Université de NANCY I

Département de Mathématiques
Appliquées et Informatique

Centre de Recherche en Informatique de Nancy

THÈSE DE DOCTORAT D' ÉTAT ES SCIENCES

présentée à l'Université de Nancy I

par

Marie-Christine HATON née AMET,
Agrégée de l'Université,

pour obtenir le grade de
DOCTEUR ES SCIENCES

Contribution à l'éducation vocale assistée par ordinateur: étude des voix et réalisation du système SIRENE

Soutenue le 19 février 1985

Composition du jury :

Président :	C. PAIR
Rapporteurs :	G. PERENNOU J.M. PIERREL
Examinateurs :	R. de MORI J.C. DERNIAME J.P. HATON R. MOREAU G. STAMON

Université de NANCY I

Département de Mathématiques
Appliquées et Informatique

Centre de Recherche en Informatique de Nancy

THÈSE DE DOCTORAT D' ÉTAT ES SCIENCES

présentée à l'Université de Nancy I

par

Marie-Christine HATON née AMET,
Agrégée de l'Université,

pour obtenir le grade de
DOCTEUR ES SCIENCES

Contribution à l'éducation vocale assistée par ordinateur: étude des voix et réalisation du système SIRENE

Soutenue le 19 février 1985

Composition du jury:

Président:	C. PAIR
Rapporteurs:	G. PERENNOU J.M. PIERREL
Examinateurs:	R. de MORI J.C. DERNIAME J.P. HATON R. MOREAU G. STAMON

à Jean-Paul,
Emmanuel et Sébastien.

AVANT-PROPOS

Le travail présenté dans ce document a été réalisé au Centre de Recherche en Informatique de Nancy où j'ai pu trouver, au fil des années, une ambiance et un environnement des plus favorables.

A son terme, je voudrais exprimer tout d'abord mes remerciements à ceux qui me font l'honneur et l'amitié de participer à mon jury,

Monsieur Claude PAIR, aujourd'hui Directeur des Lycées au Ministère de l'Education Nationale, Directeur du CRIN à l'époque de mon entrée au laboratoire, dont l'influence a marqué mes débuts de chercheur en thèse d'Etat,

Monsieur René MOREAU, Directeur du Centre Scientifique IBM-France, dont l'intérêt pour notre travail n'est pas récent et qui nous a fait bénéficier d'une aide aux Thèses il y a huit ans déjà,

Monsieur Jean-Claude DERNIAME, Professeur à l'Université de Nancy et Directeur du CRIN, attentif depuis mon entrée au laboratoire à l'évolution de mes travaux,

Monsieur Renato de MORI, Professeur à l'Université Concordia de Montréal, qui, malgré l'éloignement, a bien voulu juger mon travail,

Monsieur Guy PERENNOU, Professeur à l'Université de Toulouse, qui a suivi mon évolution personnelle au sein de la communauté des chercheurs sur la Parole,

Monsieur Jean-Marie PIERREL, Professeur à l'Université de Nancy, dont j'ai pu, avant sa propre soutenance de thèse en 1981, me sentir un peu l'aînée et qui a su m'encourager en prouvant, quelques années avant moi, que l'on pouvait à la fois assumer une Direction des Etudes et mener à bien sa recherche,

Monsieur Georges STAMON, Professeur à l'Université de Belfort, dont l'amical soutien a favorisé l'achèvement de cette thèse

et Jean-Paul HATON, enfin, Professeur à l'Université de Nancy et mon responsable de recherche, dont la présence a été fondamentale dès le début de mon travail et jusqu'à son aboutissement, avec qui rien n'est impossible.

Je voudrais associer dans une même pensée amicale et remercier

Monsieur Claude LAURENT, Professeur à l'Université de Nancy, qui fut, à l'époque de mon DES sur l'effet Hall, à l'origine de ma vocation de chercheur,

l'ensemble des chercheurs et du personnel du CRIN et en particulier l'équipe RF-IA,

les ingénieurs-système Claude SANCHEZ, Olivier MOREL et Jean-Claude JUNQUA qui se sont succédé autour du Mitra 125, lui-même compagnon souvent nocturne de mes essais en laboratoire,

Marie-Madeleine DUTEL, orthophoniste,

Mesdames Martine KUHLMANN et Danielle MARCHAND qui ont su mettre leur talent et leur soin à la frappe et à la présentation de mon texte de thèse

et mes proches, à qui ce travail appartient.

SOMMAIRE

INTRODUCTION GENERALE

L'homme possède la parole : elle est pour lui le moyen de communication le plus subtil et le plus complet. C'est aussi chez lui très certainement le comportement moteur le plus complexe dont la réalisation suppose que tout un ensemble de positionnements et de mouvements soit coordonné avec une grande précision.

La puissance de la parole est telle qu'aucun autre mode de relation avec soi-même et avec autrui ne peut l'égaler dans le cheminement de la pensée , la mise en forme de l'abstrait, l'expression du plus profond de soi, le développement de l'intelligence, etc. Bien que la relation entre la parole et la pensée soit difficile à apprécier, la parole est beaucoup plus qu'un ébranlement sonore, elle est la manifestation physique de l'expression dans le cadre de structures plus ou moins rigides mais possédant un certain degré d'organisation.

S'intéresser aux phénomènes qui régissent la parole, c'est entrer par là-même dans de nombreux domaines traitant de phonation, d'audition (sous leurs aspects psychophysiologiques, anatomiques, médicaux...), de phonétique, de linguistique ou de psychologie, c'est faire appel à l'expérience multiple de chercheurs et de spécialistes engagés dans ces différentes disciplines.

Nous présentons dans ce document les études que nous avons menées au Centre de Recherche en Informatique de Nancy dans le domaine du traitement automatique du signal de parole et de son application possible à l'éducation vocale.

Il nous a paru intéressant en effet, comme prolongement des recherches menées au CRIN dans l'équipe "Reconnaissance des Formes et Intelligence Artificielle", de mettre à profit ces techniques et le puissant outil que représente l'informatique pour définir et mettre au point un ensemble modulaire d'aides visuelles à l'éducation de la parole. L'idée première était d'apporter une assistance au rééducateur, orthophoniste ou enseignant spécialisé, dans sa tâche auprès des enfants déficients auditifs. Nos résultats nous ont ensuite orienté vers l'aide à l'apprentissage d'une langue seconde. Nous avons dans cet esprit mis au point le système SIRENE (*"Système Interactif pour la Rééducation vocale des Enfants Non-Entendants"*) dont la description, l'utilisation et les prolongements font l'objet de la quatrième et dernière partie (notée D) de ce document.

La ligne principale du travail présenté est l'analyse des voix et l'éducation ou rééducation de la parole, ce qui nous a conduit à en approfondir le contexte. De façon générale, nous avons voulu que notre exposé soit complet et souvent assez didactique, tout en restant dans cette ligne directrice.

Nous commençons la première partie (A) par des réflexions sur la communication parlée, intéressant plus spécialement l'audition, la phonation, la relation entre ces deux fonctions et les questions liées à l'acquisition du langage. Nous avons adopté un point de vue non pas vraiment de spécialiste de ces disciplines mais qui retienne dans chacune d'elles les éléments fondamentaux pour les développements ultérieurs et d'une façon que nous avons voulue personnelle.

La partie bibliographique qui vient ensuite dans la partie A est assez détaillée. Elle résulte en effet d'un noyau de base, rédigé en 1976, à l'occasion d'une table ronde spécialisée qui s'est tenue au cours des 7èmes Journées d'Etude sur la Parole à Nancy, et qui a pris de l'importance et s'est réorganisé au fil des années en suivant l'évolution des recherches. Elle a servi de point de départ de travail à un certain nombre de mémoires ou de thèses d'enseignement spécialisé, en particulier dans le domaine général "Informatique et Handicap", ce qui explique l'importance que nous lui avons laissée dans le texte.

Nous ajoutons ensuite le résultat de notre réflexion sur la définition d'un matériel spécialisé pour l'aide à la production vocale. A ce sujet, précisons que cette réflexion s'est nourrie des contacts divers que nous avons établis avec différents spécialistes : échanges de bibliographie, visites dans les laboratoires de recherche ou les écoles de malentendants, rencontres avec les enseignants spécialisés, etc.

La partie B est consacrée à la paramétrisation du signal vocal dans le souci double de l'étude des voix en temps différé et de la présentation visuelle de ces paramètres en temps réel ou, tout au moins, en ligne.

En premier lieu (chapitre B.1), nous abordons l'extraction de paramètres caractéristiques avec une classification se fondant, non pas sur la méthode utilisée, mais sur le paramètre extrait. Le chapitre B.2 traite de la réduction de données en vue de la recherche de paramètres discriminants. A l'occasion de ces développements, nous indiquons quelles sont les méthodes utilisées dans notre travail ou proposées à l'occasion

de travaux parallèles (comme le filtrage, le codage et la restitution de séquences de parole). Nous avons dans l'ensemble tenté de donner de ces questions une présentation personnelle et critique.

Dans la troisième partie, notée C, dédiée à l'étude des voix, nous décrivons en C.1 une chaîne de modules destinés à l'étude des voix en général et plus spécialement des voix pathologiques. A cette occasion, nous proposons une méthode de lissage et d'interpolation, d'une part, et une méthode d'analyse, d'autre part, de courtes portions de contours, dont nous donnons l'intérêt pour notre travail.

Le chapitre C.2, ensuite, traite de la comparaison et du traitement de formes sonores matricielles. Nous présentons les logiciels mis en place pour l'étude critique de cette question et les choix retenus à la lumière de nos résultats, avec pour objectif la recherche de commandes vocales fiables (chez des sujets I.M.C. en particulier) et l'introduction de la reconnaissance automatique en éducation de la parole.

Dans le chapitre C.3, enfin, nous proposons un plan pour l'analyse des voix qui fait référence aux développements des chapitres antérieurs ainsi qu'aux travaux en modélisation du larynx et en synthèse, dans le triple point de vue du diagnostic médical, de l'orientation à donner à la rééducation vocale et de l'évaluation des progrès.

La quatrième et dernière partie est consacrée au système SIRENE lui-même et à ses prolongements. Elle se réfère bien sûr aux trois parties précédentes de façon à ne pas reprendre les problèmes d'acquisition, d'analyse et, souvent, d'interprétation des données vocales. Nous avons

retenu, dans un premier chapitre, des réflexions sur le rapport audition-vision, la structure et le mode d'utilisation du système SIRENE, conçu comme une aide visuelle à l'éducation de la parole.

Les conditions d'une phase d'expérimentation et les conclusions et réflexions que nous en avons tirées font l'objet du chapitre C.2.

Nous terminons par les perspectives d'extensions et de développement du système, la définition d'une version de SIRENE sur microordinateur et par la façon dont notre travail, dans une certaine mesure, s'insère dans un projet d'Enseignement Assisté par Ordinateur plus vaste, développé au CRIN.

Signalons enfin l'intérêt porté à notre travail par la société IBM-France, intérêt qui s'est concrétisé en 1976 par l'octroi d'une Aide aux Thèses.

PARTIE A

COMMUNICATION PARLEE

INTRODUCTION A LA PARTIE A

Un des obstacles majeurs au développement intellectuel et à l'intégration sociale des sourds est l'incidence de la surdité sur l'acquisition de la parole et du langage et sur la communication en général. Avant de discuter en fin de premier chapitre les difficultés rencontrées sur ce plan, le type idéal d'apprentissage de la langue et les conditions pratiques de rééducation, nous allons envisager différents aspects de la communication parlée tels que l'acquisition normale du langage ou le rapport audition-phonation. Nous renvoyons le lecteur à des ouvrages spécialisés pour des informations plus détaillées sur la physiologie de la phonation et de l'audition, ainsi que pour les notions de phonétique, par exemple [FANT - 60], [MALM - 68].

Dans un deuxième chapitre, nous envisageons les aides à la communication, de façon générale d'abord, puis les aides à la perception et la compréhension et les aides à la production de la parole. Nous concluons enfin par quelques réflexions sur l'intérêt des recherches dans ce domaine, l'apport possible de l'informatique, de la technologie et des techniques de traitement du signal.

CHAPITRE 1

AUDITION - PHONATION - INTERACTIONS

I - ACQUISITION DU LANGAGE

1. Le langage et la psychophysiologie

Lors de l'apprentissage de la langue naturelle, l'enfant acquiert progressivement la maîtrise du positionnement et des mouvements de ses différents organes phonateurs : le larynx et le système respiratoire, le voile du palais (qui peut se lever ou s'abaisser pour provoquer un couplage mécanique avec le conduit nasal), le corps et la pointe de la langue, les mâchoires et les lèvres. Il finit par atteindre une précision de l'ordre du millimètre ou même moins dans l'ajustement des articulateurs et de quelques millisecondes ou dizaines de millisecondes dans la coordination et l'enchaînement des mouvements. Cette précision est acquise après une période d'apprentissage dont le processus n'est pas encore totalement élucidé à l'heure actuelle. C'est à partir de productions vocales à la fois ludiques et linguistiques plus ou moins articulées que l'enfant progresse dans l'acquisition du système phonétique. Il le dominera vers l'âge de 3 ans environ.

Ce qui est sûr, c'est que l'apprentissage est ici, comme c'est souvent le cas, de type associatif : il repose sur des liaisons entre deux événements bien différenciés satisfaisant à la condition de contiguïté temporelle, contiguïté d'une forme entendue et d'une forme émise et vice-versa. Cette contiguïté constante et fréquemment répétée est la condition la plus favorable à l'acquisition du langage ; il est vérifié que plus un mot est fréquent dans la langue, plus il est reconnu facilement à l'écoute et à la lecture. A partir de cette association de deux événements, l'imitation joue un rôle dominant. Bien que les conditions physiologiques et motrices de l'acquisition soient mal connues, on peut tenter

d'en analyser le processus [RISB - 68]. La période d'acquisition se divise grossièrement en trois phases principales :

- la phase "affective" où le langage sert de support à des manifestations d'ordre affectif de la part de l'entourage,

- la phase "ludique" où l'enfant prend conscience de sa propre production vocale et des sensations kinesthésiques et auditives qui lui sont associées,

- la phase "de construction" du langage qui lui permettra de parvenir à une autocommande de la phonation.

L'alternance sensitif-moteur et la progression par approximations successives jouent un rôle dominant. Une forme sonore perçue est mémorisée, vraisemblablement dans une mémoire à court-terme. Elle déclenche par association un processus moteur qui est la tentative d'imitation de cette forme dont le déclenchement se fait à partir des aptitudes acquises antérieurement. L'enfant produit ainsi une approximation de ce qu'il a entendu. Cette forme est comparée à la forme de référence mémorisée. Le résultat de la comparaison peut éventuellement entraîner un nouvel essai. Par essais successifs et grâce à l'intervention de l'entourage et à une autocorrection, l'enfant acquiert une nouvelle aptitude qui est alors stockée dans une mémoire à long-terme.

La détérioration inévitable de la parole des sujets atteints de surdité précoce montre que ce n'est pas seulement une question de mémoire du positionnement et des mouvements articulatoires. L'acquisition chez l'enfant normalement entendant suppose le développement interactif de ses possibilités articulatoires et de ses possibilités sur le plan de la perception auditive, en association avec les influences extérieures.

EDMONDSON [EDMO - 77] l'exprime ainsi : *"one can articulate as well as one can hear and one can hear as well as one can articulate"*, cette affirmation se fondant sur son expérience et celle d'autres chercheurs (MADISON et FUCCI) qui trouvent une corrélation significative entre l'articulation et la discrimination des sons de la parole. Cette phase s'étend à peu près jusqu'à la puberté. L'articulation devient ensuite indépendante de la perception et le contrôle de nature purement kinesthésique. Nous reviendrons sur cette question en liaison avec l'audition au paragraphe IV de ce chapitre.

2. Le langage et le linguiste

La capacité de faire l'association signifiant-signifié est une condition nécessaire au développement du langage. Inversement et parallèlement, le langage joue un rôle nécessaire, mais non suffisant, pour aider à l'intériorisation des opérations liées à l'expression verbale.

Cette idée de double alternance se rencontre dans nombre de théories en psycholinguistique. Ainsi, DE SAUSSURE [SAUS - 76] définit la langue comme un système de signes distincts correspondant à des idées distinctes. Il attribue alors une double activité au signe linguistique en fonction de la signification qui lui est attachée, c'est le "signifié", vis-à-vis de la forme qui le désigne qui devient le "signifiant". Il précise que, plus subtilement, la relation entre signifiant et signifié n'est pas biunivoque : *"il est possible d'adapter aux circonstances la quantité d'indication significative fournie au moyen du signal"*.

André MARTINET [MARTI - 66], introducteur de la linguistique "fonctionnelle", parle de double articulation de l'énoncé linguistique :

- articulation en *monèmes* sur le plan de l'expression et sur celui du contenu. Toute information à communiquer à l'interlocuteur *"s'analyse en une suite d'unités douées chacune d'une forme et d'un sens"*. Cette première articulation se rencontre dans presque tous les systèmes de codage,

- articulation en *phonèmes*, uniquement sur le plan de l'expression orale, caractéristique cette fois du langage humain. Elle concerne la succession d'unités analysables dans le signal vocal, entachées ou non de signification.

Les unités correspondant à ces deux articulations sont ainsi, d'une part, le monème (qui se divise en sémantème pour le fond et morphème pour la forme), d'autre part, le phonème. Quelques dizaines de phonèmes et quelques milliers de monèmes suffisent pour construire le système de communication orale de l'homme.

L'acte de communication, quant à lui, est un acte extrêmement complexe dans lequel le locuteur, son acquis, sa personnalité et son environnement interviennent à des degrés divers. A chacun des facteurs qui l'influencent peut être associée une fonction indépendamment des autres comme les définit, par exemple, JAKOBSON [JAKO - 63] : aux points de vue du locuteur, de l'auditeur, de l'univers, du contact, de la forme du message et du code, JAKOBSON associe respectivement les fonctions émotive, conative, référentielle, phatique, poétique et métalinguistique. On imagine aisément, sur ce plan de la communication parlée, l'importance du "bain de parole" dans lequel l'enfant normalement entendant est plongé dès sa naissance. Les travaux menés actuellement en psychologie cognitive (par exemple [LENY - 79]) apportent un éclairage nouveau à ce problème.

PETERSON [MALM - 68] représente les composantes essentielles du processus de communication orale mettant en jeu deux interlocuteurs par le schéma suivant (figure A.1) :

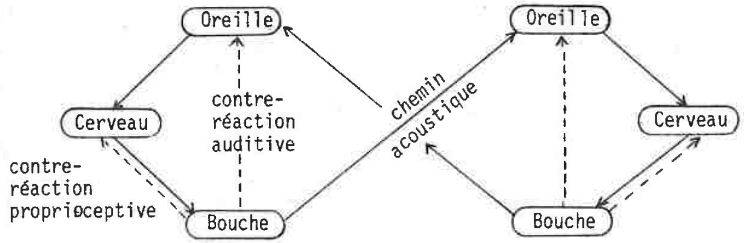


Figure A.1 : Communication "inter-locuteurs"
(d'après PETERSON)

Il est clair que l'audition joue, en plus du rôle "inter-locuteurs", un rôle "intra-locuteur" de contrôle de la phonation. Nous nous arrêtons plus spécialement au paragraphe suivant aux différents aspects de l'audition.

II - AUDITION

1. L'oreille et l'anatomie

Sans vouloir reprendre la description anatomique de l'oreille, nous proposons un schéma permettant de comprendre les différentes étapes du parcours de l'information auditive du milieu aérien extérieur jusqu'au cortex cérébral. Pour cela, nous indiquons successivement, en face de la zone anatomique concernée, les organes majeurs, leur fonction, les supports de l'information véhiculée, les mesures audiométriques généralement pratiquées pour l'examen du système auditif et enfin les surdités associées à chaque niveau (figure A.2).

Zone anatomique	Organes majeurs	Fonction	Type d'information véhiculée	Mesures audiométriques	Surdités possibles associées
Oreille externe	*Pavillon* *Conduit auditif externe*	*Concentration d'énergie* *Localisation source sonore (perception binaurale)* *Résonance (autour de 3 000 Hz)* *Amplification de pression*	*Onde de pression*	*Réponses myogéniques*	
Frontière	*Tympan*	*Conversion onde sonore en vibrations de la membrane*	*Vibration de membrane*		
Oreille moyenne	*Caisse du tympan* *Osselets et muscles*	*Transmission de l'onde du milieu aérien aux milieux liquidiens de l'oreille interne.* *Adaptation d'impédance (naideur accrue - multiplication par 30 des pressions)*	*Mouvements mécaniques des osselets*	*Réflexes stapédiens*	*Surdités de transmission*
Frontière	*Fenêtre ovale*	*Porte d'entrée vers l'oreille interne*			
	Fenêtre ronde	*Décompression*			
Oreille interne	*Cochlée* *Membrane basilaire et cellules ciliées (organe de Corti)*	*Codage de la vibration mécanique due à la propagation de l'onde en influx nerveux*	*Onde de pression en milieu liquidien* *Déformations de la membrane basilaire*	*Potentiels cochléaires*	*Surdités de réception*
Voies nerveuses	*Neurones cochléobulbaire bulbothalamique thalamocortical*	*Transmission de l'influx nerveux*	*Influx nerveux*	*Potentiels évoqués auditifs*	
Cortex auditif	*Aires de projection corticale*	*Décodage de l'information* *Identification*		*Audiométrie verbale*	*Surdités centrales : troubles de l'identification, agnosie auditive*

Fig. A-2 : L'oreille et la fonction auditive

2. L'oreille et l'audition

La cochlée, organe majeur de l'oreille interne, enroulée en limaçon, est l'organe de l'audition par excellence. Il est le siège de la conversion par les cellules ciliées (organe de Corti) de la vibration mécanique en des décharges d'influx nerveux dans leurs arborisations terminales. Ces décharges sont recueillies par environ 30 000 cellules ganglionnaires d'où partent les fibres du nerf cochléaire, l'un des composants du nerf auditif.

Diverses théories se sont affrontées au sujet du mécanisme physique d'analyse d'une onde sonore par l'oreille. Les théories de base sont dues :

- à HELMHOLTZ (1877) pour qui l'analyse se fait par des résonateurs qui, alignés le long de la cochlée, exciteraient de façon sélective les fibres du nerf auditif, chaque fibre n'étant excitée que par une fréquence (ou bande de fréquence) particulière. L'analyse serait périphérique,

- à RUTHERFORD qui émet l'hypothèse d'une analyse non pas périphérique mais centrale : sous l'action de l'onde sonore, l'ensemble de l'organe de Corti réagit et le nerf auditif, dont toutes les fibres sont excitées, réagit comme une ligne téléphonique.

Cette dernière théorie est complètement abandonnée au profit d'un affinement de la première, grâce principalement aux travaux de VON BEKESY : il existe une répartition topographique des fréquences dans la cochlée, répartition qui se retrouve probablement dans l'aire de projection corticale. Les fibres issues des cellules ciliées, suivant qu'elles sont externes ou internes, remplissent les fonctions suivantes, semble-t-il :

Cellules ciliées	Fonction de décodage	Fonction supplémentaire
.externes (fibres longues)	.excitation par ondes B.F.	.faibles stimulations .perception globale
.internes (fibres courtes)	.excitation par ondes H.F.	.fortes stimulations .localisation des sons

Malgré des variations individuelles assez marquées, les travaux réalisés dans le domaine de la psychoacoustique fournissent des précisions numériques sur les seuils différentiels de fréquence et d'intensité [ZWIC - 81].

3. La perception auditive

a) Localisation corticale

La représentation corticale de l'audition est bilatérale alors que pour le langage elle est le plus souvent localisée dans l'hémisphère dominant (bien que l'hémisphère mineur conserve dans la première enfance des possibilités d'apprentissage et de suppléance).

De nombreuses fibres d'association relient ces zones aux zones de réception des messages visuels, des mécanismes de production de la parole sur le plan purement moteur ou des mécanismes idéationnels du langage. Ces proximités jouent une part importante dans le bouclage audition-phonation que nous mentionnerons plus loin.

Sur un plan voisin, ajoutons au passage que certains auteurs [AIMA - 74] insistent sur la proximité des zones corticales motrices de la face et de la main pour expliquer les corrélations entre la langue écrite et parlée, entre la main et le langage.

b) Unités phonologiques au niveau de la perception

La question de savoir quelles sont pour le cerveau les unités phonologiques au niveau de la perception est sujette à discussion. Les conclusions tirées dans ce domaine dérivent d'une expérimentation qui fait une large place aux tests de perception.

La perception implique un acte de classification à partir d'un stimulus donné. MALMBERG [MALM - 68] établit qu'il existe une corrélation positive entre la description du stimulus et celle de la sensation auditive. KOSTER [KOST - 77], quant à lui, conclut de son expérimentation qu'il y a un manque de corrélation absolu entre le signal acoustique et le signal perçu : ou bien la description acoustique classique (en phonèmes) est insuffisante, ou bien la perception obéit à des principes autonomes qui ne se réfèrent pas aux phonèmes. Quoi qu'il en soit, il est sûr qu'il n'est possible de distinguer des sons que si l'on peut les ranger dans des classes différentes. Les auteurs situent ce processus de classification à des niveaux divers : pour les uns, l'unité est le phonème, pour les autres, les phonèmes sont intégrés dans des unités plus vastes.

LIBERMAN et al. [LIBE - 67] défendent la théorie motrice de la perception selon laquelle la capacité de discrimination est due à un lien de contreréaction entre l'articulation et la perception. A un stimulus acoustique s'associent directement les positions et mouvements articulatoires nécessaires à sa production. Ce point de vue ramène la perception au niveau du phonème. Pour appuyer cette théorie, ces auteurs avancent entre autres les arguments suivants : d'abord, les phonèmes sont des unités linguistiques connues de l'auditeur qui écoute sa propre langue (notre perception n'est pas indépendante de la langue que nous parlons). Le nombre de syllabes étant de plusieurs centaines, voire de plusieurs milliers, la perception des syllabes ou d'unités plus larges

repose sur la perception première des phonèmes, du moins d'une grande partie de ces phonèmes. Ensuite, la communication rapide (dont le débit moyen est de 10 phonèmes par seconde) est possible grâce au fait que les phonèmes sont imbriqués dans le signal de parole de façon complexe. Si chaque phonème était représenté par un son spécifique, le pouvoir de résolution temporelle limité de l'oreille ne permettrait pas de percevoir le discours continu à la vitesse d'évolution normale. Ceci n'est rendu possible que parce que l'information sur les segments phonémiques successifs est délivrée en parallèle. La perception devient en quelque sorte un travail de décodage de cette information.

JAKOBSON [JAKO - 68] estime erronée cette théorie motrice en s'appuyant sur le simple fait que, selon lui, dans les langues étrangères il est plus aisé de percevoir les phonèmes que de les prononcer. La plupart des auteurs critiquent en elle l'importance accordée à l'unité phonème. Selon NEISSER [NEIS - 67], la perception d'unités plus larges est nécessaire pour atteindre le niveau du phonème. Le contexte, la syntaxe et le sens de la phrase sont d'une grande importance. De même, MILLER [MILL - 56] estime que la décision ne peut se faire qu'après un apport d'information suffisamment grand. Pour certains, l'unité serait plutôt la syllabe. C'est le cas de MASSARO [MASS - 72] qui pense que des unités de la taille de la syllabe sont enregistrées dans une mémoire préperceptuelle puis traitées à ce niveau. La même idée est avancée par SAVIN et BEVER [SAVI - 70] ou par SEGUI [SEGU - 82] qui pensent que l'identification des phonèmes est plus lente que celle de la syllabe qui les contient.

Cette diversité d'opinions montre la difficulté qu'il y a d'explorer ce qui se passe au niveau du décodage de l'information acoustique.

Nous conclurons avec KIRMAN [KIRM - 73] pour dire que, même si la perception ne se fait pas au niveau des phonèmes, il est sûrement possible de les atteindre en dépit de la complexité de leur représentation acoustique, du moins pour nombre d'entre eux. Les phonèmes doivent posséder des traits qui peuvent être intériorisés et qui peuvent conduire à une classification malgré la variabilité du signal physique. Cette idée de traits constitue le fondement de la théorie de JAKOBSON et al. [JAKO - 63] qui introduisent la notion de traits distinctifs binaires permettant de caractériser le phonème. Il n'est pas vraiment important, en fait, que les phonèmes gardent leur identité propre ; l'essentiel est que leur contenu informatif ne soit pas perdu. L'information est alors contenue dans la forme globale dont les qualités du point de vue perceptif sont la conséquence de la présence organisée de ses éléments constitutifs.

Ces idées nous conduiront en A.2.II à discuter des problèmes de perception chez les sourds et des aides qui peuvent être mises en oeuvre pour la compréhension du message oral.

III - PHONATION

1. Organes phonatoires et anatomie

a) Présentation générale

Nous nous limitons à une présentation schématique des organes qui entrent en jeu dans la production de la parole avec l'indication de leur fonction ou de leur effet résultant sur le signal sonore (fig. A-3).

b) Le larynx

Le son laryngé résulte de la mise en vibration des cordes vocales sous influence nerveuse. Différentes théories ont été proposées depuis plus d'un siècle pour expliquer l'origine de cette vibration. Au 19ème siècle, EWALD développe la théorie myoélastique suivant laquelle la vibration des cordes vocales est due au passage de l'air entre les cordes tendues. L'amplitude du son laryngé et sa fréquence ne dépendraient respectivement que de la pression de l'air et de la tension des cordes.

En 1950, HUSSON oppose à cette idée une théorie dite neurochronaxique. Jusqu'à la fréquence de 500 Hz environ, la vibration est commandée par des salves d'influx nerveux. Mais, au delà de cette fréquence, il y a tétanisation des fibrilles de certains muscles constricteurs de la glotte (les thyro-aryténoïdiens internes) qui ne peuvent plus par leur contraction suivre l'arrivée des influx nerveux.

Après plusieurs tentatives d'explication intermédiaires, on a maintenant adopté la théorie de MAC LEOD et SYLVESTRE, développée en 1970, qui réalise une sorte de synthèse des théories neuromusculaire et myoélastique. Selon ces chercheurs, la vibration des cordes vocales est un phénomène actif d'origine musculaire. La fibre musculaire des cordes vocales possède une innervation polysynaptique qui lui permet d'accepter un nombre élevé d'excitations. Jusqu'à 500 Hz, les fibrilles répondent

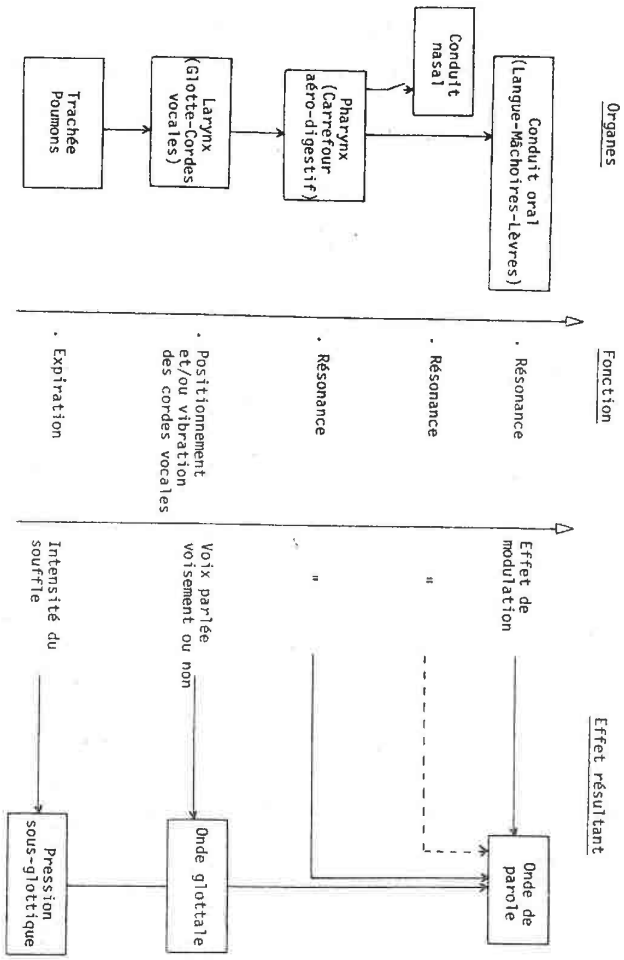


Fig. A-3 (à lire de bas en haut)
Les organes de production de la parole

toutes à chaque influx nerveux. Entre 500 et 1 000 Hz environ, elles se divisent en deux groupes qui répondent à un influx sur deux à tour de rôle : c'est le régime biphasé pour lequel on parle de voix de fausset pour l'homme et de voix de tête pour la femme (certains chanteurs ont un troisième registre correspondant au régime triphasé et, exceptionnellement, un quatrième).

On admet maintenant (théorie neurooscillatoire) que le larynx se comporte comme un système mécanique résonant. Le son laryngé est produit grâce au passage d'un courant d'air sous-glottique. Ce souffle est produit au niveau thoracique supérieur quand on parle normalement, au niveau thoracoabdominal quand on veut élever la voix. Ainsi, c'est la combinaison du passage de l'air et de la vibration des cordes vocales qui produit l'onde glottale grâce à deux systèmes d'innervation indépendants, l'un respiratoire, l'autre phonatoire (avec en plus une innervation végétative qui peut parasiter l'innervation normale, en cas d'émotion par exemple). C'est ce double aspect qui nous permet de comprendre l'origine des principaux défauts rencontrés à ce niveau dans la voix des sourds.

Les contributions relatives des systèmes laryngé et respiratoire aux variations de la fréquence fondamentale et de l'intensité du son glottal font l'objet de recherches particulières pour une meilleure compréhension de leur incidence sur la production de faits linguistiques tels que l'accentuation du mot, l'intonation de la phrase, etc. Citons en particulier les thyrogrammes, par exemple [YAKI - 76], établis à partir de la mesure des mouvements verticaux du larynx qui se révèlent très voisins des courbes d'intonation avec un facteur de déplacement de 4mm pour 40 Hz soit 10 Hz/mm à peu près. D'autres travaux sur la modélisation du larynx que nous mentionnerons plus tard (paragraphe C.3.III) contribuent à l'affinement des théories sur la production de l'onde glottale.

2. La parole et le phonéticien

Nous avons déjà mentionné le phonème comme unité de seconde articulation selon MARTINET, unité décelable dans le signal vocal, si non exactement analysable, et entachée ou non de signification. Différentes tentatives de définition du phonème ont été faites. Nous retiendrons, dans une vue pratique de rééducation vocale, celle qui, à l'idée de phonème, associe une classe dans l'ensemble des sons minimaux : des individus appartenant à des classes différentes se distinguent comme /p/ et /t/ dans les paires minimales /pa/ , /ta/ , par exemple.

