord pour sa possibilité à donner naissance à une quantité d'applications diverses. Dans ce cadre-là, les contraintes économiques ont conduit la plupart des ingénieurs et chercheurs en reconnaissance automatique de la parole à utiliser massivement les méthodes stochastiques, celles-ci donnant à court terme de bien meilleures performances. Cependant, de plus en plus d'observations laissent penser que la RAP ne pourra pas fonctionner uniquement sur la base de méthodes stochastiques fondées sur les traditionnels Coefficients Cepstraux sur une échelle Mel (MFCC). Il faut d'une part envisager de nouvelles représentations du signal de parole. D'autre part, il serait souhaitable de penser à de nouvelles techniques de RAP capables d'intégrer simultanément diverses sources d'information, surtout pour être en mesure de prendre en considération la plupart des aspects de la communication parlée, phénomène très complexe en fin de compte.

Notre travail s'est inscrit dans un courant scientifique destiné à faire progresser la recherche fondamentale pour résoudre des problèmes de RAP apparemment insolubles dans un cadre purement stochastique. Notre objectif était d'examiner dans quelle mesure un modèle à base de connaissances serait capable de décoder de façon automatique la structure phonique de la parole sans recourir aux méthodes stochastiques. Pour cela, nous avons porté notre effort tout d'abord sur la recherche de paramètres acoustiques autres que les standards MFCC. L'étape d'extraction d'information nous semble essentielle car elle représente la base du décodage, l'opération qui extrait la pulpe de la substance. Nous avons le sentiment que cette opération est encore loin d'être parfaite en général, probablement du fait de nos manques de connaissances sur le fonctionnement de la communication parlée chez l'humain. Notre deuxième effort s'est porté sur la mise en place d'une architecture originale où fonctionnent en parallèle divers processus de décodage.

Le dispositif ACHILE est opérationnel. Il fonctionne de façon indépendante du locuteur sans apprentissage, ni adaptation. Il opère sur des mots isolés contenus dans un lexique de 500 mots. Cette taille n'est pas limitative. Les résultats sont moyens comparés à des méthodes stochastiques. Toutefois, la marge de manoeuvre reste encore importante et la plupart des éléments du dispositif peut être grandement améliorée. Le mérite d'un tel système repose sur la modélisation et donc la contrôlabilité des mécanismes. On peut donc espérer faire progresser leurs performances parallèlement aux connaissances. De plus, il peut servir d'outil en recherche fondamentale pour éprouver et comparer différentes théories. Il est clair que la robustesse de l'identification est dépendante de la fusion des informations multiples apportées par les différents processus de décodage acoustico-phonétiques et par ceux de haut-niveau. Nous regrettons de n'avoir pas pu approfondir ces notions. Nous laissons à d'autres le soin de mener à bien cette tâche longue et difficile.

Enfin, pour conclure par une touche d'humour, nous vous proposons une petite planche de bandes dessinées, dont les enseignements pourraient apporter peut-être plus que de longues lignes...

Librement inspiré de la Bande Dessinée « ACHILE Talon »



BIBLIOGRAPHIE

« Le plus grand des crimes, c'est de tuer la langue d'une nation avec tout ce qu'elle renferme d'espérance et de génie (). Avec une langue, on referait un monde. »

Charles Nodier

Abry C., Perrier P. (1995), "Le contrôle des mouvements audibles et visibles dans la parole", Actes de l'Ecole Thématique "Fondements et perspectives en traitement automatique de la parole", GDR-PRC Communication Homme-Machine, H. Méloni, Marseille-Luminy, 177-196.

Altmann G.T.M. (1990), "Cognitive Models of Speech Processing: An Introduction", Cognitive Models of Speech Processing - Psycholinguistic and computational perspectives, Bradford Book, MIT Press, G.T.M Altmann, 1-23.

André-Obrecht R. (1993), "Segmentation et parole ?", Mémoire d'habilitation à diriger des recherches, Dept. Informatique et Communication, Université de Rennes I.

