Institut National
Polytechnique de Lorraine
Inria Lorraine

Centre de Recherche en
Informatique de Nancy

Outils, travaux et propositions pour le décodage acoustico-phonétique

THÈSE

présentée et soutenue publiquement le **23 mars 1990**

pour l'obtention du

Doctorat de l'Institut National Polytechnique de Lorraine

(Spécialité Informatique)

par

Yves Laprie

Composition du jury :

Président : Jean-Paul HATON

Rapporteurs : René CARRÉ
Pierre LESCANNE

François LONCHAMP
Jacqueline VAISSIÈRE
Jean-Marie PIERREL

Institut National
Polytechnique de Lorraine
Inria Lorraine

Centre de Recherche en
Informatique de Nancy

Outils, travaux et propositions pour le décodage acoustico-phonétique

THÈSE

présentée et soutenue publiquement le **23 mars 1990**

pour l'obtention du

Doctorat de l'Institut National Polytechnique de Lorraine

(Spécialité Informatique)

par

Yves Laprie

Composition du jury :

Président :	Jean-Paul HATON
Rapporteurs :	René CARRÉ
	Pierre LESCANNE
Examinateurs :	François LONCHAMP
	Jacqueline VAISSIÈRE
	Jean-Marie PIERREL

Remerciements

Cette thèse est pour moi l'occasion d'exprimer des remerciements à tous ceux qui m'ont conseillé et aidé à réaliser ce travail dans de bonnes conditions, et je suis heureux aujourd'hui de remercier plus particulièrement :

Monsieur Jean-Marie Pierrel, Professeur à l'Université de Nancy 1, qui m'a proposé un sujet de recherche passionnant, et m'a encadré et orienté tout au long de ces années de travail. Je lui suis particulièrement reconnaissant pour ses conseils, ses encouragements et sa confiance.

Monsieur Jean-Paul Haton, Professeur à l'Université de Nancy I et Directeur de Recherches INRIA au CRIN, qui m'a accueilli dans l'équipe Reconnaissance des Formes et Intelligence Artificielle du CRIN, et qui me fait l'honneur de présider ce jury. Je le remercie de sa confiance et du grand intérêt qu'il a toujours porté à mes travaux.

Monsieur René Carré, Directeur de recherches à l'Institut de la Communication Parlée à Grenoble et travaillant à l'heure actuelle à l'École Nationale Supérieure des Télécommunications à Paris et Monsieur Pierre Lescanne, Directeur de recherches au CRIN, qui ont accepté d'être rapporteurs de cette thèse et de siéger à ce jury. Qu'ils trouvent ici l'expression de ma gratitude pour l'intérêt dont ils ont fait preuve à l'égard de ce travail.

Monsieur François Lonchamp, Professeur à l'Université de Nancy II, qui m'a formé dans le domaine de la phonétique et dont les précieux conseils ont souvent permis d'orienter mes recherches, qu'il soit remercié aussi pour la grande disponibilité dont il a toujours fait preuve à mon égard.

Madame Jacqueline Vaissière, Professeur associé à l'Université d'Aix-en-Provence, qui m'a fait partager son enthousiasme pour la recherche en parole, m'a encouragé et m'a généreusement fait part de ses connaissances.

J'exprime des remerciements chaleureux à Shinji Maeda pour ses conseils d'acous -ticien, à toute l'équipe de parole du CRIN-INRIA Lorraine, notamment à Noëlle

Carbonell pour les nombreuses et fructueuses discussions que nous avons eues ensemble et de judicieux conseils de rédaction, à Anne Bonneau, à Jean-Claude Junqua, Dominique Fohr, Yfan Gong, François Charpillet pour leur savoir-faire.

Merci également à tous mes collègues et amis du CRIN, pour l'atmosphère de travail sympathique dont j'ai bénéficié, tout particulièrement à Agnes Valdenaire, Laurent Weihnard pour l'environnement technique, à mes camarades thésards Pierre Nugues, Quan Long, Marie-Odile Berger, Claude Inglebert, Jean-Jacques Bonin, Stéphane Paris, Bernard Waldmann.

Enfin, je tiens à remercier vivement toute ma famille pour ses encouragements et le soutien moral qu'ils m'ont prodigués, sans lesquels je n'aurais pas pu terminer ce travail.

Résumé

Face aux techniques globales dont il est désormais difficile d'améliorer sensiblement les performances, l'approche experte en décodage acoustico-phonétique de la parole continue multilocuteur semble très prometteuse. Ce travail présente plusieurs aspects de cette approche.

Le premier aspect abordé est celui de l'environnement logiciel offert au chercheur en parole. Il doit être à la fois convivial et puissant et c'est dans cet esprit qu'a été développé le logiciel Snorri qui fournit à la fois les outils classiques en traitement de la parole (acquisition, restitution de signaux temporels, calcul de spectrogrammes, étiquetage manuel...) mais aussi un certain nombre de fonctions qui aident l'utilisateur à compléter et à valider son expertise (exploration automatique de corpus, suivi de formants dans le plan F1-F2...).

Pour pouvoir exploiter l'expertise phonétique il est indispensable de disposer de bons détecteurs d'événements acoustiques. Le suivi de formants est sans doute le plus important d'entre eux. Après un panorama des différents problèmes liés à l'extraction des formants deux algorithmes sont proposés :

- le premier pour construire les pistes formantiques en utilisant des techniques classiques en traitement d'images (notamment la morphologie mathématique),
- le second destiné à affiner les résultats obtenus avec l'algorithme précédent en interprétant les pistes formantiques en termes de formants ; le principe de cet algorithme est de maximiser l'énergie du spectrogramme expliquée par les formants.

Les différentes approches du décodage acoustico-phonétique (DAP) sont ensuite replacées par rapport aux connaissances actuelles des phénomènes de la parole continue. À partir des faiblesses des systèmes experts en lecture de spectrogrammes une approche originale est proposée en DAP : *le triplet phonétique*, c'est-à-dire un prototype en termes d'événements et d'indices acoustiques d'un son en contexte. Les perspectives qu'ouvre l'utilisation de ce grain de connaissance sont évoquées avant de décrire un système de DAP à base de triplets en cours de développement. Ce système opère en deux étapes. Dans un premier temps on effectue un décodage local, triplet par triplet, en recherchant les meilleurs candidats sur la base des indices et des événemnets acoustiques détectés dans le signal de parole. Ensuite on accroît la consistance de la solution globale sur la phrase en utilisant des techniques de relaxation discrètes faisant appel aux contraintes acoustiques qui existent entre deux triplets.

Table des matières

Chapitre 1

Introduction

La parole est sans doute l'expression la plus manifeste de l'intelligence humaine. La toute jeune intelligence artificielle a donc relevé le défi : permettre à l'ordinateur de dialoguer avec l'homme grâce à la parole. Il serait d'ailleurs séduisant de proposer la variante suivante du fameux test de Turing [Turing 50] : un homme et un ordinateur sont dans une pièce ; un troisième sujet, l'interrogateur, est dans une autre pièce et s'entretient par la parole avec les premiers en essayant de démasquer l'ordinateur. Imaginer qu'il soit possible de concevoir une machine capable de copier à ce point l'être humain est très troublant philosophiquement. En fait si l'homme est déja capable, dans une certaine mesure, de copier la nature humaine c'est plutôt dans le domaine des sciences biologiques. Car il faut bien reconnaître que comprendre la parole (qui ne soit contrainte d'aucune manière) est tout à fait hors de portée d'une machine. Il n'y a pourtant pas si longtemps, le but a semblé un instant tout proche. Delattre n'écrivait-il pas en substance [Delattre 48] que le spectrographe [1] devait permettre aux sourds de "lire" une conversation. Ces espoirs se sont vite estompés et jamais plus depuis, sauf peut-être dans le cadre d'applications très contraintes, le but n'a paru aussi lointain.

Aujourd'hui, douce revanche de la nature humaine, la voie qui conduit aux meilleurs résultats est celle qui consiste à modéliser la parole par des "sources de Markov"... mais il s'agit tout au plus de reconnaître (et non pas de comprendre) les phrases produites à partir d'un vocabulaire limité à 1000 mots avec de sévères contraintes syntaxiques. Serait-ce à ce prix seulement qu'il est possible de reconnaître la parole continue multilocuteur[2]?

Delattre et d'autres aussi [Potter 47], ont fait émerger l'idée qu'il est sans doute possible de décoder la chaîne de sons à partir des spectrogrammes. Mais cette idée repose sur l'hypothèse que nos connaissances de l'acoustique et de l'articulatoire (et de beaucoup d'autres domaines comme la perception) sont suffisantes pour aborder ce problème. S'inspirant des phonéticiens qui parviennent à décoder assez facilement

[1] Le spectrographe est un appareil qui fournit une représentation graphique temps x fréquence x énergie de la parole, appelée spectrogramme.

[2] Parole continue multilocuteur signifie qu'il n'y a aucune contrainte ni sur la manière de parler (donc notamment sans faire de pose entre les mots) ni sur les locuteurs.

les spectrogrammes, les chercheurs en intelligence artificielle ont conçu un certain nombre de systèmes experts destinés à lire les spectrogrammes. Si ces tentatives ont prouvé la pertinence de la connaissance des phonéticiens, leurs performances sont néanmoins limitées et par les erreurs de détection d'événements acoustiques et par le formalisme adopté pour stocker la connaissance de l'expert. Malgré tout, face aux méthodes globales (notamment les modèles markoviens) dont les résultats ne s'améliorent désormais qu'en déployant des trésors d'astuces, l'approche experte représente à notre sens la voie la plus prometteuse. C'est dans ce cadre que s'inscrit, bien modestement, notre thèse.

Comme le lecteur le découvrira, cette approche conduit à aborder des sujets assez éloignés les uns des autres, mais les efforts à consentir sont nombreux pour que l'approche experte porte ses fruits. C'est d'abord dans le domaine des outils offerts au chercheur en parole que s'est dirigée notre attention. Les recherches en parole ont souvent porté sur des points très précis et il est nécessaire de les replacer dans un contexte plus général. C'est justement l'objet du logiciel Snorri que nous décrivons au cours du premier chapitre. Il doit permettre à l'utilisateur, en visualisant la parole sous toutes ses formes, de mettre au point l'expertise qu'il compte utiliser pour le décodage. Mais c'est aussi un bon outil pédagogique, entendez par là qu'il conduit l'expert à exposer sur des exemples quelles sont ses connaissances et comment il en tire parti. Il y a aussi la volonté d'offrir au chercheur un outil puissant et convivial lui permettant d'augmenter "d'un ordre de grandeur" (pour reprendre les termes utilisés par les concepteurs de Spire [Zue 86a]) sa puissance d'investigation des phénomènes de la parole.

Vient ensuite un long chapitre consacré aux formants (les fréquences de résonance du conduit vocal) et aux algorithmes de suivi de formants. Dans un premier temps nous montrons l'importance des formants en décodage acoustico-phonétique (DAP) qui sont une information capitale pour connaître les sons prononcés à partir du signal de parole. Nous décrivons ensuite plusieurs algorithmes de suivi de formants avant de proposer un nouvel algorithme faisant largement appel aux techniques de traitement d'image. Comme les résultats d'un suivi de formants sont plus souvent des lignes de pics de spectres que de véritables formants nous décrivons ensuite un algorithme qui à partir des lignes de pics fournit le meilleur suivi possible en termes d'énergie expliquée par les formants.

Après avoir montré comment il était possible d'extraire automatiquement l'un des indices acoustiques les plus importants, en l'occurrence les formants, nous abordons dans le chapitre suivant le décodage acoustico-phonétique proprement dit. Nous commençons par passer en revue quelques uns des phénomènes de la parole en montrant quels sont les points bien maîtrisés, sur lesquels peuvent donc s'appuyer les systèmes de décodage acoustico-phonétique, et ceux qui relèvent encore du domaine de la recherche. Nous présentons ensuite les approches actuelles du DAP les plus marquantes, c'est-à-dire les systèmes à base de sources de Markov, les modèles connexionnistes et les systèmes experts.

En considérant les faiblesses d'un système expert en DAP (Aphodex [Fohr 86]) constatées tant par notre expert phonéticien F. Lonchamp [3], que par D. Fohr, auteur d'Aphodex, nous proposons dans le dernier chapitre une nouvelle approche experte de DAP : les triplets phonétiques. À partir d'une idée lancée initialement par F. Lonchamp, nous proposons une approche unifiée de la compréhension de la parole s'inspirant du raisonnement hypothétique et exploitant au maximum les contraintes reliant les différents niveaux d'abstraction de la parole. Nous décrivons enfin le système de DAP à base de triplets sur lequel nous travaillons à l'heure actuelle et nous montrons comment il est possible de construire une solution globalement cohérente sur la phrase en exploitant les contraintes acoustiques entre triplets de la solution.

[3] Professeur à l'Institut de Phonétique de Nancy

Chapitre 2

Snorri : un système d'étude interactif de la parole

1 Introduction

Les progrès en décodage automatique de la parole dépendent étroitement des moyens d'analyse et d'observation mis à la disposition des chercheurs. Dans le cadre de l'approche experte du décodage acoustico-phonétique il faut disposer des outils classiques pour étudier la parole (acquisition, restitution, calcul et affichage de spectrogrammes, étiquetage manuel de la parole) mais aussi d'outils aidant à l'élaboration de l'expertise et à la mise au point des algorithmes de décodage.

L'arrivée des moyens informatiques puissants et conviviaux (la machine Masscomp 5600 en ce qui nous concerne) nous a donné l'opportunité de développer un logiciel -Snorri- répondant aux caractéristiques que nous venons d'évoquer.

Nous décrirons d'abord les outils de base de Snorri (voir aussi [Laprie 88] et [Fohr 89]). Nous accompagnerons ensuite la description des outils spécialement conçus pour Snorri d'exemples mettant en évidence l'intérêt qu'ils représentent.

Nous terminerons par une discussion portant sur la construction de Snorri, sa place parmi les autres logiciels de parole et les perspectives qu'il ouvre.

2 Configuration matérielle et logicielle

Snorri fonctionne sur un poste de travail Masscomp 5600. Il utilise une carte d'acquisition 12 ou 16 bits. Tous les calculs de traitement du signal sont effectués sur un processeur vectoriel VA-1 qui multiplie par un facteur supérieur à 10 la vitesse d'exécution. Les fichiers de parole sont stockés sur un disque de 400 Mo.

L'affichage a lieu sur un écran graphique couleur (4 ou 12 plans, 1152x910 pixels). Les copies d'écran sont faites sur une imprimante Postscript avec une résolution de 400 points par pouce.

Snorri est écrit en C et fait appel au multifenêtrage et aux primitives graphiques

de Masscomp. Toutes les primitives de traitement du signal sont programmées avec les commandes du processeur vectoriel mais une version sans appel au processeur vectoriel (sauf pour le calcul des cepstres) est aussi disponible.

3 Organisation de Snorri

Snorri est un système dont toutes les fonctions sont accessibles avec la souris. Ces fonctions sont regroupées par grandes classes. Chaque classe est liée soit à une représentation du signal de parole (signal temporel, spectrogramme), soit à une des utilisations possibles de Snorri (étiquetage de la parole, copies de fenêtres et multifenêtrage ...).

Il faut distinguer deux types d'utilisation de Snorri :

1. études d'élocutions isolées,
2. étude d'un corpus déja enregistré et commun à plusieurs chercheurs ; l'utilisateur dispose alors d'outils d'exploration du corpus.

Nous n'avons pas limité la taille des fichiers de parole contrairement à certains environnements de traitement de parole comme par exemple le projet SPAR [ea 87]. Snorri étant développé depuis le début à l'intérieur de notre équipe n'a pas eu à intégrer des modules d'origines diverses imposant des contraintes de longueur sur les fichiers manipulés. Néanmoins nous avons défini toute une classe de fonctions qui permettent au programmeur de s'affranchir de la gestion des déplacements dans le signal de parole.

Snorri exploitant largement les ressources du multifenêtrage associe une fenêtre au signal temporel, au spectrogramme, mais aussi à certaines fonctions (zoom, suivi de formants ...). L'utilisateur peut facilement modifier l'environnement dans lequel il évolue en changeant la disposition des fenêtres, en en créant de nouvelles, ou en sélectionnant une autre palette de couleurs.

Snorri dispose aussi d'une fonction de copie de fenêtres qui permet de conserver une trace écrite de toute fenêtre existante. L'utilisateur peut règler de nombreux paramètres (copie noir et blanc, en niveaux de gris, dimensions de l'image, titre, modification du contraste, affectation d'un niveau de gris à une couleur quelconque). Les images produites sont au format Postscript. Toutes les figures de ce chapitre ont été obtenues avec cette fonction.

4 Outils de base en traitement de la parole

4.1 Signal temporel

La figure 2.1 montre le signal temporel tel qu'il est acquis par Snorri [1]. La fréquence normale d'échantillonnage est 16 kHz, pour l'acquisition comme pour la restitution, mais il est possible de l'augmenter jusqu'à 330 kHz. Chaque acquisition est automatiquement sauvegardée. L'utilisateur peut redéfinir les début et fin du signal à étudier et mettre cette séquence après lui avoir donné un nom.

[1] L'unité de temps dans toutes les figures est la milliseconde.

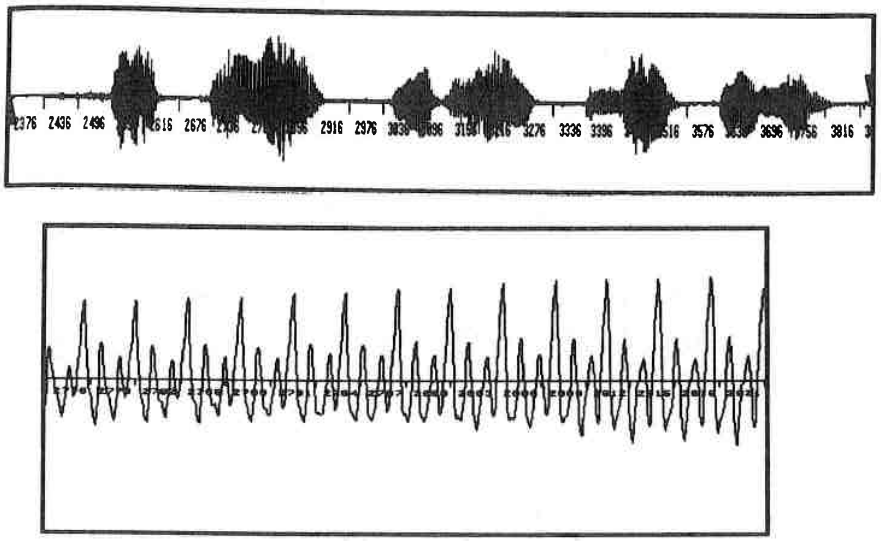


Figure 2.1 : signal temporel et zoom sur une partie du signal

L'utilisateur peut examiner avec précision une partie du signal grâce à un zoom dont il précise l'étendue avec la souris. De même, le début et la fin de phrase sont marqués automatiquement par Snorri mais peuvent être déplacés pour n'étudier que la partie la plus intéressante du signal.

Il est évidemment possible de relire des fichiers de parole déja enregistrés ; l'utilisateur n'a besoin que de la souris pour les sélectionner.

Comme la communauté scientifique de parole a été très féconde en standards (et même en sous-standards) de fichiers de parole, l'utilisateur peut décrire le format des fichiers de parole qu'il veut utiliser. Snorri peut lire tous les fichiers de parole dont il connaît le descripteur de format. Le descripteur contient les informations suivantes : nombre de bits par échantillon, fréquence d'échantillonnage, indicateur d'échange d'octets haut et bas dans les échantillons temporels, début de parole par rapport au début de fichier, taille de la phrase. Les directives du descripteur donnent soit directement la valeur de ces paramètres, soit le moyen de les extraire à partir des fichiers de parole. Aucune limite n'est imposée concernant la taille des fichiers. Le système de descripteurs permet aussi de lire les fichiers contenant plusieurs phrases (comme les fichiers de la base de données BDSONS).

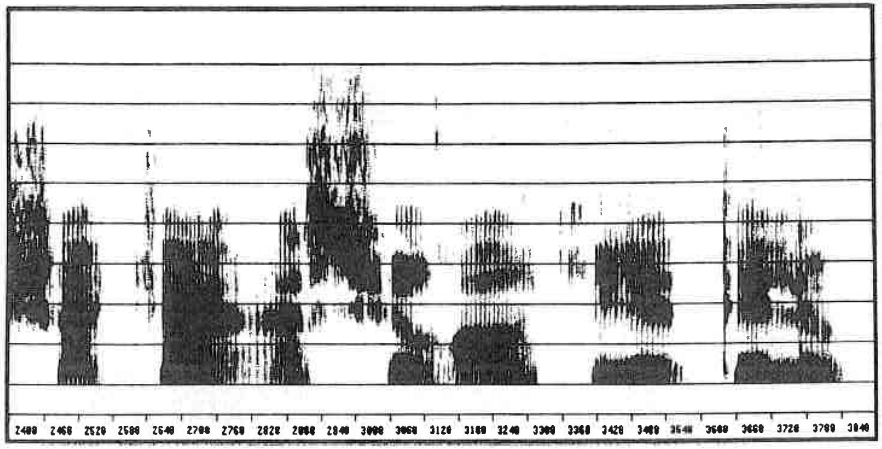


Figure 2.2 : spectrogramme bande large "Chacun assurant qu'il était le plus fort"

4.2 Spectrogrammes

Snorri calcule et affiche des spectrogrammes de très bonne qualité avec une palette de couleurs que l'utilisateur peut modifier. Le spectrogramme normal est un spectrogramme à bande large (fig. 2.2) [2] (fenêtre de Hamming de 4 ms et déplacement de 2 ms), mais Snorri peut aussi générer les spectrogrammes suivants :

- spectrogramme calculé sur les coefficients LPC et qui renforce les fréquences formantiques,
- spectrogramme à bande étroite (fig. 2.3) (fenêtre de Hamming de 32 ms) pour séparer les harmoniques de la fréquence fondamentale,
- spectrogramme lissé cepstralement (avec le filtrage de [Rabiner 69a]). L'utilisateur peut régler le lissage suivant la fréquence fondamentale du locuteur.

Les algorithmes utilisent tous le processeur vectoriel ce qui permet d'obtenir un temps de calcul très court (0.5 seconde pour le calcul d'un spectrogramme de 2 secondes de parole et une dizaine de secondes d'affichage).

Nous avons choisi de fixer la durée par défaut des spectrogrammes à 2 secondes de telle manière que l'utilisateur n'ait pas à s'adapter à plusieurs formats d'affichage. Le choix d'une palette de couleurs (du blanc au noir en passant par jaune et rouge) s'est révélé à l'usage très satisfaisant selon l'opinion unanime des utilisateurs de Snorri.

L'utilisateur peut déplacer la fenêtre de 2 secondes où il veut dans les fichiers de parole dont la longueur excède 2 secondes. Comme les affichages du signal de parole

[2] Les traits horizontaux indiquent les fréquences en kHz.

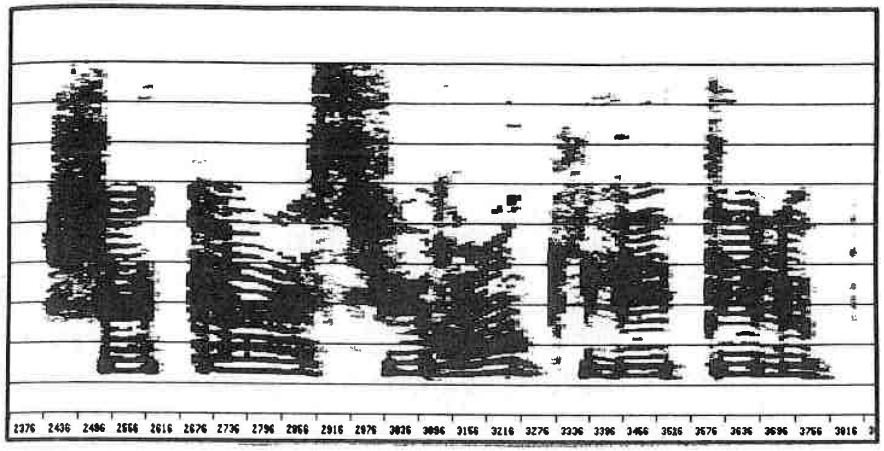


Figure 2.3 : spectrogramme bande étroite

et du spectrogramme sont synchronisés il est possible de calculer un zoom sur le spectrogramme simplement (et donc d'échapper à la contrainte de la durée fixe des spectrogrammes) en effectuant le zoom correspondant sur le signal temporel.

4.3 Etiquetage de la parole

Snorri avec les visualisations de la parole décrites précédemment est un bon outil de segmentation et d'étiquetage. L'utilisateur pose ses marques temporelles avec la souris, soit sur le spectrogramme initial (fig. 2.4), soit sur le zoom d'une partie du spectrogramme. Pour l'aider, il peut écouter un morceau de parole autant de fois qu'il le veut. D'autres options permettent, en cas d'erreur, de modifier la limite du segment ou son étiquette, et même de supprimer le segment créé.

L'utilisateur peut se déplacer dans le signal avec la souris et se positionner sur une étiquette existant déja.

Les marques temporelles et les étiquettes phonétiques sont sauvegardées dans un fichier qu'il est ensuite possible de rappeler pour le consulter ou le modifier.

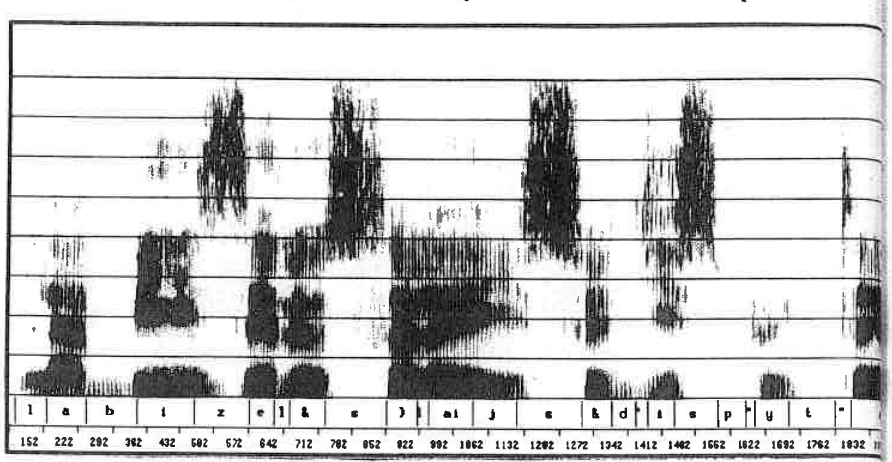


Figure 2.4 : spectrogramme et étiquetage ("La bise et le soleil se disputaient")

5 Aide à l'acquisition des connaissances acoustico-phonétiques

5.1 Utilisation de corpus de parole

Snorri est avant tout conçu comme outil d'acquisition de connaissances acoustico-phonétiques et c'est donc dans cet esprit que nous avons organisé les accès aux corpus étiquetés.

Afin de bien étudier les variations inhérentes au locuteur et celles liées au contexte (ex : plusieurs réalisations du phonème /b/) il faut pouvoir avoir accès rapidement à toutes les occurences d'un phonème d'un corpus, et pouvoir repérer facilement le contexte dans lequel ce phonème a été extrait. Il faut disposer de corpus communs à l'ensemble des chercheurs, mais aussi de corpus de test directement liés à l'étude menée.

Ces exigences nous ont donc amené à choisir l'organisation suivante :

1. L'utilisateur peut choisir soit le corpus "public" soit son propre corpus. Il lui suffit ensuite de cliquer pour choisir le fichier de parole sur lequel il souhaite travailler.

2. L'utilisateur peut extraire soit un phonème particulier (ex : les /i/), soit une séquence de phonèmes (ex : les /i/ précédés d'une labiale), soit une séquence de classes phonétiques (ex : les voyelles antérieures) dans tout le corpus étiqueté. Snorri construit alors un fichier de parole avec les séquences correspondantes présentes dans le corpus ainsi que l'étiquetage qui va avec (ex : la figure 2.5

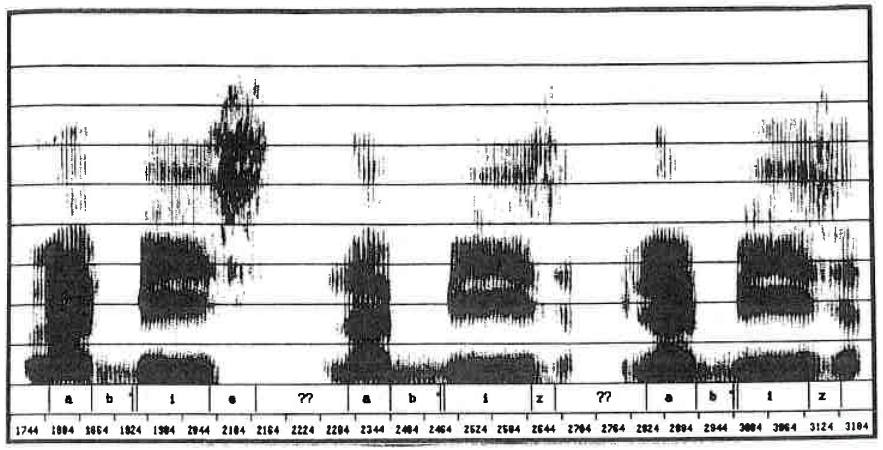


Figure 2.5 : extraction du contexte (labiale + /i/)

montre les /i/ précédés d'une labiale). Ces deux fichiers (signal et étiquetage) sont placés dans le corpus de l'utilisateur. Il est alors plus facile de découvrir les corrélats acoustiques avec ces fichiers grâce à la visualisation successive de toutes les séquences (l'écoute est également possible).

La figure 2.5 montre les spectrogrammes fins de 3 occurrences du contexte (labiale + /i/). Une observation rapide permet de tirer les conclusions suivantes :

- élévation rapide (dans le sens du temps) des formants F2 et F3 au début de la voyelle,
- apparition plus rapide de F2 que de F3 au début de /i/.

5.2 Autres outils

1. Outre les outils généraux de manipulation et d'étiquetage de parole, Snorri offre d'autres outils destinés à l'étude détaillée d'un spectre. Le spectrogramme ne faisant pas apparaître suffisamment clairement les niveaux d'énergie, Snorri calcule et affiche (fig. 2.6) pour une fenêtre choisie sur le signal temporel, une FFT bande large, une FFT bande étroite, un spectre calculé à partir des coefficients LPC et un spectre lissé cepstralement.

2. Snorri calcule et affiche le pitch calculé en utilisant une fonction d'autocorrelation [Rabiner 78].

3. Snorri calcule et affiche l'énergie dans une bande de fréquence quelconque.

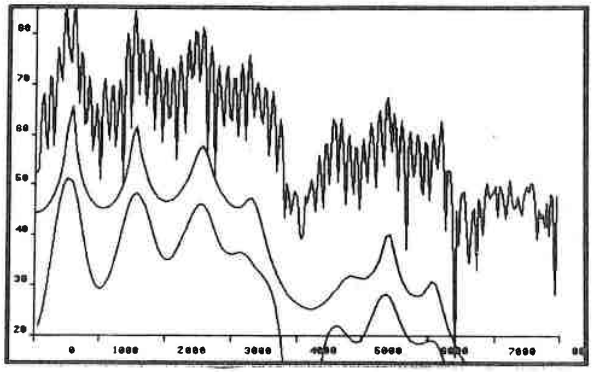


Figure 2.6: de bas en haut : spectre lissé cepstralement, spectre LPC, FFT bande étroite

4. Snorri estime et affiche les trajectoires formantiques en utilisant soit les spectres LPC soit les spectres lissés cepstralement [Fohr 88].

5. L'utilisateur a souvent besoin de décrire les phénomènes de coarticulation en étudiant le chemin que suivent les formants et en le situant par rapport aux cibles que sont les sons qui auraient du apparaître. Snorri permet de suivre dans le plan F1-F2 les deux premiers formants (estimés grâce aux pics de LPC) sur un segment de parole (fig. 2.7). Les cibles de référence des phonèmes du français pour un locuteur masculin sont indiqués et permettent de se repérer.

   La figure 2.7 montre le chemin que suivent F1 et F2 lors de la prononciation en série des voyelles /e//ɛ//a//o//u/ pour deux locutrices (sp1 sp2) et deux locuteurs (sp3 sp4). On découvre clairement sur la figure 2.7 que la position du triangle dépend très sensiblement du locuteur. Ainsi pour la voyelle /e/ les formants des locutrices sont nettement supérieurs à ceux des locuteurs (400 Hz pour F2 et 100 Hz pour F1). Le triangle vocalique apparaît donc comme un outil efficace de normalisation des références vocaliques.

6 Origine et développement de Snorri

Les outils puissants (multifenêtrage, carte d'acquisition, Unix et C) qu'offraient le Masscomp en comparaison des machines précédemment disponibles dans l'équipe ont permis à D. Fohr d'écrire une "préversion" de Snorri. Cette version, même si elle comportait déja un certain nombre des fonctions actuelles de Snorri, n'était pas très conviviale et n'utilisait pas encore toutes les ressources matérielles du Masscomp (notamment le processeur vectoriel).

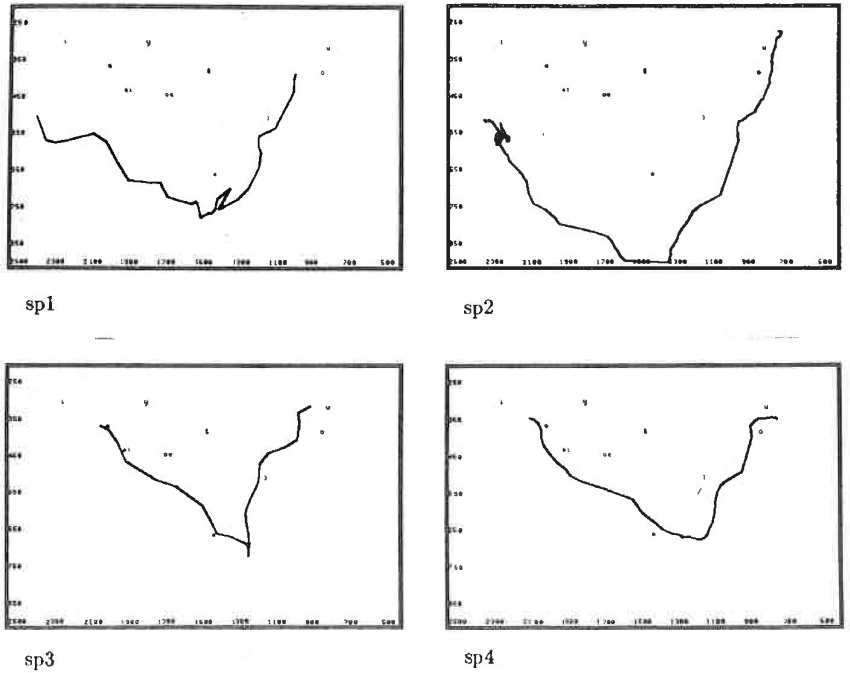


Figure 2.7 : suivis de formants dans le plan F1F2

Nos efforts ont donc d'abord porté sur la convivialité de Snorri. Tous les choix de fonctions se font grâce à la souris qui est dans la plupart des cas suffisante pour rentrer les paramètres que demande Snorri (choix des fichiers, modes de calcul ...). Nous avons aussi développé une copie d'écran adaptée au Masscomp (et générant des fichiers Postscript) qui permet au chercheur de conserver la trace de son travail. Comme Snorri est le volumineux noyau (environ 20 000 lignes de C) de nombreux programmes de notre équipe nous nous sommes efforcés de rendre le code aussi lisible et accessible que possible. C'est évidemment l'exigence la plus difficile à satisfaire mais les chercheurs ne rencontrent pas trop de difficultés pour élaborer leur propre version.

L'architecture d'une machine temps réel et les extensions ajoutées au Masscomp induisent un autre type de convivialité très important : les autres utilisateurs du Masscomp ne doivent pas désespérément attendre pour disposer d'une ressource unique (carte d'acquisition, processeur vectoriel mais aussi mémoire dans la mémoire centrale). Nous avons donc implanté une gestion de ces ressources évitant autant que possible de gêner les autres utilisateurs.

Ces aspects purement techniques et sans rapport évident avec la recherche en parole ne doivent pourtant pas être négligés : s'ils ne sont pas pris en compte le travail de recherche devient vite impossible.

7 Comparaison de Snorri aux autres systèmes destinés à la parole

Il existe un certain nombre de logiciels utilisables par les chercheurs en parole. À l'origine de chaque logiciel correspondent des motivations particulières qui impliquent un domaine d'application assez précis comme nous allons le voir avec les quelques exemples qui suivent.

ISP [Kopec 84] est conçu comme un poste de traitement du signal. Les objets abstraits de base sont le signal et l'affichage auxquels sont attachés tous les opérateurs classiques pour les manipuler. ISP gère les instances de ces types grâce à une pile auquel l'utilisateur a accès. Ce dernier peut aussi facilement (du moins pour les signaux) enrichir son système en définissant de nouveaux types et les opérateurs associés. Néanmoins, le fait que le seul type abstrait de haut niveau soit le signal limite sévèrement ISP.

ASSIA [Gong 85] repose aussi sur la définition de types abstraits mais fournit un environnement de base beaucoup moins fruste que celui d'ISP. Comme les opérateurs sont disponibles directement sous UNIX, cela permet d'utiliser les ressources d'UNIX notamment pour enchaîner les opérateurs ou accéder aux fichiers. Il reste que ces deux systèmes sont essentiellement conçus comme postes de travail destinés au traitement du signal. C'est au chercheur de faire l'effort d'ajouter ses propres types abstraits pour véritablement créer un environnement adapté à la recherche en parole, ce qui est d'ailleurs sans doute plus simple avec ISP qu'avec ASSIA car ISP est implanté sur une machine LISP (Symbolics 3600) tandis qu'ASSIA est programmé

en C sous UNIX.

Dans le domaine du traitement du signal appliqué à la parole, la démarche inverse à celle adoptée pour créer ISP ou ASSIA - définir les fonctions de traitement du signal avant de mettre à jour les types abstraits et leurs opérateurs - s'est révélée, commercialement du moins, beaucoup plus fructueuse. Le fait que le chercheur fasse d'abord porter ses efforts sur l'élaboration des algorithmes de traitement du signal, de synthèse ou de reconnaissance plutôt que sur les types abstraits manipulés explique sans doute ce succès. Ainsi ILS [Technology 86] a d'abord été conçu comme une vaste bibliothèque de sous-programmes (écrits en Fortran) de traitement de signal et de conversion de données. Par la suite (avec l'apparition des postes de travail avec multifenêtrage et souris) ILS s'est enrichi d'une "coquille" interactive permettant l'accès à toutes les fonctions existantes.

ESPS [Shore 87], quant à lui est une collection de commandes UNIX avec la librairie correspondante offrant toutes les opérations classiques de traitement du signal ainsi qu'un certain nombre de fonctions de visualisation. Il reste au chercheur à construire à partir de ESPS son poste de travail orienté vers le traitement de la parole, ce qui demande un certain travail.

Au delà d'une librairie sophistiquée de traitement du signal le chercheur en parole et notamment dans le domaine de l'approche experte du décodage automatique a besoin d'un véritable poste de travail adapté aux études qu'il mène. Reprenant les idées de Schwartz [Schwartz 76], W. Shipmann et S. Cyphers du MIT ont développé SPIRE [Zue 85] (Speech and Phonetics Interactive Research). À la différence des systèmes précédemment cités, Spire est spécialement conçu comme environnement de recherche pour l'approche acoustico-phonétique de la parole. Outre les fonctions classiques (acquisition, restitution d'un signal, calcul et affichage de spectrogrammes, étiquetage manuel ...) Spire offre la possibilité de définir de nouveaux "attributs" (paramètres acoustiques) du signal de parole et de rechercher quels sont les segments de parole vérifiant une condition exprimée avec ces attributs. En ajoutant quelques fonctions statistiques [Shipman 83] il devient possible d'explorer une base de données.

Comme Spire est implanté sur machine Lisp (Symbolics 3600) l'utilisateur peut étendre Spire avec ses propres attributs pour construire l'outil de travail adapté à ses recherches personnelles. En plus de la richesse de l'environnement logiciel, Spire dispose de moyens matériels adaptés à la parole : une carte d'acquisition-restitution avec filtre, un processeur vectoriel pour le traitement du signal et une imprimante électrostatique pour produire des spectrogrammes de bonne qualité. Spire satisfait donc parfaitement aux exigences de ses créateurs : augmenter d'un ordre de grandeur l'efficacité du chercheur en parole, aider à mettre à jour les connaissances acoustico-phonétiques pour la reconnaissance de la parole continue.

D'autres systèmes reprennent particllement le concept de poste de travail en parole, c'est notamment le cas d'Audlab [oE 86] écrit en C et tournant sur Masscomp. Audlab reprend les mêmes fonctions de base que Spire mais sans offrir la définition d'attributs acoustiques et l'exploration statistique de corpus. En fait Audlab se veut plus généraliste.