Les phonèmes français peuvent se ranger en différentes catégories suivants les traits auxquels on se réfère (vibration ou non des cordes vocales, mise en jeu ou non du conduit nasal, son tenu ou transitoire, lieu et mode d'articulation...). On peut retenir le simple classement en voyelles et consonnes qui donne à ces deux termes la valeur du langage courant.

a) Les voyelles sont caractérisées par :

- l'intervention des cordes vocales (sons voisés),
- une configuration stable des articulateurs (sons tenus),
- le passage libre de l'air,
- l'émission de tout ou partie de l'onde sonore par la bouche.

La cavité orale joue uniquement un rôle de résonateur pour ses modes de vibration contenus dans l'onde glottale.

Ces résonances, caractéristiques des sons émis, constituent les formants. Les deux premiers formants, dont les fréquences sont communément notées F_1 et F_2 , sont directement liés aux volumes des cavités limitées par la langue et au positionnement des lèvres, facteurs primordiaux dans l'articulation des voyelles.

On distingue :

- les voyelles orales, souvent représentées dans le plan de leurs deux premiers formants (fig. A-4).

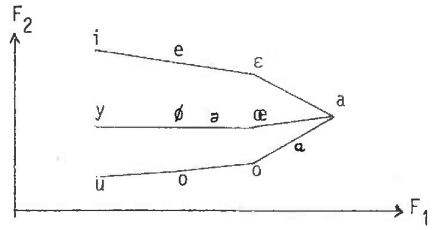


Fig. A-4 : Les voyelles orales dans le plan F_1-F_2

Cette représentation peut se rapprocher du classement suivant (fig. A-5) en fonction du lieu d'articulation et de l'aperture (ouverture minimale du conduit vocal) :

Voyelles	fermées	semi-fermées	semi-ouvertes	ouvertes
. d'avant non labiales	i	e	ɛ (ɛ̃)	a
. d'avant labiales	y	ø	ə œ (œ̃)	
. d'arrière labiales	u	o	ɔ (ɔ̃)	ɑ(ɑ̃)

opposition aigu ↑ ↓ grave

opposition diffus ⟷ compact

Fig. A-5 : Les voyelles orales (nasales) : lieu d'articulation et degré d'aperture

- les voyelles nasalisées, situées sur la figure A-5, caractérisées par la résonance du conduit nasal jouant le rôle d'un tuyau ouvert placé en parallèle sur le conduit oral. La nasalité est un trait distinctif en français, perçu immédiatement par l'oreille mais très difficile à caractériser du point de vue acoustique.

Pour une lecture plus aisée du texte, nous donnons ici le son correspondant au symbole phonétique international :

i	e	ɛ	a	:	sons	i	é	è	a
y	ø	ə	œ	:	sons	u	eux	e	oeu(f)
u	o	ɔ	ɑ	:	sons	ou	eau	o	â
ɛ̃	œ̃	ɔ̃	ɑ̃	:	sons	in	un	on	an

Les figures A-4 et A-5 mettent en évidence que le formant F_1 est directement lié au degré d'aperture alors que F_2 se rapporte au lieu d'articulation. Ces liens seront utilisés pour l'appréciation de la performance dans l'articulation des voyelles.

Les oppositions compact/diffus (F_1 et F_2 rapprochés/ séparés) et grave/aigu (concentration de l'énergie en basse fréquence/ haute fréquence) indiquées sur la figure A-5 permettent de placer les voyelles /a/ /u/ et /i/ au sommet d'un triangle, dit "triangle de HELLWAG" (fig. A-6).

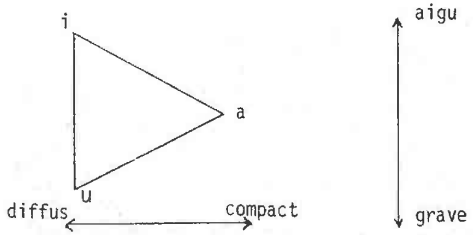


Fig. A-6 : Voyelles a i u : le triangle fondamental

b) Les consonnes comprennent les sons qui ne respectent pas l'une ou plusieurs des quatre caractéristiques des voyelles.

On distingue quatre catégories :

- les fricatives qui sont continues, provoquées par le passage forcé de l'air à travers une constriction du conduit vocal et diffusées par la bouche exclusivement. La turbulence produit une source de bruit qui excite le conduit vocal. La puissance est faible et le spectre est celui d'un bruit à large bande. Par exemple /s/ a les fréquences distribuées entre 4 et 8 ou 9 kHz alors que pour /ʃ/ les fréquences dominantes vont de 2 à 6 ou 7 kHz. On parle de "*formants de bruit*" pour caractériser les zones du spectre renforcées par la configuration des cavités vocales au moment de leur production.

On distingue :

. les fricatives voisées / v z ʒ / (sons v z j)

. les fricatives non voisées / f s ʃ / (sons f s ch)

de mêmes lieu et mode d'articulation,

- les plosives - ou occlusives - qui sont essentiellement transitoires, de faible puissance et dont la nature dépend du comportement dynamique des articulateurs. La fermeture du conduit provoque une surpression puis une relâche rapide à l'ouverture. Il s'ensuit une excitation à large bande du conduit vocal qui impose ses propriétés de transmission au spectre de fréquences des différentes sources.

On distingue :

. les plosives voisées / b d g /

. les plosives non voisées / p t k /

On peut effectuer un classement élémentaire analogue au "triangle de HELLWAG" pour les voyelles, c'est-à-dire fondé sur l'opposition compact/diffus (k/t ou k/p) et sur l'opposition grave/aigu (p/t) (fig. A-7),

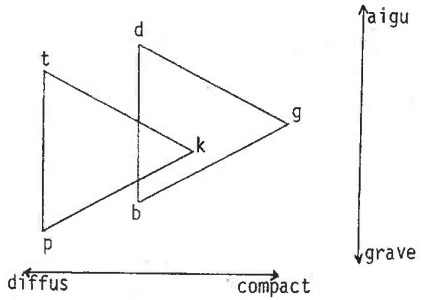


Fig. A-7 : Les consonnes /p t k/ et /b d g/ en double opposition

- les nasales, qui sont des consonnes occlusives dont la transmission et l'émission se font essentiellement par les cavités nasales.

Ce sont en français les trois consonnes / m n ɲ / (sons m , n et gn) qui, du point de vue de l'articulation, correspondent à / b d k palatal /,

- les liquides / R l / et les semi-voyelles / j ɥ w / comme dans "Yann, lui et moi" qui sont sonores et orales.
/l/ est souvent classé comme latéral : la langue entre en contact avec le point d'articulation, comme pour les occlusives, mais l'air s'échappe des deux côtés de ce point. Le spectre montre un léger bruit de friction.
/R/ postérieur peut être considéré comme vibrante : la luette provoque une série d'occlusions brèves au passage de l'air,

- il n'existe pas de diphtongues en français moderne.

Dans le tableau suivant (fig. A-8), nous tentons de placer les consonnes suivant leurs lieu et mode d'articulation, malgré le côté artificiel d'une telle classification dû aux variations phonologiques :

Lieu d'articulation \ Types de consonnes	Occlusives	Nasales	Liquides	Fricatives
Vélaire	k g*	ŋ*	R*	
Palatal				ʃ ʒ*
Alvéolaire			l*	s z*
Alvéodental	t d *	n*		
Dental				f v*
Labial	p b*	m*		

Fig. A-8 : Les consonnes : lieu et mode d'articulation ; l'astérisque (*) indique le voisement.

3. Acquisition par l'enfant du système phonétique

La figure A-9 ci-dessous, adaptée de KENT, 1962, indique l'évolution chronologique de l'acquisition des consonnes par l'enfant.

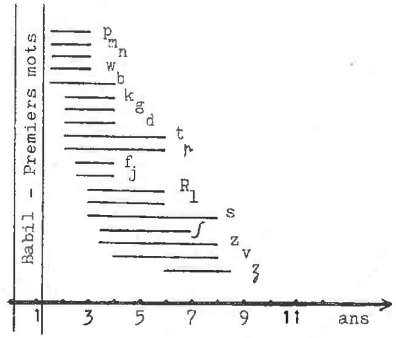


Fig. A-9 : Acquisition des consonnes (d'après KENT)

IV - RAPPORT AUDITION-PHONATION

1. Niveaux de stimulation de la voix

La commande de la voix se fait grâce à trois niveaux de stimulation [GARD - 70].

Le premier, le niveau cortical, est celui des intentions expressives ; il commande les variations de tension des muscles d'ouverture et de fermeture de la glotte et le rythme de vibration des cordes vocales. C'est à ce niveau qu'intervient le lien entre la phonation et l'audition : la suppression du contrôle auditif entraîne des perturbations de la fréquence de vibration des cordes vocales. C'est le cas de la surdité dont les répercussions sur la voix et le langage sont d'autant plus graves qu'elle est intervenue plus tôt. D'un autre côté, à une surdité totale limitée à une bande de fréquences correspond une voix chantée présentant des défaillances précisément à ces fréquences ; à moins que, grâce à une bonne mémoire auditive, si la surdité est acquise, la commande ne se fasse correctement à partir d'une représentation mentale du son voulu. C'est au second niveau, le niveau diencéphalique, qu'intervient l'influence des émotions sur la voix, la base du cerveau jouant un rôle essentiellement affectif, par opposition au rôle intellectuel de l'écorce cérébrale. Terminons enfin par l'étage bulbaire, point de terminaison des fibres sensitives des nerfs auditif et glossopharyngien, d'une part, et point d'origine en particulier des fibres motrices du nerf glossopharyngien et du grand hypoglosse, d'autre part. Cette proximité fait qu'il existe à ce niveau des liens réflectifs entre audition et phonation. Citons pour seul exemple le fait que l'on puisse influer sur le comportement des cordes vocales par simple stimulation du nerf auditif.

2. Une représentation fonctionnelle

Les remarques que nous avons faites dans ce chapitre sur les rapports entre l'audition et la phonation, soit encore entre l'oreille et le langage, nous conduisent au point de départ de notre réflexion. L'oreille joue un rôle fondamental dans l'autorégulation de la phonation, quel que soit le niveau où l'on se place. La figure A-1 qui décrit l'acte de communication orale peut se transposer de la façon suivante, du côté du locuteur (figure A-10) :

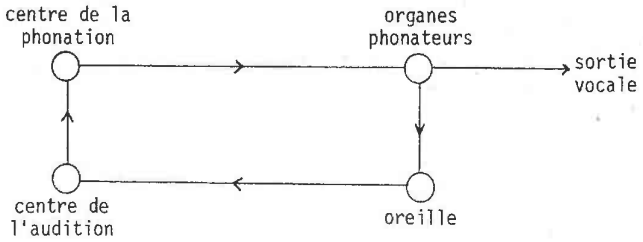


Fig. A-10 : La phonation régulée par l'oreille

On a affaire à un système asservi dont l'oreille est le régulateur. Son rôle est primordial à la fois dans la phase d'acquisition du langage et dans l'attitude générale de communication orale où elle permet en particulier au locuteur de placer sa voix, d'évaluer la valeur "informationnelle" de son discours.

Ces réflexions nous conduisent naturellement à envisager les difficultés rencontrées par les enfants sourds en ce qui concerne la parole et le langage.

V - SURDITE ET PAROLE

1. Difficultés

L'enfant sourd qui n'a pas connaissance des formes vocale émises par son entourage et par lui-même ne peut s'engager dans le processus normal d'acquisition du langage, simplement parce que toute attitude d'imitation lui est impossible. D'un côté, il ne peut savoir si ce qu'il a émis est correct, par ailleurs, il n'a pas accès aux contraintes phonologiques, lexicales ou syntaxiques de la langue. Le problème est énorme pour les personnes sourdes de naissance ou ayant perdu l'ouïe avant l'âge de cinq ans environ. Ce qui a été dit précédemment sur la phase de construction du langage explique que, dans ce cas, les déficiences soient difficiles à surmonter.

Aussi est-il important de donner à l'enfant sourd le moyen de s'intégrer dans un environnement dont la langue première ne lui est pas accessible de façon spontanée et émotionnelle. Il s'agit, d'un côté, de lui présenter un système d'information capable de le renseigner sur la façon dont doivent être prononcés les phonèmes à l'intérieur des mots, les mots et les éléments de phrases (facteurs segmentaux et facteurs prosodiques). D'un autre côté, il faut l'amener à maîtriser les structures spécifiques de la langue. Le vocabulaire acquis par l'enfant, passif par essence, devra favoriser chez lui une attitude de communication active. C'est essentiellement un problème de sociolinguistique. Le travail réside dans la nécessité de clarifier pour le sourd *"les lois spécifiques et les structures mentales de la langue"* [SCHU - 70].

2. Classification des déficiences auditives - Incidence sur le langage

On a l'habitude de définir le degré de surdité à partir de l'audiogramme tonal, courbe donnant en dB le seuil d'audibilité de l'oreille en fonction de la fréquence.

Au-dessous d'un déficit moyen de 40 dB, l'enfant est malentendant. Il présente quelques troubles articulatoires mais accède au langage parlé.

De 40 à 60 dB, le sujet est classé comme demi-sourd ; le langage peut se développer en partie malgré des défauts d'articulation plus sévères. De 60 à 80 dB, pour le demi-sourd profond, l'acquisition du langage présente d'énormes difficultés. Au delà, il s'agit de cophose ou surdité totale.

On trouve en réalité différentes définitions de la surdité profonde : moyenne des seuils pour 500, 1 000 et 2 000 Hz supérieure à 95 dB pour ERBER [ERBE - 74], entre 80 et 90 dB pour d'autres, sensibilité résiduelle en basse fréquence pour LING [LING - 64], incapacité à percevoir de petites variations en fréquence pour MARTONY [MARTO - 74] ou RISBERG [RISB - 75], etc.

D'après ERBER [ERBE - 72], pour l'anglais, ce sont les tests concernant le pouvoir de discrimination en fréquence (tests de différence de niveau de fréquence, reconnaissance de mots) qui permettent le mieux de caractériser les sourds profonds. Les différences se manifestent notamment dans la difficulté d'identifier les consonnes plosives et nasales, celle de reconnaître les voyelles autrement que par leur durée relative, leur intensité et leur "dureté" (roughness),

celle de percevoir le nombre de syllabes dans les mots contenant des consonnes voisées continues aux frontières de syllabes (difficile également avec les déficients auditifs sévères). En résumé, les sourds profonds sont ceux qui éprouvent les plus grandes difficultés à distinguer les paramètres fréquentiels et qui se servent pour la perception acoustique des variations grossières dans l'enveloppe du signal.

Ces observations conduisent à penser maintenant que les sourds profonds n'entendent pas vraiment les sons mais perçoivent les sollicitations acoustiques de façon vibrotactile au niveau de l'oreille externe ou moyenne. (Il est à noter au passage que la sensibilité de la peau caractérisée par une transmission passe-bande avec sensibilité maximale autour de 250 à 300 Hz concorde avec celle de l'oreille à audiogramme en pente à chute plus ou moins brusque vers les aigus). Nous envisagerons plus loin (A.2.II) la question des aides à la perception par lecture labiale et par des moyens tactiles.

En plus du degré de déficience auditive moyen, il convient de faire la distinction entre les types de résultats audiométriques (audiométrie par sons purs - audiogrammes phonétiques), d'une part, et entre les surdités pré et post verbales, d'autre part.

Parmi les sujets atteints depuis la naissance ou dans la période prélinguale, seul un très faible pourcentage acquiert une parole naturelle. Selon NICKERSON [ERBE - 74], rares sont ceux qui lisent correctement et, contrairement à une idée répandue, l'aptitude à la lecture labiale est faible.

Dans le cas de surdité acquise après l'apprentissage du langage, l'âge auquel elle survient est déterminant pour son évolution future : à l'âge de trois ans environ, le langage disparaît rapidement et la situation est celle du sourd prélingual. (Des observations réalisées chez de très jeunes enfants par J. VAN DER STELT aux Pays-Bas [STEL - 76] permettent de penser cependant que certaines aptitudes ont déjà été acquises à partir de la seconde année en ce qui concerne les mouvements "diadochocinétiques" des lèvres et de la langue : plus le développement normal a été long, plus les muscles ont été entraînés à un modèle rythmique et meilleure est la mémoire kinesthésique. Pourtant cette mémoire du rythme se perd avec le temps). Les altérations notables qui surviennent après sept ou huit ans démontrent la fragilité de l'édifice linguistique qui a besoin de se renforcer au fur et à mesure des échanges verbaux. Le sourd postlingual perd l'appétence à la communication orale, en même temps que dans sa voix s'estompent les nuances au niveau suprasegmental : intensité et mélodie. Le flou de l'articulation se répercute sur les phonèmes, en particulier les voyelles qui tendent vers un son moyen, vérifiant la loi du *"minimum d'effort"*, proche du /ə/ d'hésitation.

A ce niveau, il est intéressant de se demander s'il est possible de faire une séparation totale entre aptitudes intellectuelles et aptitudes verbales, ou bien si le langage est *"L'instrument majeur de la pensée"* [BRUN - 66]. Malgré la diversité des opinions du passé, il semble acquis que l'apprentissage du langage n'est pas fortement lié aux capacités intellectuelles, de même que la performance intellectuelle ne dépend pas directement du langage [FURT - 73]. L'étude des performances d'adolescents sourds et normalement entendants dans une tâche de symbolisation logique conduit FURTH [FURT - 71] aux remarques suivantes : *"Comment les opérations formelles sont-elles liées au langage ? (...). Un langage*

social n'est ni une condition nécessaire, ni une condition suffisante du développement opératoire (...). Le langage, au mieux, a un effet indirect de facilitation sur les opérations concrètes. Pour certaines opérations formelles, le langage, associé à certains facteurs sociaux, peut avoir un effet direct de facilitation ; il peut fournir à la fois l'occasion et le support figuratif pour que fonctionnent les opérations ponctuelles".
JAMART [JAMA - 81], dans la recherche d'outils mathématiques pour une pédagogie appliquée à l'adolescent déficient auditif, propose d'aider à une meilleure structuration de la pensée et un plein épanouissement du raisonnement grâce à un outil logique : l'algorithmique, s'associant à l'idée qu'il faut essayer de développer l'intelligence pour aider le déficient auditif à maîtriser le langage et non pas développer le langage pour accéder à une certaine forme d'intelligence.

Il n'en est pas moins vrai que le manque de compétence sur le plan du langage est un handicap sérieux. Si, dans le meilleur des cas, il freine l'aptitude d'une personne à communiquer avec autrui, il peut aller, disent certains, jusqu'à empêcher la communication avec soi-même.

3. Le bilan audiométrique

Devant une présomption de déficience auditive au sens large, une série de tests et de mesures permet d'en déterminer l'importance et l'origine. Ils vont de l'interrogatoire sur les antécédents familiaux, le passé du sujet, jusqu'à l'examen somatique général, en passant par une série d'examens spécifiques. Le classement simplifié suivant est tiré de P. AIMARD [AIMA - 74] :

- anamnèse,
- bilan ORL,

- examens audiométriques :
  . audioélectroencéphalographie,
  . électrocochléogramme,
  . mesures par audiomètres à sons purs,
  . audiométrie verbale (test d'intégration phonétique de LAFON, acougramme phonétique de Mme BOREL-MAISONNY...),
- examen orthophonique s'intéressant au double aspect de la parole et du langage,
- examen psychologique dans ses aspects "visuospatiaux" et psychomoteurs,
- examen neuropsychiatrique,
- examen somatique général.

4. Expériences de perception par les enfants déficients auditifs

Diverses séries de tests peuvent être retenues pour juger de l'aptitude de sujets malentendants à percevoir la parole comme par exemple :

- des tests portant sur les seuils différentiels de fréquence, c'est-à-dire les écarts minimaux autour d'une fréquence donnée qui permettent la différenciation de deux sons, et conduisant à quantifier l'aptitude à la perception de l'intonation,

- des tests portant sur l'identification de phonèmes à partir de suites de paires minimales par exemple, renseignant sur la perception au niveau segmental,

- des tests sur les seuils différentiels d'intensité, etc.

DEGUCHI et KUROKI [DEGU - 76] rapportent par exemple une expérience réalisée avec un groupe d'enfants de 14 ans environ, présentant une déficience d'au moins 60 dB à 500 Hz. Le seuil différentiel moyen de ces enfants se situe autour de 40 Hz alors qu'il n'est que de quelques unités pour des enfants normalement entendants. Après un entraînement auditif à partir de notes jouées au piano, les enfants sont plus sensibles aux écarts de fréquence mais ne peuvent dire souvent quelle est la note la plus élevée ou la plus basse. Après un entraînement mixte avec adjonction d'une aide une fois par semaine durant un mois et demi, les enfants font la différence entre plus grave et plus aigu. Le seuil différentiel de fréquence a chuté vers 20 ou 30 Hz. Un seul enfant sur sept, pourtant, est finalement capable d'ajuster sa voix au niveau requis. Les tests de discrimination phonémique par ailleurs, grâce à des séries de syllabes du type consonne-voyelle, permettent de mettre en évidence les oppositions les mieux perçues. Ce sont le plus souvent les oppositions son voisé/son non voisé, fricative/plosive, nasale/fricative. Les difficultés de discrimination sont plus grandes par exemple à l'intérieur de l'ensemble des sons voisés ou des sons non voisés que entre ces ensembles.

5. Apprentissage. Lequel, quand et comment ?

Depuis que ce sujet est étudié, différentes écoles s'opposent sur la voie à suivre pour donner un langage aux "sourds-muets".

Alfred de MUSSET [MUSS - 1844] fait mention des résultats obtenus par l'abbé de l'Epée (1712-1789) grâce à une approche de la lecture et de l'écriture fondée sur un langage gestuel, à une époque où,

"Partout, même à Paris, au sein de la civilisation la plus avancée, les sourds-muets étaient regardés comme une espèce d'êtres à part, marqués du sceau de la colère céleste. Privés de la parole, on leur refusait la pensée". Il précise : *"C'est un moine espagnol qui, le premier, au seizième siècle, a deviné et essayé cette tâche, crue alors impossible, d'apprendre aux muets à parler sans parole. Son exemple avait été suivi en Italie, en Angleterre et en France, à différentes reprises (...) mais l'intention avait été meilleure que l'effet".*

Ce langage des signes, dont l'efficacité dans la communication est certaine et que parlent entre eux les enfants, présente le défaut d'être limité, d'être marginal, ce qui ne résoud pas le problème de l'intégration sociale du non-entendant, et surtout de ne pas respecter la structure syntaxique de la langue orale. Mais il présente le grand intérêt de favoriser très vite l'aptitude à communiquer, l'acquisition des connaissances et l'établissement de constructions mentales.

A ce sujet, citons l'opinion de MARTONY [MARTO - 74] qui mentionne l'intérêt de l'application au domaine des sourds des recherches sur le bilinguisme. Selon lui, on demande généralement aux sourds d'être bilingues, le langage par signes étant souvent appris avant que ne soient entrepris les efforts vers l'acquisition d'un langage verbal. La question se pose alors d'apprécier l'influence du langage par signes sur le langage oral, leurs interactions et les effets positifs ou négatifs qu'ils peuvent exercer l'un sur l'autre.

On pense à l'heure actuelle que la langue orale doit être enseignée chaque fois que cela est possible. Selon SCHULTE [SCHU - 70], l'aptitude à communiquer suppose l'accession à un niveau du langage qui ne demande pas un trop gros effort de compréhension à l'interlocuteur.

Pour le sourd, elle ne doit pas se limiter à la communication "*in a potential ghetto of the deaf, in a subculture*", mais elle doit favoriser son intégration dans un monde imprégné d'une langue dite maternelle. Cette aptitude ne peut se faire qu'en encourageant une approche orale de la langue. La qualité de l'apprentissage dépendra de la capacité de la méthode utilisée de stimuler et commander la production vocale.

Quel qu'en soit le mode, chacun s'accorde à dire qu'il est important de stimuler très tôt chez l'enfant le désir de communiquer en le plaçant le plus souvent possible en situation d'accumuler les expériences acoustiques et d'échanges avec l'entourage. Pour la langue orale, le principe est le suivant : c'est en communiquant que l'on apprend à parler. La perception joue ici un rôle très important ; en effet, même si les conditions optimales sont réunies (milieu familial et éducatif positif, enfant d'intelligence et de comportement "normaux", rééducation bien pensée et bien dirigée...), il apparaît que l'intégration de la spontanéité et des nuances du langage est affectée par la difficulté de leur perception. Un effort important doit être fait au niveau des échanges entre l'enfant et son environnement.

Dans la phase d'approche du langage oral, il semble judicieux d'adopter une voie intermédiaire entre l'expression spontanée non dirigée et une progression rationnelle très stricte, tant au niveau de l'articulation (au sens large) que des structures syntaxiques. Dans le cas de l'enfant jeune, en particulier, il est certainement préférable de le laisser approcher naturellement la langue dans son ensemble de façon à l'encourager à s'exprimer, puis de l'amener à une approche analytique fondée sur ses essais spontanés. Nous portons notre attention sur cet aspect en D.2.II. Cette idée nous conduit à établir une hiérarchie dont les échelons ne correspondent pas à la seule maîtrise de positions et

mouvements articulatoires toujours plus évolués mais à une progression de tout le système de construction du langage, ce qui est beaucoup plus riche. Nous aurons l'occasion de souligner à nouveau que l'acquis d'une parole correcte ne trouve son sens que si elle devient le support d'un langage suffisant dans une construction parallèle des systèmes phonétique et linguistique.

Il est sûr qu'il appartient au rééducateur de déterminer quelle est la procédure à adopter en fonction de l'âge de l'enfant, ses capacités intellectuelles, son acquis, son degré de surdité et les difficultés qui lui sont propres.

Selon RISBERG [RISB - 68], l'apprentissage de la langue suit normalement trois étapes :

- une phase de prise de conscience grâce à la présentation des différents facteurs vocaux et articulatoires. C'est une étape très difficile pour les non-entendants,

- une phase de fixation où la connaissance préalable permet la répétition de formes acquises sans aide extérieure,

- une phase d'automation où la production vocale est totalement automatisée et conduit à une attitude communicative globale.

Dans la première phase, des aides venant suppléer ou compléter le sens de l'ouïe et apportant une contreréaction immédiate peuvent seconder l'enfant sourd à un stade où le petit entendant cherche à reproduire les sons qu'il perçoit. Dans les phases suivantes, il convient d'apporter une contreréaction retardée puis de tenter de la supprimer.

Nous aurons l'occasion de reprendre ces réflexions en conclusion d'une revue et d'une discussion des aides proposées au mal- ou non-entendant dans ses modes de communication divers avec le monde extérieur.

VI - TROUBLES DE LA PAROLE D'ORIGINE PHYSIQUE OU NEUROLOGIQUE

Nous mentionnons, pour clore ce chapitre concernant les problèmes de la communication parlée, l'incidence de certains troubles d'ordre physique ou neurologique sur l'émission de la parole et l'organisation du langage. Bien que ces troubles ne soient pas directement l'objet de notre propos, les efforts pour l'acquisition du langage peuvent trouver un support dans une forme adaptée des aides envisagées pour les non-entendants : aides à l'échange verbal et aides à la rééducation vocale dans lesquelles entre le système SIRENE que nous avons réalisé.

Signalons comme problème mineur les cas de raucité, fréquente chez l'enfant, sans origine pathologique, devant lesquels l'orthophoniste est souvent désarmé.

Dans les défauts d'origine physique n'affectant que la production de la parole, on peut classer :

- les malocclusions labiales qui touchent essentiellement l'émission des voyelles, des consonnes bilabiales (/p/, /b/, /m/) ou des labiodentales (/f/, /v/),

- les malformations buccofaciales qui, si elles peuvent être résolues de façon chirurgicale, nécessitent une rééducation orthophonique,

- la trachéotomie qui pose au patient un problème de communication proche de celui des sujets atteints de surdi-mutité,

- les défauts de prononciation dus à la langue ou aux dents,

- les insuffisances velaires qui provoquent une voix dite *"nasonnée"* par suite d'une fuite systématique d'air par le conduit nasal.

Dans les troubles de la parole d'origine neurologique entrent les différents types d'aphasie, acquise ou congénitale : surdités verbales, dysphasies, dyslexies qui trouvent un essai de résolution dans le travail de l'orthophoniste.

CHAPITRE 2

LES AIDES A LA COMMUNICATION

I - AIDES A LA COMMUNICATION EN GENERAL

Avec les progrès constants de la technologie se développe une gamme d'appareils jouant pour le malentendant un rôle de suppléance dans les circonstances de la vie où l'ouïe devrait intervenir : communication entre personnes, enseignement, acquisition de l'information, alarme, etc.

Le développement de ces appareillages se heurte à un certain nombre de difficultés suivant les pays :

- défaut d'une politique globale systématique en faveur des handicapés en général et en particulier des sourds, dont le handicap a une incidence mineure sur la société,

- importance limitée du marché qui peut freiner le développement commercial de nouveaux produits,

- frein dû à la technologie elle-même,

- connaissance encore incomplète du fonctionnement auditif central, des processus de compréhension du langage, etc.

Les aides mises à la disposition des déficients auditifs ou à l'étude peuvent se classer ainsi :

1 : alarmes visuelles ou vibratoires,

2 : communication à distance : les recherches s'orientent vers la création d'unités adaptables sur les lignes téléphoniques standard, le téléphone avec écran, etc., la définition de prothèses vocales à partir, par exemple, d'un tableau Bliss, d'un désigneur et d'une unité de synthèse [GRAI - 81],

3 enseignement et information :

. enseignement assisté par ordinateur essentiellement pour les langues ou les matières scientifiques,

. utilisation de récepteurs de télévision pour l'acquisition de l'information,

. adaptation possible à la transmission d'information de procédés (tels que Antiope (CCETT de Rennes)) de sous-titrage numérique des émissions télévisées pour les malentendants. Très récemment, LETELLIER [LETE - 83] a fait l'étude d'un terminal "télésigné" faisant appel aux techniques de compression d'information et de transmission d'images à bas débit,

4 perception des sons et compréhension de la parole :

. utilisation optimale de l'ouïe résiduelle : prothèses auditives,

. travaux sur la stimulation nerveuse à partir d'implants d'électrodes au niveau de la cochlée ou la stimulation directe des aires corticales correspondant à l'ouïe,

. apport d'information grâce à des méthodes gestuelles,

. à plus long terme, adaptation aux difficultés des sourds d'appareils utilisant la reconnaissance automatique de la parole et de machines à écrire à commande vocale,

5 acquisition du langage :

. en plus des projets et réalisations cités au paragraphe 4 précédent, appareillages divers destinés à assister la rééducation vocale : aides tactiles et visuelles en particulier. Nous reprenons ces deux derniers points dans la suite.

II - LES AIDES A LA PERCEPTION ET A LA COMPREHENSION DE LA PAROLE

Nous avons insisté plus haut sur l'importance de la communication orale dans les rapports entre les hommes. Le problème posé par les sourds présente un double aspect : celui de la production de la parole, et celui de sa compréhension. Dans ce paragraphe, nous abordons ce deuxième aspect qui a trait tout aussi bien à la perception des formes sonores produites par autrui qu'à celle des siennes propres.

La compréhension de la parole par l'auditeur handicapé suppose en gros deux conditions :

- que lui soit présentée une quantité d'informations suffisante sur les sons prononcés,

- qu'il ait acquis une bonne connaissance de la langue, lui permettant de lever les ambiguïtés du niveau précédent.

La deuxième condition n'est pas remplie par les enfants atteints de surdité de naissance ou acquise avant l'apprentissage du langage. Le lien est donc très serré entre les questions de production et de compréhension de la parole. C'est ainsi que des recherches comme celles que nous avons menées trouvent un prolongement, d'une certaine façon, dans la définition d'aides à la compréhension.

Le schéma ci-dessous (figure A.11) illustre les liaisons que l'on peut établir entre "locuteur" et "récepteur".

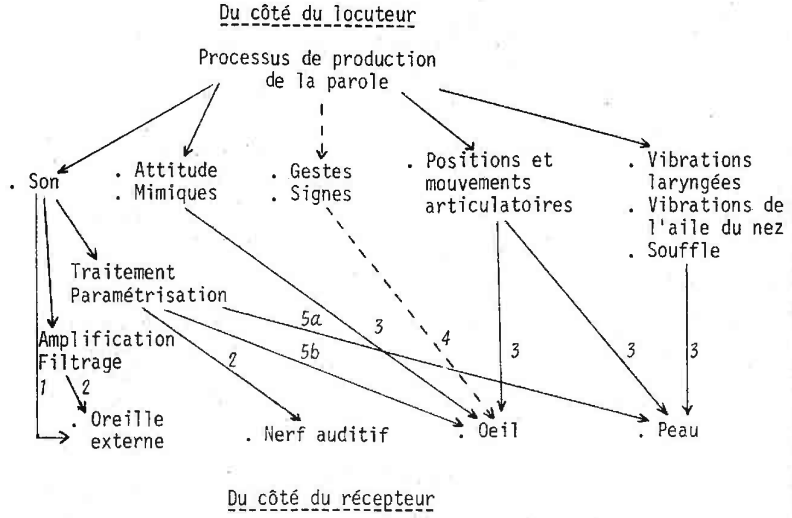


Fig. A-11 : Liaisons de communication orale du locuteur vers le récepteur

Nous avons indiqué sur la figure le numéro du sous-paragraphe qui suit où chacune de ces liaisons est plus particulièrement étudiée. La liaison notée 4, qui suppose une coopération et la connaissance d'un langage gestuel de la part du locuteur, est indiquée en pointillé.

1. Education auditive

Le diagnostic précoce des surdités actuellement pratiqué permet une éducation de l'enfant sur le plan auditif de très bonne heure. Il est fondamental de former l'enfant à utiliser ses restes auditifs de façon à ne négliger aucune chance de développer son attitude communicative orale. Ceci est réalisé au mieux lors de la phase d'éducation précoce qui précède l'intégration de l'enfant dans le milieu scolaire.