Atal B.S., Hananer (1971), "Speech analysis and synthesis by Linear prediction of the speech wave", J.Acous.Soc.Am., 50(202),637-655.

Atal B.S., Schroeder M.R. (1978), "Linear prediction of speech based on a pole-zero representation", J.Acous.Soc.Am., 64, 5.

Barras C., Caraty M.J., Montacié C., Deléglise P., André-Obrecht R., Bimbot F., Le Floch J.L. (1994), "Approche événementielle pour la reconnaissance de la parole", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Baudry M. (1985), "Synthèse automatique de la parole", Techniques de l'Ingénieur.

Béchet F. (1994), "Système de traitement de connaissances phonétiques et lexicales: application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu", Thèse de l'Université d'Avignon et des Pays du Vaucluse.

Bellanger M. (1980), Traitement numérique du signal: théorie et pratique, Masson.

Bladon A. (1985), "An auditory phonetic model", Computer Speech Processing, Fallside & Wood.

Boë L.J., Schwartz J.L., Vallée N. (1994), "The prediction of vowel systems: perceptual contrast and stability", Fundamentals of speech synthesis and speech recognition, Keller E., University of Lausanne, Suisse, 185-213.

Boite R., Kunt M. (1987), Traitement de la parole, Presses polytechniques romandes.

Bonneau A., Rossi M., Grenie Y. (1985), "Hierarchical recognition of french vowels by expert system IROISE-SERAC", Symposium franco-suédois, Grenoble, France.

Boujot C., Boyer A., Fohr D., Haton J.P. (1990), "Méthodologies pour l'évaluation phonétique", Actes des 18emes Journées d'Etudes sur la Parole, Montréal.

Bourlard H (1996), "Reconnaissance automatique de la parole: modélisation ou description ?", Actes des XXIes Journées d'Etudes sur la Parole, Avignon, 263-272.

Brancaccio A., Ceglie F., D'Acunzo G., Pelaez C., Riccio A., Rigosi F. (1992), "A comparative study of the influence of parameter processing on two different approaches for speech recognition in adverse environment", Proceedings of 'Speech processing in adverse conditions', ESCA Conference, Cannes, 93-96.

Browman C.P., Goldstein L. (1992), "Articulatory Phonology: an overview", Haskins Laboratories Status Report on Speech Research, 111/112, 23-42.

Caelen J. (1979), "Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique.", Thèse d'Etat, Université Paul Sabatier, Toulouse.

Caelen J. (1995), "Architectures logicielles en reconnaissance: parallélisme et modularité", Actes de l'Ecole Thématique "Fondements et perspectives en traitement automatique de la parole", GDR-PRC Communication Homme-Machine, H. Méloni, Marseille-Luminy, 221-234.

Caelen J., Caillaud B., Antoine J.Y. (1994), "Projet MICRO: Modélisation Informatique de la Cognition en Reconnaissance de l'Oral", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Caelen J., Tattegrain H., Méloni H., Bulot R., Mercier G., Bonneau A. (1988), "Une base de règles pour le décodage acoustico-phonétique: le cas des occlusives sourdes", Actes du séminaire GRECO-PRC, Nancy.

Caelen J., Tattegrain H., Méloni H., Bulot R., Mercier G., Bonneau A. (1991), "Expertise dans le décodage acoustico-phonétique", Actes des 2èmes journées du GRECO PRC, Communication Homme-Machine, Toulouse.

CALLIOPE (1989), La parole et son traitement automatique, CNET Masson.

Carbonnel N., Haton J.P., Fohr D., Lonchamp F., Pierrel J.M. (1986), "APHODEX, design and implementation of an acoustic-phonetic decoding expert system", Proceedings of IEEE ICASSP, Tokyo, Japon.

Carré R., Degremont J.F, Gross M., Pierrel J.M., Sabah G. (1991), Langage humain et machine, Presses du CNRS, Paris.