Situer Snorri par rapport aux systèmes que nous venons d'évoquer peut se faire

en prenant en compte les buts recherchés par Snorri :

- fournir les outils de base en parole (acquisition, restitution, spectrogramme et étiquetage manuel).
- être très puissant comparé aux moyens précédents. Ainsi grâce au processeur vectoriel (sur lequel toutes les opérations de traitement du signal sont effectuées) le calcul d'un spectrogramme de 2 secondes de parole ne demande plus une demi-seconde (cela équivaut environ à une puissance de calcul de 10 Mflop).
- être utilisable par les phonéticiens sans programmation. Ainsi F. Lonchamp et J. Vaissière nous ont beaucoup appris avec Snorri qui leur permet d'illustrer leurs idées.
- constituer le noyau acoustique des applications de parole du Crin (cf § 9).
- permettre d'enrichir l'expertise acoustico-phonétique en isolant un contexte phonétique donné pour tout un corpus (cf § 5.1). Cela permet aussi de tester les algorithmes de reconnaissance dans un contexte phonétique donné.

Ces caractéristiques rapprochent donc Snorri de Spire. La différence la plus sensible entre les deux logiciels découle sans doute de l'effort à accomplir pour enrichir le noyau de base. Ce qui est très simple voire naturel avec Spire (écrit sur une machine Lisp) l'est nettement moins avec Snorri qui oblige l'utilisateur à jouer de l'éditeur de texte, du compilateur et du "débogueur" pour ajouter une nouvelle fonction. Spire permet de plus de sélectionner les segments de parole d'un corpus vérifiant une propriété acoustique définie par l'utilisateur.

Pour résumer comparé à ISP et ASSIA qui se veulent postes de travail destinés au traitement du signal, à ILS et ESPS qui se veulent de vastes et conviviales librairies de traitement du signal, Snorri à l'image de Spire peut être qualifié de poste de travail destiné à l'expertise acoustico-phonétique.

8 Comment continuer l'effort entrepris avec Snorri?

L'expérience apportée par le développement et l'utilisation de Snorri permet, à la lumière de l'évolution de l'environnement matériel et logiciel actuel d'esquisser à grandes lignes ce que devrait être le successeur de Snorri. Appelons le Snorri2.

8.1 Aspect matériel

Snorri2 doit être quasiment indépendant de la machine sur laquelle il se trouve. L'utilisation de cartes spécialisées (acquisition et traitement du signal) semble nécessaire à moins que d'autres constructeurs suivent l'exemple de NeXT. Comme ces ressources peuvent être uniques et distantes de la machine sur laquelle tourne Snorri2, Snorri2 doit pouvoir choisir la machine sur laquelle il éxécute les entrées sorties parole et le traitement du signal.

8.2 Aspect logiciel

Une des premières vertus de Snorri a été de fixer bon nombre d'algorithmes de traitement du signal développés par plusieurs chercheurs de l'équipe. Qui plus est, cela permet à chacun de disposer d'un noyau relativement stable et fiable.

L'ensemble du noyau (Snorri représente environ 20 000 lignes) et des autres versions dépasse 100 000 lignes de C et commence à poser certains problèmes de génie logiciel (exemple : Quelqu'un a-t-il déja programmé la fonction à laquelle je travaille?). Par ailleurs il faut bien reconnaître que l'adjonction de nouvelles fonctions n'est pas très simple car malgré l'effort de structuration et de documentation le langage C est trop souple pour permettre de pouvoir réutiliser les programmes de toute une équipe.

Une solution est sans doute d'adopter un langage à objets comme Eiffel qui satisfait bien à ces critères de génie logiciel. De nombreuses classes concernant l'interface graphique sont déja disponibles et les modifications sont grandement facilitées par le mécanisme des classes qui aident à formaliser les objets manipulés par le programme. Enfin Eiffel est facilement interfaçable à C et permet donc de récupérer une partie des fonctions déja écrites.

8.3 Aspect acquisition des connaissances

Comme la vocation initiale de Snorri est l'acquisition des connaissances acoustico-phonétiques, il faut donner à Snorri les indispensables outils d'analyse de données qui permettent d'aider à la découverte de corrélats acoustiques fiables. Il faut d'ailleurs sans doute être plus ambitieux et envisager les moyens de pouvoir réaliser une certaine forme d'apprentissage automatique pour dégager l'expert phonéticien du dépouillement des essais de paramètres acoustiques. Ces quelques remarques doivent guider les réflexions qui accompagneront la poursuite de l'effort déja engagé pour développer Snorri.

9 Utilisation de Snorri au CRIN

Snorri sous la forme qui vient d'être présentée est utilisé comme outil d'étiquetage (corpus GRECO "La bise et le soleil" et un corpus de phrases météorologiques) et comme outil d'analyse acoustico-phonétique de corpus. Par ailleurs il sert de base à plusieurs projets :

1. Dominique Fohr a réimplanté le système expert en décodage acoustico-phonétique Aphodex [Carbonell 86] en utilisant les modules de Snorri.

   L'une des extensions les plus intéressantes pour l'avenir concerne l'étiquetage semi-automatique de la parole grâce à système de segmentation du projet Aphodex. C'est en effet une tâche assez fastidieuse pour un expert phonéticien de segmenter et d'étiqueter manuellement un grand nombre de phrases, d'autant plus que cela introduit d'inévitables erreurs humaines. De simples

techniques de programmation dynamique permettent de réaliser l'alignement de la segmentation produite par l'algorithme de segmentation d'Aphodex avec la transcription phonétique de la phrase à étiqueter. Durant l'alignement seules les grandes classes phonétiqués (occlusives, fricatives, voyelles et sonantes) sont prises en compte.

Après cet alignement l'étiquetage phonétique est beaucoup plus rapide puisqu'il ne reste plus qu'à corriger légèrement les limites de la segmentation sans changer les étiquettes. D'autre part cela évite la plupart des erreurs de manipulation puisqu'il n'y a plus à choisir les étiquettes phonétiques. Cette technique a été employée avec succès pour étiqueter un corpus de 340 phrases prononcées par plusieurs locuteurs.

2. Les résultats du décodage acoustico-phonétique sont exploités par le système de reconnaissance lexicale de Bernard Mangeol [Haton 86].

3. Une version de Snorri (intégrant le calcul de F0) est destinée à l'étude de la structure prosodique de la phrase et permet de localiser les fins de mots et les fins d'unités syntaxiques.

4. Une autre version de Snorri est destinée à l'étude de la parole énoncée en milieu bruité et porte plus précisément sur la reconnaissance d'un vocabulaire d'une cinquantaine de mots anglais.

10 Conclusion

Snorri a demandé environ un an de travail et continue régulièrement à être mis à jour. L'an prochain sera réalisée une version tournant sous X et permettant donc à Snorri d'être utilisé sur une large gamme de machines.

Snorri s'est révélé être un très bon outil d'étude pour l'expertise acoustico-phonétique et nous permet de faire progresser notre connaissance des phénomènes de la parole et de mieux les prendre en compte pour le décodage automatique. C'est aussi un irremplaçable outil pédagogique dans le domaine de la parole.

De nombreux visiteurs ayant montré un vif intérêt pour ce logiciel il a été donné à plusieurs équipes en France (CNET à Lannion, ICP à Grenoble, CERFIA à Toulouse) et aussi à l'étranger (l'Institut de Phonétique de Bruxelles et Matshushita au Japon). Cela renforce la conviction qu'il est nécessaire d'adopter une démarche de programmation prenant en compte plus efficacement les concepts de génie logiciel car cela permettrait sans doute de bénéficier en retour de l'effort de programmation des autres équipes utilisant Snorri [3].

[3] Ce n'était pas la peine de chercher quelle abréviation se cache sous Snorri car c'est en fait le nom d'un héros de l'une des plus célèbres sagas islandaises "Snorri le Godi" faisant partie des "Sagas islandaises" traduites par Régis Boyer et parues dans la collection de la Pléiade

Chapitre 3

Le suivi de formants

1 Introduction

Avant d'aborder le problème du suivi de formants nous allons définir le terme formant.

Parmi toutes les composantes d'un système de reconnaissance de la parole, la composante de décodage acoustico-phonétique est la seule qui doive manipuler un signal d'origine physique, le signal de parole, pour en extraire une information linguistique, une chaîne de phonèmes. Un des enjeux de la reconnaissance de parole, est donc de trouver les paramètres physiques liés le plus étroitement possible à la nature linguistique du signal de parole.

En première approche le conduit vocal peut être modélisé par un tube de forme inconnue excité par le passage de l'air entre les cordes vocales. Le fait que la forme d'un tube puisse être correctement représentée par ses fréquences de résonance, incite à rechercher les fréquences de résonance du conduit vocal. Ces fréquences portent le nom de formant.

Les formants étant de fortes concentrations d'énergie, ils apparaissent sur un spectrogramme sous la forme de bandes sombres globalement orientées horizontalement, comme le montre la figure 3.1.

Le terme de formant est réservé aux sons vocaliques pour lesquels le conduit vocal est excité par la vibration des cordes vocales. Quand le conduit vocal est excité par un bruit de friction, c'est-à-dire pour les sons fricatifs, les résonances du conduit vocal apparaissent encore comme des bandes sombres mais localisées dans une zone de bruit. Elles portent le nom de "formant de bruit" par analogie aux sons vocaliques.

La notion de formant conduit Delattre [7] à définir un triangle acoustique des voyelles du français à partir des formants F1 et F2. Par la suite Fant [Fant 60] met au point un modèle de parole à base de formants. Le terme formant prend donc comme sens supplémentaire celui d'un paramètre du modèle de production de parole. Le développement du codage par prédiction linéaire par Saito et Itakura, [Saito 66] puis par Atal et Schroeder [Atal 67] procure, à la fin des années 60 un moyen efficace pour extraire les formants.

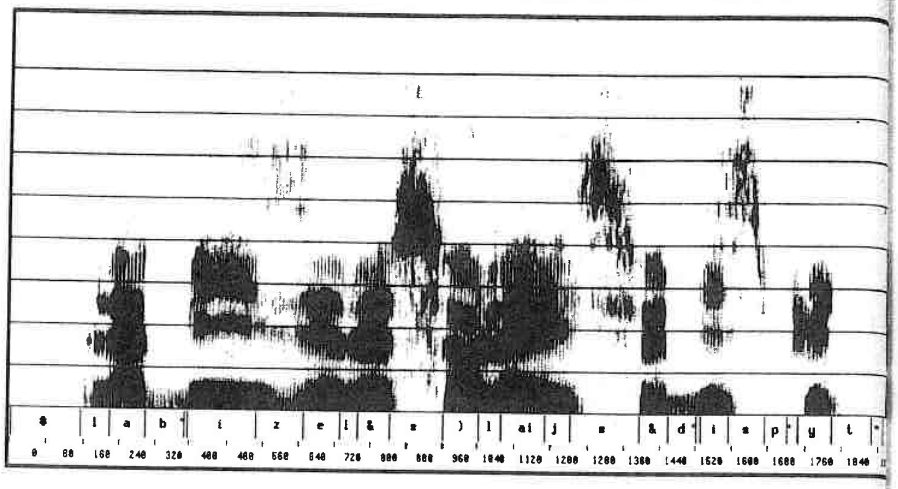


Figure 3.1 : Spectrogramme avec formants bien visibles

L'importance des formants, sans doute la notion centrale pour tous les travaux qui cherchent à lier le signal de parole à son origine physique, a rapidement été validée par les chercheurs en parole. D'une part les premiers synthétiseurs, utilisant les 3 ou 4 premiers formants, ont montré qu'ils produisaient un signal intelligible. D'autre part tous les modèles acoustiques de production de la parole font généralement appel à la notion de formant.

En ce qui concerne un système de décodage de la parole, la contribution des formants est double. Ils permettent tout d'abord de définir les cibles vocaliques grâce à leur profil formantique tel qu'il serait si ces sons étaient prononcés isolément. Par ailleurs les contraintes physiques inhérentes à la nature de l'appareil articulatoire imposent que les sons d'un énoncé s'influencent mutuellement entre voisins. Ce phénomène dit de "coarticulation" tend à écarter les trajectoires formantiques des cibles des sons pris isolément.

Estimer avec précision les trajectoires formantiques est donc un enjeu crucial pour qu'un système de décodage puisse tirer profit des connaissances acoustico-phonétiques que lui ont données les experts phonéticiens.

Ce chapitre commence par la description des formants, en mettant en évidence les points critiques pour un algorithme de suivi de formants. Ensuite nous présentons le codage de la parole par prédiction linéaire et le lissage cepstral qui sont les deux méthodes que nous avons utilisées pour évaluer numériquement les formants. Nous étudions ensuite quelques algorithmes marquants dans le domaine du suivi de formants avant de présenter l'algorithme que nous avons développé.

2 Nature des formants

Avant de présenter les techniques d'extraction et de suivi des formants, précisons comment il est possible de lier les formants à la forme du conduit vocal et comment ils apparaissent sur un spectrogramme. Cela nous donnera l'occasion de mettre à jour les difficultés que rencontrera un algorithme de suivi de formants.

2.1 Lien des formants avec l'articulatoire

La source de la parole est le passage de l'air des poumons entre les cordes vocales au niveau de la glotte. Le signal correspondant, caractérisé par la fréquence fondamentale (onde glottale), est ensuite filtré par les cavités du conduit vocal. À ces cavités correspondent des fréquences de résonance qui vont donc renforcer certaines harmoniques du fondamental.

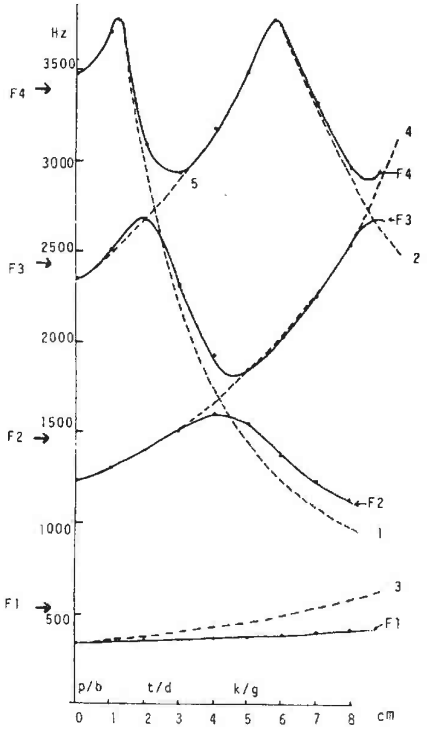
Les travaux d'acousticiens, notamment ceux de Fant [Fant 60] et de Stevens [Stevens 68] permettent de connaître l'affiliation d'un formant à une cavité résonnante et de lier une modification du profil formantique à une déformation du conduit vocal.

Le principe de ces études est d'étudier les fréquences de résonance d'un modèle simplifié du conduit vocal. Le modèle est un assemblage plus ou moins compliqué de tubes simulant les cavités, la forme des lèvres et les points de resserrement.

Les résultats, sous forme de diagrammes bidimensionnels appellés "nomogrammes", (l'axe des abscisses figurant l'un des paramètres du modèle et l'axe des ordonnées figurant les fréquences) représentent les variations fréquentielles des quatre premiers formants en fonction du paramètre étudié. De tels diagrammes donnent une explication articulatoire aux transitions formantiques qui apparaissent sur le spectrogramme et permettent notamment de mieux comprendre le phénomène de coarticulation.
Exemple :
Lors de la prononciation de "doute", noté /dut/ en API, les formants 2 et 3 du son /u/ sont très nettement décalés vers le haut. Les formants du /u/ isolé sont 320 Hz, 760 Hz et 2 020 Hz, dans le contexte /dut/ on observe couramment 320 Hz, 1 200 Hz et 2 150 Hz. En se reportant au nomogramme de la figure 3.2 tirée de [Lonchamp 84], il est possible de voir que les formants 2 et 3 s'élèvent nettement quand on impose une constriction correspondant à l'articulation de /t/ et /d/. Cette constriction subsiste lors de la production du son /u/ et déplace donc vers le haut les formants 2 et 3.

2.2 Formants et harmoniques du fondamental

Le conduit vocal est excité par la vibration des cordes vocales dont la fréquence fondamentale est de l'ordre de 120 Hz pour les hommes et de 200 Hz pour les femmes. Comme la concentration d'énergie correspondant à un formant s'étend en fait sur



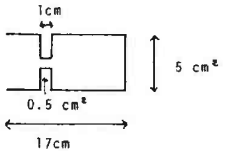


Fig. 23 - Effet acoustique d'une constriction sur un tube uniforme de 17 cm de long. constriction de 1cm /0.5 cm² - Les flèches indiquent la position des formants pour le tube sans constriction. La position de la constriction (bord gauche) est donnée en abscisse. Pour les courbes pointillées, voir le texte.

Figure 3.2 : Exemple de nomogramme

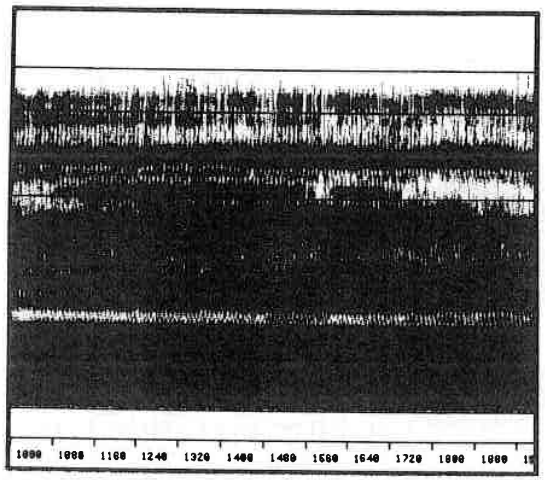


Figure 3.3 : Spectrogramme d'un /a/, avec un fondamental très élevé

un intervalle de fréquence de l'ordre de 200 Hz, elle renforce suivant la valeur du fondamental un ou plusieurs harmoniques.

Lors de la variation de fréquence d'un formant (c'est à dire de l'évolution de la forme du conduit vocal) les harmoniques renforcés changent et les formants "sautent" donc d'une harmonique à l'autre. Pour les locuteurs dont le fondamental est faible (proche de 120 Hz) l'amplitude des sauts lors de la variation de fréquence d'un formant est faible et il donc facile de "suivre" un formant. Au contraire si le fondamental est élevé (plus de 220 Hz ce qui est fréquemment le cas chez les locutrices) les sauts sont plus importants et rendent difficile le suivi de formants (voir fig. 3.3). Nous nous sommes implicitement placés jusqu'ici dans l'hypothèse favorable où la fréquence fondamentale est constante. Quand la fréquence fondamentale change, les effets des "sauts" d'harmoniques (qui viennent d'être décrits) et des variations fréquentielles des harmoniques s'ajoutent et cela complique sensiblement le suivi de formants. Sur la figure 3.4, illustrant ce phénomène, il devient difficile, même pour un expert humain, de découvrir la trajectoire des formants.

2.3 Changement d'affiliation à une cavité des formants

Un formant affilié à une cavité donnée peut changer de numéro, comme le montre Fant [Fant 60], (sa valeur fréquentielle change relativement à celle des autres formants) quand les cavités subissent de grandes modification de taille. Il se produit alors ce qu'on appelle abusivement un "croisement de formants".

Reprenons le nomogramme de la figure 3.2 pour expliquer ce phénomène. Les courbes en pointillés correspondent aux fréquences de résonance des cavités placées

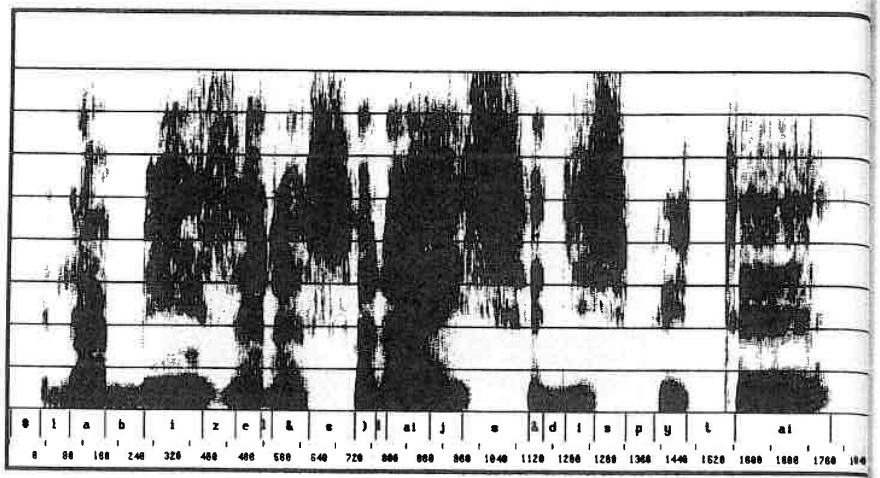


Figure 3.4 : Exemple de fondamental variant rapidement pendant /ize/ et /tai/

devant (courbes 1 et 2) et derrière (courbes 3, 4 et 5) la constriction. On constate ainsi que le formant correspondant à la première fréquence de résonance de la cavité antérieure est numéroté successivement F4, F3 et F2. Lorsque le point de constriction franchit l'abscisse du point d'intersection des courbes 1 et 4, cela se traduit par un "croisement de formants" ; abus de langage signifiant en réalité que les formants échangent leur ordre d'apparition en fréquence. Il est en général difficile de détecter ce genre d'événement ou du moins de savoir à partir du seul suivi de formants s'il a lieu.

2.4 Sons nasalisés

Pour terminer ce panorama des points d'achoppement de tout algorithme d'estimation des trajectoires formantiques il faut évoquer les problèmes de nasalisation. En effet dans le cas des sons nasalisés, le couplage des cavités nasales et orales perturbe le profil formantique en ajoutant des paires pôle-zéro dans le spectre. La grande variabilité de la configuration des cavités nasales et la pléthore d'hypothèses divergentes émises par les chercheurs en parole (voir [Lonchamp 88]) empêche d'avoir une idée précise de l'emplacement des formants d'origine nasale. Par ailleurs l'ordre d'apparition des formants dépend des fréquences de résonance des conduits vocal et nasal pris séparément.

En conclusion, il semble que les résultats d'un algorithme de suivi de formants sont intéressants là où justement ils sont le plus difficiles à obtenir.

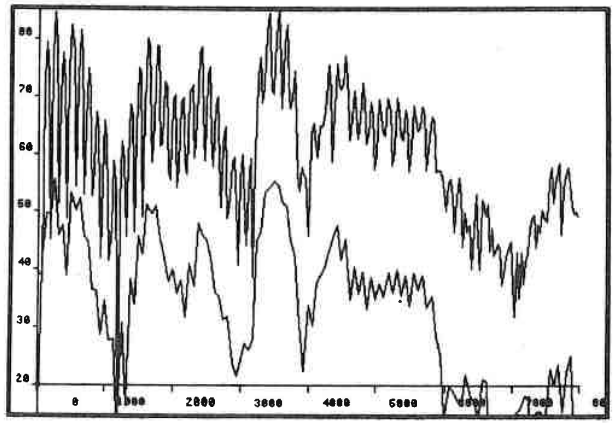


Figure 3.5 : Spectre bande étroite

3 Extraction des fréquences formantiques

Les formants peuvent être estimées directement sur le spectrogramme ; cette technique simple se heurte à deux problèmes :

- si la fenêtre choisie pour calculer la FFT est plus grande que la période du fondamental, les harmoniques du fondamental apparaissent sur le spectre et il est difficile à partir de la position et de l'amplitude des harmoniques renforcées de trouver avec précision les fréquences de résonance (voir fig. 3.5),
- si la fenêtre est plus petite que la période du fondamental, la détermination des fréquences de résonance dépend de la position de la fenêtre de calcul à l'intérieur d'une période du fondamental (voir [Cheng 87]).

Les deux méthodes que nous allons présenter au paragraphe suivant permettent de contourner ces problèmes :

- le calcul par prédiction linéaire,
- le lissage cepstral du spectre.

3.1 Prédiction linéaire

L'idée de la prédiction linéaire est d'exprimer l'onde de parole, échantillonnée par $s(n)$, en fonction des p précédents échantillons ; $\tilde{s}(n)$ étant le prédicteur :

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k * s(n-k)$$

Les coefficients α_k $1 \leq k \leq p$ sont déterminés en minimisant le signal d'erreur :

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.1}$$

L'erreur quadratique sur la fenêtre de calcul est :

$$E_n = \sum_m (s_n(m) - \tilde{s}(m))^2 \tag{3.2}$$

Nous reviendrons sur l'intervalle de sommation au §3.1.

$$E_n = \sum_m (s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k))^2$$

L'erreur minimale est obtenue pour :

$$\begin{aligned}
\frac{\partial E_n}{\partial \alpha_i} &= 0 \quad 1 \leq i \leq p \\
\frac{\partial E_n}{\partial \alpha_i} &= -2 \sum_m (s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k)) s_n(m-i)) \\
&= 2(\sum_{k=1}^{p} \alpha_k \sum_m s_n(m-k) s_n(m-i) - \sum_m s_n(m) s_n(m-i))
\end{aligned}$$

En posant :

$$\Phi(i,k) = \sum_m s_n(m-i) s_n(m-k)$$

l'équation précédente devient :

$$\sum_{k=1}^{p} \alpha_k \Phi(i,k) = \Phi(i,0) \quad 1 \leq i \leq p$$

Il est facile de résoudre ce système pour trouver les coefficients α_k.

Prédiction linéaire et modèle tout pôle

L'expression 3.1 est en accord avec le modèle de production de parole classique : filtrage par le conduit vocal du signal d'excitation glottale. En effet ce modèle s'exprime, pour les transformées en z de l'excitation ($U(z)$) et du signal produit ($S(z)$), par :

$$S(z) = H(z)U(z)$$

où la fonction de transfert $H(z)$ modélise l'effet du conduit vocal et s'exprime par :

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

Soit en passant au signal temporel :

$$s(n) = \sum_{k=1}^{p} s(n-k) + Gu(n)$$

donc :

$$Gu(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

que l'on peut rapprocher de l'expression de $e(n)$. Le modèle classique de production de parole n'étant correct que pour les sons voisés non nasalisés il faut s'attendre à ce que cette méthode conduise à des résultats erronés dans les autres cas. La prédiction linéaire donne cependant une bonne image du filtre qu'est le conduit vocal pour la plupart des sons.

Les coefficients α_k que nous avons calculés au paragraphe précédent permettent donc d'estimer la fonction de transfert du conduit vocal $H(z)$.

Comparaison des différentes méthodes de prédiction linéaire

Le mode de sommation sur m conduit à deux techniques différentes (voir par exemple [Rabiner 78]) :

- si l'on définit E_n par $\sum_{m=0}^{N-1} e_n^2(m)$ en imposant les mêmes limites de sommation pour les calculs des coefficients $\Phi(i,k)$, cela conduit à la méthode de prédiction linéaire par covariance. Cette méthode est utilisable avec un nombre de points plus petit que la période du fondamental et permet donc de réaliser des analyses synchronisées sur le fondamental. Elle n'assure cependant pas la stabilité du filtre auquel elle conduit (le filtre est dit stable si tous ses pôles sont à l'intérieur du cercle complexe unité).

- si l'on définit $s_n(m)$ par $s_n(m) = s(n+m)w(m)$, $w(m)$ étant une fenêtre de longueur finie avec $w(m) \neq 0$ pour $0 \leq m \leq N-1$, cela conduit à la méthode de prédiction linéaire par autocorrélation. Elle nécessite une fenêtre plus longue que la période du fondamental et l'erreur quadratique induite est plus grande que par la méthode précédente [1]. En revanche, cette méthode garantit que la fonction $H(z)$ est stable. Cette raison nous a conduit à la retenir pour extraire les formants.

Les considérations de complexité (en terme de nombre d'instructions et donc de temps), quant à elles, perdent de leur importance avec les machines actuelles.

[1] Notons d'ailleurs qu'elle est maximale pour une fenêtre dont le nombre de points est un multiple du nombre de points de la période du fondamental. Il ne faut donc pas tenter de synchroniser ce calcul de LPC sur le pitch.

Extraction des formants

Montrons comment obtenir le spectre de la fonction de transfert $H(z)$ et les formants. Pour cela, évaluons le logarithme de l'amplitude $H(z)$ pour N points régulièrement espacés sur le cercle complexe unité.

$$H(\exp(j\frac{2\pi}{N}k)) = \frac{E_n}{\sum_{l=0}^{p} b_l \exp(-j\frac{2\pi}{N}kl)}$$

où :
E_n est l'erreur de prédiction (cf équation 3.2)
$b_0 = 1$
$b_l = -a_p$ pour $1 \leq l \leq p$
En étendant le domaine de sommation du dénominateur de $H(k)$ nous obtenons[2] :

$$H(k) = \frac{E_n}{\sum_{l=0}^{N-1} b_l \exp(-j\frac{2\pi}{N}kl)}$$

avec :
$b_0 = 1$
$b_l = -a_p$ pour $1 \leq l \leq p$
$b_l = 0$ pour $p < l \leq N-1$

$$H(z) = E_n / B(k)$$

$B(k)$ désignant les coefficients de Fourier de la séquence $(b_l)_{0 \leq l \leq N-1}$. Par conséquent :

$$\log \mid H(k) \mid = \log E_n - \log \mid B(k) \mid$$

Un simple calcul de FFT (qu'il serait d'ailleurs possible d'élaguer puisque la plupart des coefficients b_l sont nuls [Markel 71]) suffit donc pour évaluer le spectre de la fonction H(z).

Pour estimer les fréquences formantiques à partir du spectre que nous venons de calculer, il est possible soit de retenir les pics de cette courbe, soit de calculer les racines du polynome associé.

La détection des pics sur le spectre ne conduit pas toujours à toutes les fréquences de résonance. En effet les racines sont à l'intérieur du cercle complexe unité et le spectre est évalué pour des points du cercle unité. Si deux racines sont proches, il peut donc n'y avoir qu'un seul pic. Il est possible de remédier à ce problème en évaluant le spectre sur un cercle complexe dont le rayon est inférieur à l'unité ([Markel 76] p 163).

La seconde technique consiste à extraire les racines du polynôme $1 - \sum_{k=1}^{p} a_k z^{-k}$ grâce à la méthode de Bairstow [Stoer 80]. Cette seconde approche est beaucoup plus coûteuse en temps.

[2] Par un abus de langage traditionnel, $H(\exp(j\frac{2\pi}{N}k))$ et $H(k)$ représentent la même expression.

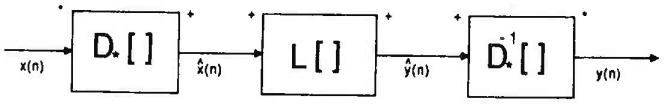


Figure 3.6 : Cascade homomorphique

3.2 Lissage cepstral

Comme les formants correspondent à des concentrations d'énergie dans le spectre, une solution pourrait être, pour les extraire, de rechercher les pics du spectre. Néanmoins la multiplicité des pics, notamment les harmoniques du fondamental, voue cette technique à l'échec. En revanche, si le spectre est suffisamment lissé, mais en respectant les pics d'origine formantique, cette technique devient applicable. Nous allons donc présenter le lissage cepstral qui élimine la plupart des pics du spectre sans rapport avec les formants.

Le signal de parole est produit par filtrage du signal glottal par le conduit vocal, c'est à dire dans le domaine temporel par un produit de convolution $s(n) = f(n) * h(n)$ ($s(n)$: signal de parole, $f(n)$: excitation, $h(n)$: conduit vocal). Afin de séparer les deux contributions on cherche un système dans lequel la superposition s'exprime par une somme. Un système obéissant au principe de superposition est appelé système homomorphique. Pour déconvoluer le signal d'origine il faut donc passer d'un système homomorphique convolutif à un système homomorphique additif où s'effectue la séparation des deux composantes, puis revenir au système homomorphique convolutif de départ. Le signal $x(n)$ subit donc les transformations suivantes (voir figure 3.6) :

- D_* passe d'une relation convolutive entre signaux à une relation additive. $x(n)$ est transformé par D_* en $\hat{x}(n)$, le signal $\hat{x}(n)$ porte le nom de cepstre.
- L sépare les deux contributions de $\hat{x}(n)$; comme $\hat{x}(n)$ est dans cet espace la somme de ses composantes, l'opérateur L transformant $\hat{x}(n)$ en $\hat{y}(n)$ est très simple.
- D_*^{-1} permet de revenir au système homomorphique convolutif. $\hat{y}_n$ est transformé en $y(n)$.

Pour obtenir un spectre lissé il semble donc qu'il faille appliquer une transformée en z à y(n). En fait D_*, comme nous allons le voir, passe par l'intermédiaire du spectre de $x(n)$; sa transformée inverse D_*^{-1} passe donc aussi par un spectre qui est précisément celui que nous recherchons. En pratique, il suffira donc de n'appliquer qu'une partie de l'opérateur D_*^{-1}.

D_* se décompose de la manière suivante :

- une transformée en z pour passer d'une relation convolutive dans le domaine temporel à une relation multiplicative. Nous choisissons comme transformée en z la transformée de Fourier.

- un logarithme pour passer d'une relation multiplicative du domaine fréquentiel à une relation additive de ce même domaine.
- une transformée en z inverse pour revenir au domaine temporel. Il s'agit ici d'une transformée de Fourier inverse.

On parle de cepstre ou encore de cepstre réel quand on ne tient compte que de l'amplitude du spectre pour le calcul du logarithme, et de cepstre complexe quand on prend aussi en compte la phase.

Pratiquement, nous obtenons donc :

- pour le calcul des cepstres réels :

$$\hat{x}(n) = IDFT(\log \mid DFT(x(n)) \mid)$$

(DFT : Transformation de Fourier discrète,
IDFT : Transformation de Fourier discrète inverse)

- Le filtrage $l(n)$, correspondant à l'opérateur$L$ mentionné plus haut, est défini par :

$$l(n) = \begin{array}{ll} 1, & \mid n \mid < n_0 \\ 0, & \mid n \mid \geq n_0 \end{array}$$

où n_0 est choisi plus petit que la période du fondamental.
exemple :
si le cepstre est calculé avec 256 points pour une fenêtre de 16 ms et un fondamental de 250 Hz (4 ms) n_0 est majoré par 64. Pour obtenir un bon lissage il faut prendre une valeur sensiblement plus petite (voir fig. 3.7).

- Comme nous voulons simplement le spectre lissé, nous calculons la transformée de Fourier discrète sur $\hat{y}_n$:

$$\hat{Y}(z) = DFT(\hat{y}(n))$$

Comme le signal d'origine est réel, il est possible d'optimiser les calculs avec uniquement des transformeés de Fourier rapides ([Markel 71]).

Problèmes liés à l'utilisation du lissage cepstral

Formant glottal Le lissage cepstral conduit dans certains cas à confondre F1 avec un formant lié à la forme de l'onde glottale, appelé "formant glottal". La forme de l'excitation est approximativement triangulaire, et la base du triangle désignée par b sur la figure 3.8, correspond à la durée de l'ouverture de la glotte.

Le spectre de l'onde glottale (estimation de Rosenberg [Rosenberg 71]) montre un premier pic d'énergie à la fréquence $2/b$ (voir fig: 3.9). Il arrive que ce pic apparaisse clairement sur le spectre. Cela se produit dans les situations suivantes (voir [Lonchamp 88] p 54-62 pour une explication plus complète) :

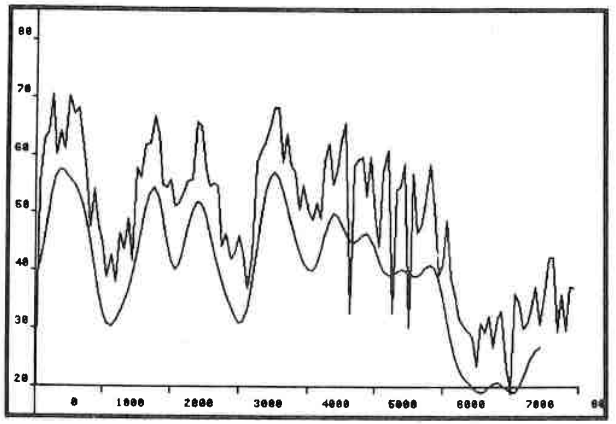


Figure 3.7: Spectre lissé cepstralement nombre de points d'une période de pitch : 110 nombre de coefficients cepstraux conservés :35

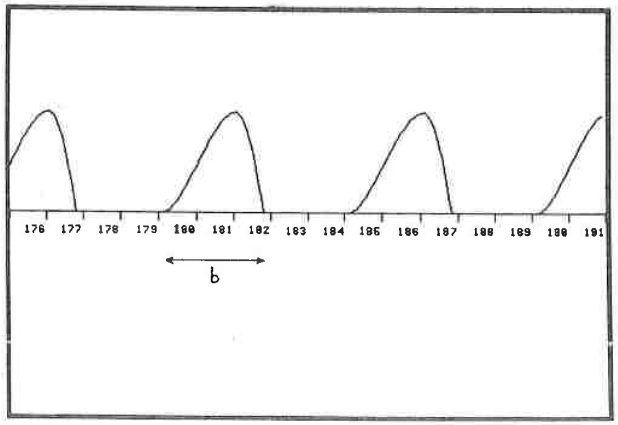


Figure 3.8 : Forme de l'onde glottale

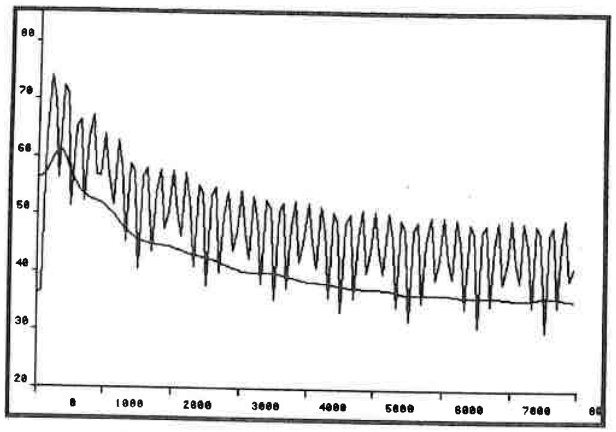


Figure 3.9 : Spectre de l'onde glottale

- voyelles nasales : F1 est affaibli par un zéro nasal voisin,
- voix faible : cela modifie la forme de l'onde glottale et renforce le formant glottal par rapport aux autres formants,
- F1 est élevé : c'est le cas de /a/ où souvent on peut nettement discerner le formant glottal (fig. 3.10).

Fréquence des pics Les fréquences des pics surtout entre 0 et 1000 Hz ne sont pas exactement celles auxquelles conduit la LPC. Cette différence peut être sensible ($\sim 100Hz$) pour /a/ car le calcul LPC a tendance à fusionner le formant glottal et F1 en donnant une valeur intermédiaire. Au contraire le lissage cepstral sépare ces deux formants et il faut donc tenir compte de ce phénomène en ce qui concerne les références formantiques des voyelles.

Pics supplémentaires À la différence du codage LPC qui fixe le nombre de résonances possibles donc le nombre de pics, le lissage cepstral ne subit pas de contrainte analogue et suit avec autant de précision que possible le spectre. Il reste donc de petits pics sans signification formantique qu'il faut éliminer.

Avantage du calcul cepstral

Le lissage cepstral, comme le codage LPC permet de séparer les contributions de la source et du conduit vocal, mais à la différence du codage LPC il ne dépend d'aucune modélisation implicite de la production de la parole. Cela est très intéressant pour les sons nasalisés qu'appréhende mal le codage LPC. La figure 3.11 illustre cet avantage sur un segment de parole nasalisé. Un algorithme de suivi de formants faisant appel au calcul LPC ne peut pas corriger l'erreur sur F1.