2. Appareillage des surdités

En ce qui concerne l'appareillage des surdités, nous distinguons trois catégories :

a) les prothèses auditives, à l'heure actuelle de plus en plus évoluées, qui se distinguent essentiellement par les caractéristiques suivantes : présentation, mode d'application, puissance et caractéristiques de sortie. Ces dernières ont trait à la compression de dynamique du signal d'entrée et plus généralement à la réponse en fréquence,

b) les amplificateurs avec écouteurs à contreréaction immédiate ou retardée associés le plus souvent à une indication visuelle du niveau sonore et utilisés en situation d'apprentissage pour aider l'enfant à s'écouter et à parler. On distingue ici aussi les amplificateurs à courbe de réponse linéaire ou non linéaire. Parmi ces derniers, citons le SUVAG II qui permet d'entraîner l'enfant à l'écoute d'un signal formé de trois composantes correspondant à un canal direct et à deux canaux "graves" et "aigus" ajustables et répondant aux principes de la méthode verbo-tonale du Pr. GUBERINA (Zagreb),

c) Les appareils transposeurs de fréquences, dont l'idée revient à PIMONOW qui déplacent vers le bas du spectre les fréquences du domaine de la parole. Cette idée repose sur le fait que, dans beaucoup de déficiences auditives, particulièrement celles d'origine neurologiques, l'ouïe est meilleure pour les sons graves que pour les sons aigus. On trouve des appareils qui transposent en bloc toutes les fréquences, d'autres qui transposent seulement une partie sans toucher aux autres. Parmi les premiers, citons le *"DIFA"* de LAMOTTE et VIGNERON [LAMO - 75] qui permet de comprimer en bloc vers le bas la zone des fréquences normalement audibles avec un coefficient de division réglable, ce qui présente l'intérêt de ne pas détruire les rapports entre fréquence de formants et de ne pas supprimer les composantes basse fréquence (comme le ferait une simple translation, par exemple).
Parmi les autres, on rencontre le transposeur de JOHANSSON [JOHA - 66] où les hautes fréquences sont divisées et superposées au spectre non filtré ; des divisions de même type sont exploitées pour la mise en évidence des fricatives [GUTT - 70], pour l'apport de sons de compensation (LAFON), ou pour la superposition à la voie directe de bruits colorés suivant la zone de concentration fréquentielle de l'énergie (système PARME : [LORA - 75]). Selon PICKETT, leur difficulté d'utilisation réside dans le fait que les formes perçues diffèrent de la parole originale. Souvent les transposeurs sont utilisés en rééducation orthophonique pour les premières prises de conscience par l'enfant de son environnement sonore. Dans le système de DRUCKER et al. [DRUC - 77], la parole est scindée en basses et hautes fréquences puis équilibrée de façon à faire ressortir les consonnes. Le traitement se fait sur microprocesseur. Un système plus flexible est étudié par STEARNS et al. [STEA - 77], dans lequel la fréquence centrale d'un filtre passe-haut peut varier de façon à obtenir l'intelligibilité optimale par le sujet.

Citons une voie nouvelle de recherche dans les cas où l'oreille interne est atteinte mais où le nerf auditif est intact : les implants cochléaires. On peut trouver une importante bibliographie sur le sujet dans [TONG - 83] ; de tels implants supposent :

- une stratégie de traitement et de codage du signal de parole [FARD - 81],
- des expériences de microchirurgie de grande précision [CHOU - 81].

Bien que les premières expériences aient été accueillies avec prudence, car le message perçu par le sujet était bien éloigné du stimulus d'origine, il est certain qu'elles eurent une action psychologique très positive, en particulier dans les cas de perte brutale de l'ouïe. Selon CHOUARD [CHOU - 78], chirurgien français et précurseur dans le domaine, il s'agit d'une première étape vers la construction d'une "oreille artificielle".

3. Les techniques classiques en orthophonie

Nous n'entrons pas dans le détail du travail de l'orthophoniste et des techniques classiques employées pour aider l'élève à percevoir la parole.

Trois catégories peuvent se distinguer :

- utilisation optimale de l'ouïe résiduelle, mentionnée ci-dessus en 1),

- observation des lèvres, du visage et des attitudes de l'interlocuteur, ce que l'on peut englober sous le terme récemment proposé de *faciolecture* ou lecture faciale (*"speech reading"*),

- prise de conscience de sensations tactiles, comme les vibrations de la glotte ou des ailes du nez, le souffle, les mouvements articulatoires.

Ces différents points peuvent être envisagés de façon parallèle dans l'aide à la production de la parole.

Dans la situation de conversation normale, la lecture faciale demeure pour le déficient auditif le moyen le plus direct d'accéder aux phrases prononcées par son interlocuteur. Elle n'apporte pourtant qu'un coefficient limité d'intelligibilité du fait surtout de la difficulté qu'il y a de faire la discrimination entre les sons de même lieu d'articulation. De plus, les caractéristiques idéales des phonèmes étant modifiées par les phonèmes environnants, il en est de même des images labiales : DISSOUBRAY (I.N.J.S., Paris) cite en exemple le /i/ de "*hôpitaux*" où les lèvres sont arrondies par la proximité des deux /o/ , alors que pour le /i/ de "*épithète*", les lèvres sont allongées sous l'influence de /e/ et /ɛ/ .

Le tableau A-12 suivant indique les six confusions majeures de consonnes en lecture labiale :

1	2	3	4	5	6
p	t	k	f	s	ʃ
b	d	g	v	z	ʒ
m	n	r			

Fig. A-12 : Confusions de consonnes en lecture labiale

Les catégories 2 et 5 correspondant à des consonnes dentales peuvent également être confondues si elles ne sont pas différenciées sur la durée.

La lecture faciale suppose de plus l'enchaînement à la cadence de l'élocution de trois performances : la perception de facteurs liées à la parole, l'identification de formes élémentaires et l'association de ces formes à des concepts signifiants.

Des expériences concernant l'aptitude à la lecture faciale par des malentendants [MARTO - 74] ont permis de dégager que, si l'effet du degré de déficience auditive se fait très nettement sentir dans la perception de stimuli mixtes (image et écoute simultanées), il n'est pas corrélé à l'aptitude à la lecture visuelle seule. Le fait plus général qu'il n'y ait pas de norme dans les différences "interlocuteurs" est une des grandes difficultés rencontrées dans l'élaboration des tests de perception. M. GENTIL [GENT - 80], à partir de tests complets, confirme l'insuffisance de la perception visuelle seule, du fait de l'existence de sosies labiaux, des phénomènes de coarticulation et de la variabilité interlocuteurs et met en évidence que les signes visuels sont peu redondants par rapport aux signes auditifs. Il n'existe pas, d'autre part, de corrélation entre la fréquence d'apparition d'un phonème et son pourcentage de reconnaissance. C'est plutôt la facilité de la réalisation du son (sourd plutôt que sonore, par exemple) qui pourra orienter et favoriser la compréhension (ce qui pourrait être en faveur de la théorie motrice de la perception).

Nous envisageons maintenant les méthodes destinées à faciliter la lecture faciale : les codes gestuels et leur extension automatique possible sous forme d'aides tactiles ou visuelles.

4. Les codes gestuels

Une perspective d'avenir se dégage à partir des techniques destinées à faciliter la lecture faciale. La nécessité de proposer au sujet *"récepteur"* des signes de complément lui permettant de distinguer les sons de même image labiale est à l'origine de codes manuels tels que :

- le *"Phonemetransmitting Manual System"* proposé par SCHULTE à Heidelberg [SCHU - 72], directement lié aux positions articulatoires : les mouvements de la main du locuteur renseignant sur lieu et mode de production. L'information véhiculée est donc d'ordre phonologique,

- le système de *"Cued Speech"* développé par CORNETT [CORN - 67], [LING - 75] dans lequel les signes manuels interviennent pour signaler les sons les plus couramment confondus ou non perçus, mais ne sont pas en relation directe avec les positions et mouvements articulatoires. Une traduction de l'ouvrage de CORNETT a été récemment mise au point à l'école d'orthophonie de Nancy [SCHI - 80] avec adaptation à la langue française.
Le *"Cued Speech"* est considéré par les auteurs comme une formule mixte se situant entre le langage signé et l'oralisme, respectant la structure linguistique. Pour un père, cet accès au langage va beaucoup plus loin : *"The gift of language is the gift of culture - with its values, attitudes, ideas and beliefs"* [Cued Speech News, vol XV, n°4, Dec. 1982],

- le système AKA,*"Alphabet des Kinèmes Assistés"* [WOUT - 76], qui répond aux mêmes motivations mais dans lequel on donne du phonème une représentation "physioplastique", c'est-à-dire liée à la façon de le prononcer, comme dans le système de SCHULTE. Le tableau A-ii vu précédemment fournissant les confusions de consonnes, donne précisément, lu horizontalement, les groupes de trois phonèmes assistés par le même geste de

la main. De cette façon, le soutien gestuel à la lecture labiale est doublé d'un élément stimulant pour la production vocale. On retrouve à ce niveau la référence à la théorie motrice de la production de la parole et au mot de LIBERMAN : *"Speech is perceived by reference to articulation"*. Les auteurs se sont de plus attachés à atténuer les problèmes posés par la coarticulation en prenant la syllabe comme élément de base.

A ces trois codes, qui font de la langue maternelle une langue signée, on peut opposer la langue des signes (voir, en France, la revue "Coup d'Oeil") dont nous avons fait mention en A.1.V et qui est souvent la langue primaire du sourd, la langue orale de la majorité linguistique étant leur seconde langue. En fait, l'utilisation de la langue des signes ne doit pas être considérée *"comme une solution de facilité conduisant le professeur ou l'éducateur à abaisser le niveau de ses exigences mais, au contraire, comme un moyen supplémentaire permettant de clarifier et d'accélérer la transmission des connaissances et, par conséquent, augmenter le savoir des élèves"* (Circulaire du Ministère de la Santé et de la Sécurité Sociale aux enseignants des I.N.J.S. du 8 juin 1977). La langue gestuelle est virtuellement incluse dans ce que l'on appelle maintenant la communication gestuelle. Cette dernière *"est conçue comme un moyen pour faciliter la perception des signes de la communication et permettre déjà ainsi un meilleur développement affectif et cognitif de l'enfant"* [BOUR - 83].

Une solution alternative peut être trouvée dans l'utilisation du langage signé. Le langage défini à partir des travaux de Madame BOREL-MAISONNY [BORE - 60] et approfondi par un groupe de spécialistes du C.H.R. de Nancy auquel appartient M.M. DUTEL, associée à nos travaux [DUTE - 79], propose l'association son-geste-"chérèmes" (traits distinctifs évoqués au niveau de la main).

Ces systèmes de communication supposent la participation volontaire et l'entraînement de l'interlocuteur, ce qui limite leur portée. L'idéal est de concevoir une aide qui apporte les informations équivalentes de façon automatique en faisant appel à un sens intact chez le sujet handicapé auditif (malgré une théorie répandue qui maintient que la parole est une sorte de code attaché de façon unique et biologique au système auditif), soit la vue, soit le toucher. Le paragraphe suivant traite de ces orientations nouvelles.

5. Les orientations nouvelles

Un prototype connu sous le nom des *"Lunettes d'UPTON"*, présenté en 1968 [UPTO - 68] par UPTON, ingénieur devenu sourd, comprend un petit analyseur qui extrait quelques paramètres de la parole et les visualise par des points lumineux à la surface de l'un des verres. Cette idée est le point de départ de recherches concernant la détection et l'indication automatique de l'information non visible sur le visage de l'interlocuteur, sorte de *"Cued Speech"* automatique dont le besoin et l'intérêt sont énormes. Deux aspects importants sont à signaler, en particulier :

- la nécessité de présenter des renseignements suffisamment complets au rythme de la parole continue sous forme visuelle ou tactile (l'extraction de paramètres en temps réel pose des problèmes que l'on retrouve en reconnaissance automatique de la parole),

- l'orientation possible vers une miniaturisation qui placerait le sujet dans une situation normale de communication.

On peut imaginer, en dernière étape, l'association d'une telle unité d'analyse à une unité de reconstruction du signal liée aux implants d'électrodes qui constituerait une *"prothèse de l'oreille"*, comprenant de

véritables calculateurs, peut-être eux-mêmes implantés...

a) Les aides tactiles

Ce problème, étudié depuis plus de cinquante ans, a été discuté par KIRMAN [KIRM - 73] ou ERBER [ERBE - 77] en particulier. Il est admis généralement que la sensibilité de la peau convient mieux au recodage des images qu'à celui de la parole. Cependant, selon VON BEKESY [BEKE - 59], il est possible d'établir un parallèle entre l'oreille et la peau du point de vue perceptif. Ce fait est noté également par PICKETT [PICK - 63] qui précise qu'une certaine similitude se rencontre dans l'aptitude à percevoir le rythme et l'évolution de paramètres dans le temps. Du point de vue fréquentiel, cependant, la discrimination pour les fréquences élevées est très limitée. D'autres auteurs, comme GOFF [GOFF - 67], signalent même une détérioration rapide au-dessus de 200 Hz et une coupure à 1 000 Hz alors que l'information linguistique pertinente se situe entre 200 et 4 000 Hz environ. D'autre part, des mesures de la résolution temporelle de la peau montrent qu'elle est bien inférieure à celle de l'oreille. Par exemple [GESC - 70], l'intervalle de temps minimal qui permettra de résoudre deux impulsions vaut un peu moins de 2 ms pour l'oreille et au mieux 10 ms pour le bout des doigts.

En fait, la peau est plus sensible à l'effet d'intensité qu'à l'effet de durée. A ces constatations s'ajoutent d'autres facteurs qui conditionnent la réalisation d'aides tactiles, comme les effets d'adaptation ou les effets de masquage simultanés (par exemple, masquages "ipsolatéral" et "contralatéral" au bout des doigts) [HAME - 83].

Ces observations ont conduit les chercheurs à étudier quel pouvait être le meilleur codage de la parole pour sa présentation sous forme tactile. Suivant les réalisations, l'information apportée concerne

la distribution spectrale, l'intensité, la fréquence fondamentale (en liaison ou non avec l'intensité), la durée... Un certain nombre de chercheurs utilisent le codage par *"vocodeur"*. L'idée de base est de moduler l'amplitude d'un jeu de vibrateurs par la sortie des canaux de l'analyseur spectral. Les fréquences des vibrateurs sont choisies dans la gamme des fréquences les mieux perçues par la peau. Les stimuli sont appliqués sur la jambe parfois, sur le poignet ou les doigts de la main le plus souvent. Citons, après 1950, l'appareil de LOVGREN et NYKVIST constitué de dix vibrateurs à conduction osseuse et les expérimentations faites par PICKETT [PICK - 63] ; le *"Vocotac"* de WITCHER et al. [WITC - 56] dans lequel les sorties de six filtres commandent une matrice à six points de type BRAILLE ; l'appareil de KRINGLEBOTN [KRIN - 63] : *"Tactus"* qui transpose les fréquences du signal de parole dans la zone 0 - 800 Hz et qui sert aussi d'aide à la production et d'indicateur de rythme. La plupart de ces auteurs jugent l'expérience encourageante dans l'identification de quelques mots, mais concluent à l'impossibilité quasi certaine de comprendre le discours continu.

Parmi les travaux les plus récents dans le domaine fréquentiel, citons l'appareil de MILLER [MILLE - 76] dérivé des travaux de PICKETT, celui de SAUNDERS [SAUN - 76] qui a imaginé une transmission électrocutanée des signaux pour les jeunes enfants, celui de ENGELMANN et ROSOV [ENGE - 75] qui en font une présentation mécanique. GOLDSTEIN et STARK [GOLDS - 76] proposent des diagrammes amplitude - fréquence variant dans le temps sous forme d'une surface vibrante évoluant sous les doigts. Ils utilisent pour cela une adaptation de la matrice vibratoire de l' *"Optacon"* conçu à l'origine comme une aide à la lecture pour les aveugles. Des données similaires (intensités quantifiées dans 36 zones de fréquence) sont présentées sous la forme d'une matrice de 288 électrodes délivrant un courant constant par l'intermédiaire d'une ceinture abdominale [SPAR - 79] dans le dispositif appelé *"MESA"*, *"Multipoint Electrotactile Speech Aid"*.

Citons également le dispositif décrit dans [SPEN - 81] ou le *"Fonator System"* de SCHULTE [SCHU - 81].

Les chercheurs semblent s'orienter depuis quelques années vers la création d'aides nouvelles où l'information apportée n'est plus uniquement de nature spectrale. EDMONDSON [EDMO - 77] présente, par exemple, la distribution spectrale associée à l'énergie totale. Le *"Single Vibrator Rythm Indicator"* de BOOTHROYD [BOOT - 72] ou le *"Multivibrator Pitch Indicator"* de WILLEMAIN et LEE [WILL - 72] transmettent les variations de la fréquence fondamentale dans le temps. ROTHENBERG et al. [ROTH - 77], considérant qu'il est possible de distinguer dix paliers entre 10 et 100 Hz, présentent la valeur du fondamental dans cette zone. Très récemment et reprenant cette idée [PLAN - 83], un équipement à un seul canal propose la transmission d'une tension alternative de fréquence proportionnelle à Fo et modulée en amplitude par l'intensité dans la zone 68-1 500 Hz. Le vibrateur est celui du *"Sentiphone"* [TRAU - 80], dispositif tenu dans la main renseignant sur spectre, intensité et durée. L'hypothèse de ROTHENBERG et al. semble vérifiée mais les effets de l'assistance à la lecture labiale se font mieux sentir pour une parole à débit plus lent que la normale. Le problème se situe maintenant dans la recherche du meilleur parti à tirer du pouvoir de résolution temporelle de la peau. ERBER, quant à lui, mentionne des tests de perception réalisés à partir de la présentation de l'enveloppe de l'onde de parole en particulier [ERBE - 77].

Il est maintenant admis que les aides tactiles conviennent mieux pour les paramètres prosodiques que les aides visuelles, sans doute du fait de la nature de la sensibilité de la peau [MARTO - 74]. Les problèmes à résoudre pour une efficacité optimale sont essentiellement d'ordre pratique : taille, nombre et points d'application des vibrateurs, utilisation de matrices ou non, mode de présentation de

l'information. (Ces derniers points sont discutés dans [SHER - 77], en particulier). Quoi qu'il en soit, ces aides requièrent une période d'entraînement assez longue et ne peuvent être efficaces qu'en complément de la lecture labiale [MILLE - 76] grâce, en particulier, à l'apport d'information sur le mode d'articulation des consonnes les plus fréquentes par l'intermédiaire surtout de la friction et du voisement. Des expériences de reconnaissance de mots sans lecture faciale sont décrites dans [BROO - 83].

Une aide tactile efficace présenterait de multiples avantages : mise en situation réaliste, information sur sa propre production vocale (l'utilisateur a l'impression d'avoir "une voix") comme sur celle de l'interlocuteur, possibilité simultanée de lecture faciale...

b) Les aides visuelles

Bien que l'idée de "Parole visible" ne soit pas nouvelle, les premières réalisations d'aides visuelles à la compréhension de la parole sont assez récentes.

Nous avons fait mention plus haut de la prothèse imaginée par UPTON pour présenter en temps réel quelques paramètres tirés aussi bien de la voix de l'utilisateur que de celle de ses interlocuteurs. Les recherches actuelles portent sur le choix de la paramétrisation optimale du signal.

Citons comme premières réalisations :

. le prototype de TRAUMULLER [TRAU - 74] comprenant dix lampes diodes disposées en arc autour de la bouche du locuteur (en association avec une aide vibrotactile) et donnant une indication sur la position du centre de gravité du spectre et sur la nasalité,

. le dispositif de GOLDBERG [GOLDB - 72] qui propose de présenter visuellement des paramètres liés aux positions articulatoires. Le système fait la distinction voyelle-consonne et fournit une information sur le lieu d'articulation et la durée des phonèmes (cette dernière renseignant, dans une certaine mesure, sur le mode d'articulation) mais ne fonctionne pas en temps réel.

On peut faire entrer dans cette catégorie le système de transcription visuelle de GUEGUEN [GUEG - 72] dans lequel une analyse par "vocodeur" est suivie d'une réduction de données par analyse factorielle des correspondances. Un mot est alors représenté par une trajectoire dans le plan des premiers facteurs obtenus par apprentissage.

CORNETT et al. [CORN - 77] ont repris un système optique analogue à celui des lunettes d'UPTON dans la définition de l'*"Autocueur"*, version automatique du *"Cued Speech"* [CORN - 67]. Des clés ("cues") dérivées du signal doivent permettre la classification des éléments de parole en groupe de phonèmes. Deux afficheurs 7-segments proposent un arrangement spatial de ces clés, après un retard dû au traitement de quelques dizièmes de seconde. De même que pour la lecture faciale avec *"Cued Speech"*, il importe pour son entraînement que le sujet soit exposé de façon intensive au dispositif.

Au Centre Scientifique IBM-France, des travaux voisins sont en cours à partir du signal numérisé pour la recherche de clés ("keys"), non plus fondées sur le calcul direct de paramètres spectraux mais sur l'optimisation du taux de clés correctement fournies par le système d'analyse [DIBE - 82] avec également la contrainte de la visualisation en temps réel.

Un projet récent, appelé VIDVOX (U.S.A.), a pour ambition de proposer une transcription du message acoustique sous forme d'un message phonétique défilant devant les yeux de la personne sourde à la cadence de l'élocution. D'après la connaissance que nous avons des difficultés d'une telle réalisation (liées aux propriétés déjà évoquées de variabilité du signal vocal), le problème de l'étiquetage phonétique "en ligne" et avec un taux d'erreurs limité est loin d'être résolu. Une réflexion sur les aspects humains de ce projet est menée par LEVITT et sur les aspects informatiques par LEA.

Enfin, signalons la mise au point d'un synchroniseur parole-texte [SARG - 82] pour l'entraînement à la lecture labiale au rythme de l'élocution qui suppose l'action conjuguée de trois modules : division du texte en syllabes, segmentation du signal vocal en noyaux syllabiques et synchronisation.

III - LES AIDES A LA PRODUCTION DE LA PAROLE

1. Introduction

Les techniques et les appareillages divers décrits dans les paragraphes précédents peuvent participer, dans une certaine mesure, aux efforts entrepris pour faciliter l'acquisition des sons et du langage, tout gain au niveau de la perception se répercutant au niveau de la production de la parole. Nous passons ici en revue les aides à la rééducation vocale faisant appel au sens de la vue pour suppléer ou compléter celui de l'ouïe, développées et testées dans le monde à notre connaissance.

L'idée des aides visuelles n'est pas nouvelle et doit revenir à BELL qui fut initialement enseignant de jeunes sourds et dont un des objectifs était *"la parole visible"* (*Visible Speech*). PICKETT laisse entendre que la mise au point du téléphone en 1876 a été favorisée par l'utilisation par BELL de visualisations d'ondes de parole. Plus près de nous, HUDGINS, en 1935, suggérait que, pour des pertes auditives moyennes supérieures à 80 dB dans le domaine de la parole, les signaux acoustiques soient accompagnés de signaux visuels ou tactiles. On peut dire que l'ère de la *"parole visible"* a débuté autour des années 1940.

En 1938, COYNE remporte un premier succès avec le *"Fo-indicator"*, indicateur de fréquence fondamentale, utilisé dans certaines écoles d'Afrique du Sud. Vers 1945, LORENTZEN et NIELSEN définissent des indicateurs de fricatives et d'intensité. En 1946, aux laboratoires Bell, STEINBERG et FRENCH [STEI - 46] proposent une des versions du *"Visible Speech Translator"*, la première visualisation du

spectre. Dix ans plus tard apparaît à Stockholm le prototype de "*Lucia*", indicateur visuel de spectre, développé sous la conduite de FANT.

Depuis, les aides se sont multipliées et, bien que ce ne soit pas général, ont tendance à s'intégrer dans les programmes des écoles spécialisées, leur but étant d'apporter au sujet une contreréaction sur sa propre production vocale dans l'espoir qu'il sera un jour capable de s'en affranchir.

Nous allons plus précisément faire la revue des dispositifs d'aide visuelle en les classant, parfois arbitrairement, en fonction du type de paramètres qu'ils tentent de rééduquer.

2. Les aides visuelles

a) Fréquence fondamentale et contour mélodique

Dès le début du siècle, on a cherché à mettre en évidence les vibrations laryngées. Le "*kymographe*" (ROUSSELOT) donnait ainsi les variations de pression à différents niveaux : nez, bouche, larynx... Pour l'évolution de tels appareils, on peut se reporter à METTAS [METT - 71].

Les techniques actuellement utilisées pour la détection de la fréquence fondamentale de la voix se distinguent par le type de capteur utilisé et les techniques de traitement (analogique et/ou numérique). Le signal électrique est capté soit au niveau du larynx ou de la trachée grâce à un accéléromètre, soit au niveau des lèvres à l'aide d'un microphone classique. L'intervalle de temps qui sépare deux vibrations des cordes vocales est ensuite déterminé grâce à différentes méthodes de filtrage ou de calcul que nous précisons dans le chapitre B.1 consacré à l'analyse du signal de parole.

Différents appareils proposent une indication instantanée de la fréquence fondamentale par déplacement d'une aiguille devant un cadran ou par illumination d'une colonne de lampes diodes [BOOT - 76].

Dans l'appareil de HANSEN [HANS - 68], la hauteur est indiquée grâce à un système de stroboscopie optique.

MARTONY et RISBERG [MARTO - 68], [MARTO - 70] se servent d'un microphone de contact appliqué sur le larynx qui fournit un signal proportionnel à la fréquence du son laryngé. Au compteur gradué en fréquences sont associés des signaux lumineux qui indiquent la déviation au-delà d'une zone jugée correcte et un signal sonore de même fréquence F_0 . MARTONY fait part des succès enregistrés dans la correction des variations anormales d'une voyelle à l'autre et dans les essais pour abaisser le fondamental naturel jugé trop haut. A cette époque pourtant, il fait remarquer que les progrès ne se sont pas transposés à la parole spontanée.

Le contour mélodique, traduisant les variations de la fréquence fondamentale dans le temps, a souvent été présenté sous diverses formes. Ainsi, le "*glottographe*" (ou "*laryngographe*"), mis au point à l'origine par FABRE [FABR - 57] est fondé sur la détection des variations d'impédance électrique résultant du mouvement des cordes vocales par des électrodes placées au niveau du larynx ou de la trachée [FOUR - 76]. De nombreux appareils dérivés sont utilisés dans les laboratoires de phonétique (on verra, à ce sujet, l'article de E. L'HOTE [LHOT - 72]).

Le "*mélographe*" MGP 2 du C.N.E.T. à Lannion utilisait un capteur de même type. CHEVRIE-MULLER et DECANTE [CHEV - 73] rapportent une expérience réalisée à l'INSERM avec une version de cet appareil. L'utilisation à Fribourg d'un détecteur de mélodie voisin est décrite par C. HOLM [HOLM - 72].

Le "*pitchmeter*" de G. FANT filtrait le signal oral de façon à obtenir des sinusoïdes de fréquence égale à F_0 . Les courbes de variation pouvaient alors être enregistrées sur un "*mingographe*". Des dérivés du "*pitchmeter*" ont été construits. Quelquefois, un laryngographe lui est associé, comme dans les expériences de VALLENCIEN [VALL - 73].

A l'Institut de Phonétique de Grenoble, BOE et RAKOTOFIRINGA [BOE - 71] ont mis au point un détecteur de fondamental par amplification, filtrage et détection au fréquencemètre puis tracé sur mingographe. Une version de cet appareil, connectée à un oscillographe cathodique et testée à l'Institut National des Jeunes Sourds (INJS) de Paris, a notamment servi de support à une étude objective de la voix des déficients auditifs par G. VOIRON [VOIR - 74].

On rencontre aussi des appareils [SONE - 60] mesurant l'intensité du rayonnement émis par une source lumineuse extérieure par une cellule photoélectrique placée derrière l'épiglotte. Les variations d'intensité lumineuse traduisent le taux d'écartement des cordes vocales.

La visualisation du contour mélodique sur écran a pris de l'extension avec le développement des aides visuelles à la production vocale. Dans ce domaine, elle est utilisée pour les problèmes de commande du pitch et aussi de rythme et de durée des sons. Citons dans ce domaine les travaux de BORRILD [BORR - 68], de LEVITT [LEVI - 73] ou de WOOD [WOOD - 71] qui effectue pour sa part une détection à partir du cepstre.

WATANABE et OKAMURA [WATA - 76] rapportent des expériences, réalisées au Japon, d'imitation de contours mélodiques par des enfants sourds. Les effets de l'apprentissage sont évalués par analyse numérique et tests subjectifs effectués sur les productions vocales des enfants.

Les conclusions sont plutôt optimistes ; ils pensent cependant qu'il conviendrait de corriger en même temps d'autres facteurs : le rythme et l'articulation, par exemple.

Les chercheurs du Département de Phonétique et Linguistique de l'"*University College London*" [FOUR - 76] utilisent le "*laryngographe*" de FABRE avec capteur sur le cartilage thyroïde (commercialisé par ailleurs par une firme anglaise) et jugent que voisement et rythme sont les premiers paramètres à travailler. Après expérimentation, les auteurs constatent que les locuteurs adultes entraînés sont capables de généraliser les formes prosodiques apprises à des phrases nouvelles, ce qui suppose que l'association entre règles syntaxiques et formes intonatives a effectivement été réalisée. Ils jugent que cette aptitude est maintenue à long terme.

Citons également trois appareils ayant atteint le stade de la commercialisation : le "*VAST*" (*Visually Assisted Speech Trainer*) fonctionnant sur écran de télévision et commercialisé par une firme du Michigan, U.S.A., le "*Visi-Pitch*" de la Kay Elemetrics Corporation, et un appareil dit "*Visual Display for Training Speech Prosody*" par une firme de Holte au Danemark, en collaboration avec P. MARTIN ("*Pitch Computer*"). L'appareil de P. MARTIN, utilisant un microprocesseur, fait la détection par décision logique après filtrage du signal oral par quatre filtres passe-bas.

Le paramètre "fréquence fondamentale" est aussi couramment associé à d'autres paramètres, en particulier l'intensité vocale, comme dans l'indicateur de SUGIMOTO et HIKI perfectionné par PORTER (et utilisé à l'Ecole d'Etat pour Sourds de Stockholm) où l'amplitude de la voix apparaît en surimpression sur le contour mélodique. De plus, la fréquence laryngée peut être divisée par un facteur constant et renvoyée sous forme de signal sonore dans des écouteurs portés par l'enfant.

Dans le *"FLORIDA" (Frequency Lowering Or Raising Intensity Determining Apparatus)* décrit par HOLBROOK [HOLB - 71], la contreréaction visuelle est apportée par deux lampes : l'une blanche qui s'allume si F_0 appartient à une plage acceptable, l'autre rouge si l'intensité est jugée trop élevée. L'appareil a d'abord été utilisé pour modifier la hauteur et le niveau de la voix chez des sujets sourds, adolescents et adultes, pendant 6 à 7 semaines. HOLBROOK rapporte qu'il a obtenu une chute de F_0 d'au moins une octave chez quatre adultes et que, trois mois après, deux d'entre eux étaient restés à ce niveau inférieur, les deux autres s'étant stabilisés à une valeur intermédiaire.

PICKETT et CONSTAM [PICK - 68] décrivent un appareillage en service au Gallaudet College de Washington, DC, U.S.A., connu sous le nom de *"Gallaudet Visual Speech Trainer"*, donnant les indications de différents contours dans le temps dont le contour mélodique. La trace peut être conservée puis effacée. La fréquence fondamentale est là aussi divisée de façon à entrer dans une zone de fréquence inférieure à 200 Hz.

Un système utilisant un calculateur, le *VSTA (Visual Speech Training Aid)*, décrit par STEWART et al. [STEW - 76], associe au contour du fondamental l'indication de la fréquence centre de gravité de l'énergie dans les hautes fréquences pour les sons fricatifs. La distinction voisé-non voisé se fait simplement par estimation des taux d'énergie inférieure à 900 Hz et supérieure à 3 500 Hz. Les auteurs rapportent des succès obtenus en ce qui concerne la durée des phonèmes ramenée à une valeur raisonnable et note le côté attrayant de l'auto-apprentissage. HOUDE [HOUD - 73] juge la technique efficace et généralisable après 700 à 1 000 essais à des formes d'allure similaire à celles qui ont servi à l'entraînement. Cependant, l'intelligibilité de la parole ne s'en trouve pas améliorée.

PHILLIPS, DOLANSKY et al. [PHIL - 68], [DOLA - 65], à Boston, Mass., U.S.A., rapportent des essais de variations du fondamental de la voix et d'imitation de contours proposés par le maître à l'aide d'une version améliorée du *"Pitch Period Indicator"* (appareil datant de 1955 [DOLA - 55]) et constatent une évolution positive pour deux sujets sur trois. L'impression générale est que les sujets se rendent compte qu'ils peuvent faire varier leur fondamental. Généralement, les enfants réussissent à passer d'un niveau bas à un niveau moyen quand on leur demande de monter plus haut. Les inflexions montantes sont plus difficiles à produire que les descendantes. Les résultats sont jugés positifs pour deux des trois enfants entraînés. Des essais ont aussi été réalisés avec des enfants entendants, munis de casques avec bruit de fond pour réduire la contreréaction auditive pendant les essais de vocalisation. Les résultats furent jugés meilleurs qu'avec un stimulus acoustique seul.

NICKERSON et al. [NICK - 73], [NICK - 74] ont proposé une visualisation originale (*"cartoon face"*) dans laquelle les variations de hauteur sont traduites par les mouvements de la pomme d'Adam d'une personnage stylisé. Lui sont associées les indications d'intensité et de nasalité. Dans un autre jeu (*"ball-game"*), l'enfant doit faire passer le tracé mélodique dans une série de chicanes réglables et, finalement, introduire un ballon dans un panier de basket-ball. Le capteur est un accéléromètre fixé sur la gorge. Ces jeux sont expérimentés à la Clarke School for the Deaf, Northampton, Mass., U.S.A. Au cours d'essais réalisés avant 1975, de très jeunes enfants se montrèrent capables de mieux commander leurs cordes vocales mais l'intelligibilité n'en était pas meilleure. La recherche des contributions relatives des différents défauts de la voix des sourds au manque d'intelligibilité s'avère un problème difficile.

BOOTHROYD [BOOT - 73], qui se sert d'une visualisation du pitch détecté par filtrage passe-bas, juge que les sujets entraînés sont capables de commander la tenue des cordes vocales à fréquence constante ou les variations de hauteur, mais sans affirmer qu'ils ne l'auraient pas été sans aide visuelle. De plus, les progrès réalisés ne se transposent pas à la communication de tous les jours.

Une expérience plus récente, rapportée par NICKERSON et al. [NICK - 76], a porté sur quarante élèves à raison de 20 mn par jour durant 7 à 14 semaines (le fondamental n'était pas le seul paramètre rééduqué). L'auto-apprentissage fut tout à fait possible pour les élèves entraînés.