Cervantès O., Sérignat J-F., Descout R., Carré R. (1986), "Définition et réalisation d'une base de données des sons du Français", Actes des 15èmes Journées d'Etudes sur la Parole, GALF, Aix-en-Provence, 213-216.

Chafcouloff M. (1994) "Nasalité et coarticulation", TIPA, 15, 101-121.

Cold B. (1966), "Word Recognition Computer Program", Technological Report, 452, Lincoln Lab., MIT, Cambridge.

Cosnier J. (1982), "Communications et langages gestuels", Les voies du langage, Dunod, 255-303.

Coste-Marquis S. (1994), "Décodage Acoustico-Phonétique à l'aide des techniques du raisonnement hypothétique: le système DAPHNE", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

D'Alessandro C., Beautemps D. (1991), "Transformation en ondelettes sur une échelle fréquentielle auditive", Actes du 13ème colloque GRETSI, Juan-les-Pins, 745-748.

Daubechies I. (1992), "Ten lectures on wavelets", Proceedings of Regional Conference Series in Applied Mathematics, CBMS-NSF.

Davis S.B., Mermelstein P. (1980), "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE ASSP, 28, 4.

De Coulon F. (1984), Théorie et traitement des signaux, Presses polytechniques romandes.

De Leeuw M., Caelen J. (1994), "Analyse de la parole par système expert mixte et modèle de cochlée", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

De Mori R. (1978), "Recent Advances in Automatic Speech Recognition", Proceedings of the 4th Int.Conf.PR, 106-124.

Delaire F. (1989), "Etude d'indices acoustico-phonétiques multilocuteurs en vue de la réalisation d'un système de reconnaissance automatique de la parole", Thèse de doctorat, Aix-Marseille III.

Dell F. (1985), Les règles et les sons, Hermann, Paris.

Di Cristo A., Haton J.P., Rossi M., Vaissiere J. (1982), "Prosodie et reconnaissance automatique de la parole", Séminaire GALF, Aix-en-Pce, France.

Dogil G., Braun G. (1988), "The pivot model of speech parsing", Verlag der Osterreichisschen Akademie der Wissenschaften, Wien, Autriche.

Dubois J., Guespin L., Giacomo M., Marcellesi C., Marcellesi J-B., Mével J-P. (1973), Dictionnaire de linguistique, Larousse, Paris.

Dubois J., Guespin L., Giacomo M., Marcellesi C., Marcellesi J-B., Mével J-P. (1994), Dictionnaire de linguistique et des sciences du langage, Larousse, Paris.

Dudley (1959), "The vocoder", Bell laboratories record, 122-126.

Duez D. (1987), "Contribution à l'étude de la structuration temporelle de la parole en français", Thèse de Doctorat d'Etat, Université de Provence, Aix-en-Provence.

Duley H., Balashek S. (1958), "Automatic Recognition of Phonetic Patterns in Speech", J.Acous.Soc.Am., 30, 721-732.

Edelmam G.M. (1992), Biologie de la conscience, O.Jacob, Paris.

El Méliani R., O'Shaughnessy D. (1996), "Gobe-tout en détection de mots nouveaux et en détection de mots-clés", Actes des XXIes Journées d'Etudes sur la Parole, Avignon, 301-304.

El-Bèze M. (1995), "Utilisation des modèles stochastiques de langage", Actes de l'Ecole Thématique "Fondements et perspectives en traitement automatique de la parole", GDR-PRC Communication Homme-Machine, H. Méloni, Marseille-Luminy, 129-138.

Eskénasi M., Liénard J-S. (1981), "Sur les notions de traits et d'indices pour les parties stables des sons du français émis par plusieurs locuteurs", Processus d'encodage et de décodage phonétiques, GALF, Toulouse, 54-69.

Espesser R. (1996), "MES: Un Environnement de Traitement du Signal", Actes des XXIes Journées d'Etudes sur la Parole, Avignon, 447.

Espesser R. (1981), "Un système de détection du voisement et de Fo", Travaux de l'Institut de Phonétique d'Aix, 8, Aix-en-Pce, France, 245-261.