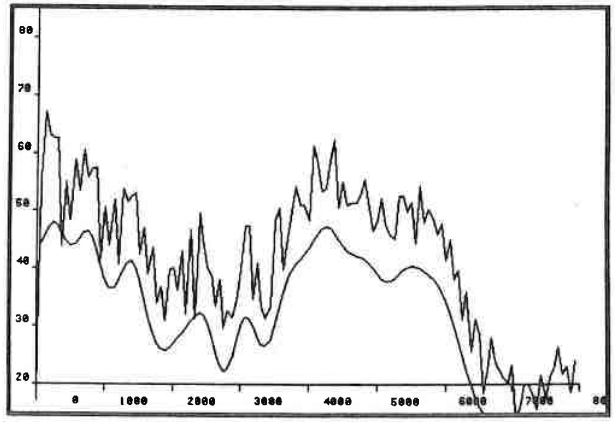


Figure 3.10 : Spectre FFT et spectre lissé cepstralement d'un /a/

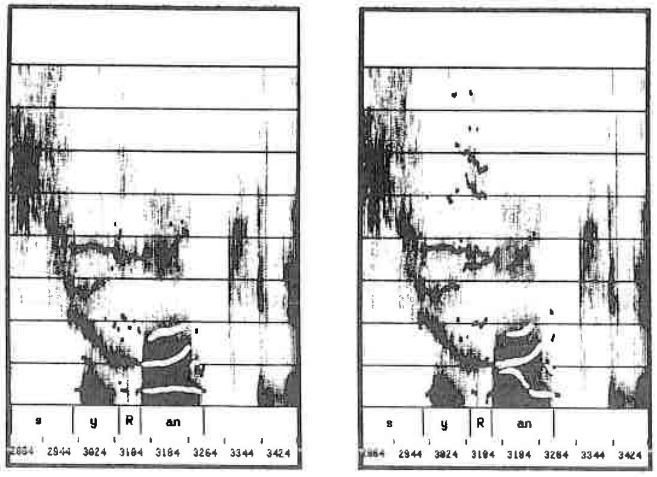


Figure 3.11: Extraction des pics du spectre lissé cepstralement (à gauche) et extraction des racines de LPC (à droite) sur le segment /uran/

4 Quelques algorithmes de suivi de formants

Nous allons maintenant passer en revue quelques techniques de suivi de formants. Avant d'aller plus loin il est bon de distinguer deux utilisations d'une analyse formantique :

la synthèse de parole Il s'agit de produire de la parole de bonne qualité. Le suivi de formants permet d'extraire certains paramètres qui permettent ensuite d'engendrer, en tenant compte de l'excitation, le signal de parole. La pertinence de l'analyse est mesurée par la qualité (en termes d'intelligibilité) de la parole produite et non par celle de l'interprétation acoustico-phonétique qu'elle permet.

la reconnaissance de la parole Dans ce cas il s'agit au contraire de disposer d'une analyse suffisamment fine pour permettre l'interprétation acoustique du suivi de formants. Pour la synthèse, l'extraction d'un nombre donné de formants (en général 3) suffit ; cette approche est impossible en reconnaissance car pour réussir une bonne analyse acoustique il faut disposer, simultanément, des formants liés à la cavité orale mais aussi de ceux liés aux cavités nasales.

Cette dichotomie ne suffit pas hélas à classer la masse des publications portant sur ce vaste sujet. Généralement un suivi de formants s'applique indifféremment à la synthèse comme à la reconnaissance.

Il reste qu'il est néanmoins possible de discerner deux directions parmi les travaux consacrés au suivi de formants. La première concerne les recherches entreprises pour élaborer de nouveaux modèles mathématiques d'extraction des formants ([Friedman 85], [Rigoll 88] ...) ou pour affiner les méthodes numériques existantes ([Willems 87], [Delsarte 86] ...). La seconde regroupe les algorithmes et les méthodes issues de l'intelligence artificielle qui permettent de construire ou d'étendre les trajectoires formantiques à partir des résultats de l'une des techniques d'extraction numérique.

Il nous semble illusoire d'espérer disposer d'une technique numérique permettant infailliblement de mettre à jour les trajectoires formantiques. En revanche, il est important de tirer parti au mieux des "morceaux" de suivi auxquels conduit directement une extraction numérique. Nous décrirons donc dans ce qui suit quelques techniques et algorithmes, faisant appel à des modèles numériques classiques, qui permettent de construire un suivi de formants.

4.1 Algorithme de McCandless

L'algorithme de McCandless [McCandless 74] travaille sur les pics du spectre LPC. L'idée générale est d'associer à un nombre donné de formants (les 4 premiers) les pics du spectre. Pour que l'appariement soit correct le suivi commence au milieu d'une zone voisée qui sert de point d'ancrage. L'appariement des pics d'une nouvelle trame est conduit de la manière suivante (EST_i désigne l'estimation fréquentielle du

formant i et est calculée grâce à la trame précédente. p_i désigne le pic i de la trame étudiée, S_i enfin désigne le pic candidat pour le formant i) :

1. extraction des fréquences et des amplitudes des 4 plus hauts pics entre 150 et 3400 Hz.
2. affectation à chaque formant i du pic le plus proche de l'estimation EST_i du formant ; un pic peut être associé à plusieurs formants.
3. si plusieurs formants partagent le même pic p, seul le couple (S_k,p) pour lequel l'estimation EST_k est la plus proche de p est conservé.
4. traiter les pics qui n'ont pas été associés à un formant ; s'il n'y en a pas l'algorithme reprend au point 5. On distingue les cas suivants :
   - si p_k n'est assigné à aucun candidat formant et S_k est vide, p_k est assigné à S_k. L'algorithme reprend au point 5.
   - si p_k n'est pas assigné et si S_k contient un pic et enfin si l'amplitude de p_k est plus petite que la demi-amplitude du pic affecté à S_k alors le pic p_k est abandonné et l'algorithme reprend au point 5.
   - si p_k n'est pas assigné et S_{k+1} est vide, le pic de S_k est affecté à S_{k+1} et p_k à S_k. L'algorithme reprend au point 5.
   - si p_k n'est pas assigné et S_{k-1} est vide, le pic de S_k est affecté à S_{k-1} et p_k à S_k. L'algorithme reprend au point 5.
5. traiter le cas de formants sans pic en calculant le spectre LPC pour un cercle complexe dont le rayon est inférieur à 1 (cf §3.1), et reprendre l'algorithme à la première étape. Le calcul du spectre n'a lieu que si le rayon est supérieur à 0,88.
6. prendre les valeurs trouvées pour chaque formant comme estimation de la trame suivante ; si un formant n'a pas de pic, son estimation est conservée.

Les estimations EST_i sont initialisées aux valeurs neutres des trois premiers formants, c'est-à-dire 500, 1 500 et 2 500 Hz.

Les traces construites doivent être lissées pour corriger les erreurs grossières, cela ne doit cependant pas dégrader les valeurs correctes et il faut donc limiter la portée du lissage. À cette fin, le lissage n'est déclenché que dans les régions où les trames incohérentes sont encadrées par au moins 4 trames (réparties autour de la trame incorrecte) dont les valeurs sont voisines.

McCandless note que les résultats sont corrects pour les sons vocaliques mais relativement imprévisibles en ce qui concerne les sons nasalisés. Cela appelle deux remarques :

- L'utilisation du calcul par prédiction linéaire est effectivement inadapté pour l'analyse des sons nasalisés ; il est donc normal que le suivi donne des résultats erronés.

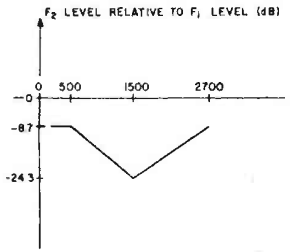


Figure 3.12 : Niveau relatif de F2 par rapport à F1

- Par ailleurs, en cas d'erreur, il semble difficile pour le niveau supérieur de corriger les formants car les solutions alternatives ont été perdues.

4.2 Algorithme de Schafer et Rabiner utilisant un lissage cepstral

L'algorithme de Rabiner et Schafer [Rabiner 69a] se place dans le cadre plus général d'un système de synthèse de parole. Rabiner et Schafer cherchent en effet à extraire le pitch (grâce à l'algorithme de Noll [Noll 67]) et les formants à partir d'un spectre lissé cepstralement. Le filtrage $l(n)$ (cf § 3.2) est cette fois de la forme suivante :

$$l(nT) = \begin{cases} 1 & si \quad \mid nT \mid < \tau_1 \\ 1/2(1 - \cos(\pi(nT - \tau_1)/\Delta\tau)) & si \quad \tau_1 \leq \mid nT \mid < \tau_1 + \Delta\tau \\ 0 & si \quad \mid nT \mid \geq \tau_1 + \Delta\tau \end{cases}$$

où $\tau_1 + \Delta\tau$ est plus petit que la plus petite période du pitch que l'on peut trouver dans la phrase.

Pour chaque formant, le domaine de fréquences est limité a priori à des intervalles déterminés expérimentalement (les valeurs suivantes ne sont valables que pour des locuteurs).

$$F1 \in [200\ Hz, 900\ Hz]$$
$$F2 \in [550\ Hz, 2\,700\ Hz]$$
$$F3 \in [1\,100\ Hz, 2\,950\ Hz]$$

La recherche des pics est contrôlée en fréquence par ces intervalles : mais elle est aussi contrôlée en amplitude. Un pic est un bon candidat si son amplitude est supérieure à une valeur dépendant du formant précédent. La figure 3.12, résultat de mesures et de calculs à partir d'un modèle de parole, représente le niveau relatif en dB du pic correspondant à F1. Elle s'interprète facilement en se rappelant que deux formants proches se renforcent mutuellement. Vers 500 Hz F1 et F2 sont très proches, donc le niveau de F2 est presque égal à celui de F1. Le minimum est atteint quand F2 est très éloigné de F1 et F3 simultanément, c'est-à-dire vers 1500 Hz. Le niveau de F2 remonte quand F2 se rapproche de F3.

La recherche des pics n'est pas toujours suffisante car il arrive que deux formants proches ne donnent lieu qu'à un seul pic sur le spectre. Pour séparer ces formants, le spectre n'est pas calculé sur le cercle unité mais à l'intérieur de ce cercle, sur une spirale ; il s'agit de la transformation Chirp-z ([Rabiner 69b]). Comme le contour que l'on choisit est plus proche des points représentant les formants, la résolution spectrale est meilleure et les pics des formants sont biens séparés.

Voici les grandes lignes de l'algorithme pour trouver le formant n, noté F_n :

- On choisit le pic le plus intense dans le domaine du formant n (D_{F_n}) ; ce domaine est éventuellement étendu à l'union $D_{F_n} \bigcup D_{F_{n-1}}$ si F_{n-1} appartient au domaine de F_n. Le pic n'est retenu que s'il est supérieur à la courbe de la figure 3.12.

- S'il n'existe pas de pic satisfaisant, on émet l'hypothèse que le formant est proche de F_{n-1} et que les deux formants ne correspondent qu'à un seul pic sur le spectre. Le spectre est donc recalculé avec la transformation Chirp-z dans l'intervalle $F_{n-1} \mp 450\ Hz$ (0-900 Hz dans le cas de F1). Quand deux pics sont trouvés, ils sont associés à F_{n-1} et F_n (dans le cas de F_1, F_{1-1} est le formant glottal). Si un seul pic est découvert, la valeur de F_{n-1} est conservée et F_n est fixé à $F_{n-1} + 200\ Hz$ (dans le cas de F_1, on donne à F_1 la valeur minimale que peut prendre F_1).

Aucune contrainte de continuité n'est imposée entre les formants de trames voisines. Une fois que l'estimation des formants sur tout le segment de parole est achevée, les trois trajectoires subissent un lissage non linéaire analogue à celui décrit dans l'algorithme de McCandless. Cela permet de corriger les erreurs grossières et de rétablir une certaine forme de continuité.

Les figures 3.13 et 3.14 montrent les résultats pour deux phrases. L'abandon de la contrainte de continuité, justifié par le fait qu'elle risque de suivre une erreur très longtemps, semble néanmoins pénaliser de manière assez sensible le suivi des trajectoires formantiques. Il faut cependant noter que certaines erreurs (notamment dans le cas de F3) sont dues à un suivi systématique, même quand il est impossible de trouver le formant, soit qu'il est trop faible, ce qui est le cas de F3 dans les exemples proposés, soit même qu'il soit tout à fait absent comme pendant les silences d'occlusives.

Outre l'abandon de la continuité, c'est aussi l'absence de prise en compte des formants nasaux (les pics qui leur correspondent ne disparaissent pourtant pas lors du lissage cepstral) qui peut expliquer quelques erreurs.

Il est possible d'améliorer la précision de la localisation des formants en adoptant une technique d'analyse par synthèse à partir du spectre lissé cepstralement comme le fait Olive [Olive 71]. Son idée est d'ajuster simultanément les trois formants du modèle qu'il a choisi en utilisant la méthode de Newton-Raphson.

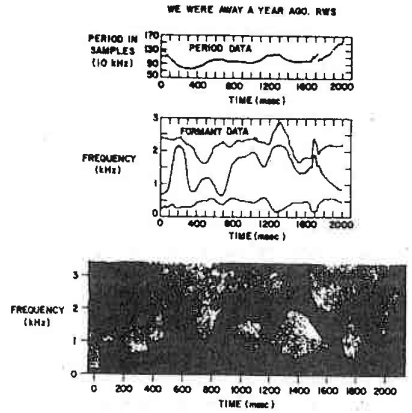


Figure 3.13 : Extraction des formants de la phrase "We were away a year ago"

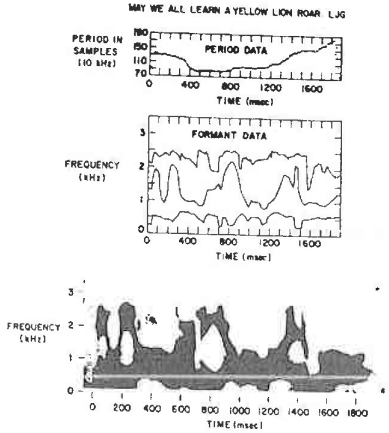


Figure 3.14: Extraction des formants de la phrase "May we all learn a yellow lion roar"

4.3 Utilisation de modèles de Markov pour le suivi de formants

Les modèles stochastiques et notamment les modèles de Markov cachés ont connu un grand succès en reconnaissance de parole et il est donc normal qu'ils aient été adaptés au problème du suivi de formants. Préconisée par Baker [Baker 75], c'est Kopec qui reprend l'idée et la concrétise (voir [Kopec 85]) sous une forme d'ailleurs un peu différente. La motivation est double : d'abord le suivi de formants que réalise un expert se prête bien à la modélisation par un processus markovien et puis les modèles de Markov éliminent les règles ad hoc utilisées dans la phase de décision de tout algorithme de suivi de formants.

Bien qu'ayant sensiblement étendu la portée de sa technique entre 85 et 86 ([Kopec 86a], [Kopec 86b]) Kopec conserve la même formulation en processus markovien que nous décrivons maintenant et qui résume [Kopec 86b].

Il est important de préciser que le modèle que nous allons présenter dans ce paragraphe est destiné à détecter et à suivre un seul formant. Cela signifie qu'il faut trois modèles pour obtenir le suivi des trois premiers formants.

Un modèle de Markov est représenté par un 5-uplet $\mathcal{M} = (\mathcal{Q}, \mathcal{V}, \Pi, \mathcal{A}, \mathcal{B})$:

- un ensemble d'états $\mathcal{Q} = \{q_i / i = 0 \ldots N-1\}$. Pour le suivi de F2, N vaut 40. Les états q_i pour $1 \leq i \leq 39$ sont des intervalles de fréquences centrés en $i \times 100 - 50$ Hz et de 100 Hz d'amplitude. q_0 est un état particulier qui représente la non détection du formant.

- un ensemble de symboles $\mathcal{V} = \{v_i / i = 0 \ldots M-1\}$ Les symboles v_i sont les vecteurs de coefficients LPC, prototypes obtenus par quantification vectorielle et constituant le dictionnaire. N est le nombre de vecteurs de ce dictionnaire.

- un vecteur de probabilités de départ pour chaque état $\Pi = [\pi_i / i = 0 \ldots N-1]$ $\pi_i = P(q_i \text{ à } t = 0)$

- une matrice de transitions d'état à état $\mathcal{A} = [a_{ij} / i = 0 \ldots N-1, j = 0 \ldots N-1]$ avec $a_{ij} = P(q_i \text{ à } t, q_j \text{ à } t+1)$

- une matrice des probabilités des vecteurs observés $\mathcal{B} = [b_{ij} / i = 0 \ldots N-1, j = 0 \ldots M-1]$ $b_{ij} = P(v_j \text{ à } t / q_i \text{ à } t)$

Dans le premier article ([Kopec 85]) Kopec utilise deux modèles de Markov, l'un du second ordre pour la détection de la présence d'un formant et le second du premier ordre pour l'estimation de la fréquence.

Dans [Kopec 86b] ces deux étapes sont fusionnées en un seul modèle du premier ordre en ajoutant aux états intervalles de fréquence, l'état particulier q_0, indiquant l'absence de formant. La détection et l'estimation d'un formant peuvent alors être formulées simultanément en utilisant $P(s_t = q_i / \mathcal{O})$ ou probabilité conditionnelle que l'état à l'instant t soit q_i étant donnée la séquence d'observations $\mathcal{O}$ générée par le

modèle, $\mathcal{O}$ est une suite de symboles appartenant à $\mathcal{V}$. La probabilité conditionnelle de la présence du formant est :

$$\tilde{\gamma}(t) = \sum_{q_i/i\neq 0} P(s_t = q_i/\mathcal{O})$$

L'estimation de la fréquence du formant quand il est détecté s'exprime par :

$$\tilde{F}(t) = \frac{\sum_{q_i/i\neq 0} F(q_i)P(s_t = q_i/\mathcal{O})}{\sum_{q_i\neq 0} P(s_t = q_i/\mathcal{O})}$$

$F(q_i)$ désigne la fréquence centrale de l'état q_i. Il faut évidemment fixer un seuil γ_{min} au-delà duquel le formant est declaré présent. La probabilité $P(s_t = q_i/\mathcal{O})$ est calculée par l'algorithme forward-backward. Nous adoptons les notations de [Levinson 83]. Par définition de la probabilité conditionnelle :

$$P(s_t = q_i/\mathcal{O}, \mathcal{M}) = \frac{P(s_t = q_i \text{ et } \mathcal{O}/\mathcal{M})}{P(\mathcal{O}/\mathcal{M})} \tag{3.3}$$

$$P(s_t = q_i \text{ et } \mathcal{O}/\mathcal{M}) = \alpha_t(i)\beta_t(i)$$

avec :

$$\mathcal{O} = O_0 O_1 \dots O_{T-1}$$

$$\alpha_t(i) = P(O_0 \dots O_t \text{ et } s_t = q_i/\mathcal{M})$$

$$\beta_t(i) = P(O_{t+1} \dots O_{T-1}/s_t = q_i, \mathcal{M})$$

$\alpha_t(i)$ et $\beta_t(i)$ sont les probabilités "forward" et "backward".
De la même façon nous avons :

$$P(\mathcal{O}/\mathcal{M}) = \sum_{i=0}^{N-1} P(O_0 \dots O_{T-1} \text{ et } s_{T-1} = q_i/\mathcal{M}) = \sum_{i=0}^{N-1} \alpha_{T-1}(i)$$

L'apprentissage est réalisé par un simple comptage des événements, à partir d'un suivi de formants manuel. Les coefficients b_{ij}, par exemple sont obtenus par :

$$b_{ij} = \frac{n_{ij} + \epsilon_i}{\sum_{m=0}^{M-1}(n_{im} + \epsilon_i)}$$

où n_{ij} est le nombre des trames étiquetées avec l'état i et le vecteur de quantification v_j.

$$\epsilon_i = 10^{-6} \times \sum_{m=0}^{M-1} n_{im}$$

ϵ_i est un biais qui évite les probabilités nulles.

À partir de cet algorithme de base, Kopec a essayé plusieurs techniques de suivi de formants.

Suivi simultané de plusieurs formants

Un état du modèle devient une configuration de formants (exemple : (F1 = 300Hz, F2 = 2 000Hz, F3 = 3 000Hz)). Cela conduit à un modèle avec un très grand nombre d'états et donc à une augmentation importante du volume de calculs. Par ailleurs, les résultats montrent peu d'améliorations, sauf en ce qui concerne la précision de la fréquence de F1.

Élagage des calculs d'estimation d'un formant

Un vecteur LPC, étant lié à la forme du filtre modélisant le conduit vocal, exclut certains états du modèle.
Exemple :
Si un vecteur LPC représente un filtre dont les pics sont à 300, 800 et 2 000 Hz, il est impossible que lors du suivi de F2, l'état du modèle puisse être q_{15} (c'est-à-dire que F2 appartienne à l'intervalle 1 450-1 550 Hz).

Cette constatation permet d'éviter le calcul systématique des probabilités conditionnelles de la formule 3.3. b_{ik} peut s'écrire en appliquant la formule de Bayes :

$$b_{jk} = P(v_k/q_j) = \frac{P(q_j/v_k)P(v_k)}{P(q_j)}$$

$P_t(q_i/O_t) = 0$ implique donc $b_{iO_t} = 0$
Comme

$$\alpha_t(i) = \sum_{j=0}^{N-1} \alpha_{t-1}(j)a_{ji}b_{iO_t}$$

on obtient donc :

$$\alpha_t(i) = 0$$

or :

$$P(\mathcal{O} \text{ et } s_t = q_i) = \alpha_t(i)\beta_t(i)$$

donc :

$$P(\mathcal{O} \text{ et } s_i = q_i) = 0$$

soit encore :

$$P(s_t = q_i/\mathcal{O}) = P(\mathcal{O} \text{ et } s_t = q_I)/P(\mathcal{O}) = 0$$

Pendant l'apprentissage, les couples (k, j) pour lesquels $p(v_k/q_j) = 0$, sont mémorisés ; lors de l'évaluation d'un formant, seuls les $P(s_t = q_i/\mathcal{O})$ pour lesquels on est assuré que $b_{iO_t} \neq 0$ sont calculés. L'évaluation des coefficients b_{ij} et a_{ij} est évidemment faite sans le biais ϵ_i.

Cette simplification du modèle ne dégrade que légèrement les résultats du suivi de formants.

Suppression des contraintes de continuité

On suppose dans ce cas que $a_{ij} = \pi_j$, c'est-à-dire que la probabilité de la transition de q_i à q_j est la probabilité a priori de q_j. Par conséquent, en appliquant la formule de Bayes nous obtenons :

$$P(s_t = q_i/\mathcal{O}) = P(s_t = q_i/O_t) = P(O_t/q_i)P(q_i)/P(O_t)$$

Le calcul de $P(s_t = q_i/O)$ ne nécessite donc plus les récurrences liées au calcul de $\alpha_t(i)$. Cette hypothèse revient à supposer qu'il n'y a aucune contrainte de continuité dans le suivi de formants.

Cette simplification grossière du modèle augmente la sensibilité en ce qui concerne la détection des formants mais augmente aussi l'erreur quadratique moyenne sur l'évaluation fréquentielle.

Extraction du formant nasal

Les travaux de Kopec ne portent que sur la détection et l'estimation des trois premiers formants oraux. Bailly et Liu [Bailly 88] proposent une extension de cette approche permettant de détecter et d'évaluer le formant nasal. L'apprentissage est effectué sur le même corpus que celui utilisé pour les formants oraux. Les résultats, hélas, ne sont pas très bons : s'il n'y a aucune fausse détection du formant nasal, il n'est détecté que dans 16% des cas où il est effectivement présent.

Discussion

L'un des principaux arguments militant en faveur des modèles de Markov pour obtenir les trajectoires formantiques est que cela évite le recours à des règles ad hoc comme c'est le cas pour les algorithmes traditionnels. Ces règles représentent en fait une somme de connaissances, concernant le suivi de formants, que le chercheur inclut dans l'algorithme. Ne disposant pas de telles connaissances, le modèle de Markov réalisant le suivi doit résulter d'un apprentissage aussi complet que possible (avec toutes les configurations formantiques possibles), d'où un énorme travail d'étiquetage manuel de corpus de parole. Par ailleurs, un tel algorithme repousse l'interprétation acoustique et les éventuelles corrections qu'elle apporte après que les trajectoires sont estimées. Or en reconnaissance de la parole, un dialogue entre plusieurs experts (dans ce cas : algorithme de suivi de formants et connaissances acoustiques) a toujours montré son intérêt.

4.4 Algorithme de Laface

L'algorithme de Laface [Laface 80] est sans doute l'un des plus élaborés parmi les algorithmes de suivi de formants. Faisant largement appel aux techniques d'intelligence artificielle, il est de plus spécifiquement conçu pour la reconnaissance de parole. Comme l'algorithme que nous avons mis au point et que nous décrirons juste

après, il opère en deux étapes : d'abord la construction d'arcs élémentaires, puis la propagation des solutions partielles à l'ensemble du segment de parole à étudier. Nous allons décrire les grandes lignes de cet algorithme et les techniques adoptées pour résoudre chacun des deux sous-problèmes.

Cet algorithme est destiné à suivre les formants sur les suites (Voyelle, Sonante, Voyelle) ; les suites sont automatiquement présegmentées V_1CV_2 et le suivi étudie séparément les trois segments avant de fournir une solution globale. Les données de départ sont des intervalles de fréquences contenant un pic de FFT et une racine de LPC s'il en existe une dans cet intervalle.

Les arcs élémentaires sont construits en respectant une contrainte de continuité : deux intervalles fréquentiels de deux trames consécutives appartiennent au même arc si l'intersection de ces intervalles n'est pas vide. Les points correspondant à la fusion de plusieurs arcs ou au dédoublement d'un arc sont mémorisés. Quand la construction des arcs est terminée, chaque point singulier (fusion ou dédoublement) est étudié et seul le meilleur couple (arc entrant, arc sortant) est conservé. Le critère de qualité repose sur la longueur, l'amplitude des pics et les discontinuités de pente. Cette démarche est comparable à celle que nous avions adoptée pour réaliser notre premier algorithme de suivi de formants, sommairement décrit dans [Fohr 88].

On construit ensuite, pour chaque formant, les deux meilleurs chemins composés d'arcs élémentaires compatibles avec le domaine fréquentiel du formant. La qualité de ces chemins est évaluée par la combinaison de critères flous faisant notamment intervenir le nombre de trous dans le chemin, son niveau de continuité en fréquence et en amplitude.

Les chemins correspondant aux trois premiers formants sont alors combinés pour fournir un triplet (F1,F2,F3) solution du suivi de formants. Les deux meilleures solutions (en combinant les scores intermédiaires) sont retenues. Cette élaboration du suivi global (F1,F2,F3) ne porte que sur les deux noyaux vocaliques V_1 et V_2. La recherche du suivi pour le segment consonantique est quant à elle guidée par la connaissance des triplets (F1,F2,F3) de V_1 et V_2 qui entourent ce segment.

Finalement les suivis des trois segments V_1, C, V_2 sont combinés en cherchant à réduire les incohérences aux frontières. Le système propose quatre solutions dont le score dépend des scores partiels et de la qualité des raccordements de formants aux frontières des segments.

À notre avis, il est dommage d'arrêter le suivi aux frontières d'une segmentation automatique car cela risque de créer des problèmes de raccordement artificiels, d'autant plus qu'une segmentation automatique est souvent imprécise justement au voisinage d'une sonante. Par ailleurs il est très difficile d'assurer que les scores définitifs, résultats de trois combinaisons successives de critères flous caractérisent correctement les solutions dont ils devraient mesurer la qualité.

5 Algorithme de suivi de formants proposé

Selon nous, le suivi de formants doit construire des trajectoires de pics d'énergie que le module de reconnaissance peut interprèter, durant son raisonnement, en termes de formants. En effet, étiqueter trop tôt les concentrations d'énergie peut orienter le décodage vers une mauvaise solution, par exemple pour les sons nasalisés.

L'algorithme que nous avons mis au point essaie donc de construire des trajectoires acoustiquement cohérentes et les plus proches possible des formants, mais sans les étiqueter.

Nous présentons d'abord les données de départ que peut traiter l'algorithme. L'étape de construction des trajectoires est scindée en deux. Nous présentons, dans un premier temps, les techniques d'image nous permettant d'extraire les trajectoires élémentaires. Ensuite nous décrivons la manière dont sont engendrés les différents types de raccordement entre trajectoires élémentaires.

5.1 Prétraitement

Notre algorithme peut travailler sur des données d'origine LPC ou cepstrales. Traditionnellement, en effet, les techniques de suivi font appel au codage LPC. Nous utilisons en plus, le lissage cepstral qui permet une approche correcte des sons nasalisés comme nous l'avons vu au §3.2.

Précisons, avant d'aller plus loin, que nous traitons de la parole échantillonnée à 16 kHz et subissant une préemphase de 6 dB par octave.

Données LPC

Toutes les 2 ms un vecteur LPC de 20 coefficients est calculé par la méthode d'autocorrélation. Nous avons adopté un déplacement temporel très fin pour être assuré que les transitions rapides ne nous échappent pas. Par ailleurs, cela permet d'avoir des résultats légèrement redondants, donc moins sensibles aux éventuels parasites de courte durée.

Les racines et leur largeur de bande sont obtenues par la méthode de Bairstow. Nous retenons les racines dont la largeur de bande est inférieure à 600 Hz ; les racines dont la largeur de bande est grande (supérieure à 300 Hz) sont généralement éliminées dans la suite de l'algorithme car elles représentent des pics isolés et sans grande signification physique.

Comme nous ne nous intéressons qu'aux trois premiers formants, nous ne considérons que les racines de fréquence inférieure à 4500 Hz.

Données cepstrales

Comme pour le codage LPC, nous calculons un spectre lissé cepstralement toutes les 2 ms. Le calcul de cepstres est effectué sur 256 points, avec une fenêtre de hamming de 16 ms.

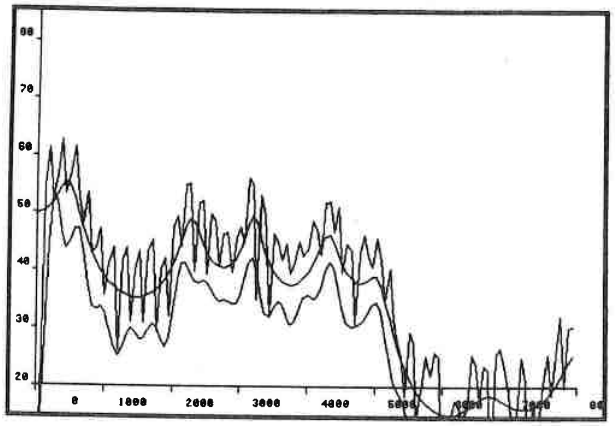


Figure 3.15: Spectre d'un /ai/ : la courbe inférieure est un spectre lissé cepstralement

Pour éviter que le spectre lissé ne porte trop de petits pics sans signification formantique comme le montre la figure 3.15, nous adaptons le lissage décrit au § 4.2 à la fréquence fondamentale de chaque locuteur. Expérimentalement nous avons trouvé qu'il convenait de choisir les coefficients $\Delta\tau$ et τ_1 comme le tableau suivant l'indique.

pitch	< 100 Hz	< 120 Hz	< 140 Hz	< 160 Hz	< 180 Hz	< 200 Hz	< 220 Hz	≥ 220 Hz
$\Delta\tau$	22	20	20	20	20	19	18	17
τ_1	22	22	20	19	18	17	16	15

Pour connaître avec plus de précision l'emplacement des pics du spectre, nous effectuons une interpolation parabolique sur trois points (le pic, et ses deux voisins immédiats). Cela permet, comme le montre la figure 3.16, de réduire les sauts de fréquence entre deux pics d'un même formant, sur deux spectres consécutifs. Enfin nous nous limitons aux pics de fréquence inférieure à 4500 Hz, d'amplitude supérieure à 5 dB et visibles sur le spectrogramme.

Comme nous utilisons par la suite des techniques de traitement d'image, les résultats sont stockés sous forme d'image binaire. L'abscisse représente le temps et l'ordonnée la fréquence. Un point vaut 1 si ses coordonnées sont celles d'une racine (dans le cas des LPC) ou d'un pic (dans le cas du lissage cepstral).

Construction des pistes élémentaires

Raccordement des points voisins L'objectif de cette étape est de connecter les points très proches et donc ne nécessitant pas de connaissances particulières.

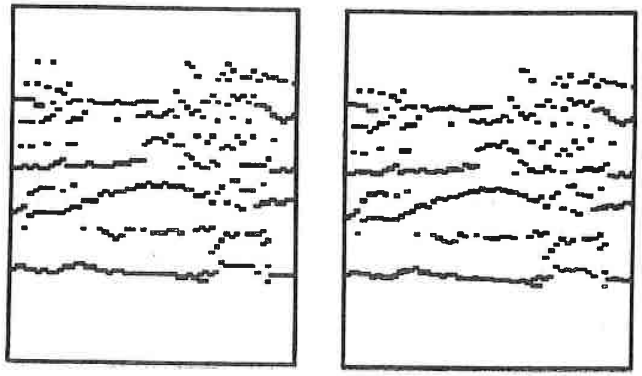


Figure 3.16 : Pics de cepstres non interpolés à gauche et interpolés à droite

Nous utilisons pour cela une technique de base de morphologie mathématique : la dilatation par un élément structurant rectangulaire. Sans entrer dans les détails (consulter par exemple [Serra 86] faisant partie d'un groupe d'articles sur le sujet dans la même revue), précisons que la morphologie mathématique est une famille de techniques de traitement d'image considérant un objet comme un ensemble de pixels et opérant donc des transformations d'images exprimées en termes d'ensembles de pixels.

La dilatation est ainsi l'une des deux transformations de base de morphologie mathématique. Étant donnés :

- un ensemble de points X, dans notre cas les points représentent une racine LPC ou un pic de spectre.
- un élément structurant $B(x)$; B est un ensemble centré autour du point x. Ici B est un rectangle de 7 points de large et de 5 points de haut ; cela correspond à 7*2ms dans le domaine temporel et 5*20.5 Hz dans le domaine fréquentiel.

La dilatation de X par $B(x)$ est l'ensemble de tous les points de l'image tels que $B(x) \bigcap X \neq \emptyset$.

L'application de cette dilatation relie les points qui appartiennent à la même trajectoire formantique comme le montre la figure 3.17. Les points isolés, issus soit d'un bruit, soit d'un pic sélectionné abusivement, le restent. L'étape suivante permet de les éliminer.

Élimination des points isolés Les formants donnent lieu, grâce à la transformation précédente, à des composantes connexes (groupes de points connectés entre eux) de taille importante. Pour éliminer les points isolés, il suffit donc de mettre en évidence les composantes connexes et ensuite de les filtrer suivant leur taille.

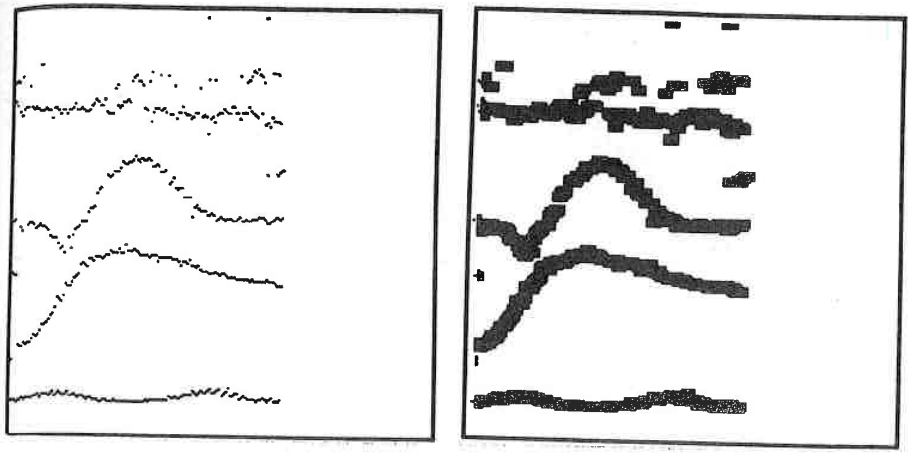


Figure 3.17 : Image des pics cepstraux et image dilatée

La recherche des composantes connexes ne nécessite qu'une seule passe sur l'image. Quand un point noir est rencontré :

- une nouvelle composante connexe est créée si c'est un point isolé,
- si le point jouxte une seule composante, il est ajouté à cette composante,
- si le point jouxte deux composantes, ces composantes sont fusionnées et le point est ajouté au résultat de cette fusion.

Il est important que les composantes connexes soient simplement connexes, c'est-à-dire qu'elles ne présentent aucun "trou". En effet, comme l'étape suivante est la squelettisation de l'image, un trou dans une composante provoque un dédoublement de la trajectoire ce qui risque d'être difficile à corriger.

Le remplissage des trous, s'effectue de la manière suivante :

- inversion de l'image,
- invasion de l'image à partir des bords et en s'arrêtant au bord des composantes connexes ; les trous de la couleur du fond de l'image ne sont donc pas atteints,
- inversion de l'image.

Squelettisation Il existe de nombreux algorithmes de squelettisation et ce sont donc essentiellement des raisons de rapidité et de facilité de manipulation des résultats qui ont motivé le choix. Nous avons retenu et implanté l'algorithme d'Arcelli

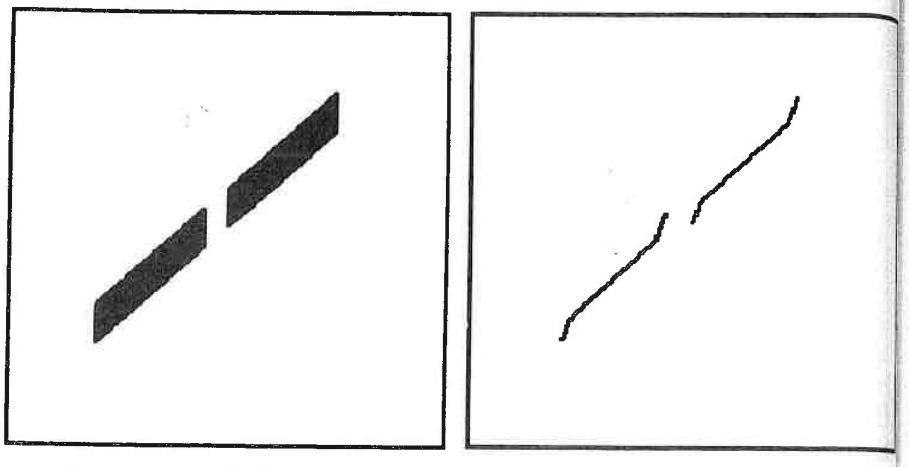


Figure 3.18: Squelettisation compliquant nettement le raccordement des deux branches

et Sanniti Di Baja [Arcelli 85] parce qu'il était beaucoup plus rapide (6 secondes au lieu de 60 pour nos images) que celui dont nous disposions dans notre équipe. De plus cet algorithme permet de mémoriser les extrémités des lignes et leurs points de jonction. Son seul défaut important est de ne pas toujours suivre la direction principale d'une branche au voisinage des extrémités (figure 3.18). Cela peut nettement compromettre le raccordement de deux branches qui devraient se raccorder très naturellement comme le montre la figure 3.18. L'algorithme récent de Baruch [Baruch 88] permettrait sans doute d'éviter ce genre de problème tout en conservant une rapidité suffisante.

Le squelette est élagué en ne conservant que les branches de longueur suffisante. Il est très important, pour la suite, que les branches du squelette soient d'épaisseur 1, ce qui permet de les parcourir très facilement et sans ambiguïté.

Construction des trajectoires La suite de l'algorithme de suivi de formants nécessite des données plus abstraites que le sont les points du squelette. Nous allons donc chercher à approcher le squelette par un ensemble de segments de droite. Les segments sont organisés en pistes qui sont par la suite interprétées en termes de trajectoires formantiques.

Nous avons donc imposé les règles suivantes lors du parcours du squelette qui accompagne la création des segments[3] :

[3]Il aurait évidemment été possible de construire un système expert pour prendre en compte ces règles de construction ; étant donné le très faible nombre de règles nous avons préféré conserver le formalisme procédural

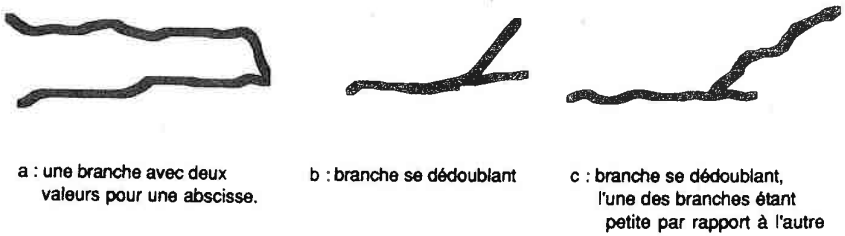


Figure 3.19 : Différentes configurations de branches

- une branche avec deux valeurs pour une seule abscisse (figure 3.19a) ne peut pas représenter un seul formant ; il faut donc la couper en deux.

- quand une branche se dédouble (figure 3.19b) elle est prolongée par la branche sortante s'écartant le moins de la direction d'entrée.

- la règle précédente ne s'applique pas si l'une des deux branches sortantes est beaucoup plus courte que l'autre comme c'est le cas sur la figure 3.19c.