On peut constater une grande variété dans les réactions des utilisateurs de ces visualisations du fondamental et du contour mélodique. L'entraînement à la production de contours intonatifs naturels devrait accroître l'intelligibilité de la parole mieux que l'entraînement à maintenir le niveau de pitch dans une fourchette imposée ou à produire des sons simples avec des inflexions ascendantes ou descendantes par exemple. Cependant, il est très difficile de juger de la qualité des contours émis et de l'intégration de l'aptitude correspondante par le sujet. La conclusion de BOOTHROYD [BOOT - 73] est la suivante : *"Persons involved in sensory aids research should realize that the development of an acceptable device is only a first step. There will remain problems in its application which may be several orders of magnitude more difficult"*.

Jean-Louis SALLES [SALL - 80], au C.T.O.P. de Fougères, France, dans un travail que nous avons encadré, se sert du *"mélographe"* du C.N.E.T. pour rééduquer la mélodie de petites phrases voisées. Ses conclusions sont d'un grand enseignement pour les utilisateurs futurs de dispositifs à contreréaction extrinsèque.

b) Rythme et durée

La plupart du temps, rythme et durée sont rééduqués par présentation des contours mélodiques et/ou d'intensité (voir le paragraphe précédent). D'autres fois, le rythme est marqué par la segmentation du signal vocal de différentes façons : KNUDSEN, à Stockholm, a proposé un indicateur de rythme associé à un indicateur de friction (*"S-indicator"*). PICKETT [PICK - 68] fonde la segmentation sur le taux de basse fréquence, celui-ci étant plus élevé pour les voyelles que pour les consonnes. Après amplification, le signal est filtré par un filtre passe-bas de 1 500 Hz de fréquences de coupure. Les consonnes produisent un tracé dans le bas de l'écran, les voyelles vers le haut avec retour au zéro si un silence les sépare. L'enfant doit tenter de reproduire le tracé du maître. La segmentation fonctionne pour l'association consonne-voyelle à condition que la consonne se trouve en tête. La généralisation à toutes les consonnes semble difficile. RISBERG [RISB - 75] fait état d'un système proche du précédent qui associe à l'apprentissage du rythme et de l'intensité celui de sons comme / r roulé/ et /s/ . BARTH [BART - 75] se sert pour la segmentation phonétique de la recherche des maxima d'une certaine *"fonction d'instabilité"* calculée à partir des valeurs de sortie des filtres d'un analyseur spectral.

c) Intensité instantanée et contour d'intensité

La visualisation du facteur intensité doit permettre à l'enfant de prendre conscience de ses possibilités vocales. Parmi les jouets commandés par la voix, citons simplement *"Teddy Bear"* (West Virginia University's Department of Electrical Engineering) dont les yeux s'allument en présence d'un son suffisamment énergétique. Différents *"vumètres"* [PRON - 47], [BORR - 68] donnent sous diverses formes l'indication de la valeur instantanée de l'intensité.

Les visualisations du contour d'intensité dans le temps donnent un analogue de la courbe enveloppe du signal [STAR - 71] (souvent utilisée dans les écoles de sourds car son tracé nécessite peu de matériel) et servent principalement d'indicateur de rythme et d'"*accentuation*" ; on retrouve dans cette catégorie les multi-indicateurs cités au paragraphe précédent : [BORR - 68], [HANS - 68], [PICK - 68], [HOLB - 71], [STEW - 76], [NICK - 74]. Dans ce dernier jeu ("*cartoon face*"), l'intensité module la taille de la bouche du personnage. BARTH [BART - 78] fait état d'une expérimentation à l'aide de différents jeux tendant à la tenue et au contrôle de l'intensité vocale : balle devant atteindre une cible, barre à maintenir à un certain niveau et apparition d'un dessin en récompense, etc.

Selon FLECHTER S.G. [FLET - 76], l'intensité moyenne et ses variations d'un instant à l'autre et, de façon générale, les paramètres prosodiques sont d'une grande importance en analyse vocale des voix pathologiques. Par exemple, certaines formes à l'attaque des sons peuvent être caractéristiques de problèmes au niveau du larynx. Les difficultés à maintenir un niveau d'énergie constant pourraient traduire des désordres neurologiques, etc. Nous reviendrons sur ces questions quand nous discuterons du développement des aides au diagnostic médical.

L'intensité I est parfois associée au fondamental F_0, soit pour en moduler le contour dans le temps, soit pour donner une représentation des sons dans le plan F_0 - I comme dans l'appareil de KISU et al. [KISU - 76] ou le "*FLORIDA*" [HOLB - 71].

d) Canal vocal

Une des premières tentatives de visualisation de la forme du canal vocal dans la production de sons tenus est celle des Anglais CRICHTON et FALLSIDE [CRIC - 74] qui en donnent une représentation en

temps réel, à l'aide d'un calculateur. Le logarithme de la fonction d'aire du conduit, estimée par prédiction linéaire, est donné en fonction de la distance à la glotte. Le schéma est normalisé à volume constant. L'enfant doit tenter de reproduire des contours type et la qualité de la production vocale (son soutenu) est estimée par calcul d'une distance entre les deux contours. La courbe mouvante de l'élève se fige lorsqu'un certain score de ressemblance est atteint. Chaque son comporte quatre ou cinq modèles différents de façon à tenir compte de l'âge de l'enfant (par l'intermédiaire de la longueur du conduit vocal). L'application de cette aide aux transitions est à l'étude.

e) Taux de friction

Les indicateurs de fricatives furent parmi les premiers appareils à être développés, sans doute à cause de leur facilité de mise en oeuvre, d'une part, et de la fréquence du défaut d'articulation des fricatives /s/ ou /ʃ/ , d'autre part. Ils se fondent sur la comparaison de l'énergie dans les hautes fréquences (supérieures à 3 000 Hz) à l'énergie totale ou sur le calcul du taux de passages par zéro du signal vocal [GUTT - 70]. Ce sont ou des indicateurs binaires [BOOT - 76], [BORR - 68] ou des indicateurs avec déplacement d'une aiguille devant un cadran (avec, en plus, une lampe s'il y a dépassement d'un niveau réglable) [RISB - 68], [MARTO - 70]. Le "*S-indicator*" est jugé particulièrement efficace, surtout par les Suédois. L'indication d'un paramètre de friction peut être liée à celle du voisement, comme dans le système "*VSTA*" [STEW - 76].

f) Nasalité

Le degré de nasalisation est un facteur difficile à détecter en temps réel. Pour les voyelles nasalisées par exemple, les indices acoustiques correspondants dus au couplage du conduit nasal avec le

conduit oral (comme le déplacement du premier formant vers les basses fréquences et son dédoublement dû à la présence d'un zéro dans la fonction de transfert des cavités en parallèle, l'accroissement de l'amplitude de certains harmoniques...) sont très difficiles à mettre en évidence et en plus variables d'une voyelle à l'autre. De toute façon, la détermination de ces indices ne permet pas de traduire simplement le taux de nasalisation.

Les indicateurs de nasalité dérivent directement de la technique qui consiste à estimer les vibrations du nez par le toucher ou du calcul du pourcentage du flot d'air émis par le nez. Dans la première catégorie, la vibration à la surface de l'aile du nez est captée par un microphone de contact, le signal est amplifié et visualisé par déplacement d'une aiguille devant un cadran, le plus souvent. On rencontre ainsi le "*N-indicator*" du KTH à Stockholm [RISB - 75] ou les appareils de HOLBROOK et CRAWFORD [HOLB - 70] ou BOOTHROYD [BOOT - 76], [STEV - 75] ou HOUDE [HOUD - 73]. Dans tous les cas, on se heurte au problème de la relativité de l'indication qui dépend du sujet et en plus de l'effort vocal indépendamment du degré de nasalisation. DE SERPA-LEITAO et GALYAS [DESE - 75] utilisent deux accéléromètres appliqués l'un sur le côté du cartilage thyroïdien, l'autre sur l'aile du nez, évaluent les temps t_L (larynx) et t_N (nez) pendant lesquels on a présence de signaux supérieurs à des seuils préfixés et estiment le degré de nasalité par le rapport : $t_N \times 100/t_L$.

Dans la deuxième catégorie où le flot d'air est mesuré aux narines, citons QUIGLEY et al. [QUIG - 64] et FLETCHER et DALY [FLET - 76]. Enfin, des chercheurs anglais (SELLEY et al.) font état d'une aide visuelle , par ailleurs commercialisée, utilisée dans le traitement de la parole hypernasale.

g) Articulation des sons

Le premier appareil de visualisation du spectre à court-terme (distribution de l'énergie sur l'échelle des fréquences) fut très certainement le "*Visible Speech Translator*" [POTT - 48]. L'analyse était réalisée par banc de filtres et la présentation faite sur écran phosphorescent. Une version modifiée du "*VST*" a été expérimentée par STARK [STAR - 71] pour la rééducation des plosives voisées et non voisées qui sont fréquemment l'objet de confusions.

Un des appareils les plus connus a été développé au KTH de Stockholm sous le nom de "*Lucia Spectrum Indicator*" [RISB - 75] : l'écran comporte une matrice de 10 x 20 lampes à incandescence, chacune des vingt colonnes correspondant à une zone fréquentielle. On rencontre maintenant des spectroscopes chez différents constructeurs.

Citons à ce propos l'expérience faite à partir du système SSD ("*Speech Spectrographic Display*") [STEW - 76] qui propose la visualisation de spectres à large-bande. La résolution temporelle est suffisante pour mettre en évidence les consonnes plosives et la résolution fréquentielle étudiée pour la mise en valeur des formants et des sons fricatifs. La difficulté rencontrée dans ce type très riche de visualisation est due au fait que l'aptitude au décodage de la parole continue est spécifique du sens de l'ouïe. On tend maintenant à limiter la représentation à celle des formants par exemple.

Les formants des sons tenus et en particulier les deux premiers, F_1 et F_2, renferment une grande quantité d'information sur le plan de la discrimination entre les sons. La relation biunivoque qui existe en F_1 et F_2 et la hauteur et la position suivant l'axe avant-arrière de la langue a suggéré la visualisation de ces paramètres pour l'aide à l'articulation des voyelles en particulier. La première d'entre

elles est apparue dans la littérature en 1948 [POTT - 48] mais ne fonctionnait pas en temps réel.

Généralement, c'est le déplacement d'un point dans le plan qui est proposé. Dans l'appareil dit *"Vowel Tongue Position Display"* de KALIKOW et SWETS [KALI - 72], les fréquences sont estimées à partir de combinaisons linéaires des sorties des 19 filtres d'un analyseur spectral avec des contraintes logiques supplémentaires. A l'origine, cette aide entrait dans le cadre d'un système destiné à l'apprentissage d'une langue étrangère (l'anglais par des Espagnols). Une version modifiée propose les trajectoires des formants dans le temps [NICK - 74] (high-low tongue position, front-back tongue position ou combinaison des deux).

WATANABE et KISU [WATA - 75] calculent le taux de passages par zéro à la sortie de deux filtres dont les paramètres sont commandés par les fréquences approchées des formants. L'utilisation de cet appareil nommé *"Articulatory Trainer"*, avec en parallèle des exercices d'intonation, a permis aux enfants d'améliorer leur articulation sans dégradation de leurs performances concernant le pitch.

THOMAS [THOM - 68] décrit une visualisation du même type sur une matrice de 12 x 12 lampes néon. Les fréquences sont estimées après filtrage passe-bande dans les zones 250-1 000 Hz et 700-3 200 Hz. L'indication du voisement et celle du bruit lui sont associées. THOMAS et SNELL [THOM - 70] tirent de leurs essais de différenciation de syllabes et de mots des conclusions positives.

GOLDBERG [GOLDB - 72] propose le *"Visual Feature Indicator"* fournissant dans les associations "Consonne + /a/" ou "/t/ + voyelle" l'indication suivante :

Consonne ↑	Avant	Milieu	Arrière
Voyelle ↓	Avant	Arrière	

BOSTON [BOST - 73] essaie d'ajuster les paramètres de façon à donner une forme réaliste à la bouche d'une figurine dans l'articulation des voyelles extrêmes /a/, /i/ et /u/ .

PICKETT et CONSTAM [PICK - 68] proposent un tracé dans le plan des fréquences moyennes en-dessous et au-dessus de 1 000 Hz. Les points représentatifs des voyelles se situent approximativement sur un demi-cercle dans l'ordre / i e ɛ a ɔ o u /. Les auteurs ne résolvent pas la difficulté d'expliciter la relation entre le résultat sur l'écran et les positions articulatoires correctes mais voient dans l'appareil un moyen d'évaluation objective des productions de l'élève.

Une autre représentation originale, insensible d'après l'auteur [COHE - 68] à la fréquence fondamentale et au niveau d'intensité, donne le taux de composantes sinusoïdales en fonction du taux de composantes en cosinus de l'enveloppe du spectre.

WOOD [WOOD - 71], quant à lui, propose une paramétrisation et une visualisation du signal sous forme de "cepstrogrammes".

Des figures de Lissajous ont été proposées assez tôt dans le *"Voice Visualizer"* [LERN - 52]. Les signaux de sortie de deux filtres passe-bande conditionnent les déviations horizontale et verticale d'un faisceau cathodique.

SCHULTE [SCHU - 70] présente les conclusions d'une étude comparant trois modes de visualisation des sons tenus : onde de parole, figure de Lissajous seule, ou modulée par l'intensité. Ce dernier mode permet, semble-t-il, une bonne discrimination des sons et est étudié dans l'optique de l'aide à la compréhension de la parole.

Un correcteur de voyelles, fondé sur une analyse dimensionnelle du spectre, a été mis au point par POVEL et évalué pour les voyelles /I/ et /i/ en contexte CVC [POVE - 74].

Pour terminer sur ce plan de l'articulation des sons, citons l'expérience de rééducation d'un locuteur adulte par FLETCHER et al. [FLET - 79] à l'aide d'un palatomètre associé à un calculateur. Les contacts de la langue avec le pseudo-palais muni de 96 électrodes sont visualisés en écho sur des diodes électroluminescentes. Les auteurs font état d'améliorations significatives, apparemment stables, dans l'articulation (lieu et mode) après huit mois d'entrainement. Des systèmes de palatographie sont maintenant souvent développés dans les laboratoires de phonétique [COUR - 81] [SAWA - 78].

h) Systèmes plus complets

La majeure partie des dispositifs décrits précédemment ont été conçus et réalisés à l'initiative des phonéticiens ou des enseignants spécialisés. L'évolution du matériel informatique, en particulier l'intégration accrue et l'abaissement des coûts, conduit maintenant à la conception de systèmes de traitement numérique centrés sur mini ou micro-ordinateurs.

Le système développé par B.B.N., aux Etats-Unis, [NICK - 73] et utilisé par BOOTHROYD (The Clarke School for the Deaf, Southampton, Mass.), fut le premier système général placé sous le contrôle de l'ordinateur. Quatre programmes permettent de visualiser :

. l'évolution dans le temps de paramètres tels que l'intensité, le voisement, la nasalité,

. le fondamental de la voix (jeu de basket-ball),

. un spectre "lissé" présenté verticalement et symétrisé pour évoquer la forme d'un "vase", rayé si le son est voisé,

. un personnage stylisé dont la forme est conditionnée par les caractéristiques du signal,

comme nous l'avons mentionné plus haut.

Nous présentons, pour notre part, le système SIRENE (partie D) se distinguant par l'usage systématique du calculateur en ligne et l'adjonction d'un système de reconnaissance automatique de la parole. Il comprend trois séries de jeux concernant respectivement :

. les paramètres prosodiques (fondamental, intensité, rythme, corrélation fondamental-intensité),

. les paramètres fréquentiels (spectres, formants, fréquences de discrimination optimales, conduit vocal),

. les performances dans l'articulation de mots ou d'éléments de phrase.

En plus de l'autoévaluation permise par le retour visuel, une évaluation quantitative des performances de l'élève permet une appréciation objective de ses possibilités et de ses progrès, dans l'optique d'une aide au bilan orthophonique et d'une aide à l'établissement d'un programme de rééducation [HATO - 75], [HATO - 81].

Depuis la définition en France de notre système SIRENE, nous pouvons citer comme réalisations informatiques :

- le système développé par la Société IBM-France [BAUD - 80] [DEST - 81] [DIBE - 81] proposant des exercices de contrôle de l'intonation et du volume de la voix sous forme visuelle. Le système actuel, testé dans différentes écoles de sourds, utilise un micro-ordinateur IBM PC,

- le système issu du M.I.T. et utilisé dans une école spécialisée de Omaha dans lequel la qualité d'élocution d'un mot est appréciée grâce à la détection du voisement, d'une part, et de la nasalité à l'aide d'un accéléromètre de contact placé sur l'aide du nez, d'autre part. L'ensemble est piloté par un micro-ordinateur LSI 11-23 [OSBE - 81],

- un ensemble de rééducation vocale, réalisé à Londres, auquel est associé un "testeur" de prothèse auditive [KING - 82],

- le FTA ("*Fricative and Timing Aid*") proposé par l'Université de Cambridge pour la rééducation des sons fricatifs et du voisement [BATE - 82].

IV - REFLEXIONS POUR LA CONCEPTION D'UN EQUIPEMENT SPECIALISE

Nous rassemblons ici un certain nombre de réflexions qui constituent la trame à partir de laquelle a été élaboré notre projet.

Le point qui nous paraît d'abord essentiel est la collaboration entre chercheurs et praticiens, dans le cadre d'une vaste politique d'aide aux handicapés. A titre de modèle, citons l'organisation "*International Society for Rehabilitation of the Disabled*" [LUND - 78] à laquelle nous appartenons et qui recouvre une grande part des efforts réalisés dans ce cadre.

Les multiples thèmes de recherche appliquée au problème des sourds et déficients auditifs peuvent se regrouper en trois grandes catégories :

- les recherches visant à une meilleure compréhension du problème général de la production et de la perception de la parole. L'apprentissage de la parole ne peut être dissocié du problème de l'apprentissage du langage, en particulier des facteurs permettant l'intégration des structures linguistiques. La parole doit en effet être considérée à la fois comme un support possible de l'information linguistique et comme un facteur déterminant dans l'acquisition du langage,

- les recherches évaluant l'impact de la surdité sur l'apprentissage de la parole et du langage et sur la perception de la parole. Sur ce point, il convient de distinguer la capacité du sujet à percevoir les sons et les paramètres prosodiques, à faire la discrimination entre sons élémentaires, son aptitude à la lecture faciale et à la perception tactile, de façon à rechercher le meilleur moyen de lui transmettre l'information contenue dans la parole,

- les recherches conduisant à la conception et la réalisation d'aides instrumentales à la communication verbale.

L'ensemble de ces recherches intéresse des domaines aussi divers que la linguistique, la phonétique, la psychophysiologie, la psychologie ou la médecine. L'apport du scientifique soutenu par la technologie intervient à chacun de ces niveaux mais de façon très imbriquée : aucun des domaines cités ne doit être négligé dans la définition de son travail.

Dans le domaine qui est le nôtre de la réalisation d'aides à la communication verbale, nous avons été guidé par les idées qui suivent.

Le terme de "contreréaction" ("*feedback*") est régulièrement employé pour désigner l'incidence a posteriori d'une influence créée par l'événement lui-même sur le locuteur. Cette idée de contreréaction est très importante en ce qui concerne l'acquisition de la parole. En effet, la prise de conscience par le locuteur normalement entendant de sa propre émission sonore n'est pas instantanée mais se fait par une sorte d'évaluation après coup, à partir de deux éléments :

- les schémas kinesthésiques provoqués par la chaîne qui crée l'événement acoustique : souffle, cordes vocales, positions et mouvements des articulateurs...,

- les schémas acoustiques dérivant des propriétés physiques des sons prononcés.

La production des sons, quant à elle, présuppose une assimilation antérieure de la suite d'activités neuromusculaires requises, cette assimilation ayant progressé et s'étant développée par apprentissage grâce à la double contreréaction auditive et sensorimotrice dont nous venons de parler.

Pour l'enfant qui ne présente guère ou pas d'ouïe résiduelle, il est nécessaire, pour que se passe le mieux possible la phase d'assimilation, qu'une contreréaction extrinsèque vienne se substituer à la contreréaction intrinsèque déficiente. Ce n'est que grâce au lien entre schémas perçus de façon extérieure et schémas sensorimoteurs associés que la progression pourra se faire vers l'acquisition d'automatismes favorables à la production de la parole.

Notre opinion est que la contreréaction extrinsèque apportée le plus souvent par l'appréciation de maîtres compétents peut être avantageusement complétée de renseignements objectifs, faisant appel aux divers facteurs dont la combinaison forme l'onde sonore particulière qu'est la parole, fournis par des équipements particuliers.

C'est à ce niveau qu'interviennent les possibilités toujours plus grandes de la technologie, qui ne peut se suffire à elle-même, mais dont l'impact sur le domaine médical et paramédical ira grandissant.

A l'heure actuelle, la difficulté est plus l'importance de la forme à donner à cet apport qu'une quelconque limitation technologique. L'aide doit s'insérer dans un programme plus vaste visant à modifier le processus de l'attitude communicative de l'enfant, ce qui requiert bien plus qu'une simple aide technique.

L'aide extérieure, enfin, doit être compatible au maximum avec la situation de communication naturelle, ce qui est l'un des points délicats en ce qui concerne les aides visuelles qui placent l'enfant dans une situation d'"*exercices*" et non de "*conversation*". Les efforts futurs doivent tendre vers une généralisation des aides à la production vocale vers des aides à la communication parlée, ce qui aurait un impact social énorme pour les déficients auditifs.

Pour la conception et le développement d'un appareillage spécialisé, l'interaction avec l'expérimentation est primordiale. Il est impératif d'avoir la possibilité de modifier, refondre, supprimer ou compléter tel ou tel module, ce qui nous a conduit, personnellement, dans la réalisation de notre système SIRENE fondé sur l'apport d'information visuelle à l'enfant sourd, à une structure modulaire très souple à partir de noyaux d'analyse de base.

Par ailleurs, la relation entre visualisation et production vocale doit être sinon biunivoque, du moins directement interprétable. Ceci n'est possible que dans les limites d'un aspect donné de la parole et nécessite une intervention directe du rééducateur pour expliquer, illustrer et critiquer les essais. La notion d'auto-apprentissage ne pourra intervenir qu'une fois cette relation assimilée par le sujet.

Deux questions de fond se posent ensuite :

- le sujet est-il capable d'aboutir à une phase de fixation qui lui permettrait de d'affranchir de l'aide extérieure dans un délai raisonnable ?

- les effets de l'apprentissage sont-ils généralisables à la parole spontanée ?

Seules l'expérience et l'observation à moyen terme peuvent fournir une réponse à ces deux questions. Les réactions relevées dans la littérature sont très diverses et souvent nuancées. Il est certain qu'il faut éviter que le sujet ne devienne trop dépendant de la contreréaction extrinsèque. L'aide visuelle présente l'avantage de placer l'enfant dans une situation de jeu dans la première phase d'apprentissage. Les perspectives apportées par les recherches sur la parole codée automatique et les aides tactiles offrent une ouverture dans la phase suivante de communication active.

Pour conclure, nous insisterons sur le rôle essentiel du rééducateur dont la tâche dans l'adaptation de ses connaissances au processus d'enseignement de la parole est multiple : identification des erreurs, diagnostic de la cause des difficultés, relation entre aptitudes motrices et développement du langage, définition du programme de rééducation adapté à l'enfant, appréciation de l'acquis et de l'apport des séances de rééducation...

Il faut alors veiller à ce que l'introduction de matériel spécialisé n'ajoute pas au travail du rééducateur une dimension technique rébarbative qui en condamnerait l'utilisation, à plus ou moins long terme. Les échanges "homme-machine" doivent prendre une forme standardisée, facilement assimilable avec cependant, pour le maître, la possibilité de choix et d'ajustement qui donne une certaine souplesse au système ; nous revenons sur ces aspects dans la partie D de cette thèse.

PARTIE B

TRAITEMENT ET PARAMETRISATION DU SIGNAL VOCAL

INTRODUCTION A LA PARTIE B

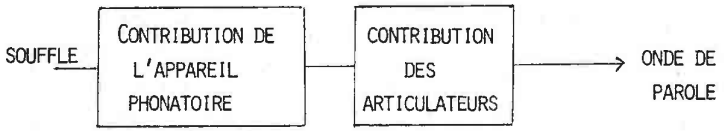
L'onde de parole peut se représenter comme la réponse d'un résonateur -le conduit vocal associé éventuellement au conduit nasal- à une suite d'impulsions quasi-périodiques émises au niveau des cordes vocales et/ou à une source de bruit dû à une constriction ou un relâchement brusque des articulateurs.

L'analyse de la parole revient à approcher tous ces éléments à partir de ses conséquences physiques immédiatement accessibles : les variations de pression au niveau de la bouche et du nez, les vibrations des ailes du nez, les vibrations au niveau du larynx. Elles sont transformées en signaux électriques grâce à des microphones pour les premières ou des accéléromètres pour les dernières. On se trouve alors en présence de signaux s(t) continus dans le temps, orientés et unidimensionnels. Ces caractéristiques (qui sont aussi celles des électroencéphalogrammes et des électrocardiogrammes par exemple) font que le traitement de la parole se distingue du traitement des images, multidimensionnelles par nature.

Une analyse complète d'un signal doit avoir deux fonctions fondamentales : en extraire les traits distinctifs sous l'angle de l'information apportée d'une part, exprimer les propriétés de transmission du système considéré du point de vue automatique à partir de la connaissance du mode de production, d'autre part.

La catégorie de représentations du signal la plus fréquemment rencontrée est celle où le signal est présenté sous forme d'une combinaison linéaire de fonctions prédéfinies. La plus courante est la résolution de Fourier en composantes sinusoïdales (la base est alors de dimension infinie), intéressante quand il s'agit de caractériser les propriétés de transmission d'un système linéaire invariant dans le temps. Une des qualités de cette décomposition est en effet d'être insensible à une translation dans le temps qui n'affecte que le spectre de phase et encore de façon simple.

Une autre façon d'aborder le problème de l'analyse du signal de parole est de remonter la chaîne :



dans le sens opposé à celui de l'onde de pression, support du message acoustique. Il s'agit alors non plus d'une représentation ou d'une caractérisation de la parole mais d'une véritable "radioscopie" où sont observés à la fois chacun des éléments qui contribuent à l'émission de la parole et la progression de cette onde de parole depuis le premier souffle infraglottique. L'intérêt de cette approche est multiple : connaissance et compréhension de l'étonnant phénomène de l'élocution, affinement des études phonologiques, classification des voix (variabilité intra et interlocuteurs), etc.

Les manières d'élaborer ce modèle, donc de définir les méthodes d'"analyse inverse", sont variées : on peut s'inspirer soit du fonctionnement du système phonatoire, soit de celui de notre système auditif, cette façon de faire étant plutôt orientée vers la reconnaissance et la compréhension du message parlé.

Les résultats fournis par les types d'analyse envisagés plus haut sont extrêmement chargés d'information. Si l'on s'intéresse aux porteurs d'information pertinents, il est nécessaire de réduire l'espace de représentation du signal suivant certains critères d'optimisation : les méthodes utilisées font alors appel aux techniques de traitement et de réduction des données.

C'est à ces deux aspects, traitement du signal vocal et analyse des données, que cette partie B est consacrée.

CHAPITRE 1

RECHERCHE DE PARAMETRES CARACTERISTIQUES

I - TECHNIQUES DE TRAITEMENT NUMERIQUE

1. Simulation numérique

La simulation est un procédé dans lequel on cherche à représenter de façon exacte ou approchée un système-source par un système-objet. Lorsque le signal-source est échantillonné dans le temps et quantifié, on parle de simulation numérique. Dans la suite, nous nous intéresserons toujours à un signal échantillonné à intervalles réguliers dans le temps. La durée d'un intervalle est la période d'échantillonnage Te , son inverse la fréquence d'échantillonnage $Fe = \frac{1}{Te}$.

Cette mise en forme du signal se fait préalablement à tout traitement numérique en vue, par exemple, de :

- l'extraction d'information pertinente,
- la diminution du bruit ou des distorsions,
- la description fréquentielle, etc.

2. Transformée en z

A un signal discret $\{x(nT_e)\}$, écrit plus simplement $\{x(n)\}$ ou $\{x_n\}$, on associe sa transformée en z, qui est l'équivalent de la transformée de Laplace dite "en s" pour un signal continu, définie par :

$$X(z) = \sum_{-\infty}^{+\infty} x_n z^{-n} = Z(\{x_n\}) \qquad (1)$$

C'est une série de Laurent à laquelle s'appliquent les propriétés classiques de convergence de ces séries [CARTA - 61]. On a par ailleurs :

$$Z(\{x_{n-1}\}) = z^{-1} \cdot Z(\{x_n\}) \qquad (2)$$

Le facteur z^{-1} apparaît ainsi comme un retard sur la suite $\{x_n\}$.

La transformation inverse donnée par :

$$x_n = \frac{1}{2\pi i} \oint_C X(z)\, z^{n-1}\, dz \qquad (3)\ ,$$

où C est un contour fermé entourant l'origine et situé dans la zone de convergence, permet de retrouver les échantillons x_n .

3. Cas particuliers de transformée en z

Dans la recherche d'une base de filtres "exponentiels orthonormés", nous avons étudié la transformée en z d'une fonction sinusoïdale amortie prenant naissance à l'instant 0 et échantillonnée à la fréquence $Fe = 1/Te$.

Soient $\{x(n) = \exp(-\sigma n\, Te) \times \cos(\omega n\, Te - \varphi)$ avec $n = 0,1 \ldots \infty\}$ l'ensemble des valeurs discrètes prises par la fonction aux instants d'échantillonnage.

Si l'on pose $\delta = \sigma Te$ et $\theta' = \omega Te$, $x(n)$ s'écrit :

$$x(n) = \exp(-\delta n) \times \cos(n\theta' - \varphi) \ .$$

La fonction analogique d'origine correspond à la valeur $s = -\sigma + i\omega$ de la variable de Laplace.

La fréquence du signal vaut $\frac{\omega}{2\pi}$ et son amortissement est caractérisé par le décrément logarithmique :

$$\operatorname{Log} \frac{x(n)}{x(n + \frac{Fe}{F})} = \frac{\sigma}{F} = 2\pi \frac{\sigma}{\omega}$$

a) Transformée en z de fonctions sinusoïdales amorties

La transformée en z de $\{x(n)\}$ vaut :

$$X(z) = \sum_{n=o}^{\infty} x(n)\, z^{-n}$$

$$= \text{partie réelle de } E(z)$$

où $$E(z) = \sum_{n=o}^{\infty} \exp(-\delta n - i\varphi + in\theta')\, z^{-n}$$

$$= \frac{\exp(-i\varphi)}{1 - \exp(-\delta + i\theta')\, z^{-1}}$$

De cette dernière expression, on tire immédiatement :

$$X(z) = \cos\varphi\, C(z) + \sin\varphi\, S(z)$$

où $C(z)$ et $S(z)$ sont les transformées en z des fonctions amorties cosinus et sinus respectivement :

$$C(z) = \frac{1 - \exp(-\delta)\cos\theta' z^{-1}}{1 - 2\exp(-\delta)\cos\theta' z^{-1} + \exp(-2\delta)\, z^{-2}}$$

$$S(z) = \frac{\exp(-\delta)\sin\theta' z^{-1}}{1 - 2\exp(-\delta)\cos\theta' z^{-1} + \exp(-2\delta)\, z^{-2}}$$

Dans la suite, nous poserons par convention :

$$\exp(-\delta) = \rho \quad , \quad \exp(-2\delta) = \rho^2 = M \quad ,$$
$$2\exp(-\delta)\cos\theta = R \ .$$

Il vient alors :

$$C(z) = \frac{1 - R/2\ z^{-1}}{1 - R\ z^{-1} + M\ z^{-2}} \qquad (4)$$

et
$$S(z) = \frac{\sqrt{M - R^2/4}\ z^{-1}}{1 - R\ z^{-1} + M\ z^{-2}} \qquad (5)$$

On peut remarquer que le dénominateur $D(z)$ s'écrit :

$$D(z) = 1 - Rz^{-1} + Mz^{-2} = (1 - zo\ z^{-1})(1 - zo^*\ z^{-1})$$

avec $zo = \rho \exp(i\theta)$ et zo^* son complexe conjugué.

Le schéma B-1 indique la position de zo dans le plan complexe par rapport au cercle de rayon unité (dit aussi cercle unité) :

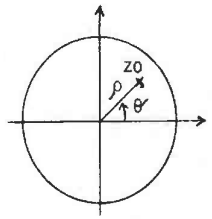


Fig. B-1 : Le cercle unité

On a :

R = 2 x Partie réelle de zo

$\rho = |zo| < 1$

et $M = zo.zo^* = |zo|^2$

b) Recherche des pôles (x) et des zéros (•) de ces fonctions et représentation de leurs affixes dans le plan z (fig. B-2)

$C(z) = \frac{1 - \rho \cos\theta\ z^{-1}}{D(z)}$ admet :

- pour pôles : zo et zo*
- pour zéros : 0 et $\rho \cos\theta$

$S(z) = \frac{\rho \sin\theta\ z^{-1}}{D(z)}$ admet :

- pour pôles : zo et zo*
- pour zéros : 0
  le deuxième
  est rejeté
  à l'infini

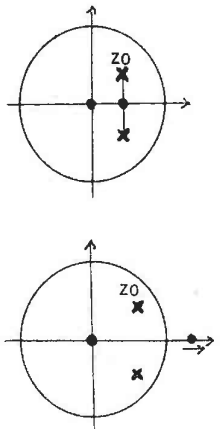


Fig. B-2

c) Cas général

Le deuxième zéro se déplace sur l'axe des réels lorsque le déphasage par rapport à la fonction cosinus varie. Sur le schéma suivant (fig. B-3), on précise ce résultat en indiquant pour des valeurs remarquables du déphasage φ à π près (notées entre parenthèses) la position du zéro (•). θ est choisi inférieur à $\frac{\pi}{2}$.