Fant G. (1973), Speech sounds and features, MIT Press.

Fant G. (1975), "Speech perception and automatic recognition", Proceedings of Speech Communication seminar, Stockholm, Suède.

Flandrin P. (1987), "Représentation Temps-Fréquence des signaux non-stationnaires", Thèse d'Etat, INPG, Grenoble.

Fodor J.A. (1983), The modularity of mind, MIT Press, Cambridge.

Fohr D. (1986), "APHODEX: un système expert en décodage acoustico-phonétique de la parole continue", Thèse de doctorat, Nancy I.

Furui S. (1989), "Digital Speech Processing, synthesis and recognition", Dekker.

G.A.R.S. (1977-1993), Recherches sur le Français parlé, Publications de l'Université de Provence, Aix-Marseille I.

Genin (1976), "Les études de synthèse de la parole au CNET", L'echo des recherches, 40-49, CNET-ENST.

Gérard C., Baudry M. (1994), "Paramétrisation centiseconde du signal de parole en milieu bruité: comparaison ", Actes du séminaire 'Reconnaissance automatique de la parole', GDR-PRC Communication Homme-Machine, J.P. Haton, Nancy.

Ghio A. (1992), "Etude et réalisations d'outils adaptés à la reconnaissance automatique de la parole", DEA de phonétique expérimentale et fonctionnelle, Aix-Marseille I.

Ghio A., Rossi M. (1993), "SYMULDEPHO: un SYStème MULtilocuteur de DEcodage acoustico-PHOnétique", Travaux de l'Institut de Phonétique d'Aix, 15, Aix-en-Pce, France, 185-214.

Ghio A., Rossi M. (1994), "Reconnaissance globale et analytique dans SYMULDEPHO, un SYstème MULti-locuteurs de DEcodage acoustico-PHOnétique", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Ghio A., Rossi M. (1995a), "Reconnaissance Analytique par règles dans SYMULDEPHO, un système multi-locuteurs de décodage acoustico-phonétique", Travaux de l'Institut de Phonétique d'Aix, 16, Aix-en-Pce, France.

Ghio A., Rossi M. (1995b), "Parallel-distributed processes for speaker-independent Acoustic-Phonetic decoding", XIIIth International Congress of Phonetic Sciences, 4, Stockholm, Suède, 272-275.

Ghio A., Rossi M. (1995c), "A knowledge-based model for speaker-independent Acoustic-Phonetic decoding", 4th European Conference on speech communication and technology, EUROSPEECH, 1, Madrid, Espagne, 807-810.

Gilles P. (1993), "Décodage acoustico-phonétique de la parole et adaptation au locuteur", Thèse de doctorat, Université d'Avignon.

Gilles P., Méloni H. (1994), "Le décodage phonétique de la parole et ses applications", Actes du séminaire "Reconnaissance automatique de la parole, GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Giraud E. (1991), "Etude et réalisations de modules adaptés à la reconnaissance automatique de la parole", Thèse de doctorat, Aix-Marseille III.

Goldman-Eisler F. (1968), "Experiments in spontaneous speech", Psycholinguistics, Academic Press, New York.

Grosjean F., Deschamps A (1975), "Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de la parole et variables composantes, phénomènes d'hésitation", Phonetica, 31, 144-184.

Grossmann A., Holshneider M., Kronland-Martinet R., Morlet J. (1987), "Detection of abrupt changes in sound signals with the help of wavelets transforms", Rapport interne CNRS - RCP 820 "Ondelettes", Centre de Physique théorique, Marseille, 18 p..

Guaïtella I. (1991), "Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée", Thèse de Doctorat, Université de Provence, Aix-en-Provence.

Harris F.J. (1976), "Trigonometric Transforms - A unique introduction to the FFT", Technical publication DSP-005 6/76, Spectral Dynamics Corporation of San Diego.

Hassall J.R., Zaveri K., Phil M. (1979), Acoustic Noise Measurements, Bruel & Kjaer.