La construction des pistes commence par un cheminement dans le squelette. Ce cheminement est destiné à obtenir des pistes aussi cohérentes que possible avec la notion de formant. Ces pistes sont ensuite approchées par une suite de segments grâce à l'approximation polygonale. Voici l'algorithme correspondant à la première partie.

```
construitPistes()
    nPiste ← 0
    pour chaque point P extrémité d'une branche de squelette faire
        si P n'appartient à aucune piste alors
            piste[nPiste] ← exploreBranche(P)
            nPiste ← nPiste + 1
        finsi
    fin pour
fin construitPistes

exploreBranche(P) /* exploration d'une branche commençant par le point P */
    p ← P
    segment ← creerSegment()
    tant que p n'est ni une extrémité ni une jonction repéter
        segment ← ajouter(segment, p)
        marquer(p)
        p ← suivant(p)
    fin tant que
    marquer(p)
    segment ← ajouter(segment, p)
    segment ← fermerSegment(segment)
    si p est un point de jonction alors
        retourner(concatène(segment, exploreBranche(meilleurPointDeSortie(p))))
        /* meilleur au sens des deux dernières règles */
    sinon retourner(segment) fin si
fin exploreBranche
```

Il faut ajouter à l'algorithme classique d'approximation polygonale le contrôle des directions pour éviter qu'une piste "fasse demi-tour" (comme c'est le cas sur la figure 3.19a) ce qui donne :


```
approchePiste(P1,P2)
    /* la piste qu'il faut approcher
    est décrite par ses points extrêmes */
    listePointsCoupure ← listeVide()
    approximationPolygonale(P1,P2)
    couperPiste(P1, P2, listePointsCoupure)
fin approchePiste

approximationPolygonale(A,B)
    C / pour tout point P de la sous-piste(A,B)
        d(P, droite(A,B)) < d(C, droite(A,B))
    si d(C, droite(A,B)) > ε alors
        /* ε est la distance au dessous de
        laquelle il n'y a pas d'approximation polygonale */
        si AC.CB < 0
            et bissectrice(A,C,B) ≠ π/2et − π/2 alors
            /* la piste retourne sur elle même */
            listePointsCoupure ← ajouter(listePointsCoupure, C)
        fin si
        insérer C entre A et B
        approximationPolygonale(A,C)
        approximationPolygonale(C,B)
    fin si
fin approximationPolygonale
```

Raisonnement géométrique

Pour que les pistes construites jusqu'à maintenant soient facilement interprétables en termes de formants, il faut :

- raccorder celles qui se prolongent naturellement,
- rechercher les éventuels croisements de formants,
- pénaliser les pistes qui ne semblent pas cohérentes avec le reste du suivi.

Nous émettons successivement, pour l'ensemble des pistes du suivi, les hypothèses correspondant à ces trois situations. Cela évite tout conflit de déclenchement des règles et respecte l'importance des hypothèses engendrées.
Exemple :
Si une configuration de pistes permet de déclencher une règle de raccordement et une règle de croisement de formants, c'est la première règle qui sera effectivement appliquée car les résultats auxquelles elle conduit sont les plus sûrs.

Raccordement de pistes Ce type d'hypothèse portent sur des pistes dont l'éloignement des extrémités est limité à 60 ms ou 400 Hz.

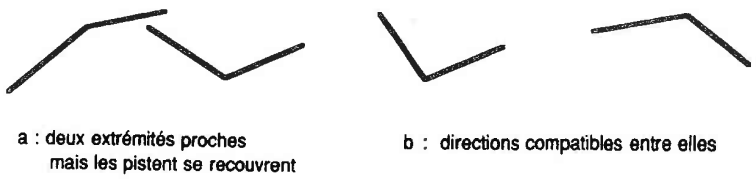


Figure 3.20 : Différentes configurations de raccordement

Notons d'abord qu'une extrémité de piste est décrite non seulement par ses coordonnées mais aussi par sa direction de sortie. Deux pistes peuvent être raccordées si elles vérifient l'une ou l'autre des conditions suivantes :

- les extrémités sont très proches ; aucune contrainte de direction n'est imposée, hormis le fait que deux pistes définies simultanément pour un même intervalle de temps ne peuvent être raccordées comme le montre la figure 3.20a.
- les directions de sortie sont compatibles entre elles. Cette compatibilité est estimée en évaluant la somme des angles de changement de direction lors du raccordement (fig 3.20).

Comme il s'agit de rétablir la continuité d'un formant perdue par une mauvaise détection de pics spectraux ou par un parasite, nous imposons que le raccordement ne coupe pas d'autres pistes.

Quand plusieurs hypothèses de cette nature peuvent être émises pour un même point, c'est l'hypothèse donnant lieu au raccordement le plus court qui est conservée.

Recherche des croisements de formants Il s'agit cette fois de prendre en compte les croisements de formants les plus manifestes, c'est-à-dire ceux pour lesquels l'une des pentes est forte (ou apparaît comme telle sur le spectrogramme). Ce type d'hypothèse porte sur des pistes dont l'éloignement en fréquence est important mais dont l'éloignement temporel est assez faible.

Comme une erreur de pente est plus fréquente avec une piste de pente forte, nous étudions comme points de raccordement les deux premiers et les deux derniers points de l'approximation polygonale d'une piste.

À la différence des hypothèses décrites au paragraphe précédent, nous admettons que le segment engendré par l'hypothèse coupe une piste existante, ce qui correspond à un changement d'affiliation des cavités résonnantes.

En cas de conflit, s'il y a deux pistes possibles, c'est celle conduisant à la plus faible déviation angulaire qui est retenue ; si les deux hypothèses portent sur la même piste, c'est l'hypothèse la plus courte qui est maintenue.

Pénalisation de pistes Les hypothèses de raccordement entre deux pistes élémentaires peuvent être considérées comme un renforcement mutuel de leur vraisemblance en termes d'information formantique. En effet les raccordements effectués respectent les contraintes acoustiques qui sont imposées aux trajectoires formantiques. Plus une piste, étendue éventuellement par l'émission d'hypothèses de raccordements, est longue, plus son pouvoir explicatif et sa vraisemblance sont importants.

Pour que cette approche soit complète, il faut donc aussi examiner les petites pistes élémentaires et éventuellement les pénaliser. Nous avons distingué les cas suivants :

- piste élémentaire isolée : elle est pénalisée
- piste élémentaire dont la direction n'est pas cohérente avec une piste voisine beaucoup plus longue : elle est pénalisée
- piste élémentaire qui renforce une hypothèse : les pistes voisines sont examinées et peuvent être corrigées pour prendre en compte la petite piste.

Nous avons choisi de ne pas utiliser de scores complexes, une piste qui doit être pénalisée reçoit une simple marque de pénalisation. Le module de décodage acoustico-phonétique peut éventuellement remettre en cause cette pénalisation et inclure cette piste dans la solution qu'il construit.

5.2 Résultats

Présentation des résultats

Pour évaluer l'algorithme que nous venons de décrire, nous l'avons appliqué à la séquence /waja/ pour 7 locuteurs et 5 locutrices. Nous avons choisi cette séquence du corpus GRECO "La bise et le soleil" pour les raisons suivantes :

- Les formants 2 et 3 échangent leur cavité d'affiliation entre /w/ et /j/. F2 est lié d'abord à la cavité du conduit vocal antérieure à la constriction puis à la cavité postérieure à la constriction.
- Les transitions formantiques sont suffisamment fortes pour qu'il ne soit pas toujours possible de suivre les formants par continuité.
- Le suivi est difficile pour les voix dont la fréquence fondamentale est élevée, c'est-à-dire les locutrices.

Pour chaque locuteur, nous présentons :

- le spectrogramme jusqu'à 6 000 Hz et la fréquence fondamentale du locuteur,
- l'extraction des racines de LPC inférieures à 4 500 Hz et le suivi réalisé à partir de ces données (à droite du spectrogramme),

- l'extraction des pics de spectrogramme lissé cepstralement et le suivi correspondant.

Les segments de suivi ont une épaisseur normale quand ils sont issus directement de l'approximation polygonale du squelette ou de raccordements de pistes décrits au §5.1. Les segments dont l'épaisseur est plus petite sont soit les segments correspondant aux croisements de formants ou aux raccordements à pente forte, soit les segments pénalisés. Il n'y a pas de risque de confusion entre ces deux types de segments puisqu'un segment pénalisé est lié au plus à un seul segment d'épaisseur normale.

Les résultats sont présentés en annexe.

Discussion

La confrontation des deux techniques d'extraction des pics (LPC et lissage cepstral) impose immédiatement deux remarques :

- Les performances du calcul LPC et du lissage cepstral sont à peu près équivalentes en ce qui concerne les locuteurs. Il apparaît clairement en revanche, que le fait d'adapter le lissage cepstral à la fréquence du fondamental donne des résultats sensiblement meilleurs pour toutes les locutrices.
- Par ailleurs, les échanges de cavité d'affiliation entre formants apparaissent plus clairement avec le lissage cepstral qu'avec le calcul LPC. Cette différence est plus marquée chez les locutrices, pour lesquelles seul le lissage cepstral permet de suivre correctement une cavité résonnante.

Comme nous l'avons déja fait remarquer, l'algorithme de squelettisation provoque quelques problèmes de raccordement. S'ajoutent à ces problèmes des erreurs d'élagage aux extrémités. Quand la piste élémentaire se dédouble à une extrémité, il est possible que la branche prolongeant "naturellement" la piste ait une longueur trop faible et soit donc détruite. La branche restante peut être mal orientée comme c'est visiblement le cas pour le locuteur BP au maximum de F3 dans le suivi cepstral.

Les raccordements entre pistes élémentaires et les pénalisations de segments qui ne s'insèrent pas dans le suivi global sont effectués correctement. Toutefois, quand le nombre de pistes élémentaires augmente, il devient presque impossible de choisir les pistes à raccorder et donc de construire un suivi global cohérent. L'algorithme de raccordement a ainsi fusionné pour le suivi LPC de LT deux petites pistes proches de F1 qui ne correspondent pas à un formant. Dans le cas du suivi LPC de JI il est de même bien difficile de savoir quelles pistes élémentaires choisir pour construire correctement F2.

Que ferait l'expert phonéticien ?

Il convient d'abord de remarquer que le phonéticien travaille sur un spectrogramme à large bande. Le filtrage inhérent à ce type de spectrogramme ne lui permet pas de donner des valeurs très précises des formants ; il lui permet en revanche

d'avoir une bonne vue générale des différents phénomènes de la parole (du moins pour les locuteurs).

Nos données de départ suffisent en général (du moins pour les pics de spectres lissés cepstralement) à un expert pour construire le suivi de formants. En se reportant aux planches de résultats (avec le lissage cepstral) il est possible de discerner deux types d'erreurs :

- les pistes détectées et qui ne correspondent pas à un formant,
- les pistes qui auraient dû être raccordées et qui ne l'ont pas été.

Même si ces erreurs ne sont pas rédhibitoires, il n'empêche que l'expert ne les aurait sans doute pas commises. Sa supériorité s'explique surtout par son approche globale : il ne fait pas que suivre un formant, mais il le replonge dans le suivi de formants qu'il élabore et même plus généralement dans le processus d'identification des sons. Nous verrons dans la partie suivante comment interpréter en termes de formants les pistes du suivi, ce qui permet notamment de supprimer les pistes qui n'apportent aucune information formantique.

6 Interprétation de pistes en termes de formants

Dans la partie précédente nous avons décrit un algorithme de suivi de formants ou plus exactement de suivi de lignes de pics (qu'il s'agisse de pics de spectres LPC ou de pics de spectres lissés cepstralement). Bien que toutes les contraintes que nous ayons imposées pour connecter les pistes élémentaires soient destinées à favoriser la construction des trajectoires formantiques, il n'est pas possible d'assurer qu'il s'agisse bien des formants. Le problème est identique avec l'autre algorithme de suivi de lignes de pics que nous avons développé [Fohr 88].Comme une partie importante des connaissancess acoustico-phonétiques fait appel aux formants, il est indispensable d'interpréter les pistes en termes de formants. La figure 3.21 met en évidence certains des problèmes d'interprétation rencontrés : [4]

1. une ligne de pics d'assez faible énergie a été détectée mais elle ne représente aucun formant,
2. la résolution du lissage cepstral ou du codage LPC n'est pas suffisante pour séparer deux formants qui ne donnent lieu qu'à une seule piste,
3. une erreur de suivi ou un parasite conduit localement à deux pistes pour un seul formant,
4. l'énergie d'un formant n'est pas suffisante pour que ce dernier soit détecté,
5. il n'est pas possible d'interpréter les pistes en termes de formants, soit qu'il y ait une erreur soit qu'il s'agisse d'un segment de parole nasalisée impossible à expliquer avec les seuls formants oraux.

[4] Les items renvoient aux configurations de pistes correspondantes sur la figure.

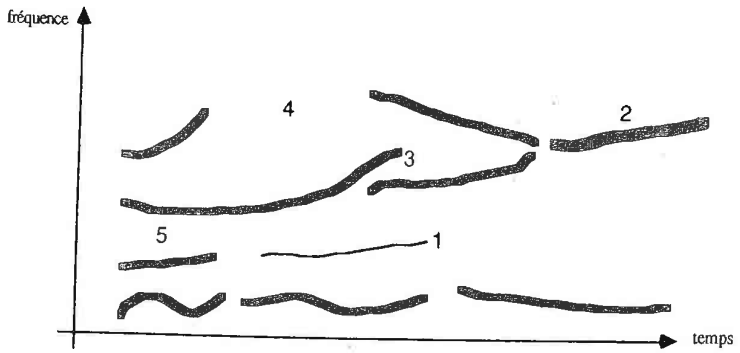


Figure 3.21: Résultat schématique du suivi de pistes (plus les pistes sont larges plus elles concentrent d'énergie)

Interpréter les pistes ne suffit pas, il faut aussi pouvoir estimer la qualité de cette interprétation : cela permet de pénaliser les déductions acoustico-phonétiques reposant sur une mauvaise interprétation formantique ; cela permet aussi de déclencher d'autres tentatives de suivi (avec des paramètres plus adaptés au segment de parole à analyser) et de ne retenir que les meilleurs résultats.

Il faut noter que "qualité" recouvre plus que performance de l'algorithme d'interprétation : interpréter un segment nasalisé en ne prenant en compte que les formants oraux doit conduire à une faible "qualité" quelle que soit l'efficacité de l'algorithme.

Ces préambules terminés, voici le principe général de l'interprétation formantique que nous allons présenter : expliquer le plus d'énergie possible des pistes du suivi avec les formants F1 F2 F3, tout en respectant les contraintes liant F1 et F2 d'une part, F2 et F3 d'autre part. L'idée sous-jacente est la suivante : s'il s'agit d'un segment de parole orale (non nasalisée) la combinaison de pistes optimale (pour représenter F1 F2 F3) est celle qui explique le plus d'énergie du suivi de pistes en respectant les contraintes imposées aux formants. Cela se traduit de la manière suivante :

- connaître à chaque instant la meilleure combinaison de pistes pour représenter F1 F2 F3 en respectant le critère d'optimalité que nous venons d'énoncer. Cette interprétation locale permettra de guider le processus d'interprétation globale ;

- le suivi global est construit à partir des pistes disponibles, en utilisant le préétiquetage précédent. Nous retenons le même critère d'optimalité pour l'interprétation globale que pour l'interprétation locale ;

- la qualité de l'interprétation est estimée en utilisant notamment la part d'énergie des pistes expliquée par les formants. Si cette qualité est trop faible, l'algorithme de suivi est relancé avec des paramètres différents.

Soulignons encore que le critère d'optimalité défini plus haut n'est valable que pour les sons oraux (vocaliques et non nasalisés). Pour les voyelles nasales ainsi que pour les sonantes /m,n/ il faudrait définir un critère un peu différent, notamment en ce qui concerne les contraintes liant les différents formants.

Nous allons maintenant revenir plus en détail sur les différentes étapes du processus d'interprétation [5].

6.1 Interprétation locale

La recherche de la combinaison de pistes optimale s'effectue de manière exhaustive pour toutes les pistes existant à un instant donné. Il n'y a aucun risque d'explosion combinatoire car le nombre de pistes à un instant donné n'est jamais très élevé (au plus 6) ; il est donc inutile d'élaguer la recherche. La combinaison optimale maximise l'énergie expliquée par les formants sous les contraintes suivantes :

- une piste doit avoir une fréquence compatible avec le domaine de fréquence du formant qu'elle représente ; ces domaines sont les suivants :
  - F1 $\in$ [180 Hz, 800 Hz]
  - F2 $\in$ [600 Hz, 2 200 Hz]
  - F3 $\in$ [1 800 Hz, 3 500 Hz]
- deux pistes représentant des formants qui se suivent doivent vérifier les contraintes liées à ces formants. Pour F1 F2 il s'agit du célèbre triangle vocalique (comme les valeurs des formants ne sont pas connues précisément, le domaine est légèrement plus étendu que le triangle comme le montre la figure 3.22). Pour F2 F3 il s'agit d'un quadrilatère (fig. 3.23).

Il n'est pas suffisant d'étudier seulement les combinaisons de pistes effectivement détectées, puisqu'il est possible d'être en présence des problèmes d'interprétation 2 ou 3 :

- Si la largeur de bande d'une piste est excessive, on émet l'hypothèse que cette piste correspond à deux formants. Pour une combinaison de pistes donnée, une seule hypothèse de fusion est autorisée.
- S'il n'existe pas assez de pistes pour construire une solution acceptable (au sens des contraintes formantiques) on émet l'hypothèse que l'un des formants

[5] Les algorithmes qui suivent ont été implantés dans Snorri et utilisent comme données les résultats de l'algorithme de suivi décrit dans [Fohr 88]. Nous avons choisi cet algorithme car il est sensiblement plus rapide que celui décrit au § 5, ce qui permet de l'utiliser en décodage sans conduire à des temps de calcul prohibitifs.

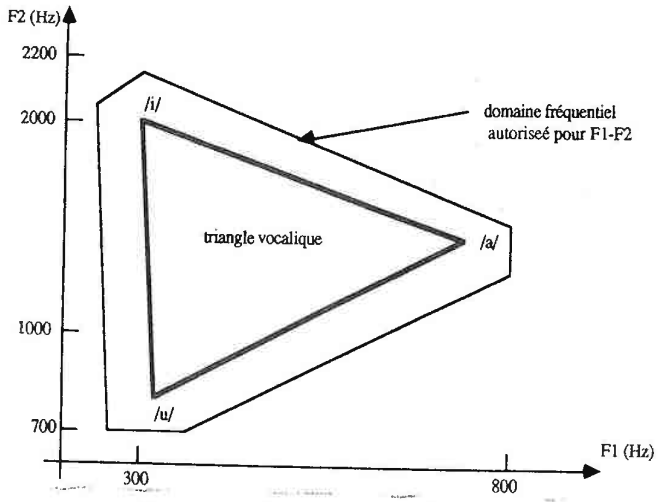


Figure 3.22 : Contrainte sur F1-F2

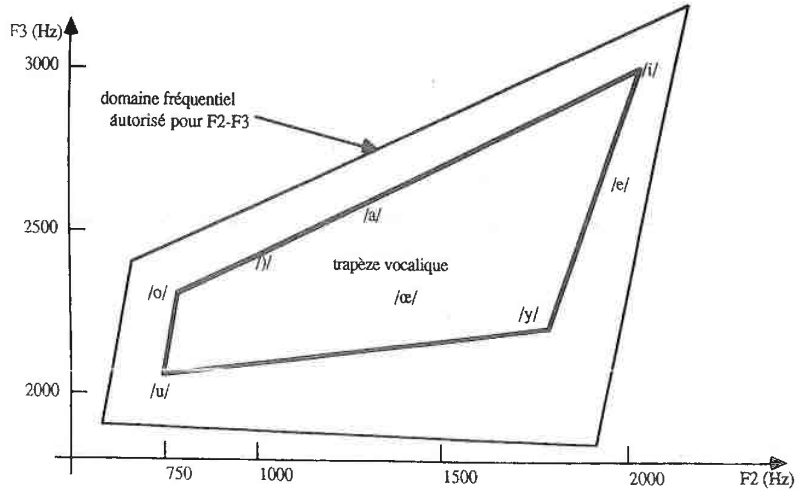


Figure 3.23 : Contrainte sur F2-F3

n'est pas suffisamment intense et qu'il n'a pas été détecté. Pour que cette hypothèse soit réaliste du point de vue acoustico-phonétique seule l'absence de pistes représentant F3 est autorisée. Cette situation se rencontre assez souvent pour les voyelles d'arrière /u o/ pour lesquelles F1 et F2 sont très proches voire fusionnés et F3 est très atténué.

Voici l'algorithme qui correspond à la recherche récursive de la meilleure combinaison de pistes.

```
combinaisonOptimale(pistes) ⟶ Liste de couples (formant, piste)
    global meilleurScore ← −∞
    global meilleureCombi ← vide
    pour chaque piste faire piste->marque ← FAUX finPour
    interprètePistes(1, 0, RIEN, RIEN, 0)
    retourner meilleureCombi
fin CombinaisonOptimale
```

```
interprètePistes(numéroFormant, score, dernièrePiste, combi, nombreFusions)
    si numéroFormant = 4 alors
        /* on vient de trouver une solution que l'on conserve seulement si
        c'est la meilleure qu'on ait trouvée jusqu'à maintenant */
        si score > meilleurScore alors
            meilleurScore ← score
            meilleureCombi ← copie(combi)
        finSi
    sinon
        Booléen auMoinsUneSolution ← FAUX

        /* exploration systématique de toutes les combinaisons possibles */
        pour chaque piste non marquée faire
            si fréquence(piste) ∈ domaineFréquentiel(numéroFormant)
              et compatibles(piste, dernièrePiste, numéroFormant) alors
                marquer(piste)
                combi ← ajouteTête(combi, couple(piste, numéroFormant))
                interPrètePistes(numéroFormant + 1, score + énergie(piste),
                                 piste, combi, nombreFusions)
                combi ← supprimeTête(combi)
                enleverLaMarque(piste)
                auMoinsUneSolution ← VRAI
            finSi
        finPour
        /* éventuellement émission d'une hypothèse de fusion de formants */
        si largeurDeBande(dernièrePiste) > 300 Hz et nombreFusions = 0 alors
            combi ← ajouteTête(combi, couple(dernièrePiste, numéroFormant))
            interprètePistes(numéroFormant + 1, score, dernièrePiste, nombreFusions + 1)
            combi ← supprimeTête(combi)
            auMoinsUneSolution ← VRAI
        finSi
        /* éventuellement émission d'une hypothèse de l'absence du formant F3 */
        si non auMoinsUneSolution et numéroFormant = 3 alors
            interprètePistes(numéroFormant + 1, score, piste, combi, nombreFusions)
        finSi
    finSi
fin interprètePistes
```

```
compatibles(A, B, formant) ⟶ Booléen
    retourner si numéroFormant = 1 alors VRAI
    sinon (fréquence(A), fréquence(B)) ∈ domaineFréquentiel(formant - 1, formant) finSi
fin compatibles
```

Remarque
Le but de cet algorithme est globalement le même que la première partie (les points 1 à 4) de l'algorithme de McCandless (cf §4.1). La solution que nous avons choisie a l'avantage d'assurer que, pour les contraintes imposées aux

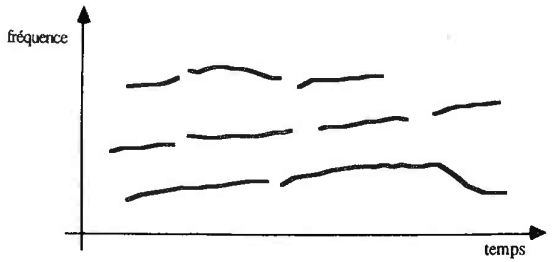


Figure 3.24 : Cas où l'interprétation globale est simple

formants, la combinaison trouvée est la meilleure solution possible. Par ailleurs, formuler la recherche avec un algorithme récursif évite la fastidieuse énumération des cas envisagés par McCandless.

6.2 Interprétation globale

L'interprétation globale est construite à partir des pistes que l'algorithme de suivi a trouvées. Le but est de proposer pour chaque formant la suite des pistes qui le représentent. Cette recherche est amorcée par les résultats de l'interprétation locale qui indique les formants auxquels peut correspondre une piste. Le cas idéal est celui de la figure 3.24 pour laquelle il n'a y clairement qu'une solution. Dans le cas général (fig. 3.21) il faut en fait rechercher la meilleure interprétation au sens du critère d'optimalité que nous avons défini ; cela conduit, comme pour l'interprétation locale, à un algorithme récursif dans lequel on cherche à minimiser l'énergie des pistes qu'on ne peut pas étiqueter comme formant. L'étiquetage est propagé du formant 1 au formant 3 et de la gauche vers la droite. L'exploration se ramifie à chaque fois qu'une hypothèse doit être émise. Nous avons pris en compte trois types d'hypothèses (illustrées par la figure 3.25) pour aborder le cas de conflit où deux pistes peuvent représenter un même formant. Étant données deux pistes A et B se recouvrant, A commençant avant B, voici les hypothèses émises :

- B [fig. 3.25.a] (respectivement A [fig. 3.25.b]) est une piste parasite qui est abandonnée, son énergie est donc ajoutée à l'énergie perdue par l'étiquetage en cours de construction.
- A [fig. 3.25.c] est la fusion de deux formants, la piste A est coupée au début de la piste B et la construction reprend à ce point sans que l'énergie perdue augmente.

Voici maintenant l'algorithme :

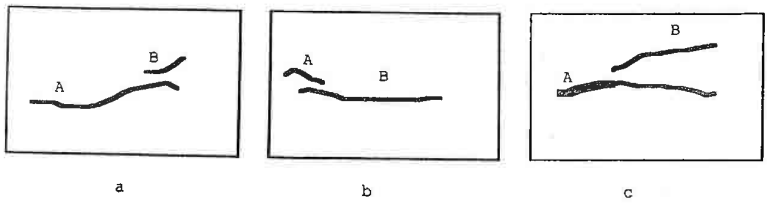


Figure 3.25 : Hypothèses d'interprétation globale

```
interprétationGlobale ⟶ liste de couples(formant, partie de piste)
    global plusPetiteEnergiePerdue ← +∞
    global meilleurEtiquetage ← vide
    global étiquetage ← vide /* étiquetage en cours de construction */
    étiquettePistes(1, 0, 0, premièrePistePossible(1))
    retourner meilleurEtiquetage
    fin interprétationGlobale
```



```
étiquettePistes(formant, énergiePerdue, début, pisteDépart)
    si formant = 4 alors      /* mise à jour éventuelle de la meilleure solution */
        si énergiePerdue < plusPetiteEnergiePerdue alors
            plusPetiteEnergiePerdue ← énergiePerdue
            meilleurEtiquetage ← copie(étiquetage)
        finSi
    sinon /* Adjonction des pistes tant qu'il n'y a pas de conflit */
        pasDeConflit ← VRAI
        A ← premièrePistePouvantEtreEtiquetée(pisteDépart, début)
        tant que A ≠ RIEN et pasDeConflit répéter
            B ← pisteSuivante(A)
            si B = RIEN ou fin(A) < début(B) alors
                /* Pas de conflit puisque les deux pistes ne se recouvrent pas */
                étiquetage ← ajoutePiste(étiquetage, A, début(A), fin(A), formant)
                début ← fin(A) + 1
                A ← B
            sinon pasDeConflit ← FAUX
        finTantQue

        si A ≠ RIEN alors
            /* Il faut explorer plusieurs possibilités pour résoudre un conflit */
            /* 1ère hypothèse: on laisse tomber la piste B (fig. 3.25.a) */
            /* Pose d'un point de reprise pour étudier cette hypothèse */
            étiquetage ← ajoutePointDeReprise(étiquetage)
            étiquetage ← ajoutePiste(étiquetage, A, début(A), fin(A), formant)
            énergieHyp1 ← énergiePerdue + énergie(B, début(B), min(fin(A), fin(B))
            étiquettePistes(formant, énergieHyp1, fin(A) + 1, B)
            /* Destruction de la partie d'étiquetage correspondant à cette hypothèse */
            étiquetage ← supprimeToutJusqu'auPointDeReprise(étiquetage)

            /* 2ème hypothèse: on laisse tomber la piste A (fig. 3.25.b) */
            étiquetage ← ajoutePointDeReprise(étiquetage)
            étiquetage ← ajoutePiste(étiquetage, B, début(B), fin(B), formant)
            énergieHyp2 ← énergiePerdue + énergie(A, début(A), min(fin(A), fin(B))
            étiquettePistes(formant, énergieHyp2, fin(B) + 1, pisteSuivante(B))
            étiquetage ← supprimeToutJusqu'auPointDeReprise(étiquetage)

            /* 3ème hypothèse: A est la fusion de deux formants (fig. 3.25.c) */
            si fusionPossible(A) alors
                étiquetage ← ajoutePointDeReprise(étiquetage)
                étiquetage ← ajoutePiste(étiquetage, A, début(A), début(B), formant)
                étiquettePistes(formant, énergiePerdue, début(B), B)
                étiquetage ← supprimeToutJusqu'auPointDeReprise(étiquetage)
            finSi
        sinon /* On ne peut plus rien faire pour ce formant, on passe donc au suivant */
            étiquettePistes(formant + 1, énergiePerdue, 0, premièrePistePossible(formant + 1))
        finSi
    finSi
fin étiquettePistes
```

Remarquons que cette recherche n'est pas absolument exhaustive :

- on pourrait systématiquement découper une piste lorsqu'elle est en conflit avec une autre piste (les deux pistes se recouvrent), mais cela reviendrait à diminuer l'importance de la continuité et ce n'est donc pas souhaitable.
- le formant $n-1$ est prioritaire devant le formant n car il est étudié avant le formant n. Cela correspond, sur le plan acoustique, au fait que si l'un des deux formants F_n et F_{n-1} est trop faible pour être détecté, il s'agit presque toujours de F_n.

6.3 Estimation de la qualité du suivi et choix du meilleur suivi

Comme nous l'avons déja souligné, il est important de pouvoir estimer la qualité du suivi et de l'interprétation. Nous disposons pour cela des mesures suivantes :

- le rapport $R_1 = \frac{\text{énergie des pistes construites}}{\text{énergie des pics détectés}}$ qui permet d'estimer la qualité du suivi de pistes.
- la rapport $R_2 = \frac{\text{énergie des pistes correspondant aux formants F1 F2 F3}}{\text{énergie des pics détectés}}$ qui permet d'estimer la qualité de l'interprétation formantique.
- le nombre d'ambiguïtés d'étiquetage (il y a ambiguïté lorsque les nombres d'étiquettes de deux formants sont comparables pour une même piste).
- le nombre de pistes qui peuvent représenter deux formants fusionnés.

L'algorithme de suivi de pistes standard utilise le lissage cepstral adapté au locuteur (cf §5.1) et un déplacement temporel de 8 ms après le calcul de chaque spectre. Lorsque les mesures précédentes incitent à penser que le suivi ou l'interprétation peuvent être incorrects (c'est-à-dire que les rapports R_1 et R_2 sont petits), l'algorithme de suivi de pistes est relancé (utilisant toujours le lissage cepstral) avec un déplacement temporel de 2 ms qui permet de capturer les fortes transitions formantiques. Si les résultats sont toujours insuffisants, nous lançons cette fois le suivi de pistes sur les données du codage LPC (qui est choisi en dernier essentiellement à cause de son temps de calcul élevé). À l'issue de ces tentatives nous retenons le suivi qui possède la meilleure qualité d'interprétation formantique.

Le nombre d'ambiguïtés et de fusions de formants permet de vérifier que le processus global de suivi et d'interprétation formantique est bien adapté au segment de parole étudié ; un échec pouvant être dû soit à de grosses erreurs dans le suivi de pistes, soit à de la parole nasalisée.

6.4 Résultats

Les figures 3.26, 3.27 et 3.28 illustrent les résultats de l'interprétation formantique .

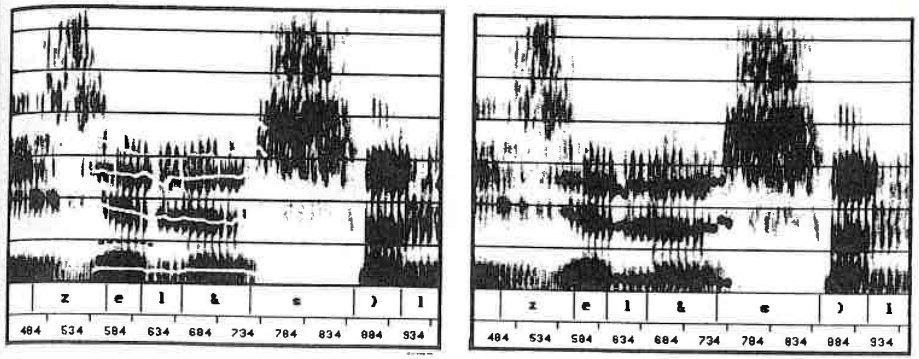


Figure 3.26 : Suivi de pistes et interprétation globale

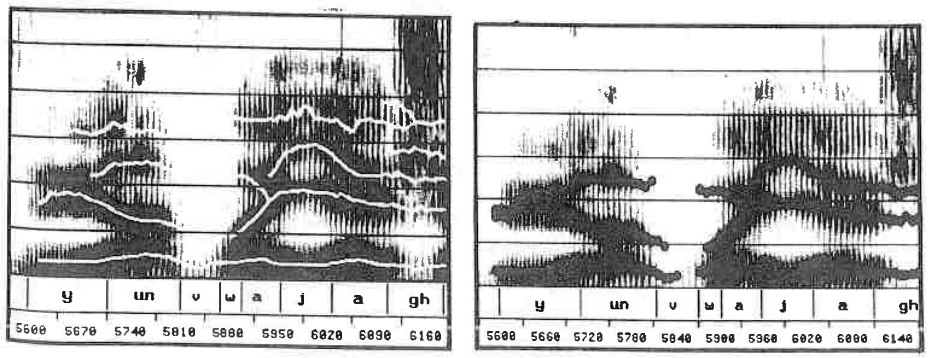


Figure 3.27 : Suivi de pistes et interprétation globale

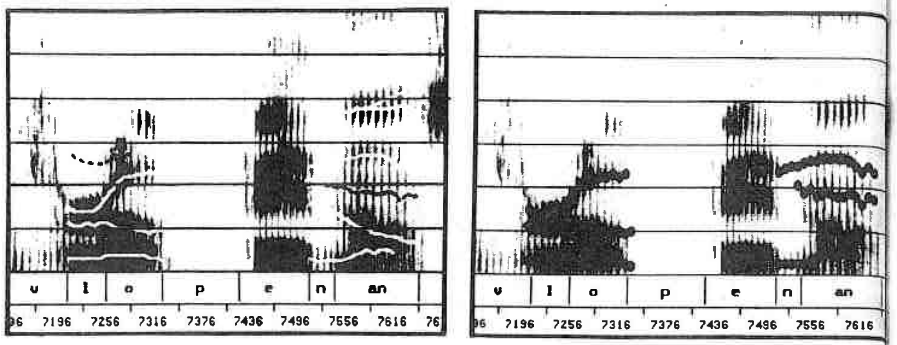


Figure 3.28 : Suivi de pistes et interprétation globale

- fig. 3.26 : une piste parasite (en pointillés) apparaît pendant une partie du segment /e/. L'interprétation formantique ne peut pas faire intervenir cette piste puisque la contrainte liant F2-F3 n'est pas vérifiée et qu'une interprétation globale s'appuyant sur cette piste n'expliquerait pas autant d'énergie que la solution retenue.
- fig. 3.27 : le suivi de piste n'a détecté qu'une seule piste pour les formants F2 et F3 de /y/. L'interprétation formantique a donc conduit à émettre l'hypothèse qu'une seule piste correspondait à deux formants ; cette hypothèse est symbolisée sur la figure par un épaississement du formant F2 pendant une partie du segment /y/.
- fig. 3.28 : il y a conflit entre deux pistes pour représenter F3 pendant le segment /lo/ :
  - * la piste correspondant réellement à F3 n'a pu être étiquetée F3 que pour /o/, puisque pour /l/ la contrainte sur F2-F3 n'est pas vérifiée.
  - * la piste en pointillés (pendant /l/) a aussi été étiquetée F3 car la contrainte sur F2-F3 est cette fois vérifiée.

  C'est la piste (correspondant effectivement à F3) qui a été retenue puisqu'elle permet de minimiser l'énergie qu'on ne peut interpréter en termes de formants.

  Le segment /an/ étant fortement nasalisé ne peut pas être expliqué par les formants F1 F2 F3 oraux (il est impossible d'avoir [$F1 = 300\ Hz$, $F2 = 900\ Hz$ et $F3 = 2\ 700\ Hz$]). Il est donc normal que l'interprétation choisie ne puisse maximiser l'énergie des pistes expliquées par les formants.

Le premier exemple de la figure 3.28 montre que l'algorithme d'interprétation formantique s'avère efficace dans un cas qui devrait lui échapper dans la mesure où le segment de parole ne remplit pas les conditions imposées aux formants. En fait c'est la présence du son voisin (/o/) qui permet d'ancrer l'interprétation, et de conduire à un résultat correct.

Pour pouvoir traiter correctement la totalité des sons il faudrait ajouter les contraintes formantiques correspondant aux sonantes /l,m,n/ et aux voyelles nasales. Le processus consisterait donc à lancer concurremment, pour un même segment de parole, les interprétations formantiques liées à ces différents types de contraintes. Sans être allé aussi loin, notre approche met en évidence l'intérêt d'inclure une forme de modèle formantique dans le processus de suivi de formant.

7 Conclusion

En fait, comme nous l'avons vu, une grande partie du problème du suivi de formants consiste à raccorder des points ou des segments représentant de petits morceaux de trajectoires formantiques. Le problème du raccordement est un problème tout à fait général dans le domaine de la vision par ordinateur et donne actuellement lieu à de nombreux travaux. Ainsi l'article de Shaashua et Ulmann [Shaashua 88] permet d'imaginer une technique de suivi tout à fait différente. Shaashua extrait des formes bruitées et incomplètes d'une image en construisant les courbes minimisant la courbure et maximisant la longueur, directement à partir des points de l'image. Dans le cas du suivi de formants, les courbes construites seraient les trajectoires fréquentielles des cavités résonnantes du conduit vocal et non pas les formants. Une telle approche combinée à la connaissance des configurations formantiques possibles permettraient sans doute d'obtenir de bons résultats.

Un bon algorithme de suivi de formants, doit commencer par l'extraction de paramètres acoustiques aussi fiables que possible. Souvent hélas, les solutions proposées ne sont pas allées beaucoup plus loin car après cette étape indispensable, il faut piloter le suivi de formants avec des connaissances acoustiques que fournissent par exemple les nomogrammes. C'est d'ailleurs ce que proposent Badin et al. dans leur article [Badin 88] consacré aux nomogrammes vocaliques. Comme l'œil humain (et surtout celui de l'expert) reste ce qui se fait de mieux comme outil de suivi de formants, il faudra sans doute, faire largement appel aux techniques de vision pour continuer à progresser dans ce domaine.

Chapitre 4

Connaissance des phénomènes de la parole et décodage acoustico-phonétique

L'étude des phénomènes physiques de la parole est ancienne (voir par exemple l'historique de Boé et Liénard [Boe 88]) puisqu'elle remonte au $XVIII^e$ siècle. Deux étapes marquantes accompagnent cette approche scientifique. La première est l'apparition de l'acoustique vers la fin du XIX^e siècle qui marque les débuts de l'interprétation acoustique de la parole et qui conduit peu de temps après à définir le formant [Hermann 89]. La seconde est sans doute l'apparition de moyens de visualisation juste après la seconde guerre mondiale et qui relance fortement l'étude de la parole dans un grand nombre de domaines : théorie acoustique, synthèse, variabilité due au locuteur, codage, reconnaissance automatique...

Ce vaste mouvement offre au chercheur une large palette de connaissances qu'il peut utiliser avec profit pour la reconnaissance automatique de la parole. Nous présenterons donc un bref panorama des phénomènes actuellement bien maîtrisés et de ceux qui au contraire suscitent encore un fort courant de recherches.

Les premières tentatives de décodage automatique sont contemporaines du regain d'intérêt pour la parole et n'ont pas, pour cette raison, pu bénéficier des recherches entreprises simultanément dans les autres domaines. Nous verrons dans quelle mesure les techniques plus récentes, à base de modèles de Markov ou encore de réseaux neuromimétiques, ont pu tirer parti des progrès réalisés dans la compréhension des phénomènes de la parole.

Les recherches portant soit sur la lecture de spectrogrammes, soit sur la théorie acoustique de la parole ont toutefois assez vite conduit au développement de systèmes de reconnaissances analytiques. Les problèmes rencontrés dans la mise en œuvre de connaissances acoustico-phonétiques et l'apparition simultanée des systèmes experts en intelligence artificielle conduisent, comme nous le verrons, à une seconde génération de systèmes de décodage qui reposent généralement sur l'expertise d'un phonéticien.