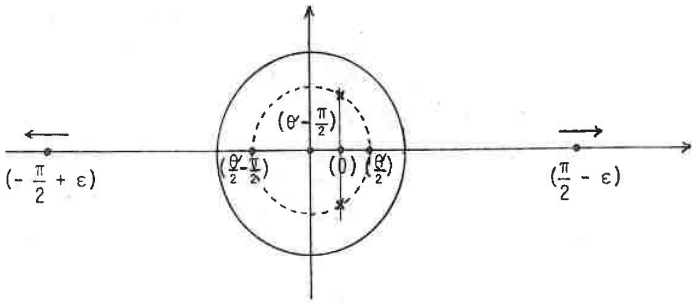


Fig. B-3 : Evolution du deuxième zéro de la transformée d'une fonction sinusoïdale amortie en fonction du déphasage φ .

d) En résumé

Le tableau B-4 résume les conclusions précédentes :

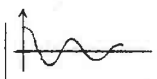
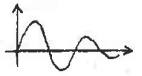
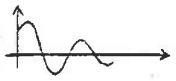
Fonction de départ			
Valeur x(n) de la fonction échantillonnée à l'instant n Te	$\rho^n \cos n\theta$	$\rho^n \sin n\theta$	$\rho^n \cos(n\theta-\varphi)$
Transformée en z des {x(n)}	$\frac{1-\rho\cos\theta z^{-1}}{D(z)}$	$\frac{\rho \sin\theta z^{-1}}{D(z)}$	$\frac{\cos\varphi-\rho\cos(\theta+\varphi)z^{-1}}{D(z)}$
Pôles de la transformée	zo et zo*	zo et zo*	zo et zo*
Zéros de la transformée	0 et $\rho\cos\theta$	0 et un zéro rejeté à l'infini	0 et $\rho\frac{\cos(\theta+\varphi)}{\cos\varphi}$

Fig. B-4 : Transformées en z de fonctions sinusoïdales exponentielles. Tableau résumé.

4. Plan s et plan z

La figure B-5 résume l'analogie qui existe entre le plan s de la variable continue de Laplace et le plan z pour des valeurs particulières des variables s_0 et z_0 .

Plan s

$s_0 = -\sigma + i\omega$

Plan z

$z_0 = \exp(s_0 T_e)$

$= \rho(\cos\theta + i\sin\theta)$

avec $\rho = \exp(-\sigma Te)$

et $\theta = \omega Te$

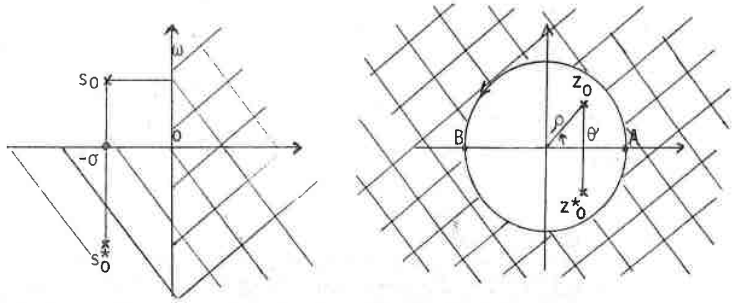


Fig. B-5 : Correspondance entre plan s et plan z

En z, le demi-cercle unité positif ($\rho = 1$ et $0 \leqslant \theta \leqslant \pi$) correspond aux vibrations sinusoïdales pures non amorties, l'intérieur aux vibrations amorties.

Les points A et B sont les points des fréquences respectivement nulle et égale à Fe/2. Dans le cas général, les fréquences sont proportionnelles à l'angle polaire θ . C'est ainsi que le point z_0 de la figure correspond à une vibration de fréquence $Fo = Fe \times \frac{\theta}{2\pi}$ et d'amortissement caractérisé par ρ . La notion de largeur de bande associée est envisagée plus loin, à propos des spectres de fréquence.

5. Filtrage numérique

a) Cas général

Considérons un filtre soumis à son entrée à la succession d'échantillons $\{x(n)\}$ ou $\{x_n\}$ et fournissant en sortie la suite $\{y_n\}$ aux instants d'échantillonnage 0 , Te, 2Te,... nTe,...

L'action du filtre peut se caractériser par sa fonction de transfert :

$$H(z) = \frac{Y(z)}{X(z)} \ ,$$

expression dans laquelle $X(z)$ et $Y(z)$ désignent les transformées en z respectives des signaux d'entrée et de sortie. $H(z)$ est choisie de façon telle que $\{y_n\}$ possède par rapport à $\{x_n\}$ certaines propriétés, généralement fréquentielles.

Si le filtre est soumis à une impulsion unité à l'instant 0 (équivalant à la fonction de Dirac en continu), on a :

$$X(z) = 1$$

donc $H(z) \equiv Y(z)$.

Il s'ensuit que la fonction de transfert du filtre numérique est aussi la transformée en z de sa réponse impulsionnelle $\{h_n\}$ ce qui peut s'écrire :

$$H(z) = \sum_{-\infty}^{+\infty} h_n \, z^{-n} \qquad (6)$$

L'action du filtre est une action de convolution du signal d'entrée avec sa réponse impulsionnelle.

L'évaluation de $H(z)$ sur le demi-cercle unité donne la réponse du filtre en fréquence et en phase. Pour ce calcul, on prend une suite de valeurs z_n de z de la forme :

$$z_n = \exp(i\, 2\pi \frac{F_n}{F_e}) \qquad \text{avec} \quad 0 \leqslant F_n \leqslant \frac{Fe}{2} \ .$$

Sauf exception, un filtre numérique doit satisfaire aux conditions suivantes :

. condition de stabilité :

La définition la plus couramment envisagée est la suivante : un système est stable si toute entrée bornée donne une sortie bornée.

Une condition nécessaire et suffisante de stabilité est qu'il existe un réel A tel que :

$$\sum_{-\infty}^{+\infty} |h_n| < A \ ,$$

. condition de causalité :

Un filtre est dit "causal" si la valeur de sortie à l'instant $n_0 \, T_e$ ne dépend que des valeurs du signal d'entrée aux instants $n \, T_e$ tels que $n \leqslant n_0$.

Dans ce cas, si le filtre est sollicité à l'instant 0 , $h_n = 0$ pour tout $n < 0$.

b) Filtres linéaires invariants dans le temps

Ils sont caractérisés par une fonction de transfert du type :

$$H(z) = \frac{\sum_{k=0}^{M} b_k \, z^{-k}}{1 + \sum_{k=1}^{P} a_k \, z^{-k}}$$

où les a_k et b_k sont des constantes réelles.

L'action de ces filtres peut être évaluée numériquement de façon récurrente suivant la relation :

$$y_n + \sum_{k=1}^{P} a_k\, y_{n-k} = \sum_{k=0}^{M} b_k\, x_{n-k} \qquad \text{pour } n \geqslant 0 \quad (7)\,,$$

avec les conditions initiales $y_\ell = x_\ell = 0$ pour tout $\ell < 0$.

Stabilité et causalité entraînent que tous les pôles de la fonction de transfert $H(z)$ d'un filtre linéaire se trouvent à l'intérieur du cercle unité dans le plan z .

c) Filtres (R,M)

Nous appellerons *couples de filtres* (R,M) l'ensemble ordonné de deux filtres numériques admettant pour fonctions de transfert les quantités $C(z)$ et $S(z)$ définies en I.3 :

$$C(z) = \frac{1 - R/2\, z^{-1}}{1 - Rz^{-1} + Mz^{-2}}$$

et

$$S(z) = \frac{\sqrt{M - R^2/4}\, z^{-1}}{1 - Rz^{-1} + Mz^{-2}}\,,$$

dans lesquelles $0 < M \leqslant 1$ et $|R| \leqslant 2\sqrt{M}$.

Par définition, leurs réponses impulsionnelles sont égales respectivement à :

$$\{\rho^n \cos n\theta\,,\ n \geqslant 0\}$$

et

$$\{\rho^n \sin n\theta\,,\ n \geqslant 0\}$$

avec $\rho = \sqrt{M}$ et $\theta = \text{Arc cos } \frac{R}{2\sqrt{M}}$.

Nous voyons au paragraphe I.6 suivant que de tels filtres permettent de caractériser les composantes sinusoïdales d'un signal et quel est leur lien direct avec la transformation de Fourier discrète dans le cas particulier où M = 1. Nous envisageons en IV.4 la projection de zones stables du signal de parole (voyelles) sur un ensemble de couples $(R_i\,,\ M_i)$ choisis pour leur pouvoir de discrimination. Nous décrivons ensuite une expérience de filtrage, codage et restitution de séquences de parole.

La figure B-6 donne les "réponses en fréquence" des filtres de fonction de transfert $H(z) = 1 / D(z) = 1 /(1 - Rz^{-1} + Mz^{-2})$ pour $\theta = \pi/3$ et différentes valeurs de ρ allant de 0.946 à 0.086. Chaque courbe est obtenue par le calcul du logarithme de $|1/D(z)|^2$ pour des valeurs de z réparties sur le cercle unité ; elle coïncide par simple translation avec la réponse du filtre S(z) de même dénominateur.

Sur la figure B-6, on remarque la déviation de l'angle α_m au maximum des courbes par rapport à l'angle θ de résonance du filtre, ceci étant d'autant plus accentué que ρ est plus faible. On vérifie en effet que :

$$\cos \alpha_m - \cos \theta = \cos \theta \times (1 - \rho)^2 / 2\rho$$

quand α_m existe.

L'écart $\theta - \alpha_m$ est du signe de cos θ et s'annule pour $\theta = \pi/2$, c'est-à-dire pour $F = F_e/4$. Cette déviation du maximum, sensible pour des amortissements importants, fausse, comme on le verra plus loin, la détermination de la fréquence de résonance par détection de pic.

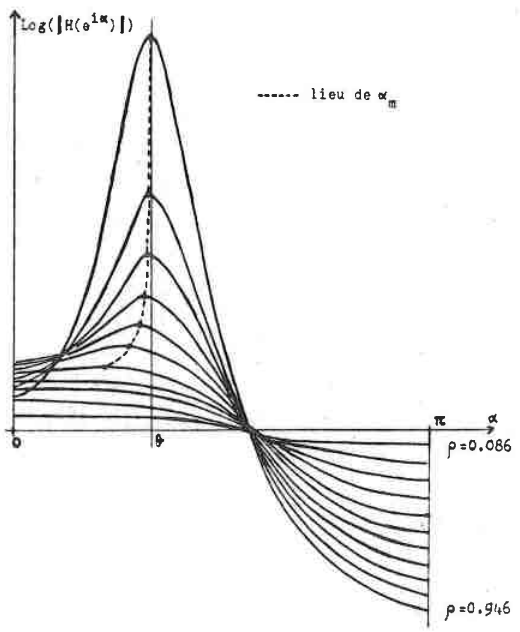


Fig. B-6 : Réponses en fréquence d'une famille de filtres de fonctions de transfert $H(z) = 1/(1-\rho z^{-1}+\rho^2 z^{-2})$ pour différentes valeurs de ρ.

6. Transformation de Fourier discrète (DFT)

Le passage au domaine fréquentiel se fait classiquement par transformation de Fourier. Sans insister sur les formulations mathématiques bien connues de cette transformation, nous étudions ici le problème pratique de son application à notre domaine et nous présentons, après avoir examiné les contraintes du filtrage numérique, une adaptation de cette notion à nos préoccupations.

a) Présentation de la méthode de calcul

Dans le cas d'un signal d'origine discret, le calcul de la transformée de Fourier (dite discrète) revient à évaluer les projections du signal sur une base finie de fonctions sinusoïdales non amorties.

La transformée de Fourier discrète d'un signal d'origine $\{x(kTe),\ k=0,1,\ldots,N-1\}$, échantillonné à la fréquence $Fe = 1/Te$ et que nous noterons plus simplement $\{x(k),\ k=0,1,\ldots,N-1\}$, est l'ensemble des $X(n)$ $\{X(n),\ n=0,1,\ldots,N-1\}$, tels que :

$$X(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(k) \exp(-i\, 2\pi \frac{nk}{N}) \qquad (8)$$

soit, si l'on pose $w = \exp(-i\,\frac{2\pi}{N})$:

$$X(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(k)\, w^{nk} \qquad (9)$$

L'ensemble des relations (9) pour n de 0 à N-1 peut s'écrire sous la forme matricielle suivante, au facteur $\frac{1}{N}$ près :

$$\begin{bmatrix} X(o) \\ X(1) \\ \vdots \\ X(i-1) \\ \vdots \\ X(N-1) \end{bmatrix} = \left[\begin{array}{c|ccc} w^o & w^o & & w^o \\ \hline w^o & w^1 & & w^{N-1} \\ w^o & w^{i-1} \quad w^{(i-1)(j-1)} & & w^{(i-1)(N-1)} \\ w^o & w^{N-1} & & w^{(N-1)^2} \end{array}\right] * \begin{bmatrix} x(o) \\ x(1) \\ x(j-1) \\ x(N-1) \end{bmatrix}$$

La matrice W de terme général $w^{(i-1)(j-1)}$ est carrée, symétrique et tous ses termes sont des racines Nièmes de l'unité. De plus, elle comprend des sous-matrices symétriques possédant entre elles des liens de symétrie. Le schéma ci-dessous illustre cette propriété pour N égal à 8. Seuls sont indiqués les exposants de $w = \exp(-i \frac{2\pi}{8})$.

0	0	0	0	0	0	0	0
0	1	2	3	4	-3	-2	-1
0	2	4	6	0	-6	-4	-2
0	3	6	1	4	-1	-6	-3
0	4	0	4	0	4	0	4
0	-3	-6	-1	4	1	6	3
0	-2	-4	-6	0	6	4	2
0	-1	-2	-3	4	3	2	1

Les termes qui n'appartiennent pas aux sous-matrices encadrées valent tous ±1.

Ces propriétés ont permis la construction de différents algorithmes de calcul de transformée rapide (FFT pour Fast Fourier Transform) comme l'algorithme de COOLEY et TUKEY [COOL - 65]. Le calcul équivaut alors à $N \log_2 N$ multiplications au lieu de N^2, ce qui pour N = 1024 correspond à un gain dans un rapport voisin de 100.

Dans [JOHN - 83], on trouve la description de l'un des algorithmes de calcul les plus rapides à ce jour.

Il nous paraît intéressant de remarquer, en rapprochant l'expression de w^n de la relation $z_n = \exp(i \frac{2\pi n}{N})$ (cf. paragraphe I.5a), que le calcul de la transformée de Fourier revient à projeter successivement le signal d'entrée $\{x(n)\}$ sur N couples de filtres du type $C(z)$ et $S(z)$ décrits au paragraphe I.3 et tels que leurs pôles soient situés sur le cercle unité. Selon nos notations, il s'agit de cas particuliers de couples de filtres (R, M) qui s'écriraient couples $(2\cos\theta, 1)$.

Soient $C_n(z)$ et $S_n(z)$ les fonctions de transfert du nième couple de filtres ; on a :

$$C_n(z) = \frac{1 - \cos\theta_n \, z^{-1}}{1 - 2\cos\theta_n \, z^{-1} + z^{-2}}$$

et

$$S_n(z) = \frac{\sin\theta_n \, z^{-1}}{1 - 2\cos\theta_n \, z^{-1} + z^{-2}}$$

avec $\theta_n = 2\pi \frac{n}{N}$.

Les valeurs de sortie des filtres $\{C_n(z) , n = 0,1...,N-1\}$ et $\{S_n(z) , n = 0,1...,N-1\}$, au temps (N-1) Te , correspondent respectivement aux parties réelles et imaginaires des $\{X(n)\}$.

Nous appelons projections d'un signal caractérisé par la suite $\{x(k)\}$ sur un couple de filtres $C_n(z)$, $S_n(z)$ les quantités réelles $X_r(n)$ et $X_i(n)$ telles que $X(n) = X_r(n) + i \, X_i(n)$.

Pour identifier le calcul de ces projections à celui de la convolution du signal d'entrée et de la réponse impulsionnelle des filtres, il est nécessaire d'entrer les valeurs d'origine dans l'ordre inverse x(N-1), x(N-2)... x(1), x(0) .

En résumé, nous pouvons donner le tableau d'équivalence suivant (fig. B-7) :

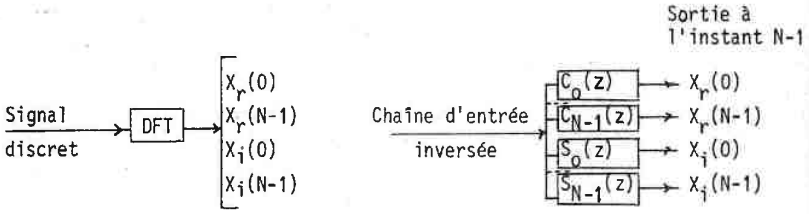


Fig. B-7 : Analogie entre le calcul de la DFT et le calcul des projections sur des filtres C(z), S(z)

On vérifie que $C_{N-n}(z) = C_n(z)$

et que $S_{N-n}(z) = -S_n(z)$,

ce qui correspond au fait que X(N-n) et X(n) sont des quantités complexes conjuguées. Il s'ensuit que, dans le cas envisagé ici d'une chaîne d'entrée de N valeurs réelles, la transformation de Fourier fournit N couples de valeurs dont seulement (N DIV 2 + 1) sont significatifs.

b) Contraintes de la DFT

- contraintes dues à la discrétisation du signal -

La discrétisation d'un signal continu à la fréquence d'échantillonnage Fe , qui équivaut à la multiplication du signal par un peigne de Dirac de période Te = 1/Fe , se traduit au niveau de la transformée par une convolution de la transformée du signal continu avec celle du peigne, qui est elle-même un peigne de Dirac de période Fe = 1/Te . Le résultat de la transformation de Fourier possède aussi cette périodicité. Il s'ensuit une limitation du nombre des valeurs utiles de la transformée, comme indiqué au paragraphe précédent.

- contrainte due à la troncature du signal -

La formulation théorique de la transformée de Fourier d'un signal continu suppose la connaissance de ce signal de $-\infty$ à $+\infty$ dans le temps. La nature du calcul numérique impose que l'on traite une durée de signal finie, ce qui conduit à des paramètres caractérisant son comportement "à court terme". Cette limitation de la tranche de signal dans le temps revient ainsi à le multiplier par une "fenêtre" de valeur nulle en dehors de la période de temps considérée.

La transformée de Fourier subit en conséquence une convolution par la transformée de la fenêtre. Le choix du type de fenêtre dépend de deux impératifs : la nécessité de réduire au maximum l'effet de troncature du signal et celle de limiter les temps de calcul supplémentaires.

- contrainte due à l'échantillonnage du spectre -

La caractérisation fréquentielle sous la forme d'un *"spectre"* est une des applications les plus courantes de la transformation de Fourier discrète à un signal réel numérisé. Ce spectre, de par la nature discrète de la transformée, est un spectre de raies séparées par un pas en fréquence

de Δf . Du fait de la dualité échantillonnage-périodisation, l'échantillonnage du spectre équivaut à une répétition avec une période de $1/\Delta f$ de la tranche de signal d'origine.

Le nombre de points sur une période de signal vaut $Fe/\Delta f$, quantité qui est aussi le nombre de points sur une période de spectre.

Il s'ensuit qu'à N valeurs initiales correspondent N raies fréquentielles. De plus, ce spectre de raies présente une symétrie par rapport à la fréquence $\frac{Fe}{2}$ du fait de la remarque faite plus haut sur le changement $n \to N-n$. Seule est utile par conséquent la zone fréquentielle $\left[0 , \frac{Fe}{2}\right]$, comme l'indique le théorème de Shannon qui impose que la valeur de Fe soit au moins égale au double de la fréquence maximale que l'on veut étudier.

La figure B-8, reprenant le cercle unité dans le plan z , illustre cette symétrie du spectre de fréquence.

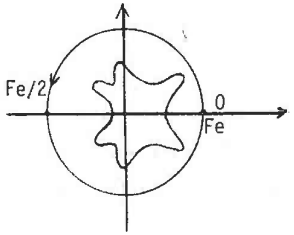


Fig. B-8 : Exemple de spectre de fréquence sous forme de diagramme polaire.

- contraintes dues au calcul numérique en général -

Pour résumer ce qui a été dit précédemment, les problèmes lors d'un filtrage numérique tel que la transformation de Fourier sont les suivants :

- chevauchement des spectres : il conviendra de prendre Fe le plus grand possible après un filtrage passe-bas, de façon à éliminer les fréquences au-delà d'une certaine limite qui dépendra du type de signal envisagé et du but recherché,

- étalement des fréquences dû à la réponse en fréquence de la fenêtre (qui ne serait une impulsion unité que pour une fenêtre en échelon, donc de durée infinie). Ceci impose le choix d'une fenêtre donc une pré-multiplication du signal. Il est particulièrement intéressant, pour un nombre de points de calcul préfixé, de conserver , en mémoire, les valeurs de la fenêtre discrète sous forme de tables, ainsi que les valeurs des fonctions trigonométriques nécessaires au calcul de la DFT. Dans le cas où il est destiné à fournir un spectre de puissance, il est possible d'envisager que ce calcul se fasse en parallèle avec l'acquisition du signal.

- concordance entre le nombre de valeurs de départ et le nombre de valeurs d'arrivée :
Il convient d'adapter le nombre de valeurs d'origine au nombre de raies recherchées dans le spectre en complétant la tranche de signal par autant de valeurs nulles qu'il est nécessaire. Il est clair qu'une plus grande résolution fréquentielle ne peut être obtenue qu'au détriment de la résolution temporelle, comme le veut le principe d'incertitude d'Eisenberg auquel on ne peut échapper.

Dans le cas où la tranche de signal d'entrée a été étendue par une séquence de zéros, il est possible de perfectionner l'algorithme de calcul de la DFT en évitant les multiplications inutiles comme proposé par MARKEL [MARK - 76]. Cette possibilité est utilisée, par exemple, pour l'obtention du spectre de prédiction linéaire (cf. IV.3), la réponse impulsionnelle du filtre dit "inverse" du conduit vocal étant réduite à un nombre fini limité de valeurs (filtre à réponse impulsionnelle finie).

c) Applications

Du fait des propriétés de la transformation de Fourier (nature de la transformée de signaux réels, décalage fréquentiel, convolutions et multiplications temporelles et fréquentielles, conservation de l'énergie totale, etc.), et grâce aux algorithmes de calcul rapide, la FFT est un outil utilisé à de multiples fins. Nous ne nous intéresserons ici qu'aux applications à notre domaine et nous indiquerons en 7) des extensions de la notion de DFT à un problème de filtrage plus adapté aux phénomènes transitoires de la parole :

- Calcul des spectres de puissance et de phase d'une tranche de signal temporel, spectres dits "à court terme" -

La FFT fournit une approximation de ces deux quantités (après les trois opérations d'échantillonnage satisfaisant le critère de SHANNON, de multiplication par une fenêtre temporelle et de périodisation) grâce au calcul de :

$$A^2(n) = X_r^2(n) + X_i^2(n) \quad , \; n = 0,1...(N \text{ DIV } 2)$$

et $$\phi(n) = \text{Arc tg} \frac{X_i(n)}{X_r(n)} \quad .$$

De façon générale, le spectre de phase ne présentera pas d'intérêt pour nous puisqu'il dépend fortement de tous les intermédiaires de la chaîne d'acquisition depuis les organes d'entrée (microphones, accéléromètres) jusqu'à l'échantillonneur-bloqueur du convertisseur analogique numérique. D'autre part, on a l'habitude de considérer que l'oreille est insensible à la phase (ce qui est à distinguer des problèmes de localisation de source sonore grâce à l'écoute binaurale dans laquelle intervient le déphasage entre les deux oreilles).

- Réponse en fréquence d'un filtre numérique -

La réponse en fréquence d'un filtre numérique de fonction de transfert $H(z)$ se traduit par le spectre de sa réponse impulsionnelle (cette dernière étant la réponse du filtre à une impulsion unité appliquée à son entrée). Dans le cas d'un système linéaire, donc d'un système tel que :

$$H(z) = \frac{NUM(z)}{DEN(z)}$$

où $NUM(z)$ et $DEN(z)$ sont des polynômes en z, la réponse impulsionnelle se calcule aisément par récurrence et la transformation de Fourier permet d'obtenir les parties réelle et imaginaire de $H(z)$ pour z évoluant sur le demi-cercle unité dans le plan z, alors que leur calcul direct se complique très vite lorsque les exposants des deux polynômes augmentent.

- Calcul de convolution -

L'intérêt réside dans le fait que la réponse des systèmes linéaires invariants dans le temps résulte de la convolution de leur réponse impulsionnelle avec le signal appliqué à l'entrée.

Comme rappelé plus haut, le calcul direct de la convolution de deux suites temporelles implique le retournement de l'une d'entre elles, une translation de l'une par rapport à l'autre et une somme de produits.

Beaucoup plus directement, la transformée de la réponse du système s'obtient par le produit simple, terme à terme, de la transformée du signal et de la réponse en fréquence du système (celle-ci envisagée comme une quantité complexe).

- Calcul de corrélation -

La fonction d'autocorrélation d'une portion de signal quasi-périodique calculée en vue de son analyse fréquentielle présente au moins trois types d'intérêt :

- elle permet, dans une certaine mesure, de détecter ce caractère de "quasi-périodicité" et de le chiffrer grâce à la détection des maxima de la fonction d'autocorrélation ; c'est ainsi que son intérêt est grand pour les portions voisées de parole auxquelles on veut appliquer des techniques d'analyse "*pitch synchrone*", c'est-à-dire en synchronisme avec l'écartement des cordes vocales (ce qui ne signifie pas pour autant que l'on travaille "en temps réel" au rythme d'émission de la parole),

- elle joue un rôle important dans le problème de la simulation du conduit vocal par un filtre numérique linéaire,

- alors que la fonction d'autocorrélation d'un signal périodique est elle-même périodique, celle d'un signal de bruit tend très rapidement vers zéro. Le calcul de corrélations permet la détection de périodicité noyée dans le bruit et l'évaluation chiffrée de la qualité de transmission d'un signal par une fonction de cohérence liée au rapport signal/bruit.

7. Extension de la notion de DFT à la définition de filtres numériques

Le calcul de la DFT d'un signal, de par son principe, fournit sa représentation dans une base de fonctions sinusoïdales non amorties.

Il est connu que le conduit vocal se comporte comme un résonateur présentant un certain facteur d'amortissement suivant les fréquences de résonance. Il est possible de lever la restriction à des fonctions sinusoïdales pures de la représentation en donnant à la variable z des valeurs $\{z_n\}$ telles que :

$$z_n = \rho_n \exp(-i\, 2\pi \frac{F_n}{F_e})$$

où ρ représente le module de z, donc la distance de son point représentatif à l'origine, dans le plan z.

Une valeur de ρ supérieure à 1 reviendrait à envisager la projection sur un filtre instable (sortie non bornée), alors que pour des valeurs de ρ inférieures à 1, il s'agit de la projection du signal sur des filtres "sinusoïdaux exponentiels", c'est-à-dire de sa projection sur une base de fonctions sinusoïdales amorties.

Nous reprenons ce point en B1.IV où l'idée de représentation fréquentielle d'un signal sera plus particulièrement envisagée.

II - DETECTION DE LA FREQUENCE FONDAMENTALE DE LA VOIX NOTEE F_0

1. Généralités

La fréquence de vibration des cordes vocales dans la production des sons voisés se traduit par la quasi-périodicité de l'onde glottale à large spectre émise au niveau du larynx.

Cette fréquence se retrouve au niveau du son produit et correspond au "pitch" (ou idée de hauteur) qui en est la manifestation subjective. Son étude permet d'observer la mélodie de la parole qui est l'évolution temporelle du fondamental et dont l'importance est grande au niveau suprasegmental, en particulier en rééducation de la parole. Nous allons donc passer en revue les principales méthodes utilisées.

Un grand nombre de techniques d'analyse permettent d'atteindre ce paramètre. Elles se fondent la plupart du temps sur le fait qu'il correspond à une oscillation en basse fréquence et que chaque montée de l'onde glottale s'accompagne d'un brusque accroissement de l'énergie acoustique émise (nous reprenons au paragraphe III cette notion d'énergie et ses modes d'évaluation).

Il s'ensuit que généralement les méthodes d'extraction du fondamental effectuent un filtrage passe-bas préalable suivi d'une détection de crêtes comme dans les extracteurs de pitch réalisés vers 1970 [GOLD - 69], [BOOT - 73] ou de la recherche de variations brusques d'énergie. Citons d'abord comme tout premier appareil l'*Electroglottographe* de FABRE où les variations d'impédance de la glotte mesurées au niveau du cartilage thyroïde étaient converties en signal basse-fréquence.

2. Méthodes de calcul

Nous classons, ci-dessous, les méthodes de calcul de fondamental suivant le caractère auquel elles font appel :

1 la quasi-périodicité de la fonction d'autocorrélation avec des maxima à intervalles de temps réguliers [DAVI - 57], [MARK - 72], [SERI - 74],

2 l'ajustement possible entre deux fenêtres temporelles distantes d'une période grâce à la recherche [ROSS - 74] de la valeur τ qui maximise la quantité :

$$\int_{\text{durée } T} |s(t) - s(t-\tau)|dt$$

où $s(t)$ désigne le signal d'origine.

Cette méthode est à rapprocher de l'application au signal d'un filtre-peigne de fonction de transfert :

$$H(z) = 1 - z^{-m} .$$

Un tel filtre possède m zéros de transmission régulièrement espacés. L'application de ce filtre à une séquence de parole et le calcul de l'énergie de sortie reviennent au calcul de la quantité :

$$\Sigma_{i=0}^{k-1} (s_{n+i} - s_{n+i-m})^2$$

dont les variations sont pratiquement celles de

$$\Sigma_{i=0}^{k-1} |s_{n+i} - s_{n+i-m}| .$$

On recherche la quantité m qui minimise l'une ou l'autre de ces expressions. La période recherchée est estimée à m fois la période d'échantillonnage,

[3] le fait que les maxima d'énergie de l'onde glottale aux instants d'ouverture des cordes vocales se retrouvent dans le signal.

MAISSIS [MAIS - 73], par exemple, évalue la quantité discrète équivalant à :

$$m(\tau) = \int_{\tau}^{\tau+T} s^2(t)dt - \int_{\tau-T}^{\tau} s^2(t)dt$$

où T est un intervalle del'ordre du quart de la valeur estimée de la période. Les pics de la fonction $m(\tau)$ donnent les instants de variation maximale d'énergie et permettent de repérer les débuts de période, ce qui est un avantage sur les méthodes précédentes.

DOURS et FACCA [DOUR - 74b], sur une idée du même type, calculent la quantité :

$$FS_i = \sum_{j=i}^{i+n_T-1} |s_j| - \sum_{j=i-n_T}^{i-1} |s_j|$$

qui peut s'évaluer par récurrence. La recherche des maxima successifs de FS se fait ensuite sur des intervalles réguliers moyennant un certain critère de régularité.

Le détecteur de mélodie de ZURCHER [ZURC - 77], dit *"mélographe"*, que nous utilisons dans nos expériences combine les avantages du principe de GOLD et de la méthode de MAISSIS. Après un filtrage passe-bas à 350 Hz, une normalisation des enveloppes de crêtes par compression de dynamique permet la détection des maxima. Comme c'est le cas couramment, un algorithme de correction intervient ensuite pour éliminer ou corriger les valeurs parasites en tenant compte des limites possibles de variation d'une période par rapport aux périodes voisines. Dans [ELMA - 77] on trouve en plus

la description d'un détecteur de voisement par préfiltrage et comptage du nombre de passages par zéro du signal,

[4] le caractère "basse-fréquence" du fondamental.
Un système contrôlé par un microprocesseur est proposé par MARTIN [MART - 77]. Le signal est traité par un banc de filtres passe-bas dont les caractéristiques sont sélectionnées par le microprocesseur en cours d'acquisition. Ce système peut être couplé à un convertisseur vidéo pour la visualisation de contours mélodiques sur moniteur TV.

Un appareil issu de dispositifs conçus spécialement pour les études en phonétique réalisé par TESTON [TEST - 77] met en jeu des filtres de BUTTERWORTH de fréquence centrale comprise entre 60 et 1 000 Hz. Le fondamental s'obtient au moyen d'un détecteur de seuil. L'intensité est déterminée parallèlement,

[5] les propriétés de l'onde glottale grâce à une atténuation de l'influence de la modulation introduite par le conduit vocal (cf. paragraphe VII) suivant différentes techniques après filtrage passe-bas. Citons par exemple :

- le calcul du signal d'erreur de prédiction linéaire. MARKEL [MARK - 76], dont l'un des centres d'intérêt est la modélisation du résonateur qu'est le conduit vocal, propose un algorithme dit *"SIFT Algorithm"* permettant la décision voisé-non voisé et le calcul éventuel de F_0 . A partir d'un modèle linéaire du conduit (filtre tout-pôles d'ordre 4), on calcule l'erreur de prédiction du modèle. Le signal est auparavant filtré par un filtre passe-bas coupant à 800 Hz. L'onde d'erreur met en évidence les instants d'ouverture de la glotte. La décision sur le voisement se fait pas comparaison du pic relatif de la fonction d'autocorrélation du signal d'erreur à une valeur limite prédéfinie. Dès que la valeur du pic dépasse le seuil, le calcul de l'intervalle $T_0 = 1/F_0$ est effectué,

- la détection de discontinuités dans l'onde glottale [DEMO - 77]. Après filtrage du signal dans cinq canaux à bande étroite, des critères de nature syntaxique permettent de faire la sélection entre ces canaux,

- l'atténuation sélective des formants. HESS [HESS - 81] propose d'approcher l'onde glottale en soumettant le signal de parole à l'action d'un filtre dit *"inverse"* constitué d'une cascade de deux filtres numériques non récursifs. Le premier effectue une transformation linéaire de façon à atténuer les formants de rang supérieur à 1 . Le second filtre élimine le premier formant après détermination de sa valeur par étude de l'histogramme des distances entre passages par zéro consécutifs. Les périodes de fondamental se déduisent ensuite de l'onde obtenue à la sortie de ce deuxième filtre,

- la méthode du *"cepstre"* proposée par NOLL [NOLL - 64] et développée ensuite [OPPE - 69], [SCHA - 72]. Un traitement logarithmique sur le module du spectre à court terme, une limitation à la zone des basses fréquences, puis une tranformation de Fourier permettent d'observer dans le cepstre ainsi obtenu, de nature temporelle, des pics marqués, séparés par la "quasipériode" du signal initial,

§ la détection du fondamental par l'oreille à partir de méthodes s'inspirant non plus d'un modèle du système phonatoire mais d'un modèle du système auditif périphérique. CAELEN et CAZENAVE [CAEL - 77] soumettent le signal à l'action d'un filtre récursif passe-bande de fréquence centrale adaptable. La fréquence fondamentale F_0 est ensuite détectée par comptage du nombre de passages par zéro du signal sur une fenêtre temporelle et évaluation des durées des écarts entre zéros consécutifs.

BASTET et al. [BAST - 77] modélisent la fonction de transfert de l'oreille moyenne par association de filtres d'ordres 1 et 2 et celle de l'oreille interne par association de filtres d'ordre 2 , les uns proches de l'amortissement critique, les autres surtendus. Après filtrage du signal, F_0 est détecté en temps réel grâce au comptage des passages par zéro.