Haton J.P. (1971), "Reconnaissance de la parole: bilan de vingt années de recherche et tendances actuelles", Annales des Télécommunications, 77-88.

Haton J.P. (1974), "Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole", Thèse de doctorat, Nancy I.

Haton J.P. (1989), "Reconnaissance automatique de la parole et décodage acoustico-phonétique", Mélanges de phonétique générale et expérimentale, Hommage à Péla, 2, Publications de l'Institut de Phonétique de Strasbourg, 425-436.

Haton J.P., Lazrek M. (1984), "Segmentation et identification des phonèmes dans un système de reconnaissance de la parole continue", Actes du 4ème congrès AFCET-RFIA, Paris, 5-21.

Haton J.P., Pierrel J.M., Perennou G., Caelen J., Gauvain J.L. (1991), "Reconnaissance automatique de la parole", Dunod.

Hermansky H. (1987), "An efficient Speaker Independent Automatic Speech Recognition by simulation of some properties of human auditory perception", Proceedings of IEEE ICASSP, 1159-1162.

Hermansky H. (1990), "Perceptual linear predictive (PLP) analysis of speech", J.Acous.Soc.Am., 87, 4, 1738-1752.

Hermansky H., Hanson B., Wakita H. (1985), "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain", Speech Communication, 4, 181-187.

Hirst D.J., Espesser R., Nicolas P. (1995), "MES (Motif Editor for speech Signals)", Rapport MULTEXT, LRE Project 62-050, Deliverable 2.6.1., CNRS, Aix-en-Provence.

Jakobson R., Fant G., Halle M. (1951), "Preliminaries to speech analysis", MIT Press, Cambridge.

Jelinek F. (1976), "Continuous Speech Recognition using Statistical Methods.", IEEE Trans. ASSP, 64, 532-556.

Junqua J.C. (1990), "Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole", Revue de traitement du signal, 7, 4, numéro spécial, GRETSI, 275-284.

Kerbrat-Orecchioni C. (1990), "Les interactions verbales", Coll. La linguistique, Colin, Paris.

King J.H., Tunis C.J. (1966), "Some experiments in Spoken Word Recognition", IBM journal, 10, 1, 65-79.

Klatt D.H. (1977), "Review of the ARPA Speech Understanding Project", J.Acous.Soc.Am., 62, 1345-1366.

Kronland-Martinet R., Morlet J., Grossmann A. (1987), "Analysis of sound patterns through wavelet transform", Journal of Pattern Recognition and Artificial Intelligence - Special Issue on Expert Systems and Pattern Analysis, 1, 2, World Scientific Publishing Compagny, Society for Industrial and Applied Mathematics, 97-126.

Kunt M. (1980), Traitement numérique des signaux, Presses polytechniques romandes.

Landercy A., Renard R. (1977), Eléments de phonétique, Didier.

Laprie Y. (1990), "Le triplet phonétique en décodage acoustico-phonétique", Actes des 18èmes Journées d'Etudes sur la Parole, Montréal.

Lea W.A. (1980), Trends in Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, USA.

Le Beux P., and Fontaine D. (1986), "Un système d'acquisition des connaissances pour systèmes experts", Technique et Science Informatique, 5, 1, AFCET, 7-20.

Levinson S., Rabiner L., Sondhi M (1983), "An introduction to the applications of the theory of probabilistic function of a Markov Process to Automatic Speech Recognition", Bell Sys. Tech. Journal, 62, 1035-1074.

Liberman A.M. (1957), "Some results of research on speech perception", J.Acous.Soc.Am., 29, 117-123.

Liberman A.M., Cooper F.S., Shankweiler D., Studdert-Kennedy M. (1967), "Perception of the speech code", Psychological review, 74, 431-461.

Liberman A.M., Delattre P.C., Gertsman J.L., Cooper F.S. (1956), "Tempo of frequency change as a cue for distinguishing classes of speech sounds", Jour. Exp. Psychology, 52, 2, 127-137.