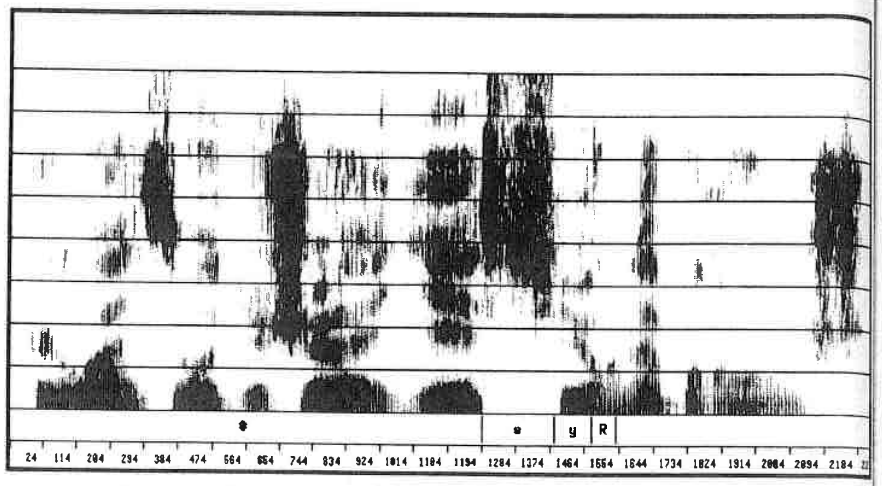


Figure 4.1 : Parole continue (L'oisillon du joailler sur la branche)

1 Les phénomènes de la parole

Les paragraphes suivants ne sont pas destinés à présenter exhaustivement les phénomènes de la parole. Il se contentent d'aborder quelques points qui intéressent particulièrement les systèmes de décodage acoustico-phonétique.

1.1 La coarticulation en parole

Un modèle naïf de la parole consistant à enfiler sur l'axe du temps les phonèmes comme on enfile des perles sur un fil est voué à l'échec. Sur la figure 4.1 il est facile de constater que la plupart des phonèmes n'ont pas de partie stable. En fait, seules les réalisations acoustiques des phonèmes apparaissent dans le signal de parole et il est tout à fait clair que le contexte joue un rôle prépondérant. Ce phénomène tout à fait essentiel en parole, porte le nom de coarticulation et nous allons voir comment il se manifeste, et comment il est possible de le modéliser ou du moins de prévoir partiellement ses effets.

La parole étant produite par le passage de l'air à travers le conduit vocal, ce sont les différents articulateurs modifiant la forme de ce conduit qui concourent au phénomène de coarticulation. Pour passer d'une configuration du conduit vocal liée à un son à une autre configuration, les articulateurs (la langue, les lèvres, le voile du palais...) se déplacent continuement et modifient les paramètres acoustiques que l'on observe.

Portée des phénomènes de coarticulation

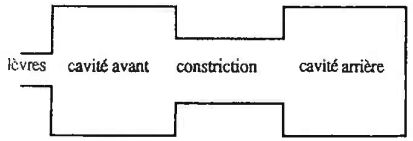
Un son n'est pas influencé de la même façon par ses voisins ; comme les articulateurs ont tendance à anticiper le geste articulatoire lié au son suivant, un son est plus influencé par son successeur que son prédécesseur. Il apparaît clairement sur la figure 4.1 que la réalisation acoustique de /y/ dont les fréquences formantiques de référence (/y/ isolé) se situent vers : F1 = 300 Hz, F2 = 1 750 Hz, F3 = 2 150 Hz est plus influencé par le son /R/ qui le suit que par /s/.

La portée des effets de coarticulation est a priori limitée aux voisins directs d'un son puisque les articulateurs se déplacent de la position qu'ils occupaient pour le son précédent à celle qu'ils occuperont pour le son suivant. Cela n'est pas toujours vrai car il est possible qu'un geste articulatoire n'implique pas tous les articulateurs simultanément. C'est notamment le cas dans le contexte VCV (Voyelle Consonne Voyelle) où la première voyelle peut être influencée par la seconde malgré la présence de la consonne. Ce type d'effet contextuel, étudié notamment par Öhman [Ohman 66] qui allait même jusqu'à supposer que les articulations de C et V étaient totalement découplées, est particulièrement sensible quand la consonne est labiale (/p b f v/). Le phénomène de labialisation laisse libres, hormis les lèvres, tous les articulateurs, c'est le cas notamment de la langue qui anticipe malgré la consonne intervocalique la seconde voyelle, modifiant ainsi son profil formantique. On constate souvent pour la séquence /abi/ de /carabine/ (d'où le nom générique d'"effet carabine" dans [Vaissiere 88]) que F2 de /a/ est "attiré" vers F2 de /i/, c'est-à-dire vers une valeur plus élevée que celui d'un /a/ prononcé isolément.

Étude et modélisation du phénomène de coarticulation

Le phénomène de coarticulation a été très largement étudié depuis l'apparition du Sonagraphe, tant du point de vue de la description des effets, que du point de vue de la modélisation acoustique. Parmi les pionniers, Delattre [Delattre 61] justifie l'importance de ces phénomènes parce qu'il est indispensable de prendre en compte les transitions formantiques (l'un des aspects de la coarticulation) pour synthétiser de la parole intelligible. Ses observations lui permettent de décrire qualitativement la forme des transitions formantiques pour les occlusives voisées. Il va même plus loin en proposant une notion acoustique correspondant à celle de lieu d'articulation et qu'il dénomme "locus" : le locus représente pour une occlusive voisée[1] et un formant donné un point de fréquence vers lequel convergent les trajectoires de ce formant pour l'ensemble des voyelles pouvant suivre l'occlusive. Fant, dans une étude consacrée aux occlusives [Fant 73], montre que la notion de locus n'est pas confirmée par les données dont il dispose à l'exception des occlusives dentales et de F2.

[1] Cette notion, héritée de celle de "hub" définie par Potter, a par la suite été étendue aux autres sons consonantiques.



Les trois paramètres sont :
la position de la constriction,
la section de la constriction,
le rapport longueur sur section des lèvres.

Figure 4.2 : Modèle à trois paramètres du conduit vocal

Ce sont les occlusives qui ont donné lieu au plus grand nombre d'études, notamment au sujet de la modélisation des transitions formantiques. Comme nous l'avons vu dans le chapitre consacré au suivi de formants, il est possible d'approcher la forme du conduit vocal par un assemblage de tubes simulant les différentes cavités. Le modèle classique à trois paramètres (fig. 4.2) permet de prévoir de manière assez fiable (voir [Ladefoged 82]) les valeurs des formants suivant la position de la langue et des lèvres.

Pour l'étude des occlusives, le paramètre le plus important est bien entendu la position de la constriction par rapport aux lèvres. En adoptant un modèle simplifié du conduit vocal, un tube séparé en deux par une constriction (fig. 4.3) il est possible [Stevens 68] de prévoir les effets de coarticulation sur les formants. Les positions approximatives des lieux de resserrement sont indiquées pour les occlusives sur la figure 3.2 du chapitre sur les formants. Par rapport aux valeurs sans constriction de la voyelle neutre /œ/ on constate qu'il y a abaissement simultané des 3 formants pour les occlusives labiales, élévation de F2 et F3 pour les dentales et enfin rapprochement de F2 et F3 pour les vélaires. Ces prédictions sont bien confirmées par l'observation de spectrogrammes.

Même s'il est assez facile de prévoir grossièrement l'allure des transitions formantiques dues aux effets de coarticulation (comme l'ont fait Öhman [Ohman 67] et Broad [Broad 87]) il reste qu'il faudrait ajouter, pour caractériser correctement les effets contextuels, un assez grand nombre d'autres indices au premier rang desquels une mesure de l'amplitude des transitions, mais aussi :

- les modifications de durée,
- les modifications des transitions dues au rythme d'élocution (sur lesquelles nous reviendrons au §1.2),
- la forme et le domaine de fréquence des bruits de friction pour les fricatives,

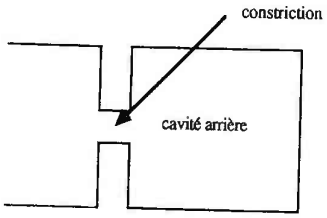


Figure 4.3 : Modèle simplifié du conduit vocal

- la forme de la barre d'explosion,
- et tous les autres indices qui n'apparaissent que dans certains contextes (par exemple le fameux éclair entre une consonne labiale et /i/ [Maeda 86]).

Une modélisation suffisamment puissante et générale pour prévoir les effets de coarticulation n'est pas encore disponible. Cela demanderait notamment une étude gigantesque à partir de données articulatoires et acoustiques et n'est donc pas envisageable avant longtemps. Cela limite donc la prise en compte de la coarticulation par les systèmes de RAPC qui doivent stocker des connaissances partielles soit sous la forme de règles de production soit directement dans les références acoustico-phonétiques.

1.2 Accentuation, vitesse d'élocution et réduction vocalique

Dans le paragraphe précédent nous avons seulement évoqué les modifications acoustiques provoquées par l'effet de coarticulation. En fait un certain nombre de paramètres peuvent amplifier ou limiter les déformations de l'image acoustique d'un phonème :

- la vitesse d'élocution du locuteur,
- l'accentuation d'une partie de la phrase,
- la position du phonème par rapport aux frontières du mot et de la syllabe auxquels il appartient.

Le travail de Lindblom [Lindblom 63] porte sur les modifications de trajectoires formantiques d'une voyelle en fonction de sa durée. La voyelle étudiée est entourée de deux occlusives identiques (ex : /bub/) et la séquence CVC

appartient à une phrase courte. Les phrases du corpus ont été prononcées par un seul locuteur.

À partir de ces données Lindblom montre que les trajectoires formantiques s'écartent de celles de la voyelle cible (voyelle prononcée isolément) quand la durée de la voyelle diminue. Ce phénomène porte le nom de réduction, on dit aussi que la voyelle perd sa "couleur vocalique". Le fait que la réduction augmente quand la durée vocalique diminue semble s'expliquer facilement : plus courte est la durée, moins les articulateurs ont le temps de se déplacer pour atteindre la position articulatoire de la voyelle attendue. Lindblom étudie aussi l'influence de l'accentuation sur la réduction vocalique avec la même séquence CVC et conclut que l'accentuation modifie très peu la réduction vocalique.

Les conclusions auxquelles est arrivé Lindblom sont en fait très contestées par de nombreuses autres études. Gay [Gay 78] étudie lui aussi les effets de l'accentuation et de la vitesse d'élocution pour des séquences CVC et CVCVC incluses dans une phrase. Aucune réduction vocalique n'est observée entre une séquence CVC énoncée lentement et la même séquence énoncée rapidement. Par contre une voyelle accentuée est plus colorée qu'une voyelle moins accentuée. En revanche l'augmentation de la vitesse d'élocution influence l'organisation temporelle de la séquence CVC de la manière suivante : plus la durée de la voyelle est faible, plus les extrémités des trajectoires formantiques se rapprochent des valeurs formantiques au centre de la voyelle. Cela correspond donc à une réduction de l'amplitude des transitions formantiques. Voici donc pour résumer les travaux de Gay, les effets de l'accentuation et de la vitesse d'élocution :

- accentuation :
  - * réduction vocalique en dehors de l'accent,
  - * augmentation de la fréquence fondamentale,
  - * augmentation de l'énergie du signal,
- vitesse d'élocution :
  - * affaiblissement des transitions formantiques.

Les conclusions de Gay sont confirmées par les travaux de Engstrand [Engstrand 88] portant sur l'étude de l'influence de l'accentuation et de la vitesse d'élocution sur le geste articulatoire. En plus des paramètres habituels (durée, valeurs des formants), le mouvement des articulateurs (langues et lèvres) a été enregistré par cinéradiographie. La séquence étudiée est, à la différence des études précédentes, une suite VCV appartenant à une phrase simple. Les conclusions de Engstrand sont tout à fait en accord avec celles de Gay en ce qui concerne les effets de l'accentuation et de la vitesse d'élocution. Engstrand conclut en effet que les caractéristiques des voyelles sont significativement altérées par l'accentuation mais pas par la vitesse d'élocution. Les enregistrements cinéradiographiques de la séquence /V_1pV_2/ permettent d'ajouter que les mouvements de la langue et ceux qui accompagnent la fermeture de l'occlusive se recouvrent plus lorsque la vitesse d'élocution augmente. L'effet est le même

lorsque $V_1 = V_2$. Cela est tout à fait en accord avec l'observation de l'effet "carabine".

L'étude des phénomènes d'accentuation et de réduction vocalique ne repose pas uniquement sur l'examen de données articulatoires (grâce à la radiocinématographie) et acoustiques (grâce au spectrogramme) mais aussi sur des expériences de perception. Ainsi Rietveld [Rietveld 87] étudie les relations entre la réduction vocalique et la perception de l'accentuation pour le Néerlandais. Les stimuli proposés aux auditeurs sont de la forme /fV_1fV_2fV_3/ ; les trois voyelles cardinales /a i u/ subissent une réduction par centralisation (les formants sont déplacés progressivement vers ceux de la voyelle centrale /œ/). L'une des voyelles est moins réduite que les deux autres, qui elles sont réduites au même niveau. Les sujets indiquent pour chaque stimulus la voyelle qui est accentuée. Cette étude confirme qu'une voyelle semble accentuée quand elle est moins réduite que les voyelles contiguës. D'autres indices renforcent la perception de l' accentuation :

- une voyelle placée au début ou à la fin d'un mot est plus facilement perçue comme étant accentuée,
- la voyelle /a/ est plus facilement perçue accentuée que /i/ et /u/.

La petite étude bibliographique qui précède permet de conclure que l'accentuation d'une voyelle se traduit acoustiquement par :

- une augmentation de la couleur de la voyelle (les formants se rapprochent de ceux de la voyelle prononcée isolément),
- une augmentation de F0,
- une augmentation de la durée de la voyelle.

L'augmentation de la vitesse d'élocution ne semble pas provoquer de réduction vocalique mais plutôt réduire les transitions formantiques.

Les conclusions précédentes ne doivent pas être considérées comme infaillibles car les conditions expérimentales sont très éloignées de la "parole naturelle". Si Lindblom n'indique pas l'origine du locuteur qu'il a utilisé pour ses études, Gay précise en revanche qu'il s'agit de sujets très entraînés, à tel point que les locuteurs ont reçu comme consigne de maintenir l'identité phonétique d'une voyelle non accentuée, ce qui est tout à fait hors de portée d'un locuteur normal. Quant aux locuteurs de Engstrand ils sont phonéticiens professionnels et les conditions expérimentales (des billes opaques aux rayons X sont placées sur la langue, le locuteur ne doit pas bouger la tête) sont tout à fait inhabituelles.

1.3 Nasalité

Les phénomènes de nasalité concernent en Français les voyelles nasales /ã, õ, ɛ̃, œ̃/ et les consonnes nasales /m, n, gn, ng/. Par ailleurs ces consonnes nasalisent

par influence les voyelles contiguës. La nasalité est donc un point important et sur lequel achoppent d'ailleurs la plupart des systèmes de DAP.

L'étude de la nasalité est sans doute l'un des exercices les plus spéculatifs que connaissent les chercheurs en parole. Hormis quelques faits tout à fait reconnus il est bien difficile de faire une synthèse des travaux les plus récents et qui ne se contredisent pas.

D'un point de vue articulatoire la nasalité ne semble pas si compliquée que cela puisqu'elle apparaît lorsque le velum s'abaisse. L'air peut donc sortir par les cavités orales et nasales. Cela se traduit du point de vue de l'acoustique par la mise en parallèle des deux conduits résonnants. Il y a donc apparition de formants supplémentaires (ceux des cavités nasales) et d'antiformants. De plus ce couplage modifie la position des formants d'origine orale. La modélisation précise des effets de nasalité se heurte à plusieurs difficultés majeures :

- il est très difficile de mesurer les cavités nasales et donc d'avoir une connaissance a priori de la position des formants nasaux,
- les importantes différences anatomiques interlocuteurs et même intralocuteur (à cause de l'état de la muqueuse nasale) sont telles qu'elles déprécieraient les éventuelles mesures,
- il faut choisir, quand on modélise la production des sons nasalisés, si l'on doit ou non tenir compte, en plus des fosses nasales, des cavités sinusales.

L'approche consistant à évaluer les formants et antiformants en interprétant le spectre d'une voyelle nasale est elle aussi vouée à l'échec [Feng 86] du moins au-dessous de 2 000 Hz. En effet les perturbations engendrées par la présence des pôles et zéros d'origine nasale sont très importantes et peuvent varier fortement d'un spectre à l'autre.

Malgré toutes les difficultés que nous venons d'exposer, les nombreuses simulations numériques réalisées par F. Lonchamp [Lonchamp 88] ainsi que la synthèse qu'il a faite des travaux précédents permettent de fournir les éléments de réponse suivants pour les voyelles nasales.

Il n'y a pas d'ordre immuable entre F'_1 et FN_1 (F'_n indique le $n^{ième}$ formant oral d'une voyelle nasale, FN_i indique lui le $n^{ième}$ formant nasal) cela s'expliquant par la grande variabilité de la configuration des fosses nasales due notamment à l'état de la muqueuse. L'ordre d'apparition de ces formants n'a que peu d'importance puisqu'il ne modifie pas l'allure générale du spectre. Le pic en très basse fréquence (vers 350 Hz) qui a souvent été pris pour le formant nasal est soit une résonance sinusale soit le premier rebond du formant glottal (cf 3.2). Voici en tenant compte des remarques précédentes les valeurs approximatives des formants pour les voyelles nasales.

- /ɛ̃/
  - FN_1 ou F'_1 entre 500 et 700 Hz,
  - F'_1 ou FN_1 vers 1 000 Hz,
  - F'_2 entre 1 400 et 1 600 Hz,
  - F'_3 entre 2 500 et 3 000 Hz,
- /ã/
  - F'_1 ou FN_1 entre 600 et 700 Hz
  - F'_2 et (FN_1 ou F'_1) peuvent être distincts ou fusionnés entre 800 et 1 000 Hz,
  - FN_2 vers 2 200 Hz,
  - F'_3 vers 2 900 Hz,
  - l'énergie du spectre est en grande partie en basse fréquence (grande compacité spectrale)
- /õ/
  - F'_1 vers 450 Hz,
  - FN_1 et F'_2 vers 800 Hz distincts ou partiellement fusionnés,
  - F'_3 entre 2 800 et 3 000 Hz,
  - le spectre est très compact en basse fréquence.

On peut déduire des descriptions spectrales précédentes quelques critères pour détecter la présence de voyelles nasales sur un spectrogramme :

- l'apparition de formants supplémentaires,
- la présence d'un "œil nasal" (c'est-à-dire un vide d'énergie au voisinage de 2500 Hz), et qui correspond à la nette élévation de F'_3 pour /ã/ et /õ/,
- la faible dynamique du spectre en basse fréquence (due aux nombreux formants et antiformants dans cette zone),
- l'instabilité des formants qui ne sont pas tous présents du début à la fin d'une voyelle nasale.

Les caractéristiques des voyelles nasales que nous venons de citer laissent entrevoir les difficultés à surmonter afin qu'un système de DAP donne de bons résultats pour les sons nasalisés.

1.4 Perception et modèle du système auditif périphérique

De la même façon que l'on étudie les phénomènes de production de la parole, on cherche à connaître les phénomènes liés à la perception de la parole. Les intérêts de telles études sont multiples et nous n'en citerons que trois :

- guider la recherche des indices acoustiques pertinents pour la reconnaissance,
- fournir un "meilleur" signal d'entrée à un système de RAPC,

- mieux comprendre les mécanismes humains de la compréhension de la parole.

Plutôt que d'offrir au lecteur un impossible panorama de l'aide qu'apportent les études perceptives voici quelques exemples concernant les occlusives : mise en évidence de l'importance des transitions formantiques [Delattre 55], de la barre d'explosion [Hoffman 58], étude des liens entre la barre d'explosion et le lieu d'articulation [Stevens 78], entre l'amplitude de la barre d'explosion et le lieu d'articulation [Ohde 83].

L'étude des phénomènes de perception et plus précisément du fonctionnement de l'appareil auditif périphérique humain peuvent dans l'avenir apporter beaucoup aux systèmes de RACP. En effet les systèmes actuels sont beaucoup plus sensibles au bruit qu'un auditeur humain, et il serait donc très intéressant de disposer d'un signal d'entrée qui renforce les indices acoustiques nécessaires à la reconnaissance de la parole.

Certains concepts comme celui de "bande critique" ont reçu une confirmation expérimentale grâce à l'étude des fibres du nerf auditif chez le chat [Kiang 65]. On sait ainsi que la résolution fréquentielle est meilleure pour les basses fréquences que pour les hautes fréquences. En fait les fibres du nerf auditif sont au centre de deux types d'études. Les premières permettent de montrer [Delgutte 84b], à partir de l'examen des réponses des fibres à des stimuli vocaliques, qu'elles détectent F0, F1, F0 et F1, F2, F0 et F1 et F2 (le stimulus d'entrée étant produit par un synthétiseur à deux formants). Ces études ne permettent à l'heure actuelle que d'émettre des hypothèses sur la prise en compte de ces informations spectrales par le système nerveux central [Delgutte 84a].

Les caractéristiques dynamiques de la parole sont un autre centre d'intérêt, et même si l'on connaît plusieurs mécanismes nerveux (adaptation, masquage, suppression) destinés à capturer les fortes variations de la parole (par exemple les barres d'explosion), il est là encore difficile de proposer un modèle de traitement des caractéristiques dynamiques [Delgutte 86].

Un certain nombre de modèles du système auditif périphérique ont été développés. Certains, très simples, ne modélisent que les bandes critiques et ne sont donc guère éloignés des vocodeurs à canaux traditionnels [Zwicker 79]. D'autres sont plutôt adaptés à la recherche de certains indices acoustiques [Caelen 79] [Wu 89], ou des formants [Carlson 82] [Seneff 85]. Par ailleurs, intégrer un modèle auditif pour obtenir un signal plus significatif à l'entrée d'un système de RACP n'est pas toujours efficace [Blomberg 83]. Le fait que ces modèles sont bien appropriés pour guider la détection de certains indices n'implique pas qu'un système de décodage sache les utiliser avec profit et donc qu'il améliore son taux de reconnaissance final. L'intérêt de tels modèles se concevrait donc mieux si l'on disposait d'un système simulant déja une bonne partie du système auditif humain, ce qui est loin d'être le cas.

2 Décodage acoustico-phonétique

Nous allons maintenant présenter les principales techniques utilisées à l'heure actuelle en décodage acoustico-phonétique. Dans cette description, nous insisterons notamment sur les connaissances propres à la parole que manipulent les systèmes de décodage.

2.1 Utilisation de modèles de Markov cachés

Les techniques à base de modèles de Markov cachés sont sans doute à l'heure actuelle les plus utilisées en reconnaissance automatique de la parole. Les références des unités de parole sont stockées sous la forme de sources de Markov composées d'états et d'arcs entre ces états. À chaque arc liant un état i à un état j est attachée une probabilité de transition a_{ij}. Pour chaque arc est aussi définie une "probabilité de sortie" $b_{ij}(k)$ représentant la probabilité d'émettre le symbole k pour l'arc liant les états i et j. Les symboles k (qui sont les seules observations dont on dispose pendant la phase de reconnaissance) sont des prototypes de l'espace de paramètres utilisé (coefficients LPC ou coefficients cepstraux).

Durant la phase de reconnaissance on cherche à maximiser la probabilité $P(M/O)$ d'être en présence du modèle M connaissant la séquence d'observations O ; $P(M/O)$ peut s'écrire en utilisant la règle de Bayes :

$$P(M/O) = \frac{P(M)P(O/M)}{P(O)}$$

$P(O)$ représente la probabilité d'apparition de la séquence d'observations O, $P(O/M)$ le modèle acoustique et $P(M)$ le modèle linguistique.

Trois problèmes fondamentaux sont liés à l'utilisation de modèles de Markov cachés :

- l'évaluation de $P(O/M)$. L'algorithme "forward" permet de calculer $P(O/M)$ en $\mathcal{O}(N^2T)$ (N étant le nombres d'états et T le nombre d'observations) ;
- la recherche de la suite d'états la plus probable pour le séquence d'observations O. L'algorithme de Viterbi très semblable aux algorithmes de programmation dynamique permet de résoudre ce problème ;
- l'apprentissge des paramètres du modèle qui peut être réalisé grâce à l'algorithme de Baum-Welch.

Lorsqu'il s'agit de reconnaître un petit vocabulaire de mots il est possible de prendre comme unité de décodage le mot et de construire des modèles les représentant. Cette démarche n'est pas possible, notamment pour des problèmes d'apprentissage lorsqu'il s'agit de parole continue. Il faut alors recourir à une unité de parole plus petite, par exemple le phonème, et modifier les algorithmes

de reconnaissance. Les principales techniques utilisées en parole continue sont les suivantes :

- l'algorithme de Viterbi pour lequel on tient compte des transitions entre états internes aux phonèmes et des transitions entre phonèmes. Comme la recherche exhaustive du meilleur chemin n'est possible que pour des applications de petite taille, elle est en général poursuivie seulement pour les états les plus probables (recherche en faisceaux) [Ney 87].
- l'algorithme de construction par niveau est une variante de l'algorithme de Viterbi, pour laquelle on tient compte aussi du nombre de mots de l'hypothèse courante [Myers 81]. Le niveau 1 représente toutes les hypothèses d'un mot commençant au début de la phrase ; on retient pour chaque instant la meilleure solution. Quand on passe du niveau $l-1$ au niveau l on étend toutes les hypothèses de niveau $l-1$ [Rabiner 88] avec un mot supplémentaire.
- l'algorithme du décodage par pile [Bahl 83] fait appel à l'algorithme "forward". Toutes les hypothèses d'un mot sont rangées dans une liste, la meilleure étant placée en tête. On continue en effaçant de la liste la première hypothèse, puis en la prolongeant ; toutes les hypothèses qui en découlent sont alors placées dans la liste avant d'être, à leur tour, effacées de la liste.

Le formalisme des modèles de Markov n'est porteur d'aucune connaissance destinée à reconnaître la parole. C'est le concepteur du système de décodage qui doit "règler" les différents paramètres à commencer par la topologie des modèles. La procédure de Baum-Welch permet ensuite d'apprendre les paramètres définissant les transitions entre états et l'émission des symboles de l'alphabet. Pour que l'apprentissage ait lieu dans de bonnes conditions il faut disposer d'un corpus très vaste. Ainsi Lee [Lee 88] dispose de 4 200 phrases prononcées par 105 locuteurs pour entraîner son système SPHINX, à base de triphones, et destiné à la parole continue multi-locuteur.

Les résultats d'un système à base de modèles de Markov sans amélioration particulière sont suffisants pour un système monolocuteur destiné à la reconnaissance d'un petit vocabulaire de mots isolés. Le taux de reconnaissance chute de manière très sensible pour la parole continue multi-locuteur. Lee indique que le pourcentage de mots reconnus par la version de base de SPHINX atteint à peine 30% pour une application de 997 mots. Cela signifie donc qu'il faut impérativement adapter les modèles de Markov au problème spécifique de la parole continue.

Le premier problème à aborder est le choix d'une unité de parole à modéliser. Le mot est tout à fait adapté lorsqu'il s'agit de reconnaître un petit vocabulaire de mots isolés. Tous les phénomènes coarticulation sont en effet directement pris en compte par le modèle. Pour des raisons de stockage et d'apprentissage, il n'est pas possible de conserver le mot comme unité acoustique pour la parole continue. S'il est naturel, le choix du phone conduit à un problème très

grave : les phénomènes de coarticulation altèrent de manière très sensible la forme acoustique d'un phone et compromettent donc la reconnaissance. Pour s'affranchir de cette limitation il est possible soit de modéliser les transitions acoustiques aux extrémités des phones soit de créer autant de modèles qu'il y a de contextes possibles. Comme il est raisonnable (cf §1.1) de limiter l'effet du contexte aux phonèmes contigus, Schwartz [Schwartz 84] propose d'utiliser le triphone comme unité de décodage. Lee [Lee 88] reprend cette idée en l'appliquant à SPHINX. Comme le vocabulaire de 997 qu'il emploie correspond à 2 381 triphones, l'apprentissage n'est pas une tâche insurmontable. Pour réduire ce nombre, les modèles représentant un phonème donné sont regroupés automatiquement selon un critère de similarité faisant appel à l'entropie mutuelle de deux modèles.

Ce type d'amélioration permet dans une certaine mesure de capturer une partie des phénomènes de coarticulation. Un second type d'amélioration est destiné à prendre en compte le fonctionnement de l'appareil auditif humain. La plus simple consiste à adopter une échelle de fréquence adaptée à la sensibilité fréquentielle de l'oreille, par exemple l'échelle Bark ou encore l'échelle Mel. Pour qu'un modèle de Markov modélise mieux les aspects dynamiques de la parole, on peut aussi changer la nature des observations acoustiques en ajoutant des coefficients exprimant l'évolution temporelle des paramètres acoustiques. Ainsi Lee a choisi de compléter le vecteur des coefficients cepstraux par le vecteur de leurs pentes calculées sur un petit intervalle de temps. Puisque le contenu d'une observation n'est pas homogène il faut définir une distance composite qui pondère les deux contributions d'un facteur choisi expérimentalement. Pour éviter de recourir à une distance composite qui n'est guère satisfaisante d'un point de vue théorique, il est possible d'étendre les modèles de Markov en émettant non pas un symbole par transition mais autant de symboles qu'il y a de types d'informations acoustiques. Pour ajouter l'évolution temporelle des coefficients cepstraux il suffit donc d'émettre un symbole supplémentaire par transition.

Les connaissances que nous venons de présenter pour compléter un modèle de Markov semblent assez pauvres comparées à celles que manipule un lecteur de spectrogrammes. Les piètres résultats enregistrés par Lee en voulant intégrer une composante experte dans SPHINX montrent à l'évidence qu'ajouter l'expertise d'un phonéticien dans un modèle de Markov est un problème très ardu voire sans solution.

2.2 Approche neuromimétique

Apparus depuis peu ou plutôt redevenus à la mode depuis peu, les modèles connexionnistes ont été appliqués à la parole. L'idée sous-jacente est d'imiter le comportement du cerveau en disposant d'un grand nombres de neurones ou de cellules connectés entre eux. Comme pour les modèles de Markov le réseau est entraîné sur un corpus avant de pouvoir reconnaître un énoncé inconnu.

La classe de modèles connexionnistes la plus répandue est celle des perceptrons

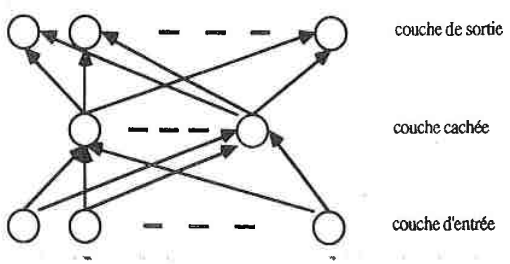


Figure 4.4 : Exemple de perceptron à trois couches

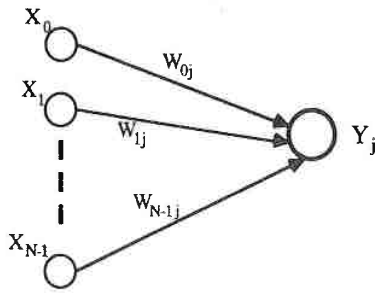


Figure 4.5 : Activation d'un cellule de perceptron

introduits par Minsky et Papert [Minsky 69]. Les perceptrons ne sont en fait intéressants que s'ils sont composés d'au moins trois couches (fig. 4.4) :

- une couche d'entrée représentant le vecteur d'entrée à classifier,
- une ou plusieurs couches d'unités cachées, chacune des cellules d'une couche étant connectée à toutes les cellules de la couche inférieure. Chaque connexion entre deux cellules porte un poids w_{ij} indiquant comment la cellule i excite ou inhibe la cellule j du niveau supérieur,
- une couche de sortie, chaque cellule représentant l'une des classes possibles du vecteur d'entrée.

Quand on applique un stimulus à la couche d'entrée, l'activité d'une cellule de la couche immédiatement supérieure (fig. 4.5) est la suivante :

$$y_j = f(\sum_{i=0}^{N-1} w_{ij} x_i - S)$$

w_{ij} étant le poids de la connexion de la cellule i de la couche d'entrée à la cellule j de la couche supérieure,
S étant un seuil de réaction,
N étant le nombre de cellules de la couche précédente,
x_i étant l'activité de la cellule i de la couche d'entrée,
f est une fonction non linéaire à valeurs dans $[-1, 1]$
si $x \ll S$ $f(x) = -1$ (inhibition)
si $x \gg S$ $f(x) = 1$ (activation)

La classe reconnue correspond à la cellule de la couche de sortie dont l'activité est la plus grande.

L'apprentissage est réalisé par l'algorithme de rétropropagation des erreurs ; il s'agit d'un algorithme d'apprentissage supervisé [Rumelhart 86] [Cun 88] qui minimise une fonction de coût égale au carré de l'erreur entre la sortie attendue (la classe du vecteur d'entrée) et la sortie réelle, en modifiant les poids des connexions de la couche de sortie vers la couche d'entrée.

Comme pour les modèles de Markov le succès d'un système connexionniste dépend de l'habilité du concepteur à construire un modèle qui capture correctement les phénomènes de parole. Cela recouvre le choix : du type de vecteur d'entrée (LPC, coefficients cepstraux, prototypes...), du nombre de couches cachées et surtout de la manière de prendre en compte le temps. En effet un perceptron ne fait que classifier un vecteur d'entrée hors de son contexte et de ce fait ne peut pas modéliser les effets de coarticulation. Plusieurs solutions ont été proposées pour pallier cette limitation ; citons entre autres :

- les perceptrons multi-couche contextuels introduits par Sejnowsky [Sejnowsky 86] traitent en entrée non pas un seul vecteur mais un vecteur avec ses voisins (3 à gauche et 3 à droite),
- les réseaux neuronaux à rétroaction ("feedback") [Rumelhart 86]. La rétroaction a lieu sur la couche d'entrée à partir de la couche de sortie ou des couches cachées, et confère donc une forme de mémoire à la machine connexionniste,
- les réseaux neuronaux à retard ("Time Delay Neural Network") [Waibel 87] prennent en entrée plusieurs trames consécutives et sont destinés à reconnaître les trois occlusives voisées de l'anglais.

Le cadre général de la programmation dynamique fournit le niveau d'intégration supérieur nécessaire pour aborder la reconnaissance de la parole continue [Bourlard 87].

Il existe une large variété de modèles neuromimétiques que nous n'avons pas présentées (notamment les cartes topologiques de Kohonen [Kohonen 84], les machines de Bolzmann [Prager 86] ou encore la colonne corticale [Guyot 89]). Au stade de développement actuel, les réseaux neuromimétiques conduisent à des résultats encourageants. Il reste néanmoins de nombreux problèmes théoriques, citons entre autres : comment être assuré que l'apprentissage conduise

à un optimum global de l'ensemble des poids de connexions pour la reconnaissance ? comment modéliser le temps ?

Selon nous, le succès actuel des réseaux neuromimétiques repose sur l'idée qu'ils sont doués de facultés d'apprentissage (et donc de modélisation puis de reconnaissance) supérieures aux autres approches actuelles. Même si des traces de conceptualisation (par exemple des indices liés au lieu d'articulation des occlusives dans [Waibel 87]) ont été observées on ignore encore dans quelle mesure l'activité du cerveau pourra être copiée. On sait en revanche très bien que le chercheur doit encore fournir de gros efforts pour adapter un réseau neuromimétique à un problème spécifique tant du point de vue de la pertinence des données que de l'architecture du système.

2.3 Systèmes analytiques

L'atout majeur des techniques de reconnaissance globale est qu'elles donnent de bons résultats pour une application spécifique. En revanche l'adaptation d'un système de reconnaissance à une nouvelle application demande un gros effort d'apprentissage des nouvelles références, qu'il s'agisse de mots ou d'unités de parole plus petites. À cela s'ajoute la difficulté d'exploiter l'expertise d'un phonéticien, même si c'est une éventualité que certains défenseurs des modèles markoviens écartent [2] avec vigueur. Les techniques analytiques, en revanche, ont souvent fait appel à la connaissance d'un expert pour améliorer les performances. Le niveau de connaissance atteint par les systèmes analytiques a suivi, pour simplifier, le développement de la puissance de calcul. Alors que le module de décodage acoustico-phonétique de Myrtille [Haton 84] ne faisait intervenir que quelques traits acoustiques grossiers calculés à partir de l'énergie, le système développé par Alinat [Gallais 89] intègre de nombreuses connaissances, notamment sur l'appareil auditif et les lieux d'articulation des occlusives. De la même façon Keal [Mercier 83][Mercier 85] dont le principe de base est d'identifier les sons avec des fonctions discriminantes linéaires possède aussi un système de règles pour la reconnaissance des consonnes.

Le processus de décodage est généralement le suivant :

Prétraitement Il s'agit de l'extraction des paramètres acoustiques. En plus du vecteur acoustique que génère le vocodeur à canaux, s'ajoutent certains paramètres destinés à la détection des différentes classes phonétiques (voyelles, occlusives, fricatives, sonantes).

Segmentation Elle peut être obtenue par regroupement de trames centisecondes pré-classées ou encore par l'étude de paramètres acoustiques (énergie vocalique, bruit de friction et explosion).

[2] "Every time we fire a phonetician linguist, the performance of our systeme goes up" F. Jelinek, IEEE ASSP Workshop 1985, Arden House

Identification Les techniques sont très variées. Pour Myrtille il s'agit de trouver la référence la plus proche de la trame inconnue ; Keal fait essentiellement appel à des fonctions discriminantes linéaires ; le système d'Alinat utilise des indices acoustiques plus élaborés comme les formants ou des traits phonétiques pour les lieux d'articulation des occlusives.

Parmi les systèmes exploitant notre connaissance actuelle des phénomènes de la parole, les systèmes analytiques procéduraux ont été relégués au second plan par le développement de "systèmes experts en décodage" que nous décrirons dans le paragraphe suivant. Les excellents résultats atteints par le module de décodage de DIAPASON [Gallais 89] (environ 95% de mots corrects) témoignent pourtant de la pertinence de la connaissance que ce système intègre.

2.4 Systèmes experts en décodage acoustico-phonétique

Le développement des systèmes de décodage analytiques a mis en évidence l'importance de disposer d'une connaissance précise des phénomènes de la parole pour améliorer les résultats. Les très bonnes performances de décodage (environ 80%) atteintes par certains phonéticiens [Cole 79] [Lonchamp 87] incitaient d'ailleurs à penser qu'un système de décodage pouvait être construit en partant de l'expertise d'un phonéticien. L'avènement des systèmes experts (Dendral en chimie organique [Lindsay 80], Mycin en diagnostic médical [Shortliffe 76]) a offert l'outil qui manquait pour concrétiser cette idée dès le début des années 80 ([Carbonell 83] [Johnson 84] [Johanssen 83] ...).

Tous les systèmes de ce type sont globalement organisés de la manière suivante :

- l'expertise formalisée par un ensemble de règles de production (Aphodex [Carbonell 86], Zue et al. [Zue 86b] [3], Stern et al. [Stern 86], Serac [Gillet 84], Mizoguchi et al. [Mizoguchi 84] [Mizoguchi 86], Méloni et al. [Meloni 86]) de la forme suivante : **si** indices acoustiques et contexte **alors** ensemble de sons ;
- une base de faits qui regroupe les mesures acoustiques et les faits déduits grâce à l'expertise ;
- un moteur d'inférences qui représente le mécanisme d'interprétation des règles et permet de déduire de nouveaux faits à partir des faits connus.

Le chaînage avant est le mode de fonctionnement adopté par la plupart des systèmes experts de décodage ; le raisonnement est donc guidé par les données. Le système Lamel et al. [Lamel 88] utilise aussi le chaînage arrière pour les informations acoustiques que le système demande à l'utilisateur ou calcule lui-même, mais surtout pour confirmer une déduction qui semble relativement sûre.

[3] Certains systèmes n'ayant pas reçu de nom nous convenons pour ce paragraphe de leur donner le nom de leurs créateurs.

À notre connaissance, tous les systèmes experts en décodage passent par une étape de segmentation du signal de parole. Cette segmentation est souvent automatique (Aphodex, Serac, Mizoguchi et al., Méloni et al.) et fait généralement appel à la courbe d'énergie en basse fréquence pour repérer les noyaux vocaliques. Le système de Mizoguchi fait exception puisque la segmentation est construite à partir de l'évolution des formants et de la courbe d'énergie.

La plupart des systèmes experts utilisent un mécanisme de scores (généralement celui de MYCIN) pour estimer le degré de confiance d'un fait déduit.

Si l'architecture générale varie peu d'un système à l'autre, il existe en revanche de nombreuses différences concernant la nature de la connaissance acoustico-phonétique. En fait ces différences se justifient par le type des données qu'exploite le système expert :

- énergies d'un vocodeur à canaux ; Serac opère à partir de la sortie d'un vocodeur à canaux représentant la bande passante téléphonique.
- fréquence fondamentale. La plupart des systèmes utilisent une information sur le voisement extraite automatiquement (Aphodex, Serac, Méloni et al.) ou rentrée directement par l'utilisateur.
- paramètres de bas niveau extraits à partir du signal ou du spectre : densité de passages par zéro, énergies dans certaines bandes de fréquence (pour détecter les voyelles), centre d'inertie des spectres... Ces indices de bas niveau sont utilisés par la plupart des systèmes (Aphodex, Méloni et al. notamment).
- événements acoustiques : barre d'explosion, bruit, limite inférieure du bruit, formants. Ces informations peuvent être évaluées automatiquement à partir du signal (Aphodex, Méloni et al.) ou être fournies par l'utilisateur (Stern et al., Zue et al.). Le système de Mizoguchi et al. utilise essentiellement les trajectoires formantiques extraites du spectrogramme grâce à un algorithme de classification.