3. Utilisation

La notion de détection en temps réel ou non joue un rôle important suivant le contexte de son utilisation. Pour notre part, nous nous sommes placé dans deux situations distinctes :

a) le but est de donner une représentation visuelle de la valeur instantanée de F_0 ou de son évolution dans une perspective de rééducation vocale. Il est fondamental que, dans un premier temps, la visualisation se fasse en temps réel. Nous utilisons un *"mélographe"* ou le système de détection d'un *"Vocoder"* CIT-Alcatel. Dans les deux cas, nous avons choisi une fréquence de 100 Hz soit de 100 valeurs délivrées par seconde. Un algorithme de correction testant le suivi des valeurs de F_0 et un algorithme de lissage précédent la visualisation,

b) la détection de F_0 est faite dans le cadre global de l'analyse d'une séquence de parole, soit pour une étude de ce paramètre associé ou non à d'autres, soit pour une étude du signal *"pitch synchrone"*. Dans ces deux cas, notre étude se fait en différé, à partir d'enregistrements sur bande magnétique (analogique) ou de segments de parole numérisée gardés en fichier sur disque. Des programmes d'acquisition (cf. C.1.I) de parole prétraitée permettent, à partir des bandes, de constituer des fichiers synchrones contenant le signal échantillonné, les données de l'analyseur spectral et le contour mélodique. Ce dernier contour, pour des raisons techniques de niveau et de qualité de l'enregistrement, nécessite un contrôle que nous faisons à partir du calcul direct de F_0 sur les portions voisées. Ce calcul est fait également lorsque les données sont le seul signal numérisé.
Pour cela, nous faisons une simple détection de crêtes négatives sur le signal temporel, définissant ainsi une succession de cycles dits *"cycles minmin"*, grâce à un algorithme de recherche des minima minimorum. Le calcul de la durée de ces cycles, avec possibilité de retour en arrière compte-tenu des valeurs plausibles de la période de fondamental, permet d'atteindre F_0 . Les ambiguïtés peuvent être levées grâce à un calcul local des variations d'énergie comme le fait MAISSIS par exemple.

Recherchons l'ordre de grandeur de la durée d'intégration pour le calcul des variations locales d'énergie. Nous faisons ce calcul sur la base d'un signal sinusoïdal non amorti, en $a.\sin\omega t$. La variation d'énergie, calculée sur deux tranches de signal de durée $T_\beta = \frac{\beta}{\omega}$ autour de l'instant $\frac{\alpha}{\omega}$, s'exprime par :

$$\Delta_\beta I(\alpha) = \frac{1}{T_\beta}\int_{\frac{\alpha}{\omega}}^{\frac{\alpha+\beta}{\omega}} a^2 \sin^2 \omega t \, dt - \frac{1}{T_\beta}\int_{\frac{\alpha-\beta}{\omega}}^{\frac{\alpha}{\omega}} a^2 \sin^2 \omega t \, dt$$

$$= \frac{a^2}{\beta} \sin^2\beta \sin 2\alpha \quad .$$

Le maximum de $\Delta_\beta I(\alpha)$ par rapport à β , qui vaut :

$\Delta I(\alpha) = 0.72\ a^2 \sin 2\alpha$, est obtenu pour $\frac{\sin^2\beta}{\beta}$ maximum soit pour

$\beta = \frac{\text{tg}\ \beta}{2}$ avec β non nul.

β vaut 1.17 radian = 67° soit un peu moins d'un cinquième de période.

Cette durée de $\frac{T_o}{5}$ environ, obtenue à partir d'une sérieuse simplification sur le signal, valide le choix d'une durée de $\frac{T_o}{4}$ généralement adoptée pour le calcul des variations locales d'énergie.

Le maximum de $\Delta I(\alpha)$ apparaîtra pour $\alpha = \frac{\pi}{4}$ soit un huitième de période après l'amorce de la montée de la sinusoïde.

Pour s'affranchir des variations de niveau du signal, il est possible de normaliser l'équivalent discret de $\Delta I(\alpha)$ par rapport à l'énergie totale des deux tranches voisines étudiées.

III - CALCUL DE L'INTENSITE

Le terme "intensité" est fréquemment employé pour caractériser l'intensité (ou puissance) acoustique du signal de parole. Au facteur temps près, elle correspond à l'énergie par unité de surface du signal acoustique intégrée avec une certaine constante de temps.

Sa manifestation subjective est connue sous le nom d'"*intensité subjective*" dont les courbes isophoniques de FLETCHER (1923) donnent l'évolution en fonction du niveau sonore objectif. Nous n'insistons pas sur cette notion que nous n'envisageons pas dans notre travail.

Une idée de l'intensité sonore est fournie par l'enveloppe du signal électrique d'un microphone dont les variations traduisent les variations de pression sonore. Cette courbe enveloppe fait d'ailleurs partie des premiers facteurs de la parole visualisés pour les déficients auditifs puisqu'elle permet d'observer la présence ou non de manifestation vocale et les écarts dans l'élocution de sons différents.

Nous emploierons le mot *intensité* du signal $s(t)$ pour désigner l'énergie moyenne par unité de temps , suivant la formule :

$$I(t) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} s^2(\tau)\, d\tau \qquad (1)$$

ou sa racine carrée (2) .

Ces formules qui exigent des multiplications peuvent souvent être remplacées sans perte d'information notable par :

$$I'(t) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} |s(\tau)|\,d\tau \qquad (3)$$

qui effectue en fait la somme arithmétique des aires des surfaces limitées par la courbe et l'axe de tension nulle.

Le principe des sonomètres et la description d'un intensimètre pour la détection objective de l'intensité de la parole sont décrits en particulier dans TESTON [TEST - 78]. Les problèmes posés par la détection analogique dans le cas particulier des sons du langage concernent la durée d'intégration, le type de capteur et l'environnement à respecter pour tenir compte du rayonnement aux lèvres sensiblement sphérique, les corrections éventuelles à partir des courbes isophoniques pour traduire la sensation auditive, etc.

L'une des questions qui se posent pour le calcul direct de l'intensité est en effet le choix de la durée T de la fenêtre temporelle. Ce choix est tributaire de l'application envisagée comme dans les deux cas suivants :

a) on veut mettre en évidence des variations rapides traduisant l'évolution locale de l'énergie (comme au paragraphe II). On prendra pour T une valeur de l'ordre de quelques millisecondes,

b) on veut, au contraire, limiter l'effet des fluctuations locales pour l'observation par exemple d'un signal stationnaire pris dans son ensemble : on calcule alors l'énergie moyenne sur une durée de quelques dizaines de millisecondes choisie en fonction du locuteur.

Recherchons la valeur T de la constante d'intégration nécessaire pour que les fluctuations dues au fondamental exercent une influence minime sur le calcul de l'intensité. Pour cela, nous referons l'hypothèse simplifiée d'un signal sinusoïdal en $a.\sin \omega t$. Effectuons le calcul de l'énergie sur une durée de signal : $T\beta = \frac{\beta}{\omega}$, entre deux instants $\frac{\alpha}{\omega}$ et $\frac{\alpha+\beta}{\omega}$:

$$I_\beta(\alpha) = \frac{1}{T_\beta}\int_{\omega t=\alpha}^{\alpha+\beta} a^2 \sin^2 \omega t \, dt = \frac{a^2}{\beta}\int_{\alpha}^{\alpha+\beta} \sin^2\theta \, d\theta$$

$$I_\beta(\alpha) = \frac{a^2}{2}\left[\, 1 - \frac{\sin\beta}{\beta}\cos(2\alpha+\beta)\,\right] .$$

La quantité $\cos(2\alpha+\beta)$ oscillant entre les valeurs $+1$ et -1 , il convient de choisir β grand devant un radian. Le choix, par exemple, d'une limite de 5% pour le rapport $\frac{1}{\beta}$ entraîne :

$\beta = 20$ radians $\cong 6\pi$ soient 3 périodes de fondamental.

Le rapport de contraste vaut alors $\frac{1 + 0.05}{1 - 0.05} = 1.1$.

Cette durée de trois périodes équivaut à une durée de 12 ms pour un fondamental de 250 Hz et de 30 ms pour 100 Hz , soit une fenêtre temporelle de respectivement $N = 120$ et 300 échantillons pour un échantillonnage à 10 kHz .

L'intensité est donnée par :

$$I_N(i) = \sum_{m=i}^{i+N-1} s^2(m) .$$

Le "vocodeur" à canaux de CIT-Alcatel avec lequel nous travaillons fournit une valeur de l'intensité dans chacune des quinze zones fréquentielles qui le composent toutes les 10 ou 20 ms .

La constante d'intégration pour le calcul temporel de l'intensité doit être du même ordre que celle des analyseurs spectraux. Ce fait se déduit de la relation de PARSEVAL qui traduit l'égalité de l'énergie d'une tranche de signal calculée à partir du spectre à court terme (obtenu, par exemple, par DFT) et de l'énergie calculée à partir de l'onde temporelle $\{s_i\}$.

Le calcul de l'intensité à partir d'une représentation spectrale présente un intérêt pratique lorsque cette représentation est directement accessible, comme c'est le cas à partir des analyseurs spectraux en temps réel, pour deux raisons :

- . l'aspect temps réel fondamental dans certaines de nos applications,
- . le fait qu'en se limitant à des bandes de fréquences particulières, on peut atteindre d'autres paramètres tels que le degré de friction, le taux de basses fréquences, des centres de gravité de l'énergie et des paramètres traduisant par exemple *"la couleur de la voix"*.

Comme dans le cas de la fréquence fondamentale d'une onde voisée, le calcul de l'intensité peut se faire par des techniques différentes suivant le but recherché :

a) dans la nécessité d'avoir l'évolution de l'intensité en temps réel, nous utilisons les données fournies par l'analyseur spectral. En admettant, ce qui est une sérieuse approximation du fait de la compression de dynamique, que les valeurs de sortie des canaux de l'analyseur

correspondent à l'énergie dans chaque bande de fréquence, nous calculons l'énergie totale en effectuant la somme de ces valeurs de sortie.

Si l'on envisageait ces valeurs comme des densités d'énergie, il faudrait les pondérer par des facteurs correspondant aux largeurs de bande. En fait, pour tenir compte de la compression de dynamique, il conviendrait d'appliquer des facteurs dépendant du canal mais variant moins vite que sa largeur de bande. Nous n'avons pas envisagé ce point de vue mais il semble possible de tenter une détermination de ces facteurs par apprentissage,

b) lors des traitements en temps différé, à partir du signal numérisé, nous calculons directement le paramètre "intensité" suivant les relations (2) ou (3).

Ces deux quantités étant de pertinence pratiquement égale, tant au niveau de la caractérisation d'une zone de parole quasi stationnaire qu'au niveau de la segmentation en noyaux élémentaires, la deuxième est retenue le plus couramment.

IV - SPECTRES DE FREQUENCE

L'analyse fréquentielle du signal vocal conduit à sa représentation sur une base de fonctions sinusoïdales dans la gamme des fréquences audibles. On se limite généralement à une plage moins étendue, jusqu'à 5 000 Hz environ, fréquence au-delà de laquelle la représentation perd son caractère pertinent.

Le diagramme amplitude carrée/fréquence, à un instant donné, constitue le spectre de puissance du signal. Sa détermination suppose le traitement d'une séquence de parole de durée non négligeable, d'où le nom de spectre à court terme plutôt que spectre instantané.

Si l'on s'intéresse à l'évolution du spectre dans le temps, on peut utiliser la représentation par spectrogramme amplitude carrée / temps/ fréquence dans un espace à trois dimensions ou temps/ fréquence dans un espace plan où l'amplitude est figurée par un symbole, un niveau de brillance ou une couleur.

Nous décrivons ici les moyens que nous utilisons pour atteindre la représentation fréquentielle du signal.

1. Le Vocodeur ou "Vocoder" à canaux

Nous disposons d'un analyseur spectral CIT-Alcatel à 15 canaux. Il fournit 50 ou 100 fois par seconde un message série que nous appelons "*prélèvement vocodeur*" comprenant :

- les valeurs de l'intensité du signal à la sortie de quinze filtres de largeurs de bandes contigues à 3dB d'atténuation et couvrant la plage 250 - 5100 Hz . Chacune de ces valeurs est codée sur 3 bits,
- la valeur de la fréquence fondamentale pour les séquences voisées, codée sur 8 bits.

2. La transformation rapide de Fourier

La FFT, algorithme rapide de calcul de la transformée discrète de Fourier (cf. paragraphe I.6 de ce chapitre), effectue le calcul de :

$$X(z) = \frac{1}{N} \sum_{k=0}^{N-1} x(k)\ z^{-k}$$

à partir d'une séquence de valeurs d'origine $\{x(k), k=0, 1 \ldots N-1\}$, pour des valeurs de la variable complexe z régulièrement réparties sur le cercle unité dans le plan z .

Si $X(z)$ représente la transformée en z du signal vocal échantillonné, donc constitué uniquement de valeurs réelles, le calcul de $|X(z)|^2$ fournit un spectre symétrique par rapport à son milieu. On se limite alors au demi-cercle unité positif qui correspond aux fréquences allant de 0 à $\frac{Fe}{2}$.

Comme pour tout traitement numérique de la parole, le calcul direct par FFT de la transformée de Fourier s'accompagne d'un prétraitement de l'onde de parole :

- Relèvement des aigus de 6 dB par octave -

Le rayonnement aux lèvres (cf. paragraphe IV.3) intervient comme un facteur d'intégration du signal. Il est nécessaire de renforcer les fréquences aigues . Pour ce faire, on calcule habituellement la pente du signal $\{x(n)\}$ suivant :

$$s(n) = x(n+1) - x(n) \quad \text{pour} \quad n \geqslant 0 .$$

Si l'on envisage à court terme (sur une période de fondamental de signal voisé) le signal $x(n)$ comme une superposition de sinusoïdes amorties :

$$x(n) = \sum_i a_i \rho_i^n \cos(\omega_i \, nTe + \varphi_i)$$

$$= \sum_i a_i \rho_i^n \cos \theta_i(n)$$

où Te désigne toujours la période d'échantillonnage et où les quantités a_i, ρ_i, ω_i et φ_i correspondent à l'amplitude d'origine, au facteur d'amortissement, à la pulsation et à la phase de la $i^{ème}$ sinusoïde, on a alors pour $s(n)$:

$$s(n) = \sum_i a_i \left[\text{Log } \rho_i \times \rho_i^n \cos \theta_i(n) - \rho_i^n \omega_i \, Te \sin \theta_i(n)\right]$$

Pour un amortissement faible tel que ρ_i est très voisin de 1 il vient :

$$s(n) \simeq Te \sum_i a_i \rho_i^n \omega_i \cos \left[\theta_i(n) + \frac{\pi}{2}\right] .$$

L'opération de différenciation revient alors à la multiplication de l'amplitude de la sinusoïde de pulsation ω_i par la quantité ω_i . Pour deux fréquences séparées par une octave, on vérifie que le rapport des amplitudes carrées rapporté en échelle logarithmique s'accroît de 6dB .

- Multiplication par une fenêtre temporelle -

Pour limiter l'effet de troncature du signal en vue de son étude à court terme, on le prémultiplie par une fenêtre temporelle dont le spectre renferme le moins possible de lobes parasites. Les choix les plus courants sont :

- *la fenêtre de HANN* donnée par :

$$f(nTe) = 0.5 - 0.5 \cos \left[2\pi \frac{n}{N-1}\right] \quad \text{pour} \quad n \in [0, N-1]$$

et $f(nTe) = 0$ ailleurs.

- *la fenêtre de HAMMING* qui garantit une valeur faible du premier lobe parasite mais une décroissance moins rapide des lobes suivants et qui est donnée par :

$$f(nTe) = 0.54 - 0.46 \cos \left[2\pi \frac{n}{N-1}\right] \quad \text{pour} \quad n \in [0, N-1]$$

et $f(nTe) = 0$ ailleurs.

Nous utilisons de préférence la *fenêtre de HAMMING* dont nous mémorisons les éléments de façon à réduire les temps de calcul dans le cas de segments de durée fixe.

- Application au cas de la parole -

Une zone de signal de parole voisée résulte de la répétition dans le temps, à intervalles à peu près réguliers, d'une superposition de sinusoïdes amorties présentant ensemble un maximum d'énergie aux instants d'ouverture des cordes vocales. Cet effet de répétition se traduit par une convolution, au niveau du spectre de Fourier, du spectre du signal non répété avec le pic correspondant à la fréquence fondamentale. Cette hachure du spectre masque les fréquences de résonance qui n'apparaissent pas nettement. Ce résultat conservant l'information sur l'onde glottale et l'effet du conduit vocal permet cependant :

- d'approcher la valeur de ces fréquences de résonance dites "*formants*" par détection des maxima maximorun du spectre,
- de déduire l'allure d'un spectre lissé, par exemple, par recherche de la courbe des maxima, de celle des minima puis interpolation,
- par suite, de calculer approximativement l'intensité globale et l'intensité dans des zones fréquentielles sélectionnées donc de simuler un banc de filtres facilement ajustable,
- de déduire une valeur approchée $\tilde{F}_0$ de la fréquence fondamentale de la voix par détection du nombre de maxima NMAX sur la zone de fréquence $[0, \frac{Fe}{2}]$. On a alors :

$$\tilde{F}_0 = \frac{Fe}{2 * NMAX} \quad .$$

La figure B-9 illustre la progression suivie dans ce paragraphe en donnant successivement :

[1] une tranche de parole échantillonnée à 10 kHz , composée d'à peu près trois "quasi-périodes" de fondamental (256 points soit 25,6 ms),

[2] le même signal différentié pour la préaccentuation des aigus et multiplié par la *fenêtre de HAMMING*, en vue du traitement fréquentiel,

[3] le spectre obtenu par FFT composé de 129 points pour des valeurs de la fréquence régulièrement espacées de 0 à 5 000 Hz . Les maxima maximorum sont mis en évidence. Le nombre des maxima étant de 38 , on peut déduire pour valeur approchée du fondamental la fréquence :

$$\tilde{F}_0 = \frac{10\ 000}{2 \times 38} = 131 \text{ Hz}$$

ce qui, sur 256 points , correspond à 3.4 périodes de fondamental,

[4] le codage sur 16 niveaux de la densité spectrale de puissance dans 16 bandes fréquentielles.

Il est clair que le simple passage au domaine fréquentiel à partir du signal d'origine, sans autre traitement, conserve l'information relative à la source. Or, le plus souvent, on attend du spectre de fréquence une information relative à l'articulation des sons. Pour ce faire, on introduit un type de filtrage faisant abstraction de la source et ne prenant en compte que le conduit vocal. Cette technique visant à rechercher la cause en étudiant les effets est dite "*technique de filtrage inverse*".

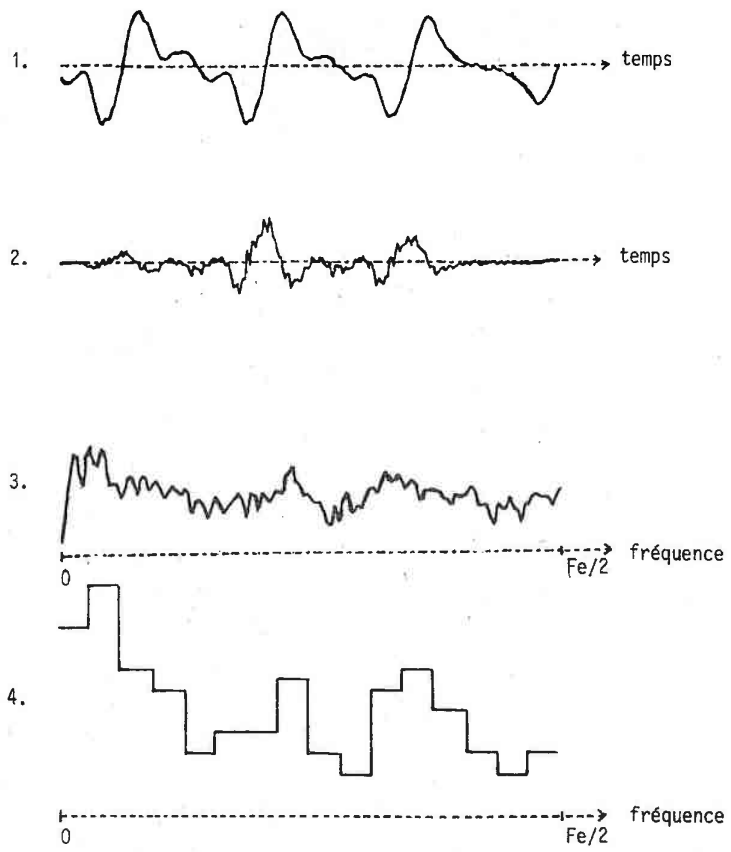


Fig. B-9 : Traitement d'une tranche de signal par transformation de Fourier discrète.

3. Le filtrage inverse

a) Généralités

L'analyse par filtrage inverse consiste à construire un modèle numérique simple du conduit vocal en le considérant comme un filtre variant dans le temps caractérisé à un instant donné par une fonction de transfert $H(z)$.

La figure B-10 donne une schématisation du canal vocal ainsi que sa transposition sous forme d'un système vu sous l'angle de l'automatique :

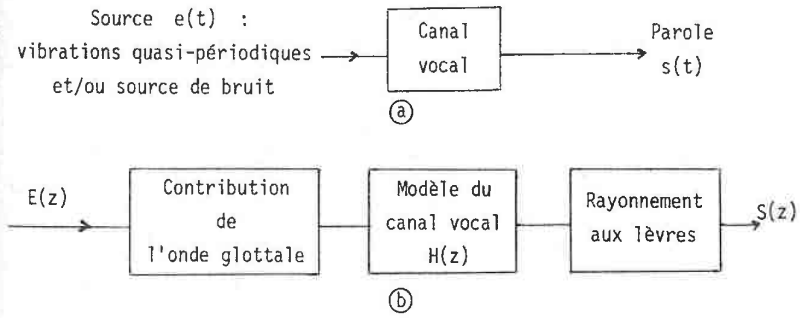


Fig. B-10 : Modèle du canal vocal
ⓐ domaine temporel
ⓑ domaine z

Le problème se pose de représenter les différentes influences qui contribuent à former l'onde de parole.

Lorsque le son est voisé, sur une période de fondamental l'onde glottale intervient par un terme $Ug(z)$ qui peut être mis sous la forme [ATAL - 71] :

$$Ug(z) = \frac{K_1}{(1 - z_a\, z^{-1})(1 - z_b\, z^{-1})}$$

où K_1 caractérise l'amplitude et où z_a et z_b sont des pôles réels de module inférieur à 1 (z_a très proche de 1).

Le rayonnement aux lèvres correspondant à peu près à la propagation d'une onde sphérique peut se représenter par le terme :

$$R(z) = K_2\, (1 - z^{-1})$$

où K_2 dépend de l'amplitude de l'onde aux lèvres et de la distance lèvres-microphone [FLAN - 72]. Ce terme correspond à une intégration du signal donc une atténuation des aigus.

Ainsi, d'après nos conventions d'écriture, il vient :

$$S(z) = R(z)\ .\ H(z)\ .\ Ug(z)\ .\ E(z)\ .$$

Dans cette expression, la quantité $R(z)\ .\ Ug(z)$ peut se réécrire, à une petite erreur près :

$$R(z)\ .\ Ug(z) = \frac{K_1\, K_2}{[1 + (1-z_a)\, z^{-1}][1 - z_b\, z^{-1}]} \qquad (1)$$

Dans un modèle linéaire pôles seulement ("*all-pole*"), on admet que le conduit vocal se comporte comme un filtre de fonction de transfert

$$H(z) = \frac{\gamma}{\prod_{i=1}^{I} (1 - R_i\, z^{-1} + M_i\, z^{-2})} \qquad (2)$$

Chacun des termes du produit au dénominateur correspond à un mode de résonance, dit aussi formant.

Considérons séparément chacun des termes de ce produit :

$$(1 - R_i\, z^{-1} + M_i\, z^{-2}) = (1 - z_i\, z^{-1})(1 - z_i^*\, z^{-1})$$

où z_i et z_i^* sont deux nombres complexes conjugués, situés à l'intérieur du cercle unité, sous l'hypothèse de stabilité du système.

Ce modèle "*all-pole*" est bien représentatif des sons voisés sans nasalité [FANT - 60]. Pour les nasales et les sons non voisés, les zéros de la fonction de transfert sont à l'intérieur du cercle unité et peuvent être approchés par des pôles au dénominateur [ATAL - 71].

Il se trouve de plus que du point de vue perceptif, la localisation d'un pôle est beaucoup plus importante que celle d'un zéro et que dans la bande de fréquence allant de 0 à 5 000 Hz environ, l'effet des antirésonances (zéros) peut être approché par un modèle "*all-pole*".

A partir des relations (1) et (2) , il vient qu'un filtre "*all-pole*" peut tenir compte globalement des contributions de l'impédance glottique, du rayonnement aux lèvres et du conduit vocal proprement dit couplé ou non avec le conduit nasal.

Ce dernier point est particulièrement intéressant ; nous avons en effet tenté de modéliser l'ensemble du conduit vocal associé au conduit nasal en parallèle, suivant la figure B-11 :

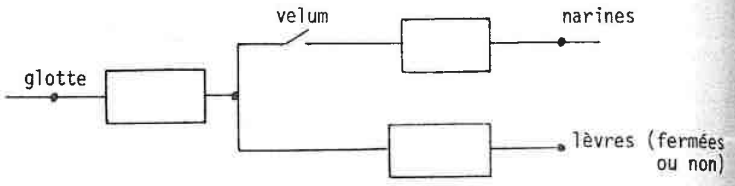


Fig. B-11 : Couplage du conduit vocal et du conduit nasal.

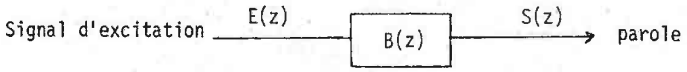
Il faut considérer, en plus, que l'impédance de sortie des deux conduits diffère suivant le type de production (consonnes nasales ou voyelles nasalisées).

Le problème mathématique conduit à la résolution d'un système d'équations non linéaires complexe pour laquelle il n'existe pas d'algorithme simple ; c'est un des problèmes posés par la détection automatique de la nasalité, qui n'est pas résolu.

Pour en revenir au modèle général *"all-pole"*, notons :

$$B(z) = Ug(z) \cdot H(z) \cdot R(z)$$

la fonction de transfert du filtre modélisant le conduit vocal comme suit :



B(z) se met sous la forme :

$$B(z) = \frac{\gamma}{\sum_{p=0}^{P} a_p z^{-p}} = \frac{\gamma}{A(z)} \qquad (3)$$

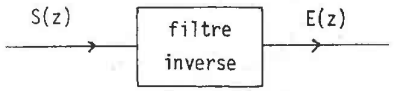
où γ est une quantité constante déterminée par le choix de a_o que l'on prend égal à 1 par commodité.

b) Le filtrage inverse linéaire à l'ordre P

Ce type de filtrage revient à rechercher la fonction de transfert du filtre dont le comportement serait l'inverse de celui du conduit vocal, en considérant en plus qu'elle est donnée sous la forme de l'expression (3) du paragraphe a) précédent, c'est-à-dire :

$$B(z) = \frac{\gamma}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{S(z)}{E(z)}$$

Le filtre inverse peut être schématisé comme suit :



Il admet pour fonction de transfert la fonction A(z) vérifiant :

$$A(z) = \frac{E(z)}{S(z)} = 1 + \sum_{p=1}^{P} a_p z^{-p}$$

au facteur γ près, ce qui dans le domaine temporel s'écrit :

$$e(n) = s(n) + \sum_{p=1}^{P} a_p \, s(n-p) \qquad (4) \quad .$$

Moyennant une erreur $e(n)$, cette expression équivaut à prédire l'échantillon $s(n)$ à l'instant nTe à partir des échantillons de sortie précédents et ceci suivant une relation linéaire. Cette approche porte le nom de *"Codage prédictif linéaire"* (ou LPC en abréviation de *"Linear Predictive Coding"*). La succession des $e(n)$ compose le signal d'erreur de prédiction linéaire.

Le filtre inverse s'obtient par exemple par minimisation de ce signal d'erreur au sens des moindres carrés, donc par la recherche du minimum sur les coefficients a_p de la quantité :

$$E = \sum_{n} \left[s(n) + \sum_{p=1}^{P} a_p \, s(n-p) \right]^2 \qquad (5) \quad .$$

Le minimum de E s'obtient pour $\frac{\partial E}{\partial a_p} = 0$, $p \in [1, P]$.

Suivant que l'on considère tous les élements de sortie $s(n)$ ou non, on a affaire à deux catégories de méthodes : les unes dites stationnaires (ou d'autocorrélation), les autres non stationnaires (ou de covariance). On trouve de nombreuses formulations de ces méthodes rassemblées dans [MARK - 76]. Une formalisation introduisant notamment la transformation de KARHUNEN-LOEVE ouvre le champ de l'analyse factorielle au traitement du signal [CARA - 76].

Toutes ces méthodes peuvent se distinguer suivant qu'elles travaillent sur une période de fondamental ("pitch-synchrones" avec, parmi elles, la méthode de PINSON [PINS - 62]) ou non ("pitch-asynchrones").

Les méthodes pitch-asynchrones ne nécessitent pas la détection des instants d'ouverture des cordes vocales mais exigent un traitement sur la durée d'au moins deux périodes de fondamental [MAKH - 75]. D'après [CHAN - 74], seule la formulation stationnaire garantit la stabilité du modèle.

Pour notre part, nous avons adopté la formulation qui nous a paru la plus simple, compte-tenu des remarques précédentes : le filtrage pitch-asynchrone stationnaire décrit par exemple dans [MARK - 73]. Ceci nous permet :

- d'une part, de travailler indépendamment de la détection des segments voisés et de la période de fondamental,

- d'autre part, d'aboutir à une résolution mathématique de type convergent et récurrent.

La minimisation du signal d'erreur (relation (5)) suivant le critère des moindres carrés à partir de N valeurs du signal de sortie $\{s(n), n=0,\ldots, N-1\}$, les valeurs de $s(n)$ en dehors de cet intervalle étant considérées comme nulles, conduit à la résolution du système d'équations linéaires suivant :

$$\sum_{p=1}^{P} a_p \, S_{|p-k|} = - S_k \quad , \; k=1,\ldots, P \qquad (6) \quad .$$

Dans ce système, $S_{|p-k|}$ et S_k désignent les valeurs prises par la fonction d'autocorrélation de la tranche de signal $\{s(n)\}$ pour les retards respectifs $|p-k|$ et k, ce qui justifie le nom de *"méthode d'autocorrélation"* donné à cette formulation.

La résolution du système (6) peut se faire grâce à une variation récurrente des coefficients a_p à partir d'une valeur $a_{p/p}$ initiale jusqu'à une valeur $a_{p/P}$ définitive à l'étape P.

Soient donc $a_{0/p}$, $a_{1/p}$, ... $a_{p/p}$ les coefficients du filtre inverse calculés à l'étape p. A l'étape suivante, on a :

$$\begin{aligned} a_{0/p+1} &= a_{0/p} = 1 \\ a_{1/p+1} &= a_{1/p} + K_p\, a_{p/p} \qquad (7) \\ &\vdots \\ a_{k/p+1} &= a_{k/p} + K_p\, a_{p+1-k/p} \\ &\vdots \\ a_{p+1/p+1} &= K_p \,, \end{aligned}$$

expressions dans lesquelles :

$$K_p = - \frac{\sum_{k=0}^{p} a_{k/p}\, S_{p+1-k}}{\sum_{k=0}^{p} a_{k/p}\, S_k} \qquad (8)$$

Les coefficients K_p ainsi calculés sont directement liés à la configuration du conduit vocal, comme il sera dit plus loin (par. VI.2).

A l'étape P, on obtient la fonction de transfert $A(z)$ du filtre inverse. On montre [WAKI - 73] que c'est aussi la fonction de transfert d'un modèle du conduit vocal en P tubes de longueurs égales si P satisfait à la condition :

$$Fe = \frac{Pv}{2L} \,,$$

où v est la vitesse du son et L la longueur du conduit (de l'ordre de 340 m/s et 17 cm respectivement). Ces valeurs conduisent à :

$$P \approx \frac{Fe}{1\,000}$$

ou bien $P \approx Fe$ si l'on exprime Fe en kHz.

Ainsi, pour une fréquence d'échantillonnage de 10 kHz qui permet d'étudier les spectres de fréquence jusqu'à 4.5 à 5 kHz, la valeur théorique à donner à P est 10. On peut augmenter la valeur de P pour tenir compte des effets de $Ug(z)$ et $R(z)$ et pour traiter le cas des nasales.

Pour revenir au problème qui nous intéresse ici de la détermination du spectre de fréquence à court terme d'un élément de parole, une approximation de ce spectre est obtenue en considérant que $\frac{1}{A(z)}$ est une approximation de la fonction de transfert du conduit vocal à une constante multiplicative près.

Il suffit alors de calculer la quantité $\frac{1}{|A(z)|^2}$ pour des valeurs de z réparties sur le demi-cercle unité supérieur.

La figure B-12 indique le résultat obtenu dans le calcul du spectre à court terme par cette méthode pour le segment de parole étudié figure B-9, dans les mêmes conditions d'échantillonnage et de pré-traitement pour un ordre de prédiction linéaire égal à 12 .
Sont présentés successivement :

[1] le spectre de prédiction linéaire superposé au spectre obtenu par transformation de Fourier,

[2] le même spectre avec, en coïncidence, les contributions individuelles des pôles de la fonction de transfert simulant le conduit vocal,

[3] la position de ces pôles dans le plan z (cf. paragraphe V.3.c),

[4] la fonction d'aire du conduit vocal (cf. paragraphe VI.2).

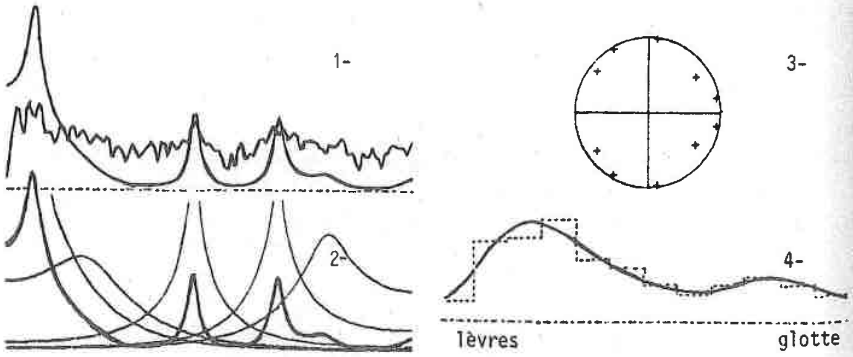


Fig. B-12 : Traitement du segment de parole de la figure B-9 par prédiction linéaire

4. Projections du signal à travers des cellules passe-bande d'ordre 2

a) Formulation

L'idée de cette méthode est d'éviter le calcul systématique donné par FFT des projections du signal sur une base de fonctions sinusoïdales aux fréquences régulièrement réparties et qui ne sont pas forcément toutes pertinentes ou caractéristiques.