Liénard J.S. (1977), Les processus de la communication parlée, Masson.

Lindblom B., Studdent-Kennedy (1967) "On the role of formant transitions on vowel recognition", J.Acous.Soc.Am., 72, 4, 830-843.

Lindsay P.H., Norman D.A. (1977), "Human information processing", An introduction to psychology, Academic Press, New-York.

Mac Aulay R.J., Quatieri T.F. (1986), "Speech analysis/synthesis based on sinusoidal representation", IEEE Transactions on Acoustics, Speech and Signal Processing, 34, 4, 744-754.

Mac Clellan J.H. (1979), "FIR Filter Design and Synthesis", Programs for Digital Signal Processing, IEEE Press, DSP Committee, IEEE Acoustics, Speech, Signal Processing Society, New York, 5.0-1/5.4-20.

Mac Clelland J.L., Rumelhart D.E. (1988), Explorations in parallel distributed processing, MIT Press.

Mac Cord Nelson M., Illingworth W.T. (1991), A practical guide to Neural Networks, Texas Instruments - Addison Wesley.

Mc Gurk H., Mc Donald (1976) "Hearing Lips and seeing voices", Nature, 264, 746-748.

Makhoul J. (1975), "Linear prediction: a tutorial review", IEEE, 63.

Malbos F., André-Obrecht R., Baudry M. (1994), "Comparaison de deux méthodes non-paramétriques pour la détection des occlusives", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Malmberg B. (1974), "Manuel de phonétique générale - Introduction à l'analyse scientifique de l'expression des langues", Collection Connaissance des Langues, 9.

Malmberg B. (1979), "La Phonétique", Coll. Que-Sais-Je, 637, Presses Universitaires de France.

Mariani J. (1990), "Reconnaissance automatique de la parole: progrès et tendances", Revue de traitement du signal, 7, 4, numéro spécial, GRETSI.

Matrouf D., Gauvain J.-L (1996), "Techniques de compensation pour la reconnaissance de la parole bruitée", Actes des XXIes Journées d'Etudes sur la Parole, Avignon, Avignon, 331-334.

Meloni H. (1982), "Etude et réalisation d'un système de reconnaissance automatique de la parole continue", Thèse de doctorat, Aix-Marseille II.

Meloni H., Bechet F., Gilles P. (1992), "Reconnaissance analytique de mots isolés d'un grand lexique", Actes des 19es Journées d'Etudes sur la Parole, Bruxelles, 195-199.

Méloni H., Bulot R. (1986), "Un système de traitement de connaissances pour le décodage acoustico-phonétique", Proc. of Symposium on Speech Recognition, Montréal, 26-27.

Meunier C. (1994), "Les groupes de consonnes - Problématique de la segmentation et variabilité acoustique", Thèse de doctorat, Université d'Aix-Marseille I.

Miclet L. (1985), "Quantification vectorielle pour le traitement de la parole", Actes du séminaire GALF, E.N.S.T.

Millikan R.G. (1984), Language, thought and other biological categories: new foundations for realism, MIT Press, Cambridge.

Minsky M. (1985), The society of Mind, Simon & Schuster, New-York.

Minsky M.L., Papert S. (1969), Perceptrons, MIT Press, Cambridge.

Mizoguchi R., Tsujino K., Kakusho O. (1986), "A continuous speech recognition system based on knowledge engineering techniques.", Proc IEEE ICASSP, Tokyo, 1221-1224.

Montacié C., Caraty M.J., André-Obrecht R., Boë L.J., Deléglise P., El-Bèze M., Herlin I., Jourlin P., Lallouache T., Leroy B., Méloni H. (1995), "Applications Multimodales pour Interfaces et Bornes Evoluées (AMIBE)", Actes de l'Ecole Thématique "Fondements et perspectives en traitement automatique de la parole", GDR-PRC Communication Homme-Machine, H. Méloni, Marseille-Luminy, 155-164.