La grande variété des données de départ se retrouve au niveau de la connaissance qu'exploite de système expert.

La connaissance de Serac porte à la fois sur le processus de décodage, c'est-à-dire les actions à accomplir pour mener à bien le décodage, et sur l'expertise phonétique en partie issue de Keal. Les règles phonétiques permettent donc d'évaluer les différents attributs du signal (trait vocalique, avant, arrière, fermé ...) et d'en déduire d'autres indices de plus haut niveau (nasalité, lieu d'articulation).

Les connaissances du système de Mizoguchi et al. portent quasi exclusivement sur la forme des trajectoires formantiques qui permettent de segmenter le signal et de reconnaître les voyelles et les consonnes. Comme ces règles ne font pas directement intervenir le contexte, un second ensemble de règles (dites de "re-reconnaissance") est destiné à capturer les phénomènes de coarticulation pour corriger les erreurs commises au niveau inférieur.

Les connaissances du système de Méloni et al., stockées sous forme de clauses PROLOG, formalisent la recherche d'événements acoustico-phonétiques et d'autre part de traits pseudo-phonétiques permettant de limiter le nombre des phonèmes candidats. Les règles sont construites de manière à faire intervenir les paramètres les plus discriminants.

Les systèmes Aphodex, Stern et al. et Zue et al. sont directement issus d'un effort de formalisation de la lecture de spectrogrammes, par un ou plusieurs experts. Il est donc normal que les connaissances fassent appel à des événements acoustiques (des attributs acoustiques qualitatifs pour Zue et al.) observables sur un spectrogramme. Les différences apparaissent plutôt dans le domaine de la structuration de la connaissance. Lamel distingue plusieurs niveaux : mesures acoustiques, attributs acoustiques qualitatifs, traits phonétiques et phonèmes. Les règles portent sur la manière de déduire les attributs acoustiques qualitatifs puis les traits phonétiques ; ces derniers définissent de manière directe les phonèmes et constituent un intermédiaire obligé. Pour Stern et al. comme pour Aphodex, les règles correspondent à une description acoustique des sons ; Stern et al. ajoutent des règles qui permettent d'émettre comme hypothèses certains sons en fonction du contexte phonétique.

Malgré certains résultats satisfaisants les systèmes experts sont loin d'égaler les phonéticiens. On ne peut pas invoquer le fait que les experts humains exploitent leurs connaissances linguistiques, car l'émission d'hypothèses lexicales a plutôt tendance à les induire en erreur. Les différences de résultats doivent donc venir des limites de l'implantation de l'expertise. La plupart des bons résultats ont été obtenus lors d'expériences limitées ; ainsi Zue et al. n'étudient que l'identification des occlusives, Stern et al. ne retiennent quant à eux que 13 phonèmes du Français (certaines classes phonétiques étant représentées par un unique phonème). Ces résultats sont par ailleurs obtenus après l'extraction manuelle des indices acoustiques sur le spectrogramme ; il ne s'agit donc pas à proprement parler de décodage automatique.

Le développement d'Aphodex au CRIN a permis de mettre à jour les limites de l'approche experte :

- Les modules de segmentation et de recherche d'indices acoustiques fournissent des résultats entâchés d'erreur (de l'ordre de 5% d'omission en ce qui concerne la segmentation).
- Traiter la totalité de la phonétique du Français demande un gros effort d'extraction de la connaissance. Ce travail a demandé environ deux ans de réunions hebdomadaires pour expliciter la connaissance et le raisonnement de l'expert.
- L'architecture du système Aphodex se prête mal à l'étude globale d'une phrase ; il est à l'heure actuelle impossible de simuler "le coup d'œil" global du phonéticien qui exploite les ressemblances et dissemblances entre les segments de parole non contigus.

- La lecture de spectrogrammes, telle qu'elle est pratiquée par le phonéticien, ne repose pas que sur la connaissance visuelle des sons mais en grande partie sur la connaissances des phénomènes articulatoires et des événements acoustiques qui les accompagnent. Ces deux niveaux de connaissance sont peu ou très mal rendus par les règles de production.

Les règles de production ne sont pas la seule voie qui ait été explorée pour stocker l'expertise. Le formalisme des prototypes ("frames") a été retenu pour plusieurs projets. Damestoy [Damestoy 86] a choisi des prototypes de phonèmes définis en termes d'événements acoustiques identiques à ceux utilisés par les règles d'Aphodex. Le raisonnement est directement contrôlé par les "frames" des niveaux supérieurs qui orientent le décodage vers les prototypes dont la classe phonétique est la même que celle de l'instance à identifier. Green et al. [Green 86] définissent leurs prototypes en termes de descripteurs de sketch acoustique. Le sketch acoustique est une stylisation en segments de droite des trajectoires formantiques qui apparaissent sur le spectrogramme [Cooke 88]. Les formants sont extraits à partir de spectrogrammes correspondant à une gamme de filtrages s'étendant depuis les filtres bande étroite jusqu'au filtres bande large. Il s'agit donc d'une analyse multi-échelle, les spectrogrammes bande étroite permettant de compléter les informations apportées par les spectrogrammes bande large [4]. Plus encore que les systèmes experts, les systèmes à base de prototypes n'ont donné lieu qu'à des tests limités :

- les phonèmes /l m n R/ pour Damestoy,
- les semi-voyelles /j w/ pour Green et al.

Il est donc bien difficile de mesurer l'intérêt de cette approche en ce qui concerne le décodage acoustico-phonétique : les connaissances étant de la même nature que celle des systèmes experts (c'est-à-dire descriptives) la supériorité éventuelle ne peut venir que de l'architecture.

3 Conclusion

La confrontation des connaissances actuelles des phénomènes de la parole continue et de celles effectivement exploitées en reconnaissance, permettent d'appréhender avec plus de précision le niveau actuel des techniques de décodage acoustico-phonétique. Globalement il apparaît que les systèmes de décodage ne savent pas ou très peu bénéficier des recherches sur les phénomènes de la parole. Deux voies s'offrent alors aux investigations des chercheurs pour pallier ces faiblesses :

[4] Cette technique est clairement apparentée au suivi de formants ; les résultats auxquels elle conduit ne sont néanmoins qu'une approximation grossière des formants puisque les pics des spectres bande étroite représentent les harmoniques du fondamental.

- Réduire le langage à décoder (un petit vocabulaire et de fortes contraintes syntaxiques par exemple) et recourir aux techniques globales qui permettent de construire un système de reconnaissance suffisamment fiable pour une application de parole continue, multilocuteur et un vocabulaire d'environ un millier de mots ; encore faut-il pour cela raffiner à l'extrême les techniques existantes.
- Tenter d'implanter l'expertise d'un phonéticien adepte en lecture de spectrogrammes. Il n'y a plus alors de limite au langage à reconnaître autres que celles imposées par l'implantation de l'expertise. C'est évidemment là que surgissent nombre d'obstacles à commencer par l'ampleur du travail pour transmettre à la machine les descriptions de tous les sons d'une langue. Plus grave, sans doute, est le manque de profondeur de cette connaissance. Elle ne peut être que purement descriptive et il n'est pas encore possible de combiner observation acoustique et analyse articulatoire au sein d'un même système de reconnaissance.

Il serait assez séduisant de pouvoir utiliser les techniques de la physique qualitative [dK 84] [Dormoy 89] pour lier les paramètres articulatoires - pour simplifier l'évolution dans le temps de la forme du conduit vocal - aux événements acoustiques que l'on observe sur le spectrogramme. Cela supposerait que l'on puisse simuler correctement les phénomènes articulatoires. L'état actuel des techniques de synthèse articulatoire montre qu'il reste encore beaucoup de progrès à faire pour entreprendre la construction d'un système de "décodage articulatoire". Les difficultés que nous avons évoquées pour estimer les paramètres de la nasalité (§1.3) ne donne qu'une faible idée des obstacles à surmonter.

Chapitre 5

Les triplets en décodage acoustico-phonétique

1 introduction

Malgré toutes les voies qui ont déja été explorées, le décodage acoustico-phonétique reste le point faible des systèmes de dialogue homme-machine. Notre conviction est que l'approche experte est la voie qui permettra à l'avenir d'intégrer le plus facilement, et pour le profit maximal, les connaissances des phénomènes de la parole. Nous décrivons dans ce chapitre l'approche experte sur laquelle nous avons travaillé, en l'occurrence les triplets phonétiques (un son en contexte).

Nous évoquons d'abord les origines de cette approche du décodage acoustico-phonétique. Nous montrons ensuite comment il serait possible d'insérer une approche à base de triplets dans le cadre plus général d'un système de compréhension de la parole continue multi-locuteur et nous proposons une solution originale pour assurer que le raisonnement poursuivi au cours de la reconnaissance repose sur un ensemble d'hypothèses compatibles entre elles.

Le décodage proprement dit se décompose en deux étapes successives. La première conduit à proposer pour chaque segment de parole la liste des triplets qui s'apparient le mieux à l'instance de triplet à identifier. La seconde étape est destinée à augmenter la cohérence de la solution globale (c'est-à-dire au niveau de la phrase) en assurant que les contraintes, qui existent entre les triplets de la base de connaissances qui ont été proposés comme solutions, sont aussi vérifiées par les instances de triplets de la phrase à décoder.

2 Origines

Parmi les sytèmes experts de décodage acoustico-phonétique, Aphodex est sans doute l'un des rares exemples de systèmes autonomes puisqu'il réalise auto-

matiquement l'extraction des toutes les informations acoustiques qu'il utilise. Cette intégration "verticale" des modules de décodage (depuis la segmentation jusqu'à la construction du treillis phonétique) permet sans doute, plus facilement qu'un système purement phonétique, d'appréhender les faiblesses de l'approche experte[1]. L'une des faiblesses (déja partiellement évoquée au chapitre précédent) est liée au volume de l'expertise elle-même et aux problèmes que l'expert a rencontrés pour la formaliser. Comme il n'est pas possible de décrire les phénomènes de coarticulation par un nombre limité de règles générales, le phonéticien est peu à peu amené à ajouter de nombreuses règles décrivant le plus grand nombre de contextes possibles pour un son donné. Outre le problème d'être assuré que les règles ne se contredisent pas entre elles, se pose la question de savoir précisément ce que cette volumineuse connaissance recouvre. Il est alors apparu au phonéticien qu'il est plus simple de regrouper tout ce qui concerne un son dans un contexte phonétique donné (par exemple : /apo/ que nous noterons désormais $/_a p_o/$) en une seule entité : le triplet phonétique.

Mais il n'y a pas que ces origines "phonétiques" pour plaider en faveur du triplet phonétique. Les expériences de systèmes de décodage acoustico-phonétique à base de connaissances qui ont été décrites au chapitre précédent montrent que le raisonnement phonétique est peu profond ; il apparaît tout au plus que deux ou trois niveaux : les événements acoustiques (parfois ces événements sont mêmes rentrés directement par l'utilisateur), un niveau correspondant grossièrement aux traits phonétiques et enfin les sons. La vocation des systèmes experts de décodage est donc plus de faire de la classification que du raisonnement. Cette fois encore, mais pour des raisons informatiques, un système à base de prototypes semble être le plus approprié. L'unité de parole correspondant aux prototypes doit être aussi insensible que possible aux effets contextuels tout en restant suffisamment petite pour être utilisée dans le cadre d'une application parole continue multi-locuteur : un son en contexte, autrement dit un "triplet phonétique" semble donc être le meilleur compromis possible.

Le triplet phonétique (que nous décrirons en détail au §4) semble donc particulièrement adapté au problème du décodage tant du point de vue du phonéticien parce que ce grain de connaissance permet de regrouper toutes les informations concernant un contexte phonétique donné, que du point de vue du cogniticien pour lequel le triplet semble tout à fait approprié au type de raisonnement utilisé en décodage.

La présentation des phénomènes de la parole du chapitre précédent permet néanmoins d'apporter quelques compléments indispensables :

- un triplet ne doit pas inclure en entier les sons encadrant le son central, sinon il devient à son tour soumis aux effets contextuels auxquels il doit précisément échapper ;

[1] En effet un système purement phonétique auquel l'utilisateur décrit le spectrogramme en termes d'événements acoustiques d'assez haut niveau (barre d'explosion, formant, bruit de friction...) permet à notre avis plus de vérifier la pertinence des connaissances que de fournir le module de décodage d'un système de dialogue.

- les effets de coarticulation ne disparaissent pas totalement ; c'est le cas dans un contexte de coarticulation labiale V_1C_2 où V_1 influence V_2 (effet "carabine" décrit au § 1.1 du chapitre 4) ;
- les effets d'accentuation ; assez sensibles au niveau des références fréquentielles ils ne peuvent pas être directement pris en compte par le triplet.

En première approche nous considérerons qu'il est possible de négliger les effets résiduels de coarticulation et les effets d'accentuation. Nous verrons dans le paragraphe suivant comment il est possible de revenir sur ce choix et plus généralement comment replonger un système à base de triplets dans le cadre plus général d'un système de compréhension de la parole.

3 Triplet phonétique et compréhension de la parole continue

Selon nous, l'utilisation des triplets en reconnaissance de la parole présente deux avantages majeurs :

- Comme le triplet phonétique est un prototype de son en contexte, il est possible de le considérer globalement et de lui faire subir des déformations sous l'effet de l'accentuation ou de la vitesse d'élocution. Lors du processus de décodage, les faits justifiant la présence d'un triplet seront complétés par l'hypothèse qui explique la déformation (par exemple l'accentuation).
- Dans un système expert, une déduction phonétique (ou encore un son reconnu) peut résulter du déclenchement de plusieurs règles. Pour conserver la trace de ce raisonnement il faut donc créer une nouvelle structure d'accueil à cette seule fin. Au contraire une instance de triplet est simultanément raisonnement et trace de raisonnement.

En extrapolant à partir de ces deux points, il est possible d'imaginer ce que pourrait être un système de compréhension de la parole continue. Nous allons limiter notre attention, pour simplifier l'exposé, aux niveaux inférieurs (depuis le signal temporel jusqu'au niveau lexical inclus). Voici une liste non exhaustive des sources d'information dont nous pourrions disposer pour décoder une phrase :

Signal temporel : C'est la phrase énoncée par le locuteur.

Segmentation : Il s'agit de découper le signal temporel en unités de parole. Le processus de segmentation fait essentiellement appel au calcul de paramètres dérivés directement du signal temporel (énergie dans diverses bandes de fréquence, densité de passages par zéro...).

Événements acoustiques : Il s'agit notamment des trajectoires formantiques, des barres d'explosion et des bruits de friction. Il existe un certain nombre

d'algorithmes pour les détecter (cf. chapitre consacré aux formants) soit directement sur le signal temporel, soit sur le spectrogramme.

Indices acoustiques : Il s'agit par exemple de la pince vélaire (rapprochement de F2 et F3 dans le contexte des consonnes vélaires) ou encore de la position relative des concentrations d'énergie de la barre d'explosion par rapport aux formants.

Micromélodie : Il s'agit de l'évolution temporelle de F0 (la fréquence fondamentale) sur un intervalle de temps voisin de la durée d'un segment de parole. Pour tenir compte des phénomènes micromélodiques au cours du décodage il faut avoir à sa disposition un détecteur de fondamental très précis.

Triplets phonétiques : Ce sont les prototypes contextuels des sons décrits en termes d'événements et d'indices acoustiques.

Accentuation et vitesse d'élocution : Ces phénomènes commencent à être suffisamment bien maîtrisés (cf. chapitre précédent) pour qu'il soit possible d'en tenir compte en décodage.

Prosodie : Il s'agit de l'évolution au cours de la phrase (à la différence de la micromélodie) de la fréquence fondamentale, de l'intensité et de la durée des segments syllabiques. Les paramètres prosodiques peuvent servir en décodage pour aider à localiser les frontières lexicales et syntaxiques.

Phonologie : Grâce aux règles d'assimilation il est possible de déduire, à partir d'une suite de phonèmes, la suite des sons correspondante, les sons étant définis en termes de traits phonétiques.

Lexique : C'est l'ensemble des mots du langage à partir desquels a été construite la phrase à décoder.

L'idée directrice du décodage que nous proposons est la suivante : expliquer de manière cohérente le signal observé grâce à toutes les sources de connaissances que nous venons d'évoquer. La cohérence est assurée lorsque toutes les contraintes propres à un niveau de connaissance ou liant plusieurs niveaux de connaissances sont vérifiées. Comme les solutions incohérentes sont écartées il ne devrait rester, à cause des erreurs de détection d'indices (segmentation, événements acoustiques, indices acoustiques, fondamental...) aucune solution, sauf cas idéal de décodage. Lorsque le raisonnement aboutit à une incohérence, il faut émettre l'hypothèse que l'un des paramètres extrait du signal est incorrect pour continuer. Une solution cohérente ne peut pas être justifiée simultanément par un fait et sa négation. Ce type de gestion du décodage acousto-phonétique est clairement apparenté au raisonnement hypothétique [dK 86]. Les travaux dans ce domaine de l'intelligence artificielle n'ont donné lieu qu'à de rares applications réelles et il est donc difficile d'évaluer a priori l'intérêt du raisonnement hypothétique pour le décodage acoustico-phonétique. Nous allons donc maintenant présenter, sur un exemple limité, ce que pourrait être le raisonnement associé au décodage.

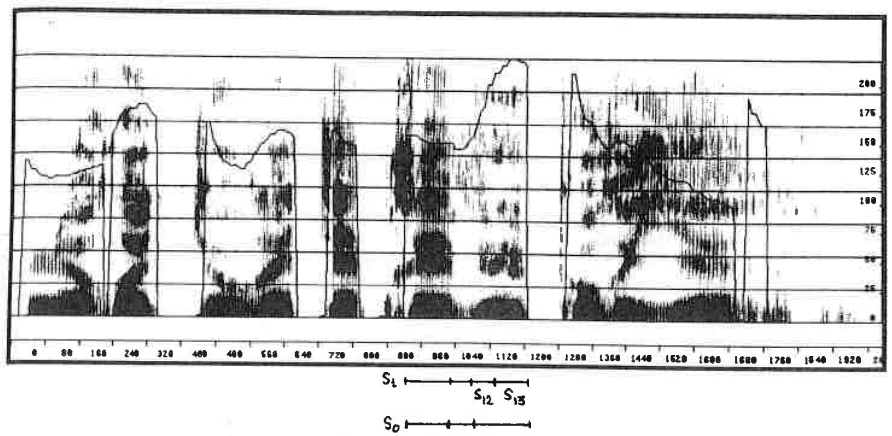


Figure 5.1 : Le retour était prévu pour hier.

Considérons le morceau de parole que nous avons délimité sur le spectrogramme (fig. 5.1). Nous n'allons pas faute de place, développer le raisonnement en entier mais insister sur quelques points. Nous supposerons qu'il y a deux segmentations possibles s_0 et s_1 indiquées sur le spectrogramme.

1. Le suivi de formants permet de conclure qu'il peut y avoir deux suivis de formants possibles et mutuellement exclusifs pour la fin du segment s_{00} :
   H_0 = {(F1 stable, F2 descend, F3 descend)} et
   H_1 = {(F1 monte, F2 descend, F3 monte)}
   les triplets possibles pour chacune des hypothèses étant :
   $D_0 = \{/_R e_v/, /_R e_b/, /_R \emptyset_b/$ pour s_{00} et $/_e v_y/, /_e b_y/$ pour $s_{01}\}$
   $D_1 = \{/_e R_y/$ pour $s_{01}\}$
   Il n'est pas possible de construire un mot avec des triplets de D_0 et D_1 puisque les hypothèses H_0 et H_1 qui justifient D_0 et D_1 s'excluent et rendent donc incohérent tout raisonnement qui repose conjointement sur H_0 et H_1.

2. Supposons que l'algorithme de segmentation puisse conduire aux deux segments vocaliques s_{12} et s_{13}. Le fondamental s'élève fortement sur les segments s_{12} et s_{13}. Une règle de prosodie assure qu'il n'est pas possible d'avoir F0 qui s'élève pendant deux voyelles contiguës. Les ensembles cohérents et mutuellement exclusifs de faits sont donc :
   $H_0 = \{s'_{12} = s_{12} \cup s_{13}$, F0 s'élève pendant $s'_{12}\}$
   ou H_1 = {deux segments s_{12} et s_{13}, F0 mal détecté (par exemple par doublement de fréquence)}
   Aucun raisonnement ne pourra être justifié simultanément par H_0 et H_1.

3. L'élévation de F0 pendant le segment /y/ est sans aucun doute une montée de continuation [Vaissiere 80] (c'est en effet la valeur maximale de F0). Ce fait n'est cohérent qu'avec une hypothèse lexicale se terminant par le segment /y/. Ainsi "vu" et la montée de continuation sont compatibles ; pour justifier la présence de "utopie" il faudrait faire l'hypothèse d'une erreur sur F0.

Les contraintes que nous venons d'évoquer ont essentiellement une portée locale ; elles relient en effet des segments de parole contigus. Il est aussi possible d'envisager, notamment au niveau des triplets, des contraintes de portée plus lointaine permettant de vérifier que les relations acoustiques liant deux triplets de référence existent aussi entre les segments de parole qu'ils représentent (nous verrons comment exploiter ce type de contrainte au §8).

Nous nous sommes placés implicitement dans le cadre d'un raisonnement ascendant du niveau acoustique vers le niveau lexical. Il est bien sûr possible de partir du niveau lexical pour répertorier les faits qui peuvent justifier un mot et chercher s'ils peuvent ou non être compatibles avec les faits déja connus.

Pour juger la qualité d'une solution globale il faut tenir compte des hypothèses d'erreur de détection qu'il a fallu émettre pour rendre la solution cohérente. Les deux extrêmes étant :

- une solution sans aucune hypothèse de mauvaise détection est une solution idéale ;
- une solution uniquement justifiée par des hypothèses de mauvaises détections doit être écartée.

Il reste évidemment deux problèmes cruciaux pour adopter ce mécanisme de raisonnement en décodage acoustico-phonétique :

- quel score donner à une solution intermédiaire entre les deux extrêmes précédents ?
- comment contrôler ce raisonnement ? c'est-à-dire : quand faut-il arrêter de construire une solution trop entâchée d'erreurs ? Les contraintes sont-elles suffisantes pour éviter l'explosion combinatoire ?

Malgré ces questions non résolues, ce type de décodage nous semble présenter deux avantages décisifs : unifier les différents niveaux de décodage, assurer que la solution proposée est cohérente.

4 Définition et structure d'un triplet

Le triplet phonétique est la description en termes d'événements et d'indices acoustiques[2] d'un son en contexte. Les sons contigus ne sont pas décrits en

[2]Il est difficile de choisir des termes qui n'aient pas pris au cours du temps plusieurs significations ; événement et indices acoustiques ont été utilisés dans de nombreux contextes. Voici le sens que nous leur attribuons dans ce qui suit. Un événement acoustique est un objet acoustique irréductible (du moins au niveau du spectrogramme), par exemple : une trajectoire formantique, une barre d'explosion. Dans un système de DAP il provient directement d'un détecteur acoustique (algorithme de suivi de formants pour les trajectoires formantiques). Un indice acoustique est une information de plus haut niveau dont le phonéticien connaît ou suspecte l'origine articulatoire, par exemple : la pince vélaire (le rapprochement de F2 et F3 au voisinage d'une frontière de consonne vélaire), la position relative d'une concentration d'énergie par rapport aux formants. En général un indice acoustique fait intervenir plusieurs événements acoustiques simultanément.

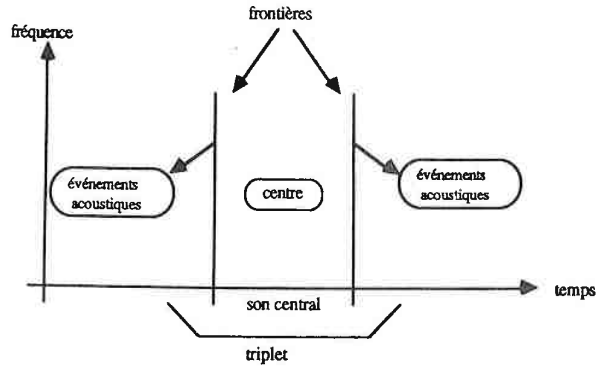


Figure 5.2 : Structure d'un triplet

entier mais uniquement au voisinage du son central.

Un triplet est structuré de la manière suivante (fig. 5.2) :

- les deux frontières du son central auxquelles est attachée la plupart des événements acoustiques,
- la partie centrale du triplet qui est destinée aux caractéristiques que le triplet ne peut pas partager avec ses voisins.

Nous ne concevons pas la frontière comme une limite précise mais comme la zone où se manifestent le plus clairement les effets de transition d'un son à l'autre. Cette structure permet de privilégier les parties dynamiques d'un son tout en conservant une partie centrale pour décrire la zone stable si elle existe. Le fait de ne pas attacher les événements acoustiques traduisant les transitions au triplet mais aux frontières permet un partage très simple de ces informations entre deux sons contigus.

Voici le liste des événements acoustiques qui peuvent être attachés à une frontière :

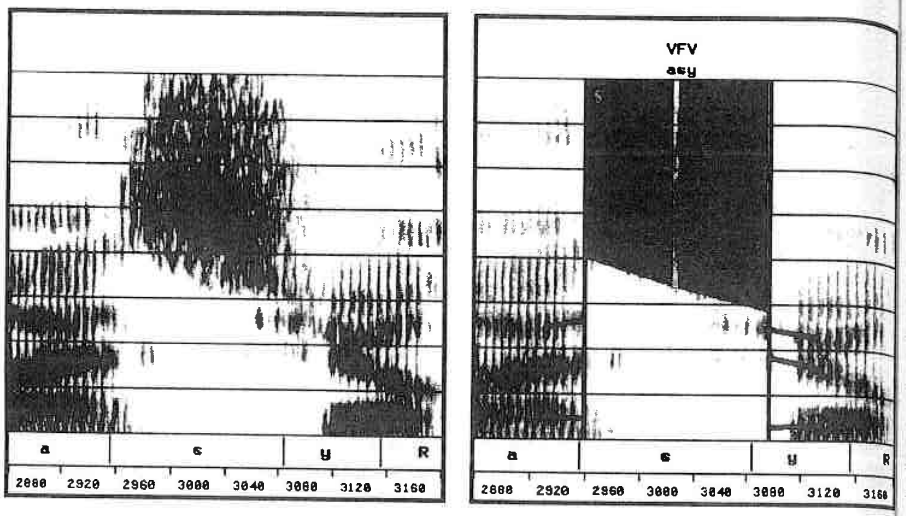


Figure 5.3 : Formants et bruit de friction d'un triplet

formants : une trajectoire formantique au voisinage d'une frontière est décrite par deux segments de droite (fig. 5.3). Il faut donc connaître la fréquence du formant à la frontière et les deux pentes. Pour définir plus complètement la trajectoire formantique nous ajoutons la valeur du formant au centre du triplet. Bien sûr, une trajectoire formantique n'est définie que si elle existe. Nous ne considérons que les formants F1, F2 et F3, éventuellement complétés dans le cas d'un son nasalisé par FN1 et FN2.

bruit de friction : Un bruit de friction est défini (fig: 5.3) par la limite inférieure de bruit et son niveau d'énergie moyen (faible, fort ou intense). La limite inférieure de bruit est définie de la même façon qu'une trajectoire formantique : la fréquence sur la frontière et la pente.
remarque : ce choix de représentation pourrait sembler curieux puisqu'il conduit à couper en deux parties un bruit de friction correspondant à un son fricatif. En fait cela permet de prendre en compte facilement d'une part la forme de la limite inférieure de bruit qui peut être montante puis descendante (par exemple pour une fricative entourée par deux labiales) et d'autre part le bruit de friction qui peut apparaître dans certains "clusters"[3] commençant par une occlusive.

[3]Un "cluster" est une suite de consonnes qu'il est difficile de partager au niveau acoustique (par exemple : /pR/, /pl/, /stR/...). Nous avons conservé le mot d'origine anglaise qu'utilisent de nombreux phonéticiens français.

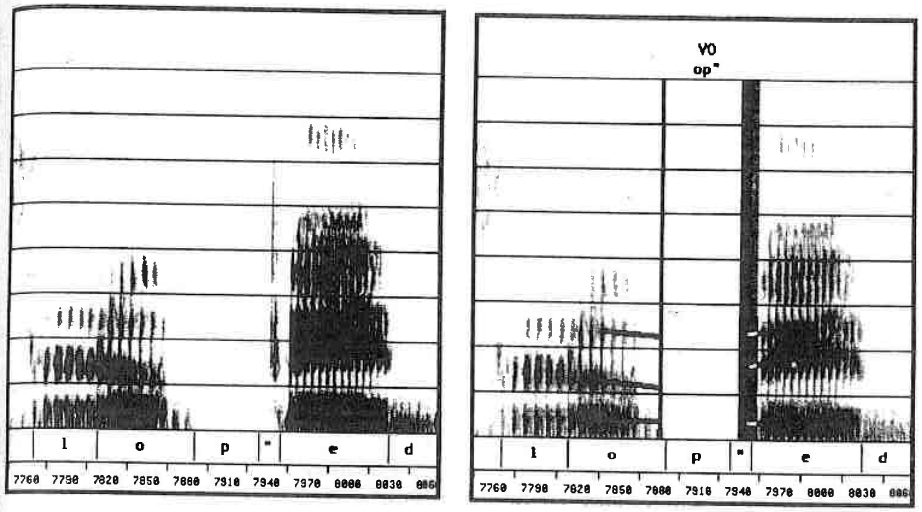


Figure 5.4 : Représentation d'une barre d'explosion

barre d'explosion : [4] Une barre d'explosion (fig. 5.4) est définie par la durée, la liste des principales concentrations d'énergie et le niveau d'énergie moyen (faible, fort ou intense).

La partie centrale du triplet est définie de la manière suivante :

durée : c'est la durée du son central du triplet ;

formant : à chaque trajectoire formantique existante correspond la valeur du formant au milieu du son central, ce qui permet de compléter les informations concernant l'allure de la trajectoire au voisinage des frontières ;

profil micromélodique : Initialement, nous voulions ajouter l'allure de la courbe du fondamental au cours du triplet. Étant donné qu'il est très difficile de disposer d'un algorithme de détection du fondamental très précis nous avons renoncé, pour l'instant, à ajouter une information de ce type mais nous pensons pouvoir le faire à l'avenir.

La description précédente est construite à partir d'informations acoustiques de bas niveau. Cela ne signifie pas qu'elles apparaissent directement dans le signal temporel ; en revanche elles sont tout à fait spécifiques à la parole sans être suffisantes pour décoder une phrase. L'expertise du phonéticien repose sur ces informations acoustiques en les structurant sous la forme d'indices qui sont l'image acoustique d'événements articulatoires très significatifs quant à

[4]ou encore burst qui est très largement utilisé par les phonéticiens.

la nature des sons énoncés. Comme il n'est pas possible à l'heure actuelle de pouvoir déduire d'un événement articulatoire son image acoustique, nous avons ajouté à la description d'un triplet les indices acoustiques reconnus pertinents au niveau articulatoire par le phonéticien. Ces indices sont représentés par des relations liant plusieurs événements acoustiques de base ; jusqu'à présent nous avons défini les relations suivantes :

rapproche(formant k, formant k±1, face) signifie que les formants se rapprochent (dans le temps), face désignant la partie gauche ou droite d'une frontière ;

écarte(formant k, formant k±1, face) c'est le contraire de la relation précédente ;

parallèleDescendant(formant k, formant k±1, face) signifie que les pentes des deux formants sont négatives ;

parallèleMontant(formant k, formant k±1, face) signifie que les pentes des deux formants sont positives ;

burstVoisin(formant k, face) signifie que la principale concentration d'énergie de la barre d'explosion est proche du formant k ;

burstEntre(formant k, formant k+1, face) signifie que la principale concentration d'énergie est entre les formants k et k+1.

L'existence simultanée de ces deux facettes permet, comme nous le verrons au §7.2, d'orienter le décodage soit vers la recherche de triplets en fonction d'indices quand ils apparaissent clairement dans le signal, soit vers la recherche de triplets en fonction de leur réalisation acoustique quand peu d'indices ont été découverts.

La description que nous venons de faire est valable pour les triplets de référence comme pour les instances de triplets d'une phrase à décoder. Il existe néanmoins quelques différences, notamment en ce qui concerne les indices acoustiques, qui se justifient par des contraintes d'implantation (cf §9).

5 Organisation des connaissances

Le Français compte 34 phonèmes (en ne considérant qu'une seule voyelle pour /œ̃/, /ɛ̃/, pour /a/ et /ɑ/ et en ne comptant pas le phonème d'origine anglaise /ŋ/). Cela permet donc potentiellement de construire $34^3 = 39304$ triplets. Outre les problèmes d'apprentissage que nous décrirons au paragraphe suivant cela semble représenter de sérieux problèmes d'organisation de la connaissance. En fait nous allons voir tout de suite que de nombreux triplets ne se rencontrent pas en Français. Boé et Tubach ont analysé [Tubach 85] un ensemble de conversations très variées, transcrites phonétiquement et représentant un volume total de 304000 sons. En tenant compte des poses, les différents corpus ne comptent plus que 274000 triplets, ce qui conduit finalement à 12019 triplets

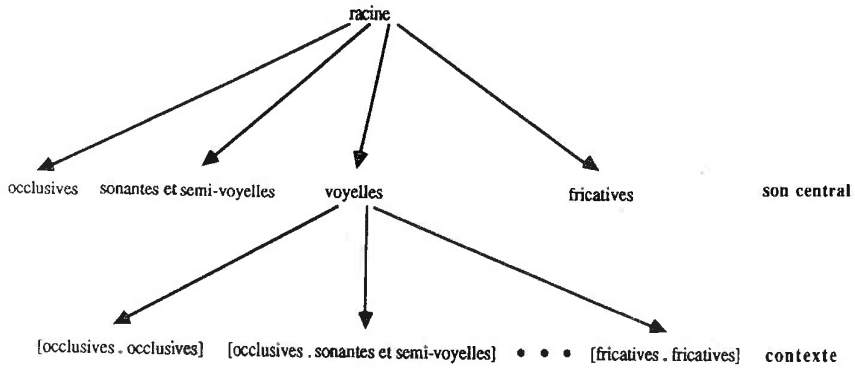


Figure 5.5 : Hiérarchie des triplets

possibles, c'est-à-dire à peine 30% des triplets potentiels. Comme ce nombre se stabilise assez rapidement au cours de l'exploration statistique, cela laisse penser qu'il est très proche du nombre de triplets existant en Français. Qui plus est, les 1650 triplets les plus fréquents représentent 75% des occurrences rencontrées dans les corpus. Les auteurs font d'ailleurs remarquer que parmi les triplets les plus fréquents il y a manifestement deux classes : ceux qui le sont parce que leurs constituants le sont également (par exemple /tRa/ /pas/ /aRi/) et ceux qui le sont par eux-même (par exemple /sjõ/ /bjɛ̃/ /âfɛ̃/).

Il peut aussi être intéressant de se placer dans le cadre d'une application limitée de la reconnaissance de la parole continue. Ainsi Lee [Lee 88] dénombre 2381 triplets possibles pour une application d'interrogation de base de données navales en anglais.

Bien que le nombre de triplets ne soit pas aussi considérable qu'on pourrait le craindre il est toutefois nécessaire de structurer les triplets de référence. Il existe a priori deux solutions classiques pour organiser la base de connaissances :

- les traits phonétiques (par exemple : labial , vélaire, dental),
- les classes phonétiques (voyelles, occlusives, fricatives, semi-voyelles, sonantes)

Souvent d'ailleurs, ces deux solutions sont combinées ([Caelen 88]) ; on distinguera ainsi les occlusives vélaires, labiales ou dentales.

Nous avons choisi de structurer les triplets en utilisant seulement le profil phonétique (c'est-à-dire la classe phonétique des constituants du triplets) et le voisement (fig. 5.5). Nous n'avons pas fait intervenir les traits phonétiques car cela présente le risque d'assimiler abusivement trait phonétique et indice acoustique. Le nombre de triplets potentiels d'une classe reste donc relativement important ; la classe [Voyelle Sonante Voyelle] compte ainsi $13 \times 6 \times 13 = 1014$ prototypes (en fait sans doute entre le tiers et la moitié de ce nombre étant donnés les nombreux triplets impossibles).

6 Apprentissage

Comme dans toute approche de décodage, la construction des formes de référence requiert une attention extrême. En revanche à la différence des techniques globales (modèles de Markov, programmation dynamique) nous devons concevoir les outils d'apprentissage de manière à exploiter au mieux le jugement phonétique de l'expert. L'environnement d'apprentissage doit donc être aussi convivial que le permettent les postes de travail actuels et tirer parti des corpus de parole accessibles. Nous avons donc développé et intégré dans Snorri un environnement de travail destiné à l'approche par triplets et en particulier un éditeur de prototypes. Il permet de créer et de modifier les triplets avec ou sans le support du spectrogramme qui apparaît par transparence derrière les triplets. L'utilisateur se sert de la souris pour "dessiner" les formants, barres d'explosion et autres événements acoustiques. Comme il n'y a pas de différence entre la description en termes d'événements acoustiques d'un triplet de référence et une instance, l'utilisateur peut utiliser les détecteurs automatiques (notamment de trajectoires formantiques décrits à la fin du chapitre sur les formants) pour amorcer son travail et corriger ou compléter les triplets ainsi construits.

Cet environnement de travail offre donc deux solutions pour l'apprentissage :

1. l'expert peut créer ex nihilo les prototypes en utilisant ses propres connaissances si elles lui semblent suffisantes ;
2. quand il s'agit de construire un triplet plus difficile, l'utilisateur peut d'abord le rechercher dans les corpus de parole dont il dispose (cf. chapitre consacré à Snorri). Ensuite, à partir des occurrences présentes dans le corpus il peut alors dégager un prototype qui lui semble modéliser correctement le triplet à construire.

Recourir à un expert ne se justifierait pas si son travail se limitait à décrire acoustiquement les triplets. Son travail est surtout de donner les bons indices acoustiques, c'est-à-dire ceux qui découlent d'un unique événement articulatoire (ou du moins d'événements articulatoires très dépendants), et non pas de plusieurs, et qui ne sont pas liés au locuteur. L'expert est indispensable dans la mesure où lui seul peut assurer la pertinence du modèle en faisant le lien entre

acoustique et articulatoire. Ceci est d'autant plus indispensable d'ailleurs que l'on connaît certains sons qui présentent une grande variabilité articulatoire pour un même contexte. Ainsi $/_a l_a/$ donne t-il lieu à deux types de réalisations acoustiques [Lonchamp 84] : l'une avec une trajectoire formantique très intense et correspondant à la fusion de F3 et F4 et l'autre avec disparition de la trajectoire de F3 neutralisée par un zéro, et une très faible trajectoire pour F4. F. Lonchamp suppose que la langue est en position dentale pour la seconde réalisation et en position prépalatale (c'est-à-dire nettement plus en arrière) pour la première. Cela signifie donc qu'il faut sans doute construire deux triplets correspondant aux deux types d'articulation... et qu'un bon amateur en phonétique se serait complètement fourvoyé en construisant un modèle "moyen" peu significatif. Ce type de problème se rencontre rarement (a priori pour certains triplets avec /l/ comme son central) mais il permet de mesurer l'importance de disposer de vastes corpus qui permettent à l'expert de déceler les gestes articulatoires et donc les bons indices.

Jusqu'à maintenant nous avons travaillé (notamment pour les algorithmes de suivi de formants) sur le corpus "La bise et le soleil" (extrait de BDSONS [Carre 84]) qui contient 30 répétitions par des locuteurs d'origines assez diverses d'une petite histoire d'environ 30 secondes. Malheureusement ce corpus comporte de nombreuses répétitions des mots "bise" et "soleil" et ne contient finalement que 400 triplets. En considérant tous les autres corpus dont nous disposons (10 phrases de Combescure [Combescure 81] et 17 phrases du corpus de Mella [Mella 89]) nous arrivons tout au plus à 1040 triplets, ce qui est encore fort peu. En fait il est même impossible de composer des phrases de test avec ces triplets sans réutiliser les mots dont ils sont extraits. Nos test porteront donc sur les occlusives en contexte présentes dans les corpus ; ce problème ayant été abordé à de nombreuses reprises cela nous permettra de mieux apprécier les forces et les faiblesses de notre approche.