Nous avons alors recherché le moyen de représenter le signal sur une base qui serait adaptée au corpus étudié et qui répondrait à certaines exigences de pertinence ou de discrimination [HATO - 76].

Une cellule élémentaire passe-bande peut être simulée par l'équivalent électrique suivant (fig. B-13) :

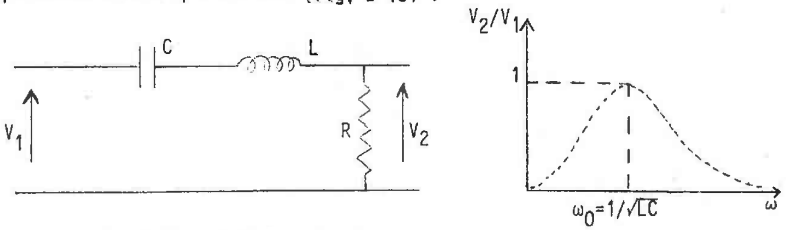


Fig. B-13 : Cellule élémentaire passe-bande

Si l'on pose $\omega_0 = \frac{1}{\sqrt{LC}}$, fréquence de résonance du circuit

RLC série et $Q = \frac{1}{RC\omega_0}$, coefficient de surtension du circuit,

la fonction de transfert en s du quadripôle s'écrit :

$$H(s) = \frac{s/Q\omega_0}{s^2/\omega_0^2 + s/Q\omega_0 + 1}$$

$$= \frac{As}{(s-s_1)(s-s_1^*)}$$

En se plaçant dans le plan $z = e^{sTe}$, il vient, au coefficient A près :

$$H(z) = \int_C \frac{H(s)}{(1 - e^{sTe} z^{-1})} \frac{ds}{2\pi i}$$

où C est un contour englobant les singularités de H(s) et elles seules.

D'après le théorème des résidus, on a :

$$H(z) = \frac{s_1}{(s_1 - s_1^*)(1 - e^{s_1 Te} . z^{-1})} + \frac{s_1^*}{(s_1^* - s_1)(1 - e^{s_1^* Te} . z^{-1})}$$

ou, si l'on pose $s_1 = -\sigma + i\omega_1$:

$$H(z) = \frac{1 - \exp(-\sigma Te) \left[\frac{\sigma}{\omega_1} \sin \omega_1 Te + \cos \omega_1 Te\right] z^{-1}}{(1 - z_1 z^{-1})(1 - z_1^* z^{-1})} ,$$

soit encore, avec nos conventions d'écriture (paragraphe I.3 a de ce chapitre) qui sont : $z_1 = \rho \exp(i\theta)$

$$|z_1|^2 = M = \rho^2$$

$$\text{et } Re(z_1) = \frac{R}{2} = \rho \cos \theta ,$$

$$H(z) = \frac{1 - \left[\frac{\sigma}{\omega_1} \sqrt{M - R^2/4} + R/2\right] z^{-1}}{1 - Rz^{-1} + Mz^{-2}}$$

La correspondance entre s_1 et z_1 est celle de la figure B-5.

On peut remarquer que, lorsque l'amortissement est faible (σ petit devant ω_1), cette expression se ramène à :

$$C(z) = \frac{1 - R/2\, z^{-1}}{1 - Rz^{-1} + Mz^{-2}}$$

qui est, comme nous l'avons vu (paragraphe I.3 a) , la transformée en z de la fonction cosinus amortie. (On démontre facilement que C(z) équivaut, dans le domaine s , à : $\frac{s + \sigma}{(s - s_1)(s - s_1^*)}$).

De façon générale, dans les fonctions de transfert du type :

$$\psi_\mu(z) = \frac{1 - \mu z^{-1}}{1 - Rz^{-1} + Mz^{-2}}$$

μ détermine la phase φ par rapport à la fonction cosinus suivant :

$$\mu = \rho (\cos \theta - \sin \theta \operatorname{tg} \varphi) .$$

b) Application

Nous nous sommes donné la possibilité de représenter le signal par les valeurs de ses projections sur une base de filtres exponentiels de ce type [HATO - 76], ce qui revient à effectuer une sorte de "*Chirp-z transform*" passant par des points choisis à l'avance pour leur pertinence.

Un tel développement est proposé dans [FLAN - 72] pour le domaine s . Dans [YOUN - 62], le critère d'orthogonalité de deux fonctions $\psi_{\mu 1}(z)$ et $\psi_{\mu 2}(z)$ est défini par la condition suivante :

$$\oint_{\text{cercle unité}} \psi_1^*(1/z) \cdot \psi_2(z) \cdot \frac{dz}{2\pi i\, z} = 0 \quad ,$$

ce qui conduit les auteurs à imposer une condition qui équivaut, dans nos notations, à :

$$(1 + \mu_1 \mu_2)(1 + M) - (\mu_1 + \mu_2).R = 0$$

et à choisir simplement :

$$\mu_1 = 1 \quad \text{et} \quad \mu_2 = -1 \quad .$$

Pour notre part, il ne nous paraît pas fondamental que la condition d'orthogonalité définissant la base soit vérifiée sur le cercle unité, la notion d'amortissement conduisant précisément à s'écarter de ce cercle. Nous choisissons de construire notre base à partir de fonctions exponentielles orthogonales dans le domaine temporel, en $\cos(n\theta - \varphi)$ et $\cos(n\theta - \varphi - \frac{\pi}{2})$ par exemple.

μ_1 et μ_2 ont alors pour valeur :

$$\mu_1 = \rho(\cos\theta - \sin\theta\, \mathrm{tg}\varphi)$$
$$\mu_2 = \rho(\cos\theta + \frac{\sin\theta}{\mathrm{tg}\varphi})$$

et vérifient la relation :

$$2 \cdot (\mu_1 \mu_2 + M) - (\mu_1 + \mu_2) \cdot R = 0 \quad .$$

Le cas particulier :

$$\mu_1 = \rho \cos\theta = \frac{R}{2}$$

et μ_2 rejeté à l'infini

nous ramène aux filtres $C(z)$ et $S(z)$ définis en I.3 a). On vérifie aisément que leurs normes sont très voisines et deviennent égales quand l'amortissement tend vers 0 (ρ tend vers 1).

Le calcul des projections sur de tels filtres se fait comme indiqué au par. I-6 après retournement de la chaîne d'entrée. Le signal d'entrée peut être le signal d'origine, après mise en forme pour le traitement numérique, ou bien la succession des coefficients a_p du dénominateur de la fonction de transfert du filtre inverse vu au paragraphe précédent (HATO - 76).

Dans l'idée de faire une discrimination optimale entre classes de sons, le processus suivant peut être adopté :

- apprentissage d'un ensemble d'occurrences de sons appartenant aux classes concernées,
- analyse de façon à repérer les filtres optimaux au sens de la meilleure séparation de ces classes (cf chapitre B.2),
- projection de nouvelles occurrences sur ces filtres qui servent alors de support à des sortes de "squelettes fréquentiels".

5. Expérience de filtrage numérique : codage et restitution d'éléments de parole

Le but visé dans notre expérience était la définition d'une carte de reconnaissance vocale [RUCH - 82] avec *"resynthèse"* du mot prononcé comme mécanisme de validation de l'acquisition, le retour auditif étant, du point de vue ergonomique, bien supérieur au retour visuel (par visualisation du spectrogramme, par exemple). Le marché visé est ici celui des automatismes industriels à commande vocale, la saisie de données, la messagerie vocale, etc. et non le grand public. Les limites d'une telle carte sont essentiellement son fonctionnement en " unilocuteur" et l'obligation actuelle de faire une saisie par clavier de l'orthographe et de l'étiquette associées à chacun des mots prononcés. Des études sont menées actuellement dans l'espoir de lever cette dernière contrainte grâce à la dictée phonétique.

Le tableau B-14 suivant donne la chaîne de calcul de la simulation effectuée sur MITRA 125 en Fortran :

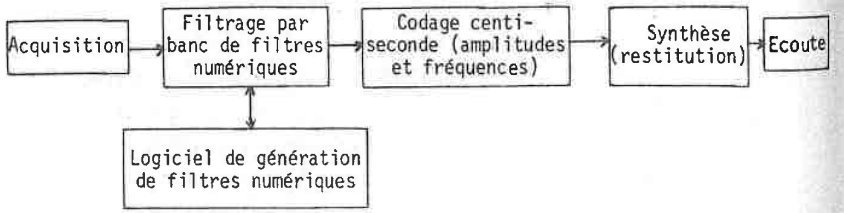


Fig. B-14 : Acquisition, codage et restitution d'éléments de parole.

Nous donnons dans la suite les grandes lignes concernant ces différents modules :

a) l'acquisition numérique est faite à 12 kHz entre 80 et 5000 Hz sur 10 bits,

b) le filtrage est effectué par une batterie de filtres numériques de Butterworth (présentant l'avantage d'avoir en s des pôles faciles à calculer) suivant une échelle "Mel" ou tiers d'octave. Pour ce faire, nous avons mis au point un logiciel permettant d'engendrer des filtres numériques à partir de leur ordre NORDR (pair) et des fréquences-limites de leur bande passante F_B et F_H. Chaque filtre numérique est constitué d'une cascade de cellules d'ordre 2 en z obtenue selon la progression suivante :

- établissement de l'expression $H_N(s)$ de la fonction de transfert du filtre de Butterworth analogique passe-bas normalisé (c'est-à-dire possédant une pulsation de coupure égale à 1 rad/s), d'ordre égal à $N = NORDR/2$:

$$H_N(s) = \frac{1}{\sum_{m=0}^{N} \alpha_m s^m} \qquad (1) .$$

Le dénominateur de $H_N(s)$ est en réalité mis sous la forme d'un produit de facteurs de degré 1 ou 2 à coefficients réels,

- calcul de la quantité :

$$\omega = tg\left[\frac{2\pi}{Fe}(F_H - F_B)\right] \qquad (2) ,$$

- calcul des coefficients des fonctions de transfert des cellules constituant le passe-bas ayant ω comme pulsation de coupure, ce qui donne, en remplaçant dans (1) s par $\frac{s}{\omega}$:

$$H(s) = \frac{1}{\sum_{m=0}^{N} \frac{\alpha_m}{\omega^m} s^m} \qquad (3) ,$$

- calcul de la quantité :

$$h = \cos\left[\frac{\pi}{F_e}(F_H + F_B)\right] \quad / \cos\left[\frac{\pi}{F_e}(F_H - F_B)\right] \qquad (4)\ ,$$

- calcul de H(z) par substitution dans l'expression (3) de s par l'expression :

$$\frac{z^2 - 2hz + 1}{z^2 - 1}\ .$$

Le calcul a posteriori des pôles de H(z) permet un contrôle de la stabilité des filtres.

La figure B-15a indique la réponse en fréquence de tels filtres pour des ordres du filtre passe-bande numérique égaux à 2, 4 et 6. On représente le carré de l'amplitude en fonction de la fréquence,

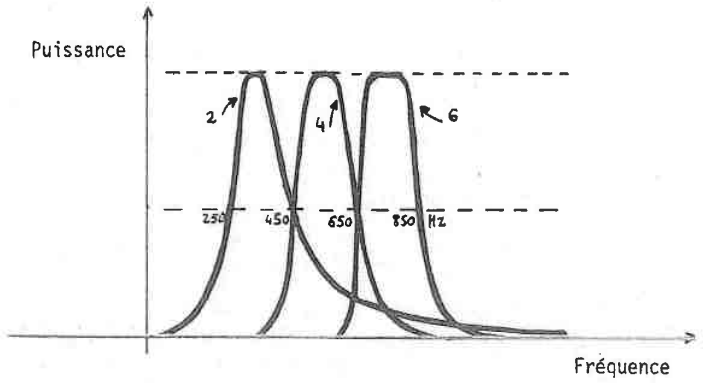


Fig. B-15a : Réponse en fréquence de filtres numériques d'ordre 2, 4 ou 6

c) le codage centiseconde est effectué après étude des signaux de sortie des filtres numériques. Chacun de ces signaux étant sensiblement périodique, on détecte sa fréquence dominante et son amplitude. Cette dernière est codée sur 4 bits. Pour la fréquence, on retient la fréquence la plus proche dans une échelle 1/12ème d'octave. L'analyse étant effectuée sur une bande de 5 octaves, 60 valeurs de fréquence sont possibles ; elles peuvent être codées sur 6 bits. Le codage envisagé correspond ainsi à un débit d'information de 1 000 bauds par filtre.
Les tableaux de la figure B-15b donnent des exemples de spectres ainsi obtenus (locuteur masculin E.R.) :

- en 1, pour le mot /ʃapo/ , sur une échelle de fréquences Mel,
- en 2, pour une élocution du mot /ʃa/ , sur une échelle 1/3 d'octave à l'analyse. Sous les amplitudes sont indiquées les fréquences retenues (échelle 1/12ème d'octave) pour la restitution,

d) la restitution du message se fait enfin par superposition de signaux sinusoïdaux ayant pour amplitudes et pour fréquences les quantités codées dans la phase précédente.

Différents essais en simulation numérique ont permis d'apprécier la qualité du son restitué et de retenir les meilleures conditions d'analyse et de " resynthèse ", notamment en ce qui concerne :

- le choix de la durée des fenêtres d'analyse, des conditions de préfiltrage, de l'ordre des filtres,
- le choix de la largeur des filtres les plus bas pour rendre compte de la fréquence fondamentale s'il y a lieu,
- le choix de l'échelle de codage des amplitudes (linéaire ou logarithmique),

1-

4	4	2	1	1	2	2	3	10	14	21	11	6
5	3	1	1	1	1	2	5	9	14	32	15	8
3	4	3	1	1	1	1	5	16	19	24	14	8
15	5	4	1	1	1	2	7	13	14	20	16	8
4	3	4	3	2	1	2	9	9	14	24	14	7
10	3	6	2	3	3	3	6	7	10	8	10	5
8	15	5	3	2	1	2	2	3	5	5	4	1
105	127	54	29	12	17	24	17	14	13	11	11	3
107	159	171	44	27	47	75	29	39	32	35	48	13
171	102	223	84	34	58	81	25	37	25	27	39	9
156	145	193	89	37	67	79	18	33	15	18	21	5
159	121	175	89	44	66	45	17	40	12	13	17	5
135	118	120	28	23	50	17	18	19	4	4	6	2
58	40	11	4	7	12	3	2	3	1	1	1	0
32	18	5	2	1	2	1	1	1	1	1	1	0
15	4	4	1	1	1	1	1	1	1	1	1	0
3	3	2	2	1	1	1	1	1	1	1	1	1
1	2	3	2	1	1	1	1	1	0	1	1	0
4	2	2	1	1	1	1	1	1	0	1	1	1
3	3	2	1	1	1	1	1	1	0	0	0	0
4	2	2	1	1	1	1	1	1	1	1	1	0
2	2	3	1	0	1	1	1	1	1	1	1	0
4	3	1	2	2	1	1	1	1	1	0	1	1
2	1	3	2	1	1	1	1	0	0	1	1	1
2	2	3	3	1	1	1	1	1	1	1	1	1
17	9	5	4	3	2	2	3	2	1	1	1	0
7	3	3	2	1	1	1	1	1	1	1	1	1
77	81	35	16	14	5	3	3	4	2	3	3	1
158	197	84	31	59	12	3	7	12	3	4	6	2
99	228	115	31	57	14	5	8	11	2	5	5	1
154	199	89	21	45	8	2	4	7	1	3	1	1
143	171	104	26	27	10	3	5	5	1	3	4	1
143	151	93	21	30	8	3	4	6	2	3	3	1
135	111	62	26	21	5	3	5	8	2	4	4	1
127	160	58	28	22	5	2	3	5	1	4	3	1
125	143	55	23	17	3	1	2	3	1	2	3	1
118	99	58	27	15	5	2	3	4	1	3	3	1
107	126	49	25	13	3	1	2	3	1	2	3	1
101	121	36	17	12	3	1	2	3	1	2	2	1
85	77	41	25	11	3	1	2	3	1	1	1	0
87	78	31	19	10	2	1	1	2	1	1	1	0
83	73	23	12	5	1	1	2	1	1	1	1	0
82	61	21	6	2	1	1	2	1	1	1	1	0

2-

4	2	2	1	1	1	1	1	1	2	6	16	22	10	15
161	206	258	345	408	578	691	875	1156	1556	1958	2063	2937	3700	4125
4	6	4	1	1	2	1	1	2	2	6	16	19	28	28
195	219	258	389	438	578	647	975	1219	1381	1958	2188	2763	3388	4125
6	10	8	6	4	2	3	4	4	5	13	25	27	23	25
195	219	258	389	413	689	734	925	1156	1381	1050	2188	3112	3500	4625
7	9	7	2	2	2	2	2	2	3	5	28	29	45	39
161	231	258	323	408	547	734	925	1219	1294	1958	2188	2937	3300	4125
8	3	3	3	4	4	4	3	4	8	18	21	40	27	26
161	219	289	367	413	689	734	875	1219	1381	1958	2437	2737	3300	4125
2	2	3	3	5	4	3	2	2	4	6	18	16	13	14
183	219	273	323	438	547	734	875	1219	1381	1958	2063	3112	3300	4125
15	13	32	23	58	58	21	28	17	27	22	25	25	31	61
195	244	258	389	462	578	691	975	1219	1556	1958	2188	2763	3300	4125
15	65	39	78	74	82	49	36	47	53	48	32	30	60	27
161	231	258	389	413	516	778	975	1219	1294	1958	2437	2763	3300	4075
28	60	29	58	88	95	91	91	39	56	38	33	34	83	32
183	231	258	345	408	516	734	825	1156	1294	1750	2063	2588	3700	4075
42	33	17	12	38	77	85	67	59	62	48	48	38	59	25
195	219	289	323	438	578	778	875	1219	1556	1958	2437	2763	3300	4125
25	31	25	19	38	82	73	66	50	56	38	35	27	28	13
183	206	258	345	462	547	691	825	1031	1556	1750	2312	2763	3900	4875
35	55	30	33	35	63	100	58	65	59	27	18	20	15	40
195	219	303	367	413	516	734	825	1156	1294	1650	2063	3112	3300	4125
37	39	11	26	48	77	54	45	30	27	18	11	6	3	3
170	219	304	345	462	547	647	925	1031	1294	1850	2063	2580	3300	4875
23	18	6	7	13	20	19	11	8	8	7	8	7	7	15
195	206	304	389	462	547	691	825	1031	1556	1750	2312	2588	3300	4125
10	10	11	3	6	6	6	5	6	6	5	5	5	6	7
161	231	258	309	462	689	778	975	1094	1556	1750	2437	2763	3300	4125
9	4	4	3	4	4	4	4	4	4	3	4	5	5	5
161	206	273	389	438	689	778	875	1219	1381	1950	2188	2763	3300	4075

Fig. B-15b : Spectres résultant du codage des amplitudes (1) et des fréquences (2)

- le mode de raccordement des signaux synthétisés à chaque modification des paramètres, l'amortissement des signaux,
- l'introduction de bruit pour traduire les phénomènes de friction, etc.

Cette simulation a ensuite permis la définition et la réalisation industrielle d'une carte d'analyse avec retour auditif utilisant les circuits utiles à la reconnaissance.

V - FREQUENCES ET LARGEURS DE BANDE DES FORMANTS

Le canal vocal agit comme un filtre sur le spectre de fréquence à large bande provenant des sources d'excitation (cordes vocales en vibration pour les sons voisés et/ou partie contractée du canal pour les sons fricatifs...) en lui imposant ses propriétés de transmission. Il se caractérise alors par ses fréquences de résonance de transmission qui, associées à leur amortissement, constituent les formants.

Si l'on considère que pour les sons voisés la plus grande partie de l'information sur le signal de parole est portée par les formants, en plus du fondamental de la voix, on peut tenter d'en faire une analyse temporelle suivant la méthode de PINSON [DOUR - 74a] en minimisant l'écart entre la réponse impulsionnelle du conduit vocal et une somme de sinusoïdes amorties. La difficulté de cette méthode réside dans la détection des instants de fermeture de la glotte qui ne peut se faire que dans une étude parallèle de l'onde glottale.

Le passage au domaine fréquentiel permet de s'affranchir de cette difficulté mais d'autres problèmes que nous évoquons par la suite sont à considérer.

La présence d'un formant (mode de résonance du conduit vocal) se traduit généralement dans le spectre par un maximum local. Il se peut cependant que la présence voisine d'une antirésonance le fasse disparaître ou que deux pics voisins soient confondus en un seul. Aussi la détection et le suivi dans le temps des formants sont-ils plus compliqués qu'une recherche de pics dans le spectre. Différentes méthodes peuvent être envisagées suivant le but poursuivi, aucune ne résoud complètement le problème. Nous indiquons ici celles auxquelles nous faisons appel.

1. Taux de passage par zéro

La méthode convient pour des segments de parole voisée. Le signal est d'abord filtré par des filtres passe-bande convenablement choisis. On compte ensuite dans le signal temporel de sortie de chacun des filtres le nombre N de passages de la courbe par zéro pendant un intervalle de temps τ. Sur chacune des plages, la fréquence moyenne est estimée par :

$$\tilde{F} = \frac{N}{2\tau} .$$

Les atténuations dans chaque zone par rapport au signal brut constituent aussi des paramètres intéressants ; elles sont calculées à partir de l'intensité totale et de celle du signal filtré.

2. Moments spectraux

Une méthode particulièrement simple et rapide du même type que la précédente revient à calculer les fréquences dominantes dans différentes zones de fréquences à partir du spectre. Par exemple, à partir des valeurs a_i de l'intensité aux sorties d'un analyseur spectral, ce calcul est celui du moment spectral de premier ordre :

$$\tilde{F} = \frac{\Sigma_{zone}\ a_i\ F_i}{\Sigma_{zone}\ a_i}$$

3. Traitement fréquentiel numérique

Le calcul de la transformée de Fourier sur le cercle unité dans le plan z (après un éventuel traitement de prédiction linéaire) donne un spectre dans lequel le sommet et la forme de chacun des pics sont influencés par plusieurs facteurs :

- la proximité d'un autre pôle ou d'une antirésonance (paire pôle-zéro),
- la position relative du pic par rapport aux autres,

- le déplacement du sommet à cause du facteur d'amortissement de la résonance.

Sur la figure B-12[2] on avait, superposées au spectre résultant, les contributions individuelles des différentes composantes du signal, après traitement de prédiction linéaire.

Pour chacune de ces courbes prises séparément, il est possible de relier directement l'amortissement à la "largeur de bande" à 6 dB d'atténuation en puissance, c'est-à-dire la largeur en unités de fréquence de la courbe à l'ordonnée où l'amplitude carrée vaut la moitié de celle du maximum.

On montre que la largeur de bande ainsi définie est reliée au facteur d'amortissement (notations du par. I.3), au second ordre près, par :

$$B = \frac{\sigma}{\pi}$$

ou $$B = \frac{Fe}{\pi} \text{Log}\left(\frac{1}{\rho}\right) \quad ,$$

alors que la fréquence F est donnée par :

$$F = \frac{\theta}{2\pi} Fe \quad .$$

Cette quantité B, que l'on ne peut pas extraire du spectre résultant sans erreur, peut être atteinte par la détermination de la position du pôle correspondant (module et angle polaire).

Pour ce faire, différentes méthodes peuvent être proposées, ainsi que décrit dans la suite.

a) Calcul de la transformée de Fourier sur un cercle de rayon inférieur à l'unité

Nous nous sommes inspiré, dans un premier temps, de la transformation dite *"Chirp z Transform"* sur un contour elliptique à l'intérieur du cercle de rayon unité. Un cas particulier simple en est le calcul de la transformée en z sur des demi-cercles de rayon inférieur à l'unité.

Plus le pôle est voisin du cercle envisagé, plus le pic apparaîtra comme accentué. Le rayon du cercle se relie directement à l'amortissement donc à la largeur de la bande.

Cette représentation présente un intérêt immédiat : des pôles voisins dont les effets se confondent dans le spectre se différencient au fur et à mesure de la variation du rayon ρ .

Un exemple est donné sur la figure B-16, à partir de l'élocution du message /a i/, échantillonné à 7.5 kHz et traité par prédiction linéaire à l'ordre 10. On présente successivement :

1. le signal d'origine numérisé,
2. la succession des spectres calculés sur des fenêtres consécutives avec recouvrement partiel, pour des valeurs de ρ égales respectivement à 1., 0.95 et 0.9 .

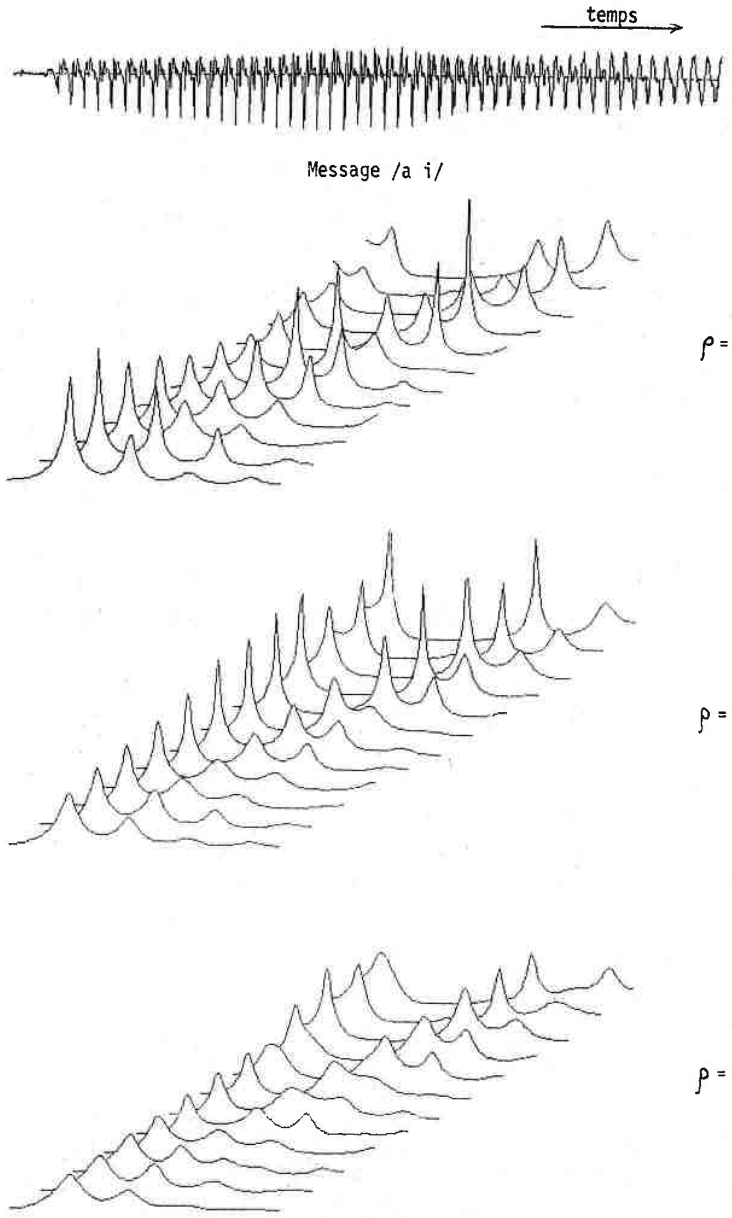


Fig. B-16 : Traitement de prédiction linéaire et calcul de spectre par transformation de Fourier sur des demi-cercles de rayon ρ

Du point de vue mathématique, un tel calcul revient à une simple prétransformation du signal avant l'application de la FFT.

En effet, notons $\mathcal{F}_r(\{x_n\})$ la transformée d'une suite de valeurs réelles $\{x_n , n = 0, 1..., N-1\}$ calculée sur un demi-cercle de rayon r . $\mathcal{F}_1(\{x_n\})$ désigne alors la transformée de Fourier classique. On vérifie aisément que :

$$\mathcal{F}_\rho(\{x_n\}) = \mathcal{F}_1\left(\{\frac{x_n}{\rho^n}\}\right)$$
$$= \mathcal{F}_1(\{x'_n\}) \quad \text{si l'on pose} \quad x'_n = \frac{x_n}{\rho^n} .$$

Il est évident que le signal $\{x'_n\}$ peut diverger. Cependant, le caractère fini dans le temps de la tranche de signal considérée, d'une part, et l'identité en module de la transformée de Fourier discrète d'un signal et du même signal retourné dans le temps, d'autre part, font que la prétransformation proposée n'entraîne pas de difficulté théorique.

b) Accentuation directe des formants

Il vient à l'esprit, à partir de l'étude des voyelles orales du français en particulier, de proposer une "Chirp z Transform" adaptée à la recherche directe des formants.

D'après nos notations, rappelons que la fréquence et la largeur de bande associées aux pôles complexes $z = \rho \exp(i\theta)$ et $z^* = \rho \exp(-i\theta)$ ont pour expressions, en désignant par Fe la fréquence d'échantillonnage :

$$F = Fe \frac{\theta}{2\pi} \qquad (1)$$

et $$B = \frac{Fe}{\pi} \text{Log}\left(\frac{1}{\rho}\right) \qquad (2)$$

ce qui entraîne pour le rapport de la largeur de bande à la fréquence l'expression :

$$\frac{B}{F} = -2 \frac{\text{Log } \rho}{\theta} \qquad (3)$$

L'étude du rapport B_i/F_i pour les formants i des voyelles françaises pourrait permettre de trouver l'équation polaire $\rho = \rho(\theta)$ de la courbe passant par les pôles.

En particulier, il semble que, pour les formants d'indice plus grand que 1, le rapport B_i/F_i soit sensiblement égal à une quantité r_i ne dépendant que de l'indice i. En faisant abstraction de cet indice, on obtient, à partir de (3) :

$$\rho = e^{-r\theta/2} \qquad (4),$$

qui est l'équation d'une spirale.

L'angle θ variant de 0 à π, ρ peut être approché, lorsque r est petit devant l'unité (il est en réalité de l'ordre de quelques %), par le développement parabolique :

$$\rho = 1 - r\theta/2 + r^2\,\theta^2/8 .$$

La transformation proposée revient, tout comme à l'alinéa a précédent, à une prétransformation des données.

Une telle méthode, idéale du point de vue théorique, n'a de valeur que si les traitements imposés au signal depuis sa sortie aux lèvres conservent l'information sur la largeur de bande. Dans la réalité, comme il sera vu à l'alinéa suivant, ce n'est pas vraiment le cas.

c) Calcul des formants après traitement par prédiction linéaire (LPC)

Dans la modélisation du conduit vocal par la méthode de prédiction linéaire (cf. paragraphe IV.3.c de ce chapitre), les pôles de la fonction de transfert du modèle permettent d'approcher les résonances du conduit, c'est-à-dire les formants.

Dans un modèle tout-pôle, la fonction de transfert peut s'écrire :

$$B(z) = \frac{Y}{A(z)} \quad \text{où} \quad A(z) \text{ est un polynôme en } z^{-1} .$$

C'est la résolution mathématique de l'équation : $A(z) = 0$ qui va permettre d'atteindre les formants.

Bien que la méthode que nous avons retenue de LPC par autocorrélation garantisse dans son principe la stabilité du modèle, on se heurte à plusieurs difficultés dues au calcul numérique et qui entraînent :

- le risque d'avoir des pôles non situés à l'intérieur du cercle unité, ce qui, physiquement, est contraire à la réalité,
- des variations des résultats suivant l'ordre de prédiction retenu.

Afin d'apprécier les difficultés d'interprétation des résultats, nous avons recherché systématiquement les déviations de ces résultats à partir de signaux synthétiques composés d'un nombre variable de sinusoïdes amorties caractérisées par :

- l'amplitude,
- la fréquence et l'amortissement,
- la phase à l'instant d'origine,
- la présence ou non de fondamental.

Nous avons pour cela suivi la chaîne de calcul ci-dessous :

```
Synthèse d'un signal numérique
p:=p0
Tant que non condition d'arrêt faire
     Traitement de prédiction linéaire à l'ordre p
     Détermination de la chaîne des (p+1) coefficients ai
     Calcul des pôles de 1 / Σ_{i=0}^{p} ai z^{-i}
     Calcul des fréquences et largeurs de bande correspondantes
     p:=p+1
fin tant que
Interprétation des résultats
```

Sur l'interprétation des résultats, nous pouvons faire quelques remarques :

- malgré l'écriture théorique idéale de la fonction d'autocorrélation du signal, l'énergie résiduelle et les résultats dépendent de la troncature du signal "à gauche". Nous en donnons comme illustration l'étude d'un signal sinusoïdal amorti sans fondamental avec F = 2 700 Hz et B = 400 Hz isolé par une fenêtre rectangulaire et présentant, par rapport à la fonction sinus, une phase à l'origine φ variable. La figure B-17 indique les résultats obtenus à l'ordre 2 en fonction de φ .
La variation des résultats met en évidence l'influence de la phase à l'origine dans une analyse "pitch-asynchrone" et l'importance dans ce cas de la durée du message retenue pour l'analyse. Une durée de plusieurs périodes est nécessaire pour limiter cette influence.

- l'ordre de prédiction retenu influe sur les résultats qui, par ailleurs, dépendent nettement des proximités des formants entre eux.