Montrésor S. (1991), "Etude de la transformée en ondelettes dans le cadre de la restauration d'enregistrements anciens et de la détermination de la fréquence fondamentale de la parole", Thèse de 3ème cycle, Le Mans.

Nishinuma Y., Kitazawa S., Shinmura T. (1993), "Réseau neuromimétique identifiant les traits distinctifs des voyelles du japonais", Travaux de l'Institut de Phonétique d'Aix, 15, 185-214.

O'Shaughnessy D. (1990), Speech communication, human and machine, Addison Wesley.

Paliwal K.K. (1983), "Effect of pre-emphasis on speech recognition performance", Actes du 11ème Congrès International d'Acoustique, 4, A, Paris, France.

Perennou G., De Calmes M. (1985), "Segmentation en événements phonétiques et en unités syllabiques", Actes des 14ème Journées d'Etudes sur la Parole, Paris.

Pérennou G., De Calmès M. (1986), "BDLEX: une base de données et de connaissances du français parlé", Actes du séminaire GRECO-GALF - Lexique et traitement automatique des langages, Toulouse.

Peterson G., Barney H. (1952), "Control methods used in a study of the vowels", Journal of the Acoustic Society of America, 24, 175-184.

Pierrel J.M. (1987), Dialogue oral homme-machine, Hermes.

Poirier F. (1990), "Reconnaissance multilocuteur de voyelles par un réseau connexionniste auto-organisateur", Actes des 18èmes Journées d'Etudes sur la Parole, Montréal, 196-200.

Pousse L., De Calmès M., Pérennou G., (1996), "Apports d'une composante phonologique à la reconnaissance automatique de la parole continue", Actes des XXIes Journées d'Etudes sur la Parole, Avignon, Avignon, 277-280.

Rabiner L.R. (1979), "Fast Convolution", Programs for Digital Signal Processing, IEEE Press, DSP Committee, IEEE Acoustics, Speech, Signal Processing Society, New York, 3.0-1/3.1-9.

Rabiner L.R. (1988), "Mathematical foundations of H.M.M.", Recent advances in speech understanding and dialog systems, NATO ASI Series, Niemmann, Lang & Sagerer.

Reddy D.R. (1966), "Segmentation of Speech Sounds", J.Acous.Soc.Am., 40, 307-312.

Rodellar V., Naharro F., Garcia C., Martin S., Munoz M.L., Gomez P. (1991), "A neural network for the extraction and characterization of the phonetic features of speech", Proceedings of International Conference on Neural Networks and their Applications, Nîmes, France.

Rossi M. (1972), "Le seuil différentiel de durée", Papers in linguistics and phonetics to the memory of Pierre Delattre, Albert Valdman, Indiana University, 435-450.

Rossi M. (1975), "Les contraintes phonologiques dans un système de reconnaissance automatique de la parole", Actes des Journées d'Etudes sur la Parole.

Rossi M. (1977), "Les traits acoustiques", La linguistique, 13, 1, 63-82.

Rossi M. (1980), "Le rôle de la phonologie dans la reconnaissance automatique de la parole", Revue d'Acoustique, 53, 102-105.

Rossi M. (1981), "De la physiologie à la perception phonémique", Modèles Linguistiques, 3, 2, 5-22.

Rossi M. (1985), "De la quiddité des variables", Actes du séminaire variabilité et spécificité du locuteur, Société Française d'Acoustique, Marseille-Luminy, 11-27.

Rossi M. (1990), "Segmentation automatique de la parole: pourquoi ? Quel segments ?", Revue de traitement du signal, 7, 4, numéro spécial, GRETSI, 315-326.

Rossi M. (1995), "Quelles connaissances utiliser pour la reconnaissance de la parole ?", Actes de l'Ecole thématique sur les "Fondements et persepectives en Traitement Automatique de la Parole".

Rossi M., Nishinuma Y., Grenie Y. (1983), "Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole", Speech communication, 2.

Sabah G. (1988), L'Intelligence Artificielle et le langage - Représentations des connaissances, 1, Hermes.