Étant donnée l'ampleur de l'apprentissage que nécessiterait une langue comme le Français, on ne peut manquer d'évoquer la possibilité de recourir aux techniques de l'apprentissage automatique pour construire les prototypes à partir d'un certain nombres d'exemples. Comme le mode de description des triplets est très symbolique cela suggère d'utiliser des techniques d'apprentissage symbolique, par exemple la classification conceptuelle [Michalski 84] que P. Nugues utilise, en dehors du cadre simpliste des exemples d'école, pour classifier des gels d'électrophorèses bidimensionnelles [Nugues 89]. En ce qui nous concerne, il s'agirait, à partir de la description acoustique d'une série d'instances d'un ou de plusieurs triplets, de dégager grâce à la classification conceptuelle les meilleurs représentants acoustiques. S'il s'avérait qu'il n'y a pas d'obstacle technique insurmontable (concernant la masse de données, la taille des formules logiques ou encore l'évaluation de la partition construite) il resterait encore à définir un mode d'interaction judicieux entre l'apprentissage automatique et le phonéticien qui assure le lien avec l'articulatoire.

7 Décodage

Le système de décodage acoustico-phonétique à base de triplets que nous avons conçu opère de la manière suivante :

- segmentation de la phrase à décoder en instances de triplets,
- détection des événements acoustiques (formants, barres d'explosion...),
- appariement des instances de triplets de la phrase aux triplets de référence,
- augmentation de la cohérence globale de la solution ; cette dernière étape est destinée à assurer que les contraintes acoustiques, qui existent entre les triplets de la base de connaissances qui ont été proposés comme solutions, sont aussi vérifiées par les instances de triplets de la phrase à décoder.

Nous allons maintenant revenir en détail sur chacun de ces points.

7.1 Segmentation et détection d'indices

Segmentation

Le processus de segmentation du signal de parole est toujours une étape particulièrement délicate en DAP :

- trop segmenter conduit à des segments de parole parasites que l'on peut éventuellement regrouper avec les segments auxquels ils devraient appartenir s'il n'y a pas trop de différences acoustiques ;
- trop peu segmenter est plus grave car il n'est pas possible de revenir sur l'erreur.

Segmenter la parole est d'ailleurs d'autant plus difficile qu'il n'y a bien sûr pas de frontière précise entre deux sons mais plutôt un certain laps de temps pendant lequel il n'est pas possible d'affirmer qu'il s'agit de l'un ou de l'autre des deux sons. En fait la segmentation se justifie surtout parce qu'elle permet de guider la reconnaissance ce qui simplifie beaucoup l'architecture d'un système de DAP. En effet, lorsqu'on connaît les limites et la nature d'un son, cela permet de savoir quels sont les détecteurs d'événements acoustiques à déclencher pour l'identification. Dans certains cas (par exemple une suite de sonantes et de semi-voyelles) il est très difficile de segmenter le signal de parole et la segmentation est même un handicap.

Nous avons repris l'algorithme conçu et développé par D. Fohr dans le cadre du projet Aphodex et qui fournit les résultats suivants :

type de son	taux de détection (%)	taux d'insertion (%)
occlusives	86	2
fricatives	85	10
noyaux vocaliques	91	4
autres sons	67	19

La localisation des segments repose sur les paramètre suivants :

- f_1 = *somme de l'énergie entre 250 et 3 500 Hz et de l'énergie totale* est destiné à la détection des noyaux vocaliques ;
- f_2 = - *somme de l'énergie entre 600 et 8 000 Hz* est destiné à la détection des occlusives ;
- f_3 = *centre d'inertie du spectre* est destiné à la détection des fricatives.

Un segment d'un type donné correspond à un maximum local bien marqué de la courbe associée au type de son recherché. L'étendue d'un segment correspond à l'intervalle temporel $[t_g, t_d]$ contenant $t_{maximum}$ (l'instant correspondant au maximum) et tel que $f_i([t_g, t_d]) \subset [f_i(t_{maximum}), \infty[$. Dans le cas d'un noyau vocalique, il faut en plus qu'il y ait voisement à l'instant du maximum local, c'est-à-dire que le détecteur de fondamental retourne une valeur non nulle (cf § suivant).

Après avoir détecté les segments des trois types il faut supprimer les conflits de segmentation (recouvrement de deux segments de types différents ou morceau de parole ne correspondant à aucun des trois types recherchés). Voici, cas par cas, la stratégie adoptée :

recouvrement important de deux segments : on crée un segment en fusionnant les deux segments de départ qui a les deux types simultanément (par exemple vocalique et fricatif) ;

faible recouvrement de deux segments : dans ce cas on supprime le recouvrement en cherchant la meilleure frontière possible entre les deux segments. On procède par différence spectrale en partant de la zone centrale de chacun des deux segments, la nouvelle frontière correspondant à l'instant où il y a égalité entre les différences spectrales issues des deux segments. Quand il y a plusieurs points de ce type on retient celui qui est le plus proche de noyau vocalique si l'un des deux segments en conflit est vocalique, sinon celui qui est le plus à gauche ;

un petit espace libre entre deux segments : on supprime cet espace en étendant les segments voisins et en ajustant leur frontière par différence spectrale ;

un grand espace libre entre deux segments : on ajoute un segment de type inconnu et on ajuste les frontières par différence spectrale. Si le segment intercalaire est voisé on considère qu'il s'agit d'une sonante ou d'une semi-voyelle.

Détection du fondamental

L'estimation de la fréquence fondamentale (la fréquence à laquelle vibrent les cordes vocales) est l'un des paramètres décisifs pour le décodage acoustico-phonétique. Elle permet en effet de savoir s'il s'agit d'un son voisé (voyelles, sonantes, semi-voyelles, occlusives et fricatives voisées) ou d'un son sourd (occlusives ou fricatives sourdes). Il y a deux grandes classes d'algorithmes de détection du fondamental (se reporter éventuellement à l'ouvrage de Hess qui fait référence en la matière [Hess 83]) :

- les algorithmes opérant dans le domaine temporel (principalement par autocorrélation),
- les algorithmes qui opèrent dans le domaine fréquentiel.

J.J. Bonin [Bonin 87] a implanté au CRIN un algorithme par autocorrélation [Dubnowski 76]. Comme cet algorithme a été adapté pour les besoins spécifiques de la prosodie (nécessitant la connaissance de l'évolution temporelle de la fréquence fondamentale au cours de la phrase) il ne permet pas de détecter les variations très brèves du fondamental et très significatives en DAP.

Détecteur de formants

C'est bien sûr celui que nous avons intégré à Snorri. Les trajectoires formantiques sont approchées par des segments de droite de part et d'autre de la frontière du triplet.

Détection et analyse des barres d'explosion (burst)

Détection Sur un spectrogramme une barre d'explosion correspond à une bouffée d'énergie brève et soudaine (fig. 5.4) qui apparaît après le silence d'une occlusive. Ses caractéristiques énergétiques (position et intensité des principales concentrations d'énergie, étendue des concentrations) représentent les indices qui associés aux formants permettent d'identifier les occlusives. La recherche des barres d'explosion s'effectue de la manière suivante. On calcule l'énergie dans 8 bandes de fréquence (fig. 5.6) et l'on considère qu'il y a un burst quand un pic apparaît simultanément dans au moins trois bandes de fréquence.

Analyse L'analyse est faite à partir du spectre (obtenu par FFT ou mieux encore par lissage cepstral) du burst. Les concentrations d'énergie sont les pics du spectre les plus marqués.

Limite inférieure de bruit

La forme du bruit de friction et son intensité sont les principaux indices pour l'identification des fricatives. Il est en général assez délicat de détecter la limite

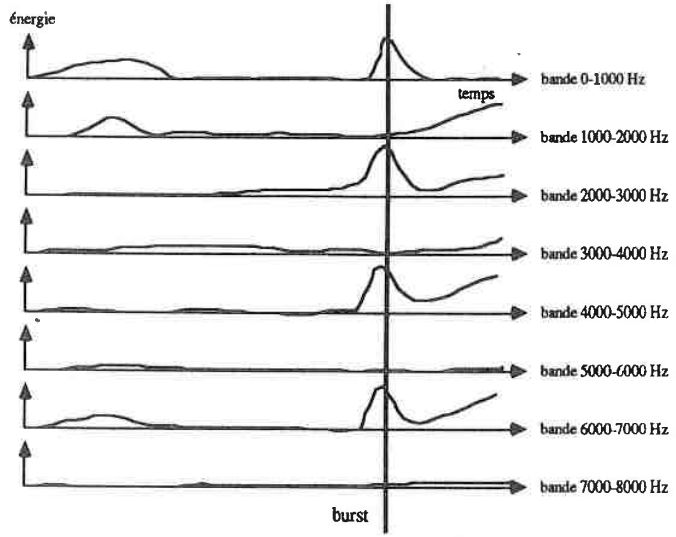


Figure 5.6 : Recherche d'une barre d'explosion

inférieure de bruit car les formants de bruit peuvent être assimilés au bruit de friction et risquent donc "d'attirer vers le bas" la limite inférieure de bruit. Pour pallier ce problème l'algorithme développé par D. Fohr élimine les pics au-dessous de 4 000 Hz.

7.2 Décodage par appariement des instances de triplets

Le but est de trouver quels sont les prototypes les plus proches de l'instance de triplet à reconnaître. Il existe de nombreux systèmes, voire de langages de programmation, à base de prototypes. Ce sont les travaux de Minsky sur la perception dans les systèmes de vision qui sont à l'origine de ces langages. Minsky considère que l'être humain devant une situation inconnue, essaie de retrouver une situation analogue parmi celles qu'il a déja rencontrées. Une situation typique, dénommée *frame*[5] est une structure de données contenant un certain nombre d'"attributs" décrivant cette situation (par exemple rouge pourrait être la valeur de l'attribut couleur du frame camion)[6]. Ces frames sont organisés en une hiérarchie à la racine de laquelle se trouve le frame le plus général ; un sous-frame représente la spécialisation du frame dont il descend. Damestoy [Damestoy 86] s'inspire de KRL (langage de frames conçu par Bo-

[5] Le Français "informatique" semble avoir adopté ce mot d'origine anglaise.

[6] Le lecteur peut se reporter à l'excellent livre de Masini et al. [Masini 89] qui fait référence dans le domaine des langages à objets.

brow [Bobrow 77]) pour son système de décodage, le frame le plus général est un phonème[7]. Au cours du décodage il s'agit de se déplacer dans la hiérarchie, autrement dit de rechercher la "spécialisation" du frame qui s'apparie le mieux avec l'instance à décoder ; on passe ainsi du frame phonème aux frames occlusive, fricative, voyelle ou sonante ; au plus bas niveau on mesure la qualité de l'appariement entre l'instance à décoder et les prototypes acoustiques de sons. Le score est simplement le rapport du nombre d'attributs du prototype compatibles avec ceux de l'instance sur le nombre total d'attributs du prototype.

La représentation hiérarchique des connaissances (de la forme la plus générale à la forme la plus particulière) est très classique en reconnaissance des formes. C. Granger [Granger 85] adopte la même solution pour étiqueter les objets d'une vue aérienne. Le critère de construction de la hiérarchie arborescente, dans laquelle sont stockés les objets à reconnaître, est la forme de l'objet (polygonale, circulaire, linéaire puis au niveau inférieur : coin , rectangle, cercle...). Le score d'appariement est la somme des distances entre les attributs de l'instance et ceux du prototype.

Dans notre cas, il est un peu plus difficile de s'appuyer sur le type de hiérarchie que nous avons choisi. En effet, il arrive que le module de segmentation commette une erreur sur la classe phonétique d'un son. Voici les confusions les plus fréquentes :

- occlusives voisées avec sonantes ou semi-voyelles,
- la fricative voisée /v/ avec les sonantes ou les semi-voyelles.

En fait la différence la plus notable entre les deux systèmes que nous venons d'évoquer et le nôtre porte sur l'appariement d'une instance et d'un prototype : alors que tous les attributs des phonèmes de Damestoy ou des formes de Granger sont au même niveau, les attributs d'un triplet sont organisés en deux couches :

- la description acoustique qui ne repose sur aucune expertise,
- la description en termes d'indices acoustiques qui modélise l'expertise du phonéticien.

Notre système exploite cette particularité de la manière suivante : dans un premier temps l'appariement s'effectue au niveau des indices acoustiques qui sont des informations plus significatives que les simples événements acoustiques. S'il n'y a pas de solution l'appariement ne porte plus que sur les événements acoustiques. À ces deux types d'appariement correspondent bien sûr deux types d'évaluation.

[7]Nous conservons le terme employé par Damestoy même s'il s'agit plutôt de son.

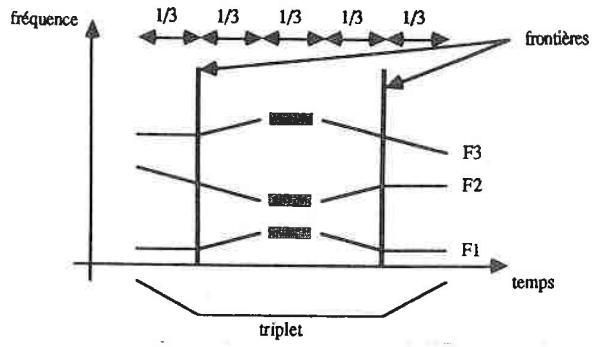


Figure 5.7 : Normalisation en durée d'un triplet

Évaluation de l'appariement par événements acoustiques

Cette évaluation est une estimation grossière de la distance acoustique entre l'instance et le prototype. Afin de pouvoir comparer ces deux formes, nous leur donnons arbitrairement la même durée (fig. 5.7).

Voici comment nous évaluons les distances acoustiques pour chacun des événements :

formants : Étant donnés deux arcs de trajectoire formantique, A_0 appartenant au triplet de référence, A_1 appartenant à l'instance, chaque arc est défini par sa valeur à la frontière v_f et sa pente p. La distance acoustique est

$$d(A_0, A_1) = max(\int_0^{1/3} \mid v_{1f} - v_{0f} + (p_1 - p_0)t \mid dt, \int_0^{1/3} \mid p_1 - p_0 \mid tdt)$$

c'est-à-dire (fig. 5.8) la plus grande des deux surfaces s_v et s_p, ce qui permet de pénaliser les arcs de trajectoire dont les directions sont tout à fait incompatibles. Si l'un des deux arcs de trajectoire n'est pas défini nous attribuons à $d(A_0, A_1)$ la plus grande des distances inter-arcs obtenue pendant la comparaison du triplet référence et de l'instance.

bruit de friction : C'est la même distance mais en prenant en compte les limites inférieures de bruit de friction.

formants au centre du triplet : Étant données c_0 et c_1 les valeurs d'un formant au centre, comme nous supposons que la valeur du formant est stable au centre du triplet cela donne $d = \mid c_1 - c_0 \mid /3$.

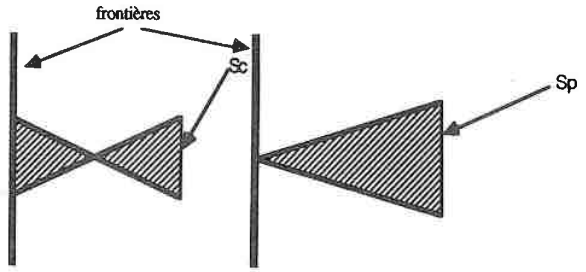


Figure 5.8 : Distance acoustique entre deux arcs de trajectoire formantique

barre d'explosion : Étant données les deux barres d'explosion b_0 et b_1, avec durée(b_1) > durée(b_0) (fig. 5.9), $(c_{0i})_i$ et $(c_{1j})_j$ les concentrations d'énergie de b_0 et b_1

$$d(b_0, b_1) = \sum_i \mid surface(c_{0i}) - \sum_{j/c_{0i} \cap c_{1j} \neq \emptyset} surface(c_{0i} \cap c_{1j}) \mid$$

Toutes les distances sont homogènes en dimension (Hz x s). La distance acoustique entre les deux formes est donc simplement la somme des distances élémentaires.

Évaluation de l'appariement par indices acoustiques

Considérons par exemple l'indice acoustique *écarte(F1, F2, face0)*. Cet indice peut prendre une valeur dans l'ensemble {FAUX, INCERTAIN, VRAI} :

$$écarte(F1, F2, face0) = \begin{cases} FAUX & si\ Seuil_0 < pente(F2) - pente(F1) \\ INCERTAIN & si\ Seuil_1 \leq pente(F2) - pente(F1) \leq Seuil_0 \\ VRAI & si\ pente(F2) - pente(F1) < Seuil_1 \end{cases}$$

L'appariement par indice acoustique est possible si :

- aucun indice ne prend la valeur FAUX ;
- les indices prenant la valeur VRAI sont plus nombreux que ceux prenant la valeur INCERTAIN.

Les triplets qui peuvent s'apparier avec l'instance sont triés suivant leur distance acoustique par rapport à l'instance.

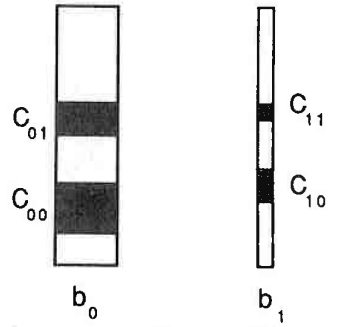


Figure 5.9 : Distance acoustique entre deux barres d'explosion

8 Cohérence globale de la solution

Le résultat du décodage précédent est la juxtaposition des solutions propres à chacun des segments de la phrase. Sur la figure 5.10 nous avons représenté quelques triplets possibles pour les instances $/_a b_i/$ et $/_y t_e/$. Les formants de ces deux segments sont liés par les relations suivantes où FI_{iD} représente la valeur du formant I du triplet i sur la frontière droite, PI_{iD} la valeur de pente de ce formant, i le segment 1 ou 2 (cf fig. 5.10) :

$$\begin{aligned} F2_{1D} &> F2_{2D} \\ F1_{1D} &< F1_{2D} \\ P2_{1G} &\ll P2_{2G} \\ P3_{1G} &\ll P3_{2G} \\ F1_{1G} &> F1_{2G} \\ F2_{1D} - F1_{1D} &\gg F2_{2D} - F1_{2D} \\ F3_{1G} - F2_{1G} &\gg F3_{2G} - F2_{2G} \end{aligned}$$

Si la solution globale est cohérente ces relations doivent se retrouver entre les triplets qui ont été proposés comme solutions pour chacun des segments. Éliminer les incohérences qui existent permet donc de proposer une solution consistante sur toute la phrase et donc en somme de modéliser l'unicité du locuteur dans le processus de décodage.

Plus formellement le problème est le suivant. Étant donnés n nœuds (les segments de parole) notés i et les ensembles d'étiquettes L_i proposées pour chaque

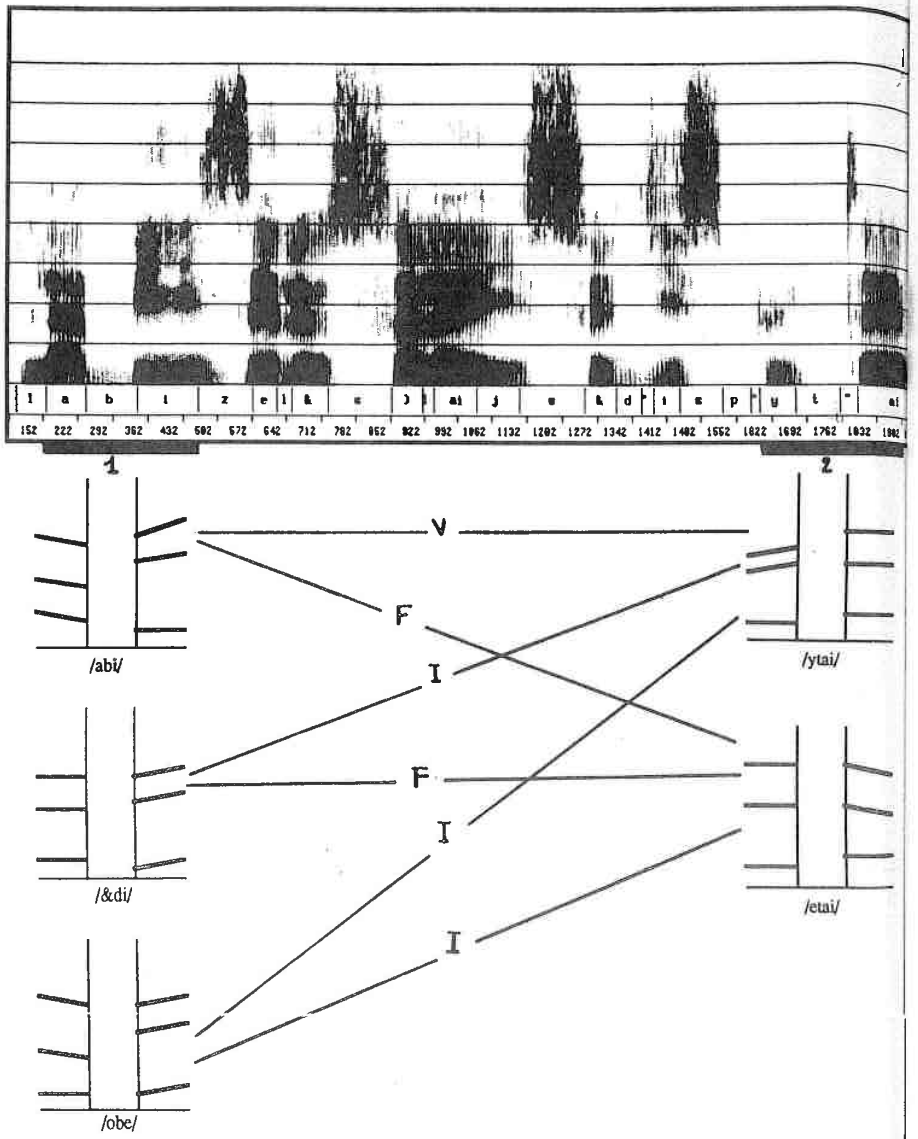


Figure 5.10: Les triplets proposés pour les segments du spectrogramme qui sont soulignés et numérotés 1 et 2

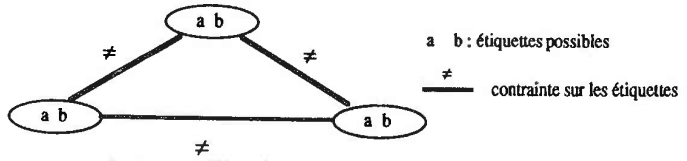


Figure 5.11 : Réseau de contraintes sans cohérence globale

nœud (l'ensemble des étiquettes étant l'ensemble des triplets proposés pour le segment i) comment éliminer les étiquettes incompatibles avec celles des autres nœuds de la phrase ? Il s'agit là du schéma classique de la relaxation discrète, connue aussi sous le nom de filtrage de Waltz [Waltz 75]. Il existe en fait deux types de cohérence :

cohérence globale : Les contraintes sont satisfaites simultanément pour l'ensemble des étiquettes de tous les nœuds.

cohérence sur les arcs : Cette cohérence est plus faible que la précédente et assure que les étiquettes de deux nœuds contraints l'un par l'autre sont compatibles.

La figure 5.11 montre que ces deux types de cohérence ne sont pas équivalents. La cohérence sur les arcs est bien vérifiée alors la cohérence globale est impossible à assurer car il n'est pas possible d'affecter trois étiquettes différentes aux trois nœuds.

Dans notre cas nous nous contenterons du second type de cohérence, celle pour laquelle il existe de nombreux algorithmes. Les contraintes (par exemple celle que nous avons définie entre les segments 1 et 2 de la fig. ??) sont construites à partir des informations suivantes :

- les positions relatives des trajectoires formantiques aux frontières du triplet,
- les relations entre pentes des trajectoires formantiques,
- les écarts entre formants.

Pour que les contraintes soient significatives, seules les relations bien marquées (fortes pentes ou grands écarts entre les valeurs des formants) sont retenues. Les relations qui constituent une contrainte peuvent prendre, comme les indices acoustiques, les valeurs FAUX, INCERTAIN ou VRAI.

Voici pour la contrainte que nous avons définie plus haut, les valeurs des différentes relations (F représentant FAUX, I représentant INCERTAIN, V représentant VRAI) :

relation	${}_ab_i$ vs. ${}_yt_\varepsilon$	${}_ab_i$ vs. ${}_ct_\varepsilon$	${}_ob_c$ vs. ${}_yt_\varepsilon$	${}_ob_\epsilon$ vs. ${}_ct_\varepsilon$	${}_\&d_i$ vs. ${}_yt_\varepsilon$	${}_\&d_i$ vs. ${}_{e'_i}$
$F2_{1D} > F2_{2D}$	V	V	I	I	I	I
$F1_{1D} < F1_{2D}$	V	V	I	V	V	V
$P2_{1G} \ll P2_{2G}$	V	I	V	I	I	F
$P3_{1G} \ll P3_{2G}$	V	I	V	I	I	F
$F1_{1G} > F1_{2G}$	V	I	I	F	I	I
$F2_{1D} - F1_{1D} \gg F2_{2D} - F1_{2D}$	V	V	I	I	V	I
$F3_{1G} - F2_{1G} \gg F3_{2G} - F2_{2G}$	V	F	V	I	I	F
valeur de la contrainte	**V**	**F**	**I**	**F**	**I**	**F**

Le tableau précédent permet de voir que le triplet ${}_ct_\epsilon$ n'est compatible avec aucun des triplets du premier segment, que le triplet ${}_ab_i$ est compatible avec ${}_yt_\varepsilon$, et qu'on ne peut rien dire pour le triplet ${}_ob_\epsilon$.

Le principe général d'un algorithme de relaxation est de supprimer une étiquette a du nœud i (notée (i,a)) contraint par le nœud j pour lequel aucune étiquette (j,b) n'est compatible avec (i,a) et de propager récursivement ces modifications dans tout le réseau. L'algorithme AC4 de Mohr [Mohr 86] est particulièrement efficace car sa complexité est linéaire dans le nombre de contraintes. Une première étape consiste à calculer pour chaque étiquette a du nœud i contraint par le nœud j le nombre d'étiquettes de j acceptant a : $C(i,a,j)$ et de déterminer les ensembles $E_{i,a}$ des couples (k,c) (étiquette c du nœud k) compatibles avec (i,a). Lors de la seconde étape on décrémente de 1 $C(i,a,j)$ lorsqu'une contrainte n'est pas vérifiée et on étudie récursivement les ensembles $E_{i,a}$. En fait AC4 est conçu pour des contraintes binaires alors que les nôtres prennent leurs valeurs dans l'ensemble {FAUX, INCERTAIN, VRAI}. De plus, le fait que certains segments de parole soient bruités risque de conduire à un étiquetage global vide. Nous avons donc retenu la version floue de AC4 qui permet de pondérer les contraintes entre les nœuds et de ne pas forcément supprimer une étiquette (i,a) qui n'est "supportée" par aucune des étiquettes d'un nœud j contraignant i.

Nous utiliserons le mécanisme de relaxation uniquement pour des triplets de même type ([Voyelle Occlusive Voyelle] dans notre exemple).

9 Choix d'implantation et tests

Pour l'instant nous avons développé tout l'environnement "triplets" (intégré dans Snorri) sur le poste de travail Masscomp en C ; il intègre les aspects suivants :

- édition de triplets de référence,
- génération automatique de la suite d'instances de triplets d'une phrase avec les événements et indices acoustiques détectés,

- décodage des instances de triplets d'une phrase sans intégrer l'augmentation de la cohérence globale de la solution. Cette partie n'est pas encore suffisamment testée pour que nous donnions des résultats qui seront présentés lors de la thèse dans l'annexe 1.

Nous avons conçu cet environnement de façon à pouvoir à tout moment examiner et modifier très simplement les références ou les instances de triplets. Cela permet notamment d'étudier l'influence respective des détecteurs d'événements acoustiques et de la construction des triplets. Bien sûr, le langage C n'est pas le plus approprié à ce genre d'approche, mais c'est le seul langage disponible sur notre poste de travail. Nous sommes donc en train de concevoir une version écrite dans le langage orienté objet Eiffel fonctionnant sur un poste de travail Sun.

Comme nous l'avons déja dit nous ne testerons pas notre système pour l'ensemble du Français, ce qui demanderait notamment un effort trop grand pour la construction des prototypes, mais seulement sur les occlusives présentes dans le corpus de "la bise et le soleil".

10 Conclusion

L'approche à base de triplets est toute jeune, il reste donc encore de nombreux problèmes à aborder. Il faut notamment valider sur de vastes corpus cette approche tant du point de vue phonétique que du point de vue de l'intelligence artificielle. Après que cette question sera résolue il faudra aussi intégrer le système de DAP à base de triplets dans le cadre plus général d'un système de compréhension de la parole continue. En effet, le système que nous avons décrit est essentiellement destiné à tester l'approche proposée sans tenir compte des autres niveaux de la reconnaissance de la parole. C'est d'ailleurs sans doute un des points qui a longtemps joué à l'encontre des approches expertes qui, souvent, n'ont pas été réellement intégrées par leurs créateurs dans un système global de reconnaissance de la parole.

Chapitre 6

Perspectives

Cette thèse laisse entrevoir la grande diversité des problèmes à aborder dans le domaine de la reconnaissance automatique de la parole continue multi-locuteur. Cette diversité est très attrayante pour le chercheur mais il faut pouvoir la maîtriser pour préserver l'unité du problème de départ :

Comprendre **UNE** phrase [1] produite par **UN** locuteur.

L'unité de la phrase à décoder doit influencer l'architecture d'un système de reconnaissance automatique de la parole : il faut privilégier la construction de solutions cohérentes pour "expliquer" le signal temporel représentant la phrase. Selon nous, cela signifie qu'il faut aller plus loin que l'architecture consistant à faire "dialoguer" les modules correspondant aux différents niveaux d'abstraction et considérer que chaque niveau impose des contraintes qui limitent l'espace des solutions à son propre niveau mais aussi aux autres niveaux (cf. §3). Lorsqu'une contrainte ne peut être vérifiée il faut faire l'hypothèse que l'un des paramètres extraits du signal a été mal détecté (que ce soit à cause d'un algorithme pas assez performant, d'un parasite ou même du locuteur). L'une des voies de recherche que nous explorerons prochainement est donc de répertorier le plus grand nombre de contraintes qui interviennent en parole continue et d'étudier comment il est possible d'en tirer parti dans le cadre d'un système de reconnaissance automatique de la parole.

L'unicité du locuteur est sans doute l'un des aspects qui a été le moins étudié, du moins en ce qui concerne l'approche experte du DAP. Nous avons proposé une solution consistant à vérifier, grâce à un algorithme de relaxation, que les contraintes reliant plusieurs triplets de référence se retrouvent aussi entre les instances de triplets de la phrase à reconnaître. Pour l'instant cette modélisation de l'unicité du locuteur est limitée au seul niveau du décodage acoustico-phonétique. Mais il est possible de l'étendre au niveau lexical : à partir des hypothèses de mots qui ont été

[1] Bien sûr cette phrase peut faire partie d'un dialogue.

émises, on peut déduire les sons qui doivent être présents à plusieurs endroits de la phrase et s'assurer que les contraintes acoustiques qui les relient sont bien vérifiées.

Dans le cas du suivi de formants, il faut aussi préserver l'unité du problème. En effet, Il y a manifestement deux aspects à concilier :

- l'extraction des lignes de pics à partir de spectres lissés cepstralement ou de spectres LPC ; de nombreuses techniques de traitement d'image existent pour aborder ce problème ;
- l'interprétation les lignes de pics en termes de formants.

On pourrait sans doute lier plus étroitement ces deux aspects en imitant le comportement de l'expert qui connaît la plupart des configurations formantiques possibles et peut donc les reconnaître sur la base de trajectoires formantiques incomplètes. Voici donc une autre voie de recherche qui s'ouvre et dont les points clés sont les suivants :

- il faut d'abord recenser et modéliser les configurations formantiques que l'on peut rencontrer ; leur nombre est a priori assez limité et ce travail ne devrait pas être trop lourd ;
- on peut envisager que le suivi de formants se décompose en deux étapes, la première destinée à sélectionner les configurations formantiques possibles sur la base de fragments de trajectoires obtenus grâce à un algorithme classique, la seconde destinée à trouver la meilleure configuration de référence, c'est-à-dire celle qu'il faut le moins déformer pour la recaler sur le segment de parole à analyser. Il serait intéressant d'utiliser des contours actifs [Kass 88] qui se déforment sous l'effet du champ de potentiel créé par les concentrations d'énergie du spectrogramme (principalement les formants).

Il faut enfin préserver l'unité de l'environnement logiciel dans lequel travaille le chercheur en parole. Tout notre travail, jusqu'ici, a été réalisé avec Snorri que nous avons spécialement conçu pour nos recherches. Il est hélas parfois difficile pour un chercheur ne connaissant pas du tout Snorri de savoir comment le modifier pour l'adapter à ses besoins spécifiques. Cela provient bien sûr du fait que Snorri est programmé en langage C, langage mal adapté aux contraintes du génie logiciel. Nous avons donc choisi le langage orienté objet Eiffel [Meyer 88][2] qui nous permettra à l'avenir de construire les classes d'objets que nous sommes amenés à manipuler en parole... et parmi eux bien sûr les triplets.

Nous venons d'évoquer quelques directions vers lesquelles nous comptons orienter notre recherche future... mais la route est encore longue avant qu'une machine puisse réussir la variante du test de Turing que nous proposions au début de cette thèse.

[2]En plus des facettes génie logiciel et objets, Eiffel offre une vaste bibliothèque de classes prédéfinies parmi lesquelles de nombreuses classes graphiques.

Bibliographie

[Arcelli 85] C. Arcelli et G. Sanniti Di Baja. A width-independent fast thinning algorithm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-7(4):463–474, July 1985.

[Atal 67] B. S. Atal et M. R. Schroeder. Predictive coding of speech signals. *Proc. Conf. Commun. and Process.*, pages 360–361, 1967.

[Badin 88] P. Badin, L.J. Boe, P. Perrier, et L. Abry. Vocalic nomograms: acoustic considerations upon formant convergence. *Bulletin du Laboratoire de la Communication Parlée*, (2A):65–94, 1988.

[Bahl 83] L. R. Bahl, F. Jelinek, et R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. PAMI.*, PAMI-5(2):179–190, 1983.

[Bailly 88] G. Bailly et J.P. Liu. Détection de formants par quantification vectorielle et réseaux markoviens. *Actes du séminaire "Décodage acoustico-phonétique"*, Nancy, 1988.

[Baker 75] J. K. Baker. Stochastic modeling for automatic speech understanding. D. R. Reddy, éditeur, *Speech Recognition*, Academic Press, New York, 1975.

[Baruch 88] O. Baruch. Line thinning by line following. *Pattern Recognition Letters*, (8):271–276, November 1988.

[Blomberg 83] M. Blomberg, R. Carlson, K. Elenius, et B. Grandstrom. *Auditory Models and Isolated Word Recognition.* Rapport, STL-QPSR 4, 1983.

[Bobrow 77] D.G. Bobrow et T. Winograd. An overview of KRL, a knowledge representation language. *Cognitive Science*, 1(1):3–46, 1977.

[Boe 88] L.J Boé et J.S. Liénard. La communication parlée est-elle une science ? *Actes des 17èmes Journées d'Etudes sur la Parole*, pages 79–92, Nancy, 1988.

[Bonin 87] J.J Bonin. *Détection d'indices prosodiques linguistiquement significatifs.* Mémoire de DEA informatique, Université de Nancy 1, 1987.

[Bourlard 87] H. Bourlard et C. J. Wellekens. Multilayer perceptrons and automatic speech recognition. *Proc. IEEE First Annual Int. Conf. on Neural Networks*, San Diego, California, June 1987.

[Broad 87] D.J. Broad. A methodology for modeling vowel formant contours in CVC. *J. Acoust. Soc. Am.*, 81(1):155–165, January 1987.

[Caelen 79] J. Caelen. *Un modèle d'oreille. Analyse de la parole continue.* Thèse d'État, Univeristé de Toulouse, 1979.

[Caelen 88] J. Caelen et H. Tattegrain. Le décodeur acoustico-phonétique dans le projet DIRA. *Actes des 17^{ièmes} Journées d'Etudes sur la Parole*, pages 115–121, Nancy, 1988.

[Carbonell 83] N. Carbonell, J. P. Haton, F. Lonchamp, et J.M. Pierrel. Élaboration d'un système expert pour le décodage phonétique de la parole. *Speech Communication*, (2-3):231–233, 1983.

[Carbonell 86] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, et J. M. Pierrel. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986.

[Carlson 82] R. Carlson et B. Granstrom. Towards an auditory spectrograph. R. Carlson et B. Granstrom, éditeurs, *The Representation of Speech in the Peripheral Auditory System*, Elsevier Biomedical, Amsterdam, 1982.

[Carre 84] R. Carré, R. Descout, M. Eskénazi, J. Mariani, et M. Rossi. The french language database : defining, planning and recording a large database. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing 1984*, San Diego, 1984.

[Cheng 87] Y.M. Cheng et B. Guerin. Study of the source - filter interactive concept and its application to male and female speech synthesis. *Bulletin du Laboratoire de la Communication Parlée*, (1A):29–66, 1987.

[Cole 79] R.A. Cole, A.I. Rudnicky, et V.W. Zue. Performance of an expert spectrogram reader. *J. Acoust. Soc. Amer.*, 65, Supp. 1:S81, Paper presented at the 97th meeting of the ASA, 1979.

[Combescure 81] P. Combescure. Vingt listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 14(56), 1981.

[Cooke 88] M.P. Cooke et P.D. Green. On finding objects in spectrograms: a multiscale relaxation labelling approach. H. Niemann, M. Lang, et G. Sagerer, éditeurs, *Recent Advances in Speech Understanding and Dialog Systems*, pages 129–133, Springer Verlag, 1988.

[Cun 88] Y. Le Cun. Learning process in an assymetric threshold network. E. Bienenstok, Fogelman-Soulié, et G. Weisbuch, éditeurs, *Disordered Systems and Biological Organization, NATO ASI Series in Systems and Computer sciences, F20*, pages 233–240, Springer Verlag, 1988.

[Damestoy 86] J.P Damestoy. Réalisation d'un système à base de prototypes pour le contrôle du décodage acoustico-phonétique. *Thèse de Doct. Univ. de NANCY 1*, 1986.

[Delattre 48] P. Delattre. Un triangle acoustique des voyelles orales du Français. *The French Review*, XXI(6):477–484, may 1948.

[Delattre 55] P. C. Delattre, A. M. Liberman, et F. S. Cooper. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Amer.*, 27(4):769–773, 1955.

[Delattre 61] P. Delattre. Le jeu des transitions de formants et la perception des consonnes. *Proceedings of the Fourth International Congress of Phonetic Sciences*, pages 407–414, Helsinki, 1961.

[Delgutte 84a] B. Delgutte et N.Y.S. Kiang. Speech coding in the auditory nerve: ii. processing schemes for vowel-like sounds. *J. Acoust. Soc. Amer.*, 75(3):879–886, 1984.

[Delgutte 84b] B. Delgutte et N.Y.S. Kiang. Speech coding in the auditory nerve: i. vowel-like sounds. *J. Acoust. Soc. Amer.*, 75(3):866–878, 1984.

[Delgutte 86] B. Delgutte. Analysis of french stop consonants using a model of the peripheral auditory system. J. S. Perkell et D. M. Klatt, éditeurs, *Problem of Variability in Speech Recognition and in Models of Speech Perception*, Lawrence Erlbaum Assoc., Hillsdale, N.J., 1986.

[Delsarte 86] P. Delsarte et Y.V. Genin. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-34(3):470–477, June 1986.

[dK 84] J. de Kleer et J.S. Brown. A qualitative physics based on confluences. *Artificial Intelligence*, 24(1-3), December 1984.

[dK 86] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, (28):127–162, 1986.

[Dormoy 89] J.L. Dormoy. Méthodologies pour des systèmes experts de seconde génération. *Actes des 11èmes Journées Francophones sur l'Informatique*, pages 219–236, Nancy, janvier 1989.

[Dubnowski 76] J. J. Dubnowski, R. W. Schafer, et R. W. Rabiner. Real-time digital hardware pitch detector. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:2–8, Feb. 1976.

[ea 87] M.A. Huckvale et al. The Spar speech filing system. *European Conference on Speech Technology*, Edinburgh, 1987.

[Engstrand 88] O. Engstrand. Articulatory correlates of stress and speaking rate in swedish VCV utterances. *J. Acoust. Soc. Am.*, 83(5):1863–1875, May 1988.

[Fant 60] Gunnar Fant. *Acoustic Theory of Speech Production.* The Hague: Mouton & Co., 1960.

[Fant 73] G. Fant. Stops in CV-syllables. *Speech Sounds and Features*, pages 111–139, Cambridge MA, MIT Press, 1973.

[Feng 86] G. Feng. *Apport de la modélisation au traitement du signal de parole : le cas des voyelles nasales et la simulation des pôles et des zéros.* Thèse de doctorat, Université de Grenoble, 1986.

[Fohr 86] D. Fohr. APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue. *Thèse de Doct. Univ. de NANCY I*, 1986.