Saussure F. (1968), Cours de linguistique générale, 3e édition.

Schafer, Rabiner (1974), "Parametric representations of speech", Actes du symposium IEEE.

Schwartz J.L. (1995), "Perception de la parole: des représentations sensori-motrices à l'émergence des systèmes linguistiques", Actes de l'Ecole Thématique "Fondements et perspectives en traitement automatique de la parole", GDR-PRC Communication Homme-Machine, H. Méloni, Marseille-Luminy, 9-21.

Schwartz R., Chow Y., Dunham M., Kinball O., Krasner M., Kubala F., Makhoul J., Price P., Roucos S. (1988), "Acoustic phonetic decoding of speech", Recent advances in speech understanding and dialog systems, NATO ASI Series, Niemmann, Lang & Sagerer.

Spriet T. (1992), "Un superviseur intelligent pour la gestion des connaissances linguistiques en reconnaissance de la parole", Actes des 19èmes Journées d'Etudes sur la Parole, Bruxelles.

Stern P.E., Eskenazi M., Memmi D. (1986), "An expert system for speech spectrogram reading", Proc. ICASSP, 23.1.1-23.1.4.

Stevens S.S. (1957), "On the psychological law", Psychological review, 64.

Straka G. (1965), "Album phonétique", Québec.

Teston B. (1983), "Description d'un analyseur de sonie de la parole continue", Actes du Congrès International d'Acoustique, 4, Paris, 115-118.

Titze I.R., Liang H. (1994), Principles of voice production, Prentice Hall, Englewood Cliffs.

Traissac L. (1992), Réhabilitation de la voix et de la déglutition après chirurgie partielle ou totale du larynx, Sté Française d'ORL et de pathologie cervico-faciale, Arnette.

Tubach J.P. (1970), "Reconnaissance automatique de la parole", Thèse d'Etat, Université de Grenoble.

Tubach J.P., Chollet G., Choukri K., Montacie C., Mokbel C., Valbret H. (1990), "Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multicouches", Revue de traitement du signal, 7, 4, numéro spécial, GRETSI, 285-292.

Vallée N., Boé L.J. (1992), "Vers des prototypes acoustiques et articulatoire des 37 phonèmes vocaliques d'UPSID", Actes des 19èmes Journées d'Etudes sur la Parole, Bruxelles, 53-58.

Véronis J. (1994), "Distance entre chaînes: extension aux erreurs phono-graphiques", Travaux de l'Institut de Phonétique d'Aix, 15, 219-233.

Vintsyuk T.K. (1968), "Speech recognition by dynamic programming", Kybernetika, 4, 1.

Waibel A., Hanazawa T., Hinton G.H., Shikano K., Lang K. (1987), "Phoneme Recognition using Time-Delay Neural Networks", TR-I-0006, ATR Interpreting Telephony Research Laboratories.

Walker D.E. (1978), Understanding Spoken Language, North Holland.

Wesfreid E., Wickerhauser M.V. (1994), "Traitement de la parole par ondelettes de Malvar", Actes du séminaire "Reconnaissance Automatique de la Parole", GDR-PRC Communication Homme-Machine, CRIN/INRIA, Nancy.

Yong G., Mason J.S. (1987), "A comparison beetween vocal tract and auditory feature analysis in ASR", Proceedings of Eurospeech, Edinburgh, 132-135.

Zue V.W. (1982), "Acoustic Phonetic Knowledge representation: implications from spectrogram reading experiments", Automatic Speech Analysis and recognition, Haton J.P., D.Reidel, Dordrecht, Holland.

Zue V.W., Cole R.A. (1979), "Experiments on Spectrogram reading", Proc. ICASSP, 116-119.

Zwicker E., Therhart E. (1980), "Analytical expression for critical bands rate and critical bandwidth as a function of frequency", J.Acous.Soc.Am., 68, 1523-1525.

Zwicker E., Feldkeller R. (1981), Psychoacoustique de l'oreille, MASSON.