[Fohr 88] D. Fohr, J.P. Haton, Y. Laprie, F. Lonchamp, et J.M.Pierrel. Paramétrisation acoustique et décodage phonétique fondé sur des connaissances, pour la parole continue multi-locuteurs. *Actes du séminaire "Décodage acoustico-phonétique"*, Nancy, 1988.

[Fohr 89] D. Fohr et Y. Laprie. Snorri: an interactive tool for speech analysis. *Proceedings of European Conference on Speech Technology*, Paris, France, September, 1989.

[Friedman 85] D. H. Friedman. Instantaneous-frequency distribution vs. time : an interpretation of the phase structure of speech. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing 1985*, Tampa, Florida, 1985.

[Gallais 89] E. Gallais, P. Alinat, G. Souvay, et J.M. Pierrel. Intégration d'un système de reconnaissance analytique de la parole dans une console sonar : vers un dialogue naturel. *article proposé à la Revue de Traitement du Signal*, 1989.

[Gay 78] T. Gay. Effect of speaking rate on vowel formant movements. *J. Acoust. Soc. Am.*, 63(1):223-230, January 1978.

[Gillet 84] Gillet et al. SERAC : Un système expert en reconnaissance acoustico-phonétique. *Actes du 4ème congrès AFCET-RFIA*, Paris, Janvier 1984.

[Gong 85] Y. Gong et J. P. Haton. Assia, Un éditeur "intelligent" pour la manipulation et l'analyse du signal vocal. *Actes des 14èmes Journées d'Etudes sur la Parole*, Paris, 1985.

[Granger 85] C. Granger. Reconnaissance d'objets par mise en correspondance en vision par ordinateur. *Thèse de Doct. Univ. de Nice*, 1985.

[Green 86] P.D. Green et A.R. Wood. A representational approach to knowledge-based acoustic-phonetic processing in speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1205–1208, Tokyo, 1986.

[Guyot 89] F. Guyot, F. Alexandre, et J.P. Haton. Toward a continuous model of the cortical column: application to speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Glasgow, May 1989.

[Haton 84] J.P. Haton et M. Lazrek. Segmentation et identification des phonèmes dans un système de reconnaissance de la parole continue. *Actes du 4ème congrès AFCET-RFIA*, pages 5–21, Paris, Janvier 1984.

[Haton 86] J.P. Haton, B. Mangeol, et J.M. Pierrel. Organisation et fonctionnement d'une composante lexicale dans un système de dialogue homme machine. *Actes du séminaire "Lexique et traitement automatique des langages"*, pages 259–267, Toulouse, 1986.

[Hermann 89] Hermann. Phonophotographische Untersuchungen. *Archive für die gesamte Physiologie, Pflüger*, XLV:582, 1889.

[Hess 83] W. J. Hess. *Pitch Determination of Speech Signals - Algorithms and Devices.* Springer Berlin, 1983.

[Hoffman 58] H. S. Hoffman. Study of some cues in the perception of the voiced stop consonants. *J. Acoust. Soc. Amer.*, 30(11):1035–1041, 1958.

[Johanssen 83] J. Johanssen, J. Mac Allister, T. Michaele, et R. Ross. A speech spectrogram expert. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 746-749, 1983.

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

[Kass 88] M. Kass, A. Witkin, et D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, ():321–331, 1988.

[Kiang 65] N.Y.S. Kiang, T. Watanabe, E.C. Thomas, et L.F.Clark. *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, Research Monograph No 35.* Rapport, The M.I.T. Press, Cambridge, MA, 1965.

[Kohonen 84] T. Kohonen, K. Mäkisara, et T.Saramäki. Phonotopics maps - insightful representation of phonological features for speech recognition. *Proc. 7th ICPR*, pages 182-185, Montreal, 1984.

[Kopec 84] G. E. Kopec. The integrated signal processing system ISP. *IEEE Trans. on Acoust., Speech, Signal Processing*, ASSP-32:842–851, August 1984.

[Kopec 85] G. E. Kopec. Formant tracking using hidden Markov models. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing 1985*, pages 1113–1116, 1985.

[Kopec 86a] G. E. Kopec. A family of formant trackers based on hidden Markov models. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing 1986*, pages 1225–1228, 1986.

[Kopec 86b] G. E. Kopec. Formant tracking using hidden Markov models and vector quantization. *IEEE Acoust., Speech, Signal Processing*, ASSP-34:709–729, August 1986.

[Ladefoged 82] P. Ladefoged et R.A.W. Bladon. Attemps by human speakers to reproduce FANT's nomograms. *Speech Communication*, 1:185–198, 1982.

[Laface 80] P. Laface. A formant tracking system toward automatic recognition of speech. *Signal Processing*, 2(2):113–129, April 1980.

[Lamel 88] L.F. Lamel. *Formalizing Knowledge used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops, PhD thesis.* Massachussets Institute of Technology, May 1988.

[Laprie 88] Y. Laprie. Snorri, un système d'étude interactif de la parole. *Actes des 17èmes Journées d'Etudes sur la Parole*, pages 71–76, Nancy, 1988.

[Lee 88] K.F. Lee. *Large-Vocabulary Speaker-Independant Continuous Speech Recognition: The SPHINX System, PhD thesis.* CMU-CS-88-148, Carnegie-Mellon University, April 1988.

[Levinson 83] S. E. Levinson, L. R. Rabiner, et M. M. Sondhi. An introduction of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4), 1983.

[Lindblom 63] B. Lindblom. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.*, 35(11):1773–1781, November 1963.

[Lindsay 80] R. Lindsay, B. G. Buchanan, E. A. Feigenbaum, et J. Lederberg. *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project.* New York, McGraw-Hill, 1980.

[Lonchamp 84] F. Lonchamp. *Les Sons du Français – Analyse acoustique descriptive.* Cours de Phonétique, Institut de Phonétique, Université de Nancy II, 1984.

[Lonchamp 87] F. Lonchamp. Reading spectrograms: the view from the expert. J.P. Haton, éditeur, *Fundamentals in Computer Understanding: Speech and Vision*, Cambridge University Press, 1987.

[Lonchamp 88] F. Lonchamp. Études sur la production et la perception de la parole. *Thèse de Doctorat d'état en Lettres et Sciences Humaines, Université de Nancy II*, 1988.

[Maeda 86] S. Maeda. *On The Generation of Sound in Stop Consonants.* Rapport, Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1986.

[Markel 71] J. D. Markel. FFT pruning. *IEEE Trans. Audio Electroacoustic.*, Au-19(4):305–311, December 1971.

[Markel 76] J.D. Markel et A.H. Gray. Automatic formant trajectory estimation. *Linear Prediction of Speech*, chapitre 7, Springer-Verlag, Berlin Heidelberg New York, 1976.

[Masini 89] G. Masini, A. Napoli, D. Colnet, D. Léonard, et K. Tombre. *Les Langages à objets.* InterEditions, 1989.

[McCandless 74] S. McCandless. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-22(2):135–141, April 1974.

[Mella 89] O. Mella et M.C. Haton. Méthodologie d'étude de la pertinence de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur. *Actes du séminaire sur la varibilité du locuteur*, Luminy, juin 1989.

[Meloni 86] H. Méloni et R. Bulot. Un système de traitement de connaissances pour le décodage acoustico-phonétique. *Proc. of the Montreal Symposium on Speech Recognition*, pages 26–27, Montreal, 1986.

[Mercier 83] G. Mercier. Le décodage acoustico-phonétique de la parole. *Bulletin de liaison de la recherche en Informatique et automatique, INRIA*, (84):15–22, 1983.

[Mercier 85] G. Mercier, M. Gilloux, C. Tarridec, et J. Vaissière. From KEAL to SERAC: a new rule-based expert system for speech recognition. R. De Mori et R. Suer, éditeurs, *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, Springer Verlag, Berlin Heidelberg New York, 1985.

[Meyer 88] B. Meyer. *Object-Oriented Software Construction.* Prentice-Hall, 1988.

[Michalski 84] R. S. Michalski et R. E. Stepp. Learning from observation: conceptual clustering. R. S. Michalski, J. G. Carbonell, et

T. M. Mitchell, éditeurs, *Machine Learning – An Artificial Intelligence Approach*, Spring-Verlag, 1984.

[Minsky 69] M. L. Minsky et S. Papert. *Perceptrons*. MIT press, Cambridge, 1969.

[Mizoguchi 84] R. Mizoguchi et O. Kakusho. Continuous speech recognition based on knowledge engineering techniques. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 638–640, San Diego, 1984.

[Mizoguchi 86] R. Mizoguchi, K. Tsujino, et O. Kakusho. A continuous speech recognition system based on knowledge engineering techniques. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1221-1224, Tokyo, 1986.

[Mohr 86] R. Mohr et C. Henderson. Arc and path consistency revisited. *Artificial Intelligence*, (28):225-233, 1986.

[Myers 81] C. Myers et L. R. Rabiner. A level building dynamic time warping algorithm for connected word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(2):284, April 1981.

[Ney 87] H. Ney, D. Mergel, A. Noll, et A. Paeseler. A data-driven organisation of the dynamic beam search for continuous speech recognition. *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, pages 833-836, 1987.

[Noll 67] A. M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, 41:293-309, 1967.

[Nugues 89] P. Nugues. Interprétations de gels d'électrophorèses bidimensionnelles. *Thèse de Doct. Univ. de NANCY 1*, 1989.

[oE 86] University of Edinburgh. *The AUDLAB Interactive Speech Analysis System*. Rapport, 1986.

[Ohde 83] R. N. Ohde et K. N. Stevens. Effect of burst amplitude on the perception of stop consonant place of articulation. *J. Acoust. Soc. Amer.*, 74(3):706-714, 1983.

[Ohman 66] S.E.G. Öhman. Coarticulation inVCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.*, 39(1):151-168, 1966.

[Ohman 67] S.E.G. Öhman. Numerical model of coarticulation. *J. Acoust. Soc. Am.*, 41:310-320, 1967.

[Olive 71] J.P. Olive. Automatic formant tracking by a Newton-Raphson technique. *J. Acoust. Soc. Am.*, 50:661-670, 1971.

[Potter 47] R.K. Potter, G.A. Kopp, et H.C. Green. *Visible Speech*. Van Nostrand, New York, 1947.

[Prager 86] R. W. Prager, T. D. Harrison, et F. Fallside. Bolzmann machines for speech recognition. *Computer, Speech and Language*, 3–27, 1986.

[Rabiner 69a] L. R. Rabiner et R. W. Schafer. System for automatic formant analysis of voiced speech. *J. Acoust. Soc. Amer.*, 47:261–275, 1969.

[Rabiner 69b] L.R. Rabiner, R.W. Schafer, et C.M. Rader. The chirp z-transform algorithm and its application. *The Bell System Technical Journal*, 48:1249-1292, 1969.

[Rabiner 78] L. R. Rabiner et R. W. Schafer. *Digital Processing of Speech*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

[Rabiner 88] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Proceedings*, 1988.

[Rietveld 87] A.C.M. Rietveld et F.J. Koopmans-van Beinum. Vowel reduction and stress. *Speech Communication*, 6(3):217–229, 1987.

[Rigoll 88] G. Rigoll. Formant tracking with quasilinearization. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1988*, pages 307–310, New York City, April 1988.

[Rosenberg 71] A.E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, 49:583–590, August 1971.

[Rumelhart 86] D. E. Rumelhart, G. E. Hinton, et R. J. Williams. Learning internal representation by error propagation. D. E. Rumelhart et J. L. McClelland, éditeurs, *Parallel Distributed Processing. Exploration of the Microstructure of Cognition*, page ., MIT Press, 1986.

[Saito 66] S. Saito et F. Itakura. *The Theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density*. Rapport no. 3107, Electrical Communication Laboratory, N.T.T., Tokyo, 1966.

[Schwartz 76] R. M. Schwartz. Acoustic-phonetic experiment facility for the study of continuous speech. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1976*, Philadelphia, April 1976.

[Schwartz 84] R. Schwartz, Y.L. Chow, S. Roucos, M. Krasner, et J. Makhoul. Improved hidden Markov modeling of phonemes for continuous speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1984.

[Sejnowsky 86] T. J. Sejnowsky et C. R. Rosenberg. *NETtalk: A Parallel Network that Learns to Read Aloud.* JHU/EESC-86/01, JHU, 1986.

[Seneff 85] S. Seneff. *Pitch and Spectral Analysis of Speech based on an Auditory Synchrony Model, PhD thesis.* Massachussets Institute of Technology, Cambridge, MA, 1985.

[Serra 86] J. Serra. Introduction to mathematical morphology. *Computer Vision, Graphics, and Image Processing*, 35:283–305, 1986.

[Shaashua 88] A. Sha'ashua et S. Ulmann. Strutural saliency: the detection of globally salient structures using a locally connected network. *Proc. Second International Conference on Computer Vision*, Tampa, Florida, 1988.

[Shipman 83] D.W. Shipman. Spirex: statistical analysis in the spire acoustic-phonetic workstation. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, 1983.

[Shore 87] J. Shore et A. Parker. *Introduction to the Entropic Signal Processing System.* Rapport, e^{n} Entropic Processing, Inc., 1987.

[Shortliffe 76] E. H. Shortliffe. *Computer-based Medical consulation: MYCIN.* American Elsevier, New York, 1976.

[Stern 86] P.E. Stern, M. Ezkénazi, et D. Memmi. An expert system for speech spectrogram reading. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1193–1196, Tokyo, 1986.

[Stevens 68] K. N. Stevens. *Acoustic Correlates of Place of Articulation for Stop and Fricative consonants.* Rapport no. MIT-RLE-QPR (89), Massachussets Intitute of Technology, Cambridge, 1968.

[Stevens 78] K. N. Stevens et S. E. Blumstein. Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Amer.*, 64(5):1358–1368, 1978.

[Stoer 80] J. Stoer et R. Burlirsch. Bairstow's method. *Introduction To Numerical Analysis*, chapitre 5, pages 285–287, Springer-Verlag, New York, 1980.

[Technology 86] Signal Technology. *Introduction to ILS Interactive Laboratory System.* Rapport, Signal Technology, Inc., 1986.

[Tubach 85] J.P. Tubach et Louis-Jean Boé. *Un corpus de transcriptions phonétiques : constitution et exploitation statistique.* Rapport no. ENST-85D001, ENST, 1985.

[Turing 50] A.M. Turing. Computing machinery and intelligence. *Mind*, LIX(236), 1950.

[Vaissiere 80] J. Vaissière. The search for language-independent prosodic feature. *Proceedings of First International Congress on the Perception of Speech*, pages 1–27, Florence, Italy, 1980.

[Vaissiere 88] J. Vaissière, M. Eskénazi, et F. Lonchamp. *Cours sur les indices acoustiques du Français.* GRECO-CNRS Communication parlée, 1988.

[Waibel 87] A. Waibel, T. Hanazawa, G. H. Hinton, K. Shikano, et K. Lang. *Phoneme Recognition Using Time-Delay Neural Networks.* TR-I-0006, ATR Interpreting Telephony Research Laboratories, 1987.

[Waltz 75] D. Waltz. Understanding line drawinggs of scenes with shadows. *The Psychology of Computer Vision*, chapitre 2, pages 19–91, McGraw-Hill, New York, 1975.

[Willems 87] L. F. Willems. Robust formant analysis for speech synthesis applications. *Proceedings of European Conference on Speech Technology*, pages 250–253, Edinburgh, U.K., September, 1987.

[Wu 89] Z. L. Wu, J. L. Schwartz, et P. Escudier. A theoretical study of the neural mechanisms specialized in the detection of articulatory-acoustic events. *Proceedings of European Conference on Speech Technology*, Paris, France, September, 1989.

[Zue 85] Victor W. Zue et D. Scott Cyphers. *The MIT Spire System.* Rapport, 1985.

[Zue 86a] V.W Zue, D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M. Randolph, S. Seneff, J.E. Unverferth, et T. Wilson. The development of the MIT lisp-machine based speech research workstation. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986.

[Zue 86b] V.W. Zue et L.F. Lamel. An expert spectrogram reader: a knowledge-based approach to speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 23.2.1–4, Tokyo, 1986.

[Zwicker 79] E. Zwicker, E. Terhardt, et E. Paulus. Automatic speech recognition using psychoacoustic models. *J. Acoust. Soc. Amer.*, 65(2):487–498, 1979.

ANNEXE 1

Première évaluation du système de décodage à base de triplets

1 État du système

Le système s'organise en deux parties principales. La première est destinée à la création des triplets de référence. Elle est entièrement opérationnelle et permet donc de créer des triplets d'une classe quelconque. Pour l'instant l'utilisateur ne dispose que du suivi de formants comme détecteur automatique d'événements acoustiques. La seconde partie destinée au décodage n'est que partiellement réalisée. Comme le suivi de formants est le seul détecteur d'événement acoustique disponible, l'appariement et l'étape de relaxation n'ont été testés que pour les triplets centrés sur des voyelles. La prochaine étape sera d'implanter la détection automatique des barres d'explosion et de leurs caractéristiques.

2 Évaluation des performances

Afin d'avoir une idée des performances de l'approche par triplets nous avons choisi un petit corpus de 25 voyelles choisies dans deux phrases du corpus "La bise et le soleil" prononcées par 10 locuteurs masculins (soit au total 250 voyelles). Les phrases sont :

- La bise et le soleil se disputaient ...
- Alors la bise s'est mise à souffler de toute sa force ...

Nous avons adapté le processus d'appariement étant donnée la très faible taille de ce corpus. L'appariement ne porte que sur la partie centrale des triplets : cela ne signifie pas que nous supprimons la nature contextuelle du triplet, cela signifie seulement que nous limitons la prise en compte des phénomènes de coarticulation au centre du triplet.

La table 1 donne les résultats de l'appariement (sans relaxation) :

Comme nous n'avons pris en compte qu'une partie du contexte et qu'il ne s'agit que d'une première version, ces résultats (91% de voyelles bien reconnues) sont tout à fait encourageants.

38%	le triplet correct en première position
36%	le triplet correct parmi les trois meilleurs candidats
17%	la voyelle correcte (sans le bon contexte) parmi les trois meilleurs candidats
9%	erreur

Table 1 : Performances sur les voyelles

Comme l'étape de relaxation (qui utilise la version floue d'AC4) est à notre sens destinée à localiser les erreurs possibles ou les ilots de confiance nous ne l'avons pas prise en compte dans ces résultats. La fig. 1 montre les triplets identifiés grâce à l'appariement puis ceux que conserve l'étape de relaxation.

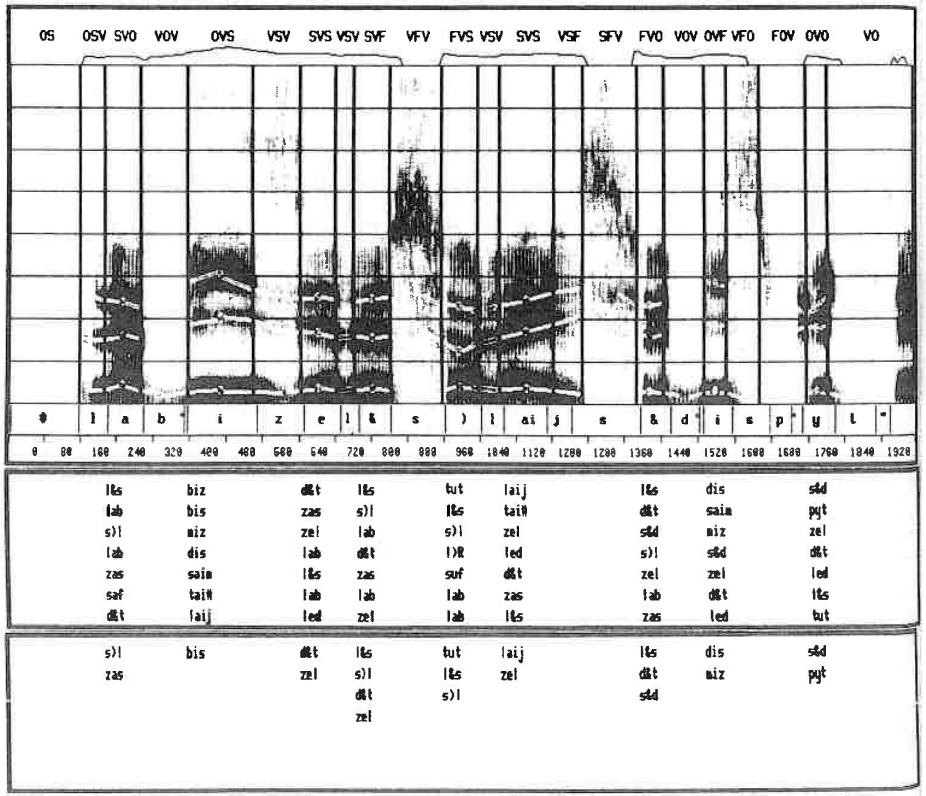


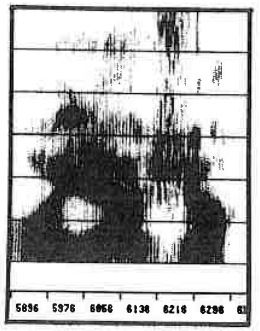
Figure 1: De haut en bas : spectrogramme avec les instances de triplets, résultats de l'appariement et résultats après relaxation

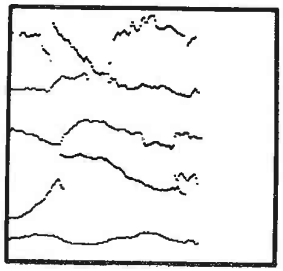
ANNEXE 2

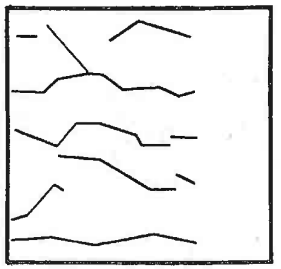
Dans cette annexe nous présentons les résultats du suivi de formants. Pour chaque locuteur masculin (M) ou féminin (F) nous donnons :

- le spectrogramme,
- le sexe du locuteur (M ou F),
- la fréquence fondamentale,
- l'extraction des racines de LPC à droite du spectrogramme,
- les pistes formantiques construites avec les racines LPC et après le raisonnement géométrique (en haut à droite),
- les pics des spectres lissés ceptralement (en bas à gauche),
- les pistes formantiques construites avec les pics des spectres lissés cepstralement et après le raisonnement géométrique (en bas à droite).

Les segments des pistes ont une épaisseur normale quand ils sont issus directement de l'approximation polygonale du squelette ou de raccordements de pistes décrits au paragraphe 5.1. Les segments dont l'épaisseur est plus faible sont soit les segments correspondant aux "croisement de formants" soit les raccordements à pente forte.

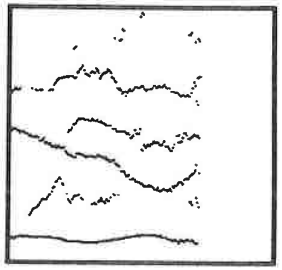


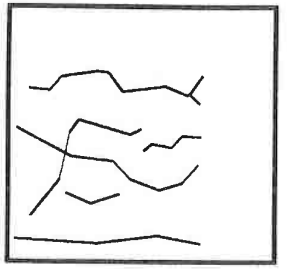


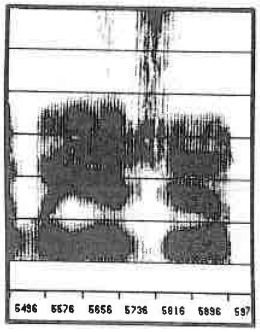


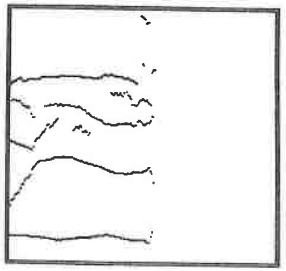
BT (M)

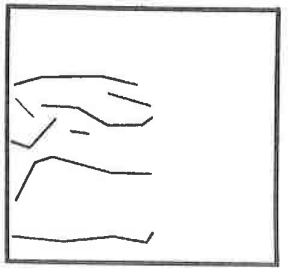
150 Hz





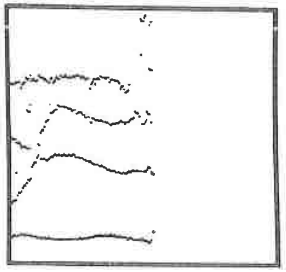


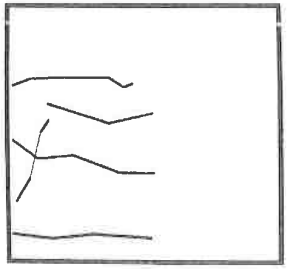


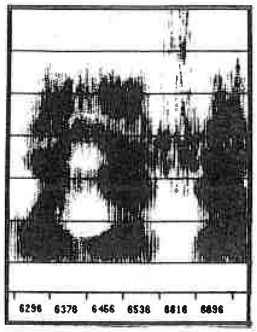


BP (M)

140 Hz

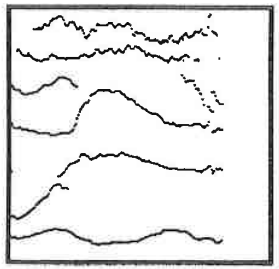


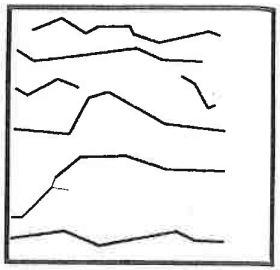


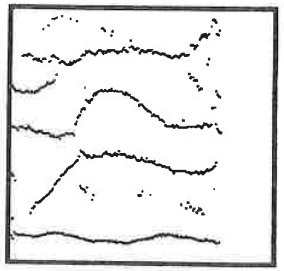


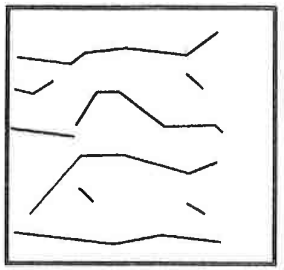
MF (M)

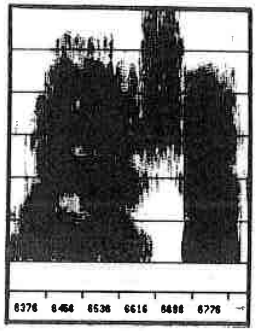
160 Hz





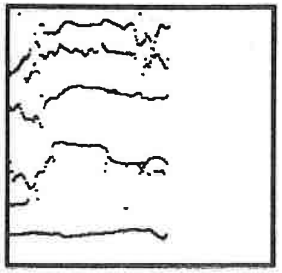


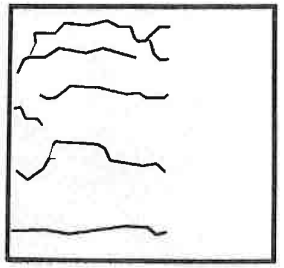


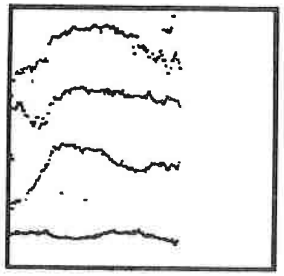


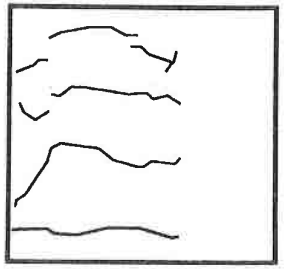
RS (F)

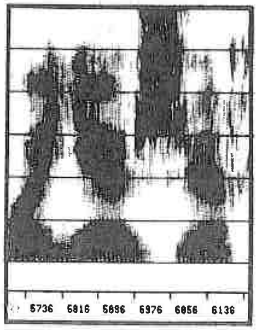
250 Hz





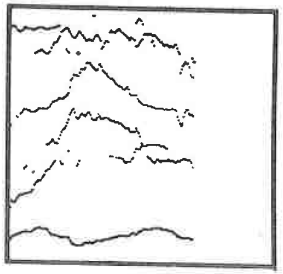


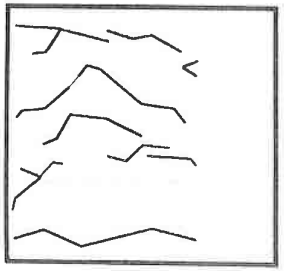


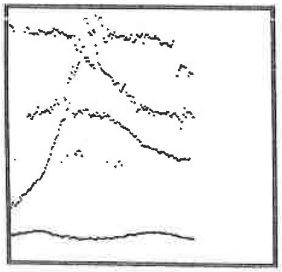


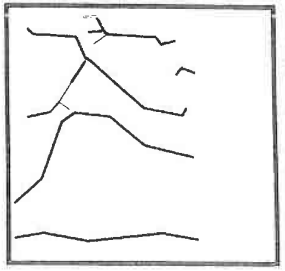
JI (F)

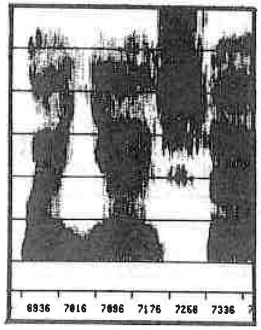
225 Hz





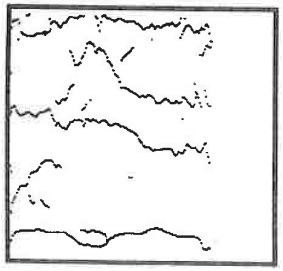


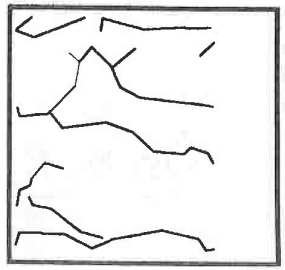


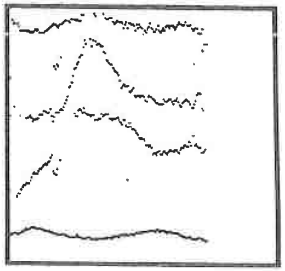


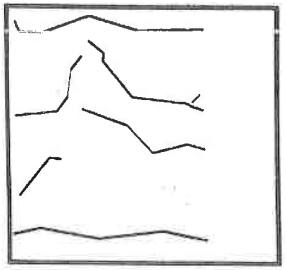
LT (F)

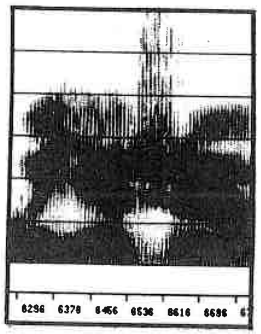
240 Hz





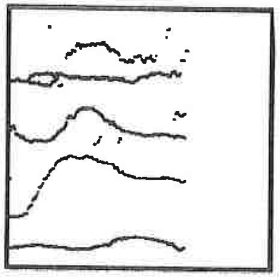


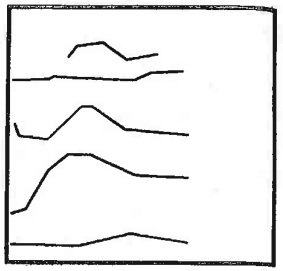


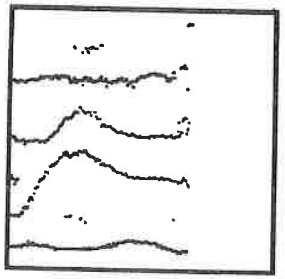


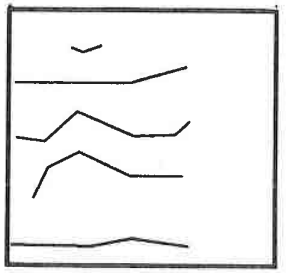
GM (M)

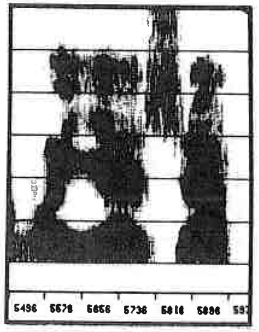
125 Hz





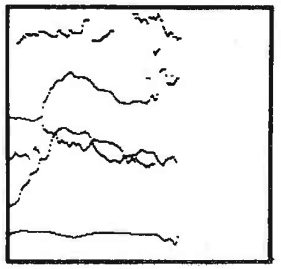


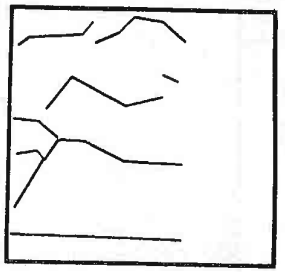


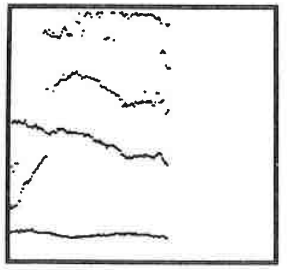


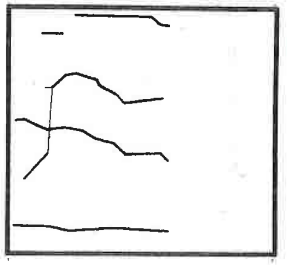
DP (F)

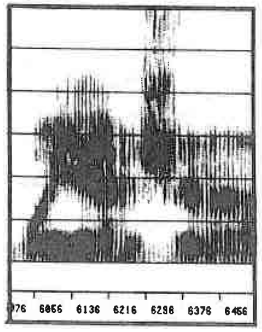
230 Hz





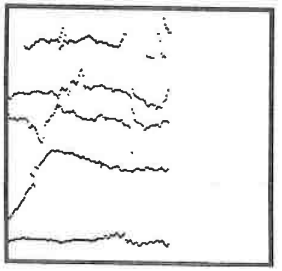


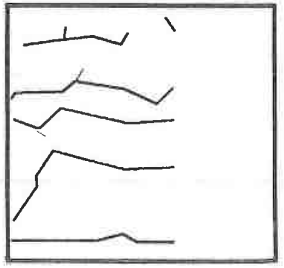


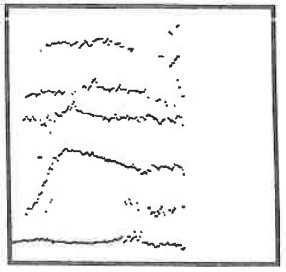


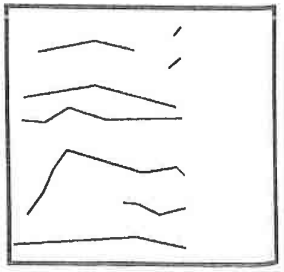
BG (M)

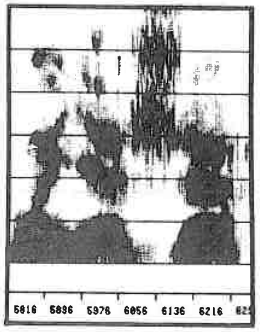
110 Hz





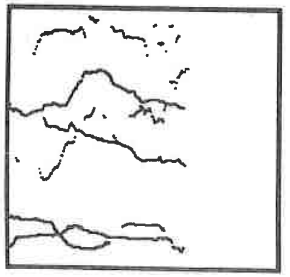


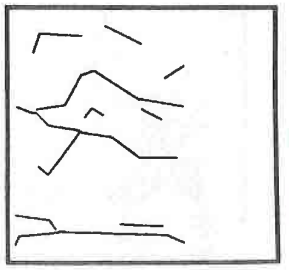


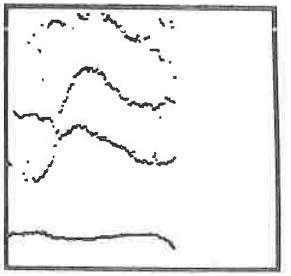


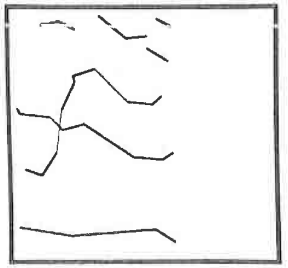
JF (F)

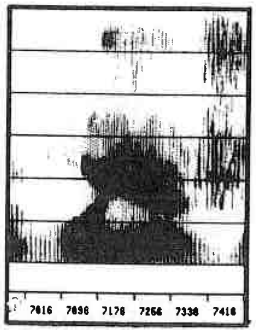
250 Hz





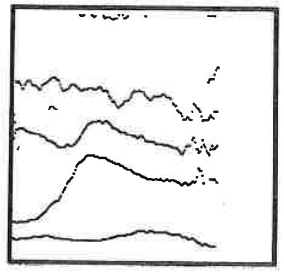


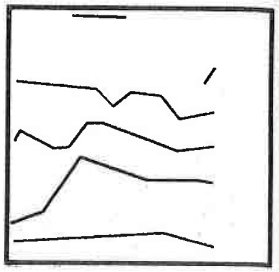


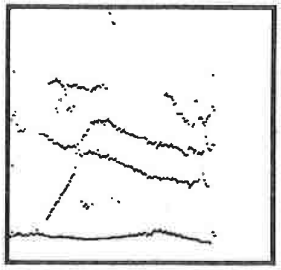


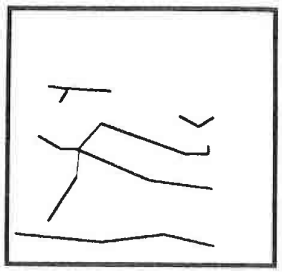
SA (M)

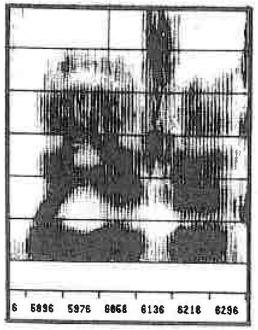
150 Hz





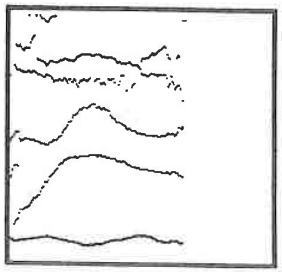


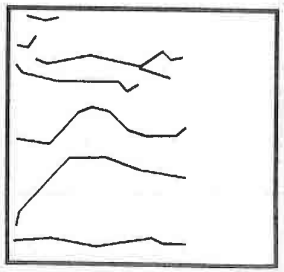


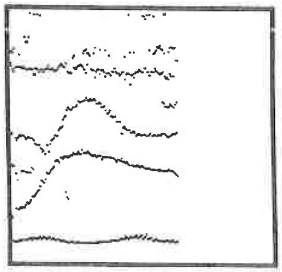


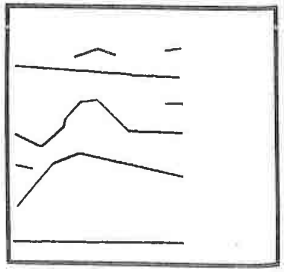
FL (M)

130 Hz









Résumé

Face aux techniques globales dont il est désormais difficile d'améliorer sensiblement les performances, l'approche experte en décodage acoustico-phonétique de la parole continue multilocuteur semble très prometteuse. Ce travail présente plusieurs aspects de cette approche.

Le premier aspect abordé est celui de l'environnement logiciel offert au chercheur en parole. Il doit être à la fois convivial et puissant et c'est dans cet esprit qu'a été développé le logiciel Snorri qui fournit à la fois les outils classiques en traitement de la parole (acquisition, restitution de signaux temporels, calcul de spectrogrammes, étiquetage manuel...) mais aussi un certain nombre de fonctions qui aident l'utilisateur à compléter et à valider son expertise (exploration automatique de corpus, suivi de formants dans le plan F1-F2...).

Pour pouvoir exploiter l'expertise phonétique il est indispensable de disposer de bons détecteurs d'événements acoustiques. Le suivi de formants est sans doute le plus important d'entre eux. Après un panorama des différents problèmes liés à l'extraction des formants deux algorithmes sont proposés :

- le premier pour construire les pistes formantiques en utilisant des techniques classiques en traitement d'images (notamment la morphologie mathématique),
- le second destiné à affiner les résultats obtenus avec l'algorithme précédent en interprétant les pistes formantiques en termes de formants ; le principe de cet algorithme est de maximiser l'énergie du spectrogramme expliquée par les formants.

Les différentes approches du décodage acoustico-phonétique (DAP) sont ensuite replacées par rapport aux connaissances actuelles des phénomènes de la parole continue. À partir des faiblesses des systèmes experts en lecture de spectrogrammes une approche originale est proposée en DAP : *le triplet phonétique*, c'est-à-dire un prototype en termes d'événements et d'indices acoustiques d'un son en contexte. Les perspectives qu'ouvre l'utilisation de ce grain de connaissance sont évoquées avant de décrire un système de DAP à base de triplets en cours de développement. Ce système opère en deux étapes. Dans un premier temps on effectue un décodage local, triplet par triplet, en recherchant les meilleurs candidats sur la base des indices et des événemnets acoustiques détectés dans le signal de parole. Ensuite on accroît la consistance de la solution globale sur la phrase en utilisant des techniques de relaxation discrètes faisant appel aux contraintes acoustiques qui existent entre deux triplets.