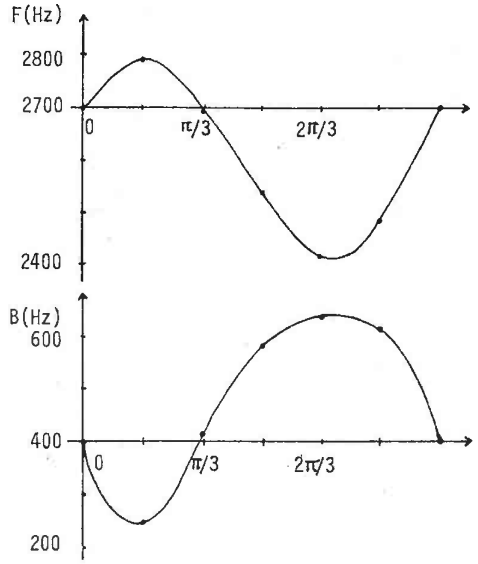


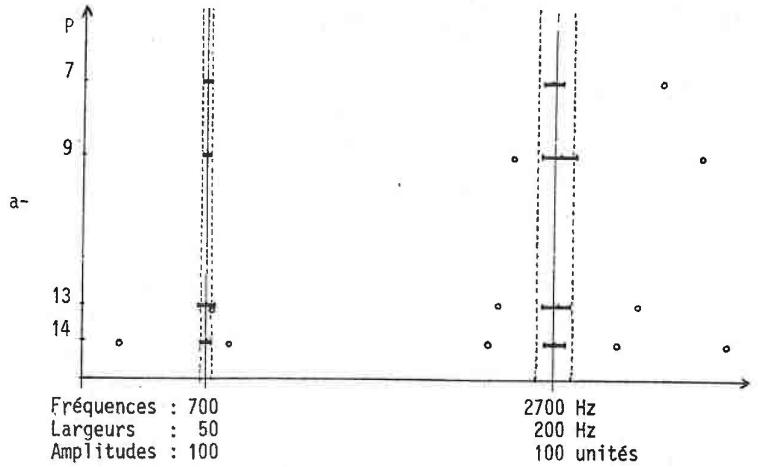
Fig. B-17 : Influence de la troncature "à gauche" du signal temporel sur la détermination de la fréquence et de la largeur de bande

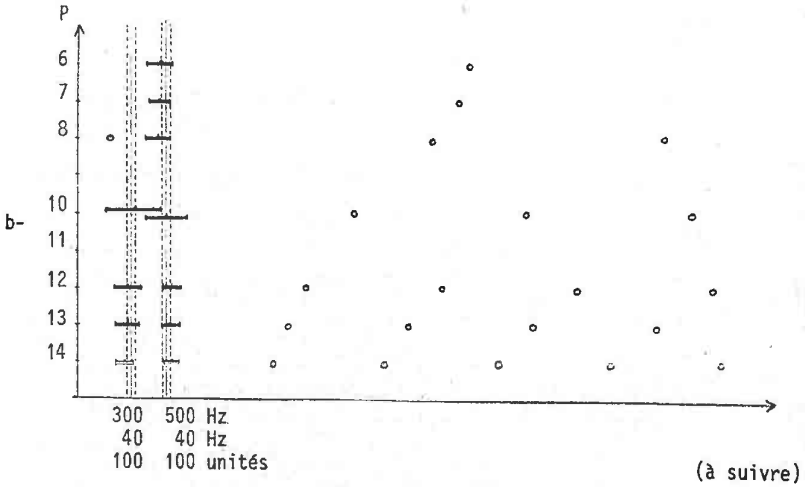
La figure B-18 correspond à quatre situations différentes (Fe = 8 kHz) :

a. les formants sont éloignés (caractère diffus) et d'amplitudes à l'origine égales,

b. les formants sont rapprochés (caractère compact). La différenciation ne se fait qu'à l'ordre 10 après l'apparition de pôles parasites (notés o),

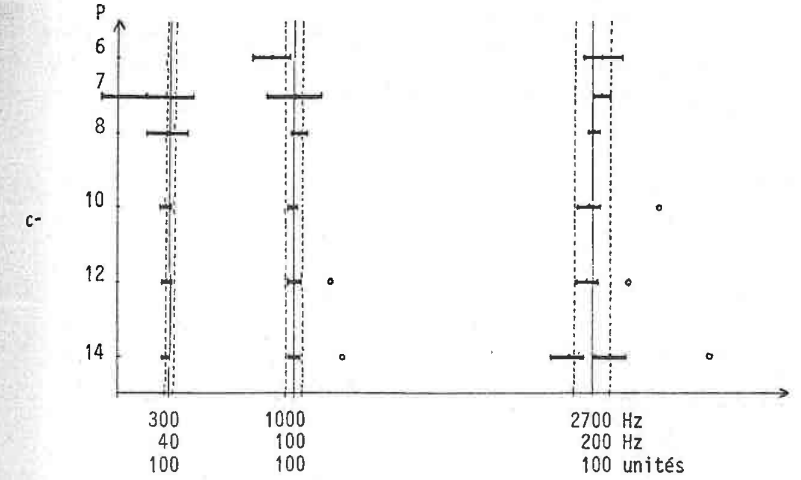
c et d. les deux signaux sont composés de trois sinusoïdes amorties et ne diffèrent que par les amplitudes relatives de ces sinusoïdes. On peut remarquer que pour de faibles amplitudes en relatif (d), les résultats obtenus pour la largeur de bande sont très éloignés de la réalité.





(à suivre)

Fig. B-18 : Influence de l'ordre de prédiction linéaire sur les résultats de l'analyse
a : formants éloignés



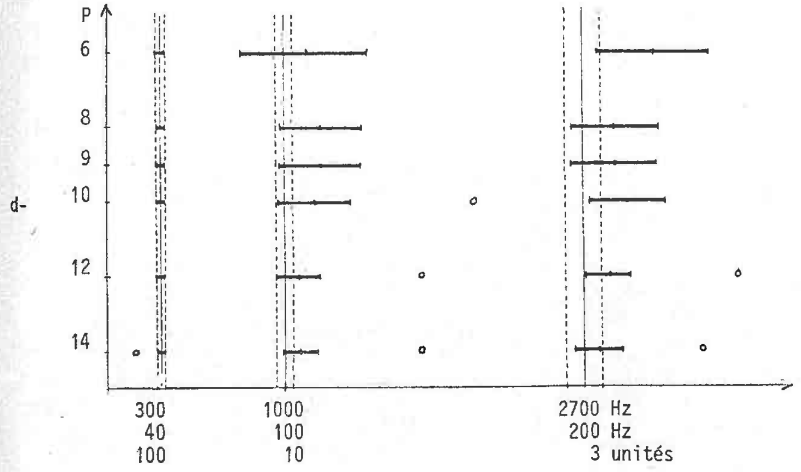


Fig. B-18 (suite)
c : formants d'égales amplitudes
d : formants d'amplitudes différentes

Dans le cas où les amplitudes des composantes sont égales, malgré des résultats convenables pour la largeur de bande, on remarque, en plus des pôles parasites, un dédoublement de la fréquence haute pour $P = 14$.

La différence entre c et d s'explique par le fait que le modèle "tout-pôle" ne rend pas compte des amplitudes relatives des formants dans leur individualité.

Les remarques précédentes nous ont conduit à rechercher des conditions d'arrêt qui limitent l'ordre de prédiction linéaire à une valeur compatible avec le signal étudié. Nous avons retenu deux critères d'arrêt :

. le nouveau coefficient $a_{p+1/p+1}$ qui apparaît à l'ordre $p+1$ est petit devant 1 ,

. l'énergie résiduelle ne diminue plus notablement.

Dans le cas de signaux de parole cependant, sauf cas particulier, nous respectons la condition qui impose que P soit égal ou légèrement supérieur à Fe exprimé en kHz. (Pour les voyelles nasalisées, il est nécessaire d'aller jusque $P = 16$ pour $Fe = 12$ kHz). Il est par ailleurs nécessaire de forcer le degré jusqu'à une valeur prédéterminée dans le cas où l'on veut visualiser le modèle en tubes du conduit vocal après lissage unique qui suppose que le nombre de paliers est fixe, comme nous le verrons plus loin.

Les inconvénients majeurs de la méthode sont toutefois les suivants :

- l'information sur les amplitudes relatives des formants à l'origine est perdue,

- il y a introduction de pôles parasites si l'on donne à P une valeur trop grande par rapport à la richesse du message étudié. Cependant, la plupart du temps, les largeurs de bande anormalement élevées permettent d'éliminer ces artefacts.

La figure B-19 illustre l'évolution des résultats du traitement de prédiction linéaire après préaccentuation et limitation par la fenêtre de Hamming dans la situation réelle ($Fe = 10$ kHz) de l'élocution de la voyelle neutre /ə/ . On a :

. le signal de durée 25.6 ms,

. la succession des spectres dans le temps pour $P = 12$,

. un tableau indiquant les valeurs des fréquences et largeurs de bande des formants suivant l'ordre P pour une fenêtre prise au milieu du message,

. les résultats équivalents obtenus pour /ɸ/ .

Les largeurs de bande sont figurées par des segments en trait accentué de part et d'autre de la valeur des fréquences de formants, à la même échelle.

Pour terminer, nous donnons sur la figure B-20 la succession des spectres de prédiction linéaire pour une élocution du message /a i/ ($F_e = 10$ kHz) pour $P = 10$, 12 et 14, à rapprocher de la figure B-16.

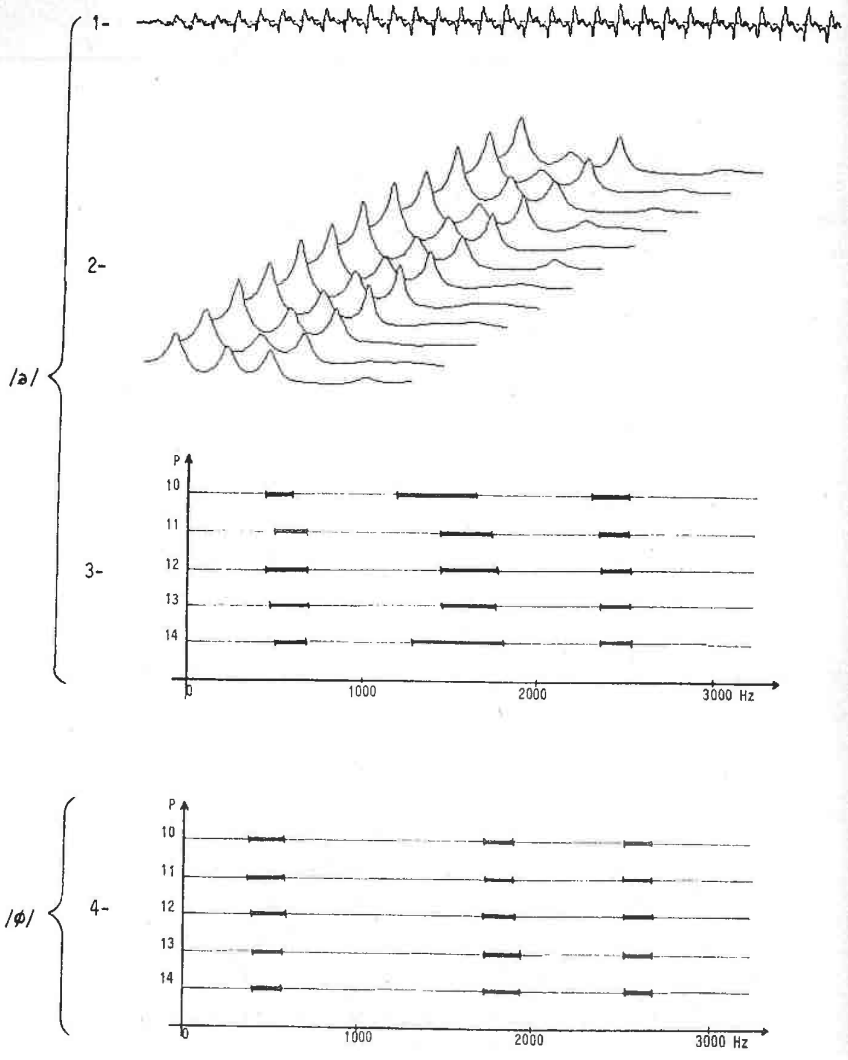


Fig. B-19 : Traitement de prédiction linéaire :
1, 2 et 3 : message /ə/ dans le temps, successions des spectres pour P = 12, évolution suivant l'ordre P,
4 : message /φ/

Rappelons, à l'occasion de l'interprétation des résultats fournis par la méthode de prédiction linéaire, les limites de la méthode elle-même, indépendamment du calcul numérique :

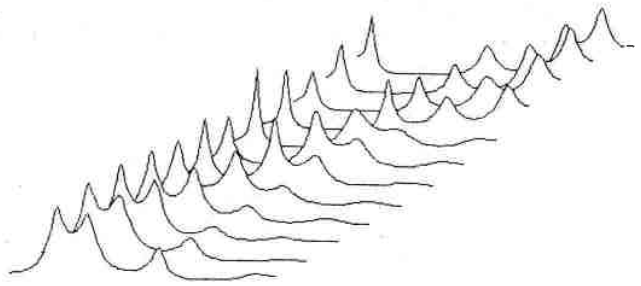
- elle se fonde sur l'hypothèse de linéarité du processus de production de la parole,
- si les résultats sont crédibles pour les sons voisés, elle ne convient pas chaque fois qu'intervient le conduit nasal ou dans le cas de constrictions étroites du conduit vocal,
- elle n'est adaptée qu'aux zones quasi stationnaires du signal.

Enfin, comme chaque fois que l'on étudie le passage au domaine fréquentiel, il faut établir un compromis entre la résolution temporelle (étude d'une fenêtre étroite, filtres à large bande) et la résolution fréquentielle (filtres plus étroits, appliqués à une fenêtre de plusieurs périodes de fondamental).

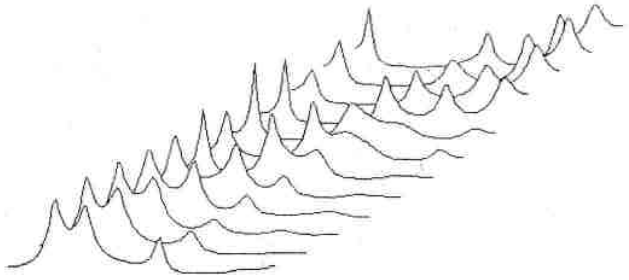
d) Cepstre

De même qu'un cepstre dont on ne conserve que la partie "haute période" permet d'approcher la fréquence du fondamental (parag. II), la transformation de Fourier de sa partie "basse période" fournit un spectre débarrassé de l'influence glottale. Nous avons toujours jugé les résultats obtenus par cette méthode comparable à la méthode de prédiction linéaire sans avantage particulier.

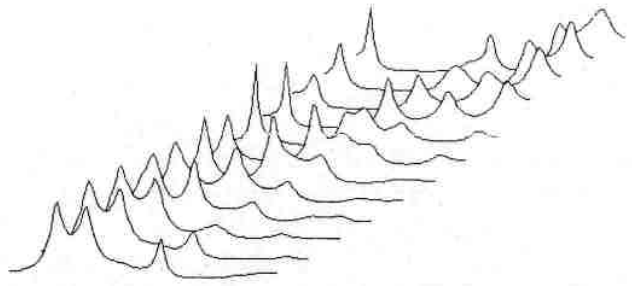
message /a i/



P = 10



P = 12



P = 14

Fig. B-20 : Message /a i/ : signal et successions de spectres

VI - LE CONDUIT VOCAL

La recherche de la configuration du conduit vocal à partir de l'onde de pression sonore rayonnée aux lèvres est un problème de mécanique physique qui fait l'objet de développements toujours plus sophistiqués.

1. Lien entre configuration du conduit et modes de vibration

Le point de départ du calcul de la fonction d'aire $\$(x)$ à la distance x de la source sonore est l'équation de Webster qui relie $\$(x)$ à la pression de l'onde sonore $p(x, t)$ dépendant de x et du temps (donc aux fréquences composant l'onde sonore) :

$$\frac{1}{\$(x)} \frac{\partial}{\partial x} \left[\$(x) \frac{\partial p}{\partial x} \right] = \frac{1}{v^2} \frac{\partial^2 p}{\partial t^2} \quad ,$$

v désignant la vitesse de propagation de l'onde.

Dans le cas du conduit vocal, x désigne la distance de la section considérée à la glotte et varie entre 0 et L, longueur totale du conduit.

La résolution de l'équation de Webster exige que soient précisées les conditions aux limites et c'est du choix toujours plus élaboré de ces conditions que dérivent les méthodes d'analyse fondées sur sa résolution [MERM - 67], [JOSP - 79].

Dans une première étape, on peut approcher le conduit vocal par un tube de section uniforme fermé à la glotte et sans rayonnement aux lèvres. La transmission est alors caractérisée par les modes de vibration de pulsations :

$$\omega_n^{\,0} = \pm (2n-1) \frac{\pi}{2} \frac{v}{L} \quad ,$$

donc par les fréquences :

$$f_n^{\,o} = (2n-1)\,\frac{v}{4L}$$

Pour une valeur de v prise égale à 340 m/s (air sec à 15° C) et une valeur moyenne de L égale à 17 cm pour un locuteur masculin, les fréquences de résonance valent :

$$f_n^{\,o} = (2n-1) \times 500 \text{ (Hz)}$$

soit la succession 500, 1 500, 2 500, ... Hz qui caractérise assez bien la voyelle neutre dite "*schwa*" correspondant à une position de repos des articulateurs.

A partir de ce premier modèle, il est possible d'approcher la configuration réelle du canal vocal en affinant les points suivants :

- liaison entre variations géométriques du conduit et articulation,
- déplacement des formants dû au rayonnement aux lèvres,
- présence de conduction thermique du canal et de viscosité dynamique de l'air à son contact...

2. Modèle acoustique en tubes du canal vocal

Les techniques de prédiction appliquées au conduit vocal dans un état stationnaire reviennent en son approximation par une succession de P sections cylindriques mises bout à bout, à partir du signal recueilli à sa sortie et sous les hypothèses restrictives suivantes :

- la dimension transversale de chaque section est suffisamment petite devant les longueurs d'onde composantes pour que l'onde acoustique qui se propage puisse être considérée comme plane,

- le tube est supposé rigide ; les pertes dues aux frottements visqueux et aux échanges de chaleur sont négligées [WAKI - 73].

Si l'on choisit pour fonction de transfert du modèle une expression de la forme :

$$H(z) = \frac{Y}{A(z)} = \frac{Y}{1 + \sum_{p=1}^{P} a_p\, z^{-p}}$$

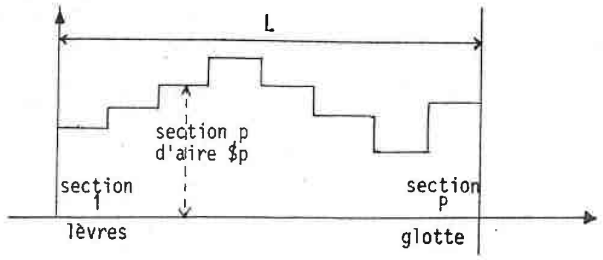
(cf paragraphe IV.3 de ce chapitre), on montre que la relation de récurrence permettant de passer d'un filtre inverse $A_m(z)$ d'ordre m à un filtre inverse $A_{m+1}(z)$ d'ordre $m+1$ traduit également l'adjonction d'un $(m+1)$ème tube à un modèle en m tubes.

Cette analogie entre le filtre numérique et le modèle en tubes à section variable mais cylindriques introduit des liens directs entre deux catégories de paramètres :

- les coefficients K_p définis au paragraphe IV.3 et les coefficients de réflexion μ_p aux jonctions entre les tubes p et $p+1$,

- la période d'échantillonnage Te du signal et le temps de transfert de l'onde d'origine glottale à travers les différentes sections du conduit vocal.

Ces relations, compte-tenu des grandeurs introduites sur le schéma suivant :



s'écrivent :

$$\mu_p = \frac{\$_p - \$_{p+1}}{\$_p + \$_{p+1}} = K_{p-1}$$

et $$Te = \frac{2L}{Pv}$$

où P est le nombre de sections et v la vitesse de propagation du son.

Les coefficients K_p étant directement calculés grâce aux relations (8), les valeurs des $\$_p$ peuvent être déduites en cours de calcul à partir de conditions aux limites convenables. L'ensemble $\{\$_p\}$ constitue une approximation de la fonction d'aire $\$(x)$ du conduit vocal.

La figure B-21 donne l'évolution de la fonction d'aire $\{\$_i, i = 1, \ldots, P\}$ pour des valeurs de P allant de 6 à 11 dans le cas d'un message composé de trois formants d'amplitudes égales, échantillonné à 10 kHz . Le volume du conduit oral n'est pas normalisé, pour les besoins du schéma. En vue de la visualisation de la configuration du conduit vocal dans un état stationnaire, nous effectuons un lissage sur les valeurs des aires des P tubes envisagés. Divers essais nous ont conduit à retenir le lissage trigonométrique présenté au paragraphe C.1.II avec P points de départ, un nombre de points d'arrivée supérieur à P et un ordre de lissage voisin de (P DIV 2) + 1 .

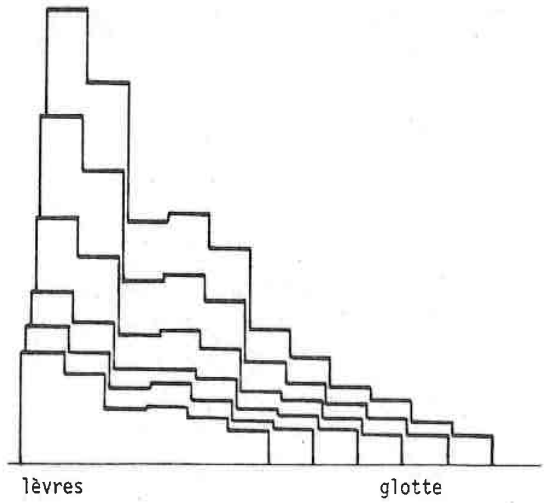


Fig. B-21 : Evolution de la fonction d'aire pour des valeurs de P allant de 6 à 11 (signal synthétique).

La figure B-22 suivante résume les possibilités du traitement de la parole par codage prédictif linéaire avec mention des paragraphes précédents correspondants.

Ⓐ Détermination des coefficients a_p du filtre inverse $A(z) = 1 + \Sigma_{p=1}^{P} a_p z^{-p}$ (par. IV.3.b) par :

- Numérisation du signal → suite $\{x_n\}$
- Différenciation et multiplication par une fenêtre temporelle de durée N échantillons → suite $\{s_n, n = 0, 1, \ldots, N-1\}$
- Calcul de la fonction d'autocorrélation de s → suite $\{S_k, k = 0, 1, \ldots, P\}$
- Calcul des coefficients du filtre inverse → suite $\{a_p, p = 0, 1, \ldots, P\}$

Ⓑ Approximation de la fonction d'aire du conduit vocal (par. VI.2) par :

- Calcul des aires relatives des sections du modèle en tubes,
- Normalisation éventuelle (à volume constant par exemple).

Ⓒ Détermination des premiers formants (par. V.3.c) par :

- Résolution de $A(z) = 0$ ou "*Chirp-z transform*" adaptée,
- Correction éventuelle (problème de stabilité du modèle),
- Calcul des fréquences et largeurs de bande.

Fig. B-22 : Possibilités du traitement par codage prédictif linéaire.

VII - L'ONDE GLOTTALE

En dehors des méthodes de simulation [GUER - 78] de la glotte par analogie électrique ou des mesures effectuées à partir d'exploration visuelle directe de la glotte, il est possible d'approcher l'onde glottale sur une période de vibration des cordes vocales par filtrage inverse du signal émis aux lèvres [ROTH - 73]. Ce filtrage doit supprimer au mieux l'influence du conduit vocal responsable de l'amplification des moyennes et hautes fréquences dans le spectre.

L'enveloppe spectrale de l'onde glottale est assimilable, en première approximation, à une ligne brisée avec une pente moyenne correspondant à une chute de 12 dB par octave et présentant de nombreux minima, comme l'indique la figure B-23 tirée de [FLAN - 72].

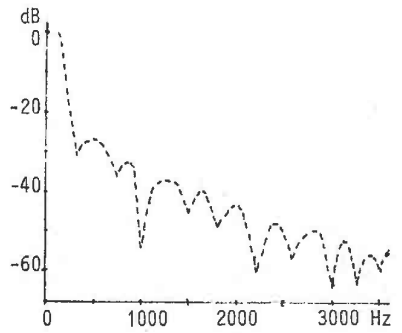


Fig. B-23 : Spectre de l'onde glottale de la fig. B-24 (d'après Flanagan)

La forme de l'onde glottale pour les fréquences inférieures à 1 000 Hz serait caractéristique du locuteur et indépendante du "vocoïde" prononcé [LAND - 77] . Cette voie que nous n'avons pas approfondie est intéressante dans le domaine de l'identification de locuteurs et pour nous plus encore dans la perspective de définition d'aides au diagnostic médical, à partir de"l'objectivisation"et la classification des voix pathologiques. Les renseignements sur l'onde glottale et son spectre seraient un apport d'information considérable, venant s'ajouter aux résultats de l'analyse de la fréquence de vibration des cordes vocales.

Nous reprenons cette idée en C.3.IV à propos de l'extension des études des voix à l'aide au diagnostic médical.

La figure B-24, d'après Flanagan, donne un exemple d'onde glottale dans l'élocution de /æ/ par un homme.

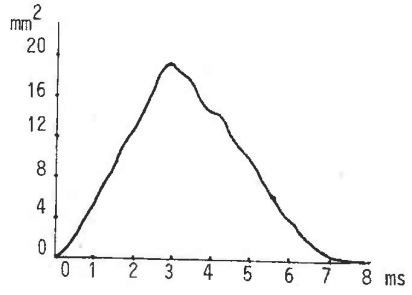


Fig. B-24 : Une période d'onde glottale (aire)
(d'après Flanagan)

CHAPITRE 2

REDUCTION DE DONNEES VOCALES. RECHERCHE DE PARAMÈTRES DISCRIMINANTS

I - GENERALITES

Le signal de parole échantillonné à 12 kHz et codé sur 10 bits renferme une quantité d'information considérable de 120 kbauds. Pour des besoins autres que l'étude fine de l'onde temporelle (nombre de passages par zéro, nombre d'extrema, longueur de courbe, "microstructure") ou de son équivalent fréquentiel (fréquences et largeurs de formants, etc.) qui se justifient en segmentation automatique par exemple [SANC - 78], il est suffisant de conserver l'information obtenue par une première réduction de données, l'analyse spectrale par banc de filtres, par exemple. Ainsi, une analyse à 50 prélèvements par seconde par un banc de 15 filtres, les densités d'énergie étant codées sur 3 bits avec information sur le fondamental, réduit-elle le taux d'information à 3 000 bauds maximum. Cette première phase de réduction de données correspond à l'objectif évoqué au chapitre B.1 précédent de recherche de paramètres caractéristiques avec compression d'information.

Nous nous plaçons dans la suite de cette partie dans le cas particulier d'une caractérisation du signal vocal par des spectrogrammes numériques tels que ceux que nous obtenons à partir des "vocodeurs" utilisés au laboratoire. Cette caractérisation possède en elle-même un point de vue descriptif d'ailleurs classique. La figure B-25 montre les spectrogrammes de sons tenus /a/ , /i/ et /u/ avec un codage mettant en évidence les niveaux d'énergie.

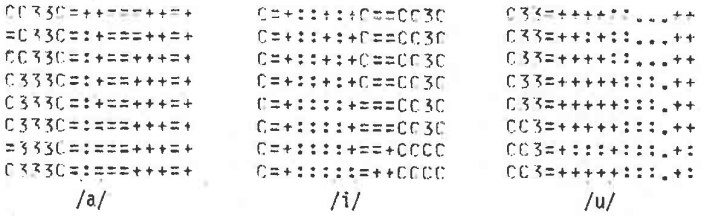


Fig. B-25 : Spectrogrammes de sons tenus

Nous présentons dans la suite un travail effectué dans le but de proposer une description de "formes-spectres" qui puisse :

- être saisie visuellement en temps réel,
- être interprétée au sens de la proximité à une forme-témoin

et qui réponde par conséquent à des critères de description et de classification.

II - CHOIX DE LA METHODE ET SITUATIONS D'APPLICATION

1. Choix de la méthode

Dans la suite, un individu (soit une observation) x_i correspondra à l'apparition d'un prélèvement-vocodeur ou spectre, vecteur ayant pour dimension n_p, nombre de canaux de l'analyseur spectral.

Devant la quantité d'information contenue dans un ensemble $\{x_i, i = 1, \ldots, n_T\}$, les questions suivantes peuvent se poser : l'information est-elle redondante ? Tous les paramètres sont-ils significatifs ou peut-on supprimer certains d'entre eux ou les remplacer par une de leurs combinaisons linéaires ? Ceci revient en quelque sorte à se demander s'il n'existe pas une façon d'observer l'ensemble x_i dans l'espace des n_p paramètres qui mette en évidence l'appartenance de chacun de ses éléments à une classe prédéfinie.

Une réponse à de telles questions est fournie par des méthodes classiques en analyse des données :

- l'analyse en composantes principales à but essentiellement descriptif,
- l'analyse factorielle discriminante à but à la fois descriptif et décisionnel.

Pour aborder ces deux méthodes, il est nécessaire d'introduire les matrices de covariance totale T, de covariance "interclasses" E et de covariance "intraclasse" D des données de départ dont on suppose connue l'appartenance à l'une ou l'autre de n_c classes. Si l'on retient le même nombre d'individus $n = n_T / n_c$ par classe, les matrices se définissent ainsi :

$$T = \frac{1}{n_T} \sum_{k=1}^{n_c} \sum_{i=1}^{n} (x_i^k - \bar{x}).(x_i^k - \bar{x})^T \qquad (1)$$

$$E = \frac{1}{n} \sum_{k=1}^{n_c} (\bar{x}^k - \bar{x}).(\bar{x}^k - \bar{x})^T \qquad (2)$$

D = Somme des k matrices de covariance intraclasse D_k

$$= \sum_{k=1}^{n_c} \left[\frac{1}{n_T} \sum_{i=1}^{n} (x_i^k - \bar{x}^k).(x_i^k - \bar{x}^k)^T \right] \qquad (3)$$

Dans les relations (1), (2) et (3) :

x_i^k désigne le $i$ème individu de la classe k,

$\bar{x}^k$ désigne l'individu moyen de la classe k,

et $\bar{x}$ la moyenne générale.

Le thèorème de Huygens permet d'écrire :

$$T = D + E \ .$$

Les deux méthodes d'analyse évoquées précédemment ont pour caractéristique commune de donner la même importance aux n_p paramètres pour caractériser une observation. Les différences sont résumées dans le tableau B-26 ci-dessous.

	a) Analyse en composantes principales	b) Analyse factorielle discriminante
But	. essentiellement descriptif	. à la fois descriptif et décisionnel
Idée	. pas de regroupement des données en classes	. partition a priori des données en classes que l'on ne remet pas en cause
	. on cherche à ajuster au mieux le nuage des points individus dans l'espace $\mathbb{R}^{n_p}$ des n_p paramètres	. on cherche à séparer au mieux les classes tout en agrégeant au mieux les points à l'intérieur de chaque classe
Principe	. recherche des composantes principales du nuage des n_T points i.e. des vecteurs u qui maximisent $u^T T u$	. recherche des vecteurs u qui maximisent la variance interclasses tout en minimisant la variance intraclasse i.e. recherche des premiers u qui maximisent $\frac{u^T E u}{u^T D u}$ ou $\frac{u\ E u}{u^T T u}$
Méthode	. diagonaliser T et retenir les vecteurs propres associés aux p plus grandes valeurs propres	. diagonaliser $T^{-1}E$
Résultats	. les vecteurs propres sont les composantes principales du nuage	. chaque valeur propre (toujours < 1) chiffre le pouvoir discriminant du vecteur associé
	. les valeurs propres associées chiffrent l'énergie du nuage	
Distance	. l'espace $\mathbb{R}^{n_p}$ est muni de la métrique unité	. l'espace $\mathbb{R}^{n_p}$ est muni de la métrique T^{-1}

Fig. B-26 : Caractères distinctifs de l'analyse en composantes principales et de l'analyse factorielle discriminante.

Nous donnons ensuite quelques précisions sur la méthode d'analyse factorielle discriminante (s'intéressant à la matrice $T^{-1}E$) :

- On suppose que la loi de dispersion des données est la même à l'intérieur de toutes les classes. Nous avons considéré en plus un nombre égal $n = n_T/n_c$ d'individus par classe,

- Les valeurs respectives de n_T , n_p et n_c ne sont pas indifférentes. En effet, si toutes les données sont indépendantes, on a :

$$\text{rang } (T) = \inf(n_p\ ,\ n_T - 1)$$

$$\text{rang } (D) = \inf(n_p\ ,\ n_T - n_c)$$

$$\text{rang } (E) = \inf(n_p\ ,\ n_c - 1)\ .$$

Dans notre cas, $n_p = 15$ et n_c vaut quelques unités. La matrice T est toujours inversible dès que l'on prend plusieurs représentants (indépendants entre eux) par classe. Par ailleurs, le rang de la matrice $T^{-1}E$ vaut $n_c - 1$, ce qui signifie qu'il existe $n_c - 1$ axes discriminants. En particulier, si l'analyse ne concerne que deux classes, il existe un seul axe discriminant ; la valeur propre unique donne alors la distance généralisée des deux classes (D^2 de Mahalanobis).

- Au niveau algorithmique, la matrice $T^{-1}E$ n'étant pas symétrique, on peut utiliser une technique classique qui ramène sa diagonalisation à celle d'une matrice symétrique C , ce pour quoi il existe des algorithmes éprouvés. En effet, E et T peuvent s'écrire respectivement :

$$T = \frac{1}{n_T}\ X_c^T\ .\ X_c \qquad (4)$$

$$\text{et} \quad E = \frac{1}{n_c}\ X_{MC}^T\ .\ X_{MC} \qquad (5)\ ,$$

où X_c désigne le tableau par lignes $(n_T\ ,\ n_p)$ des données centrées et X_{MC} le tableau $(n_c\ ,\ n_p)$ des moyens de classes centrées par rapport à la moyenne générale. On montre alors aisément que :

$$u = T^{-1}\ .\ X_{MC}\ .\ v \qquad (6)$$

$$\text{et} \quad v = \text{vecteurs propres de } C = \frac{1}{n_c}\ X_{MC}\ .\ T^{-1}\ .\ X_{MC}^T \qquad (7)\ .$$

- Un raffinement consisterait à réduire la variable j pour un individu de la classe k par son écart-type à l'intérieur de la classe. La variabilité des spectres qui constituent nos données ne justifie pas ce traitement.

- Un individu y extérieur au corpus d'apprentissage se projettera dans l'espace des L premiers axes discriminants $\{u_\ell, \ell = 1, \ldots, L\}$ en un point de coordonnées :

$$p_\ell = u_\ell^T \cdot (y - \bar{x}) \quad , \quad \ell = 1, \ldots, L \qquad (8) .$$

D'autre part, il se situera par rapport aux barycentres de classes à partir du calcul des distances :

$$d_k^2 = (y - \bar{x}^k)^T \cdot T^{-1} \cdot (y - \bar{x}^k) \quad , \quad k = 1, \ldots, n_c \qquad (9) .$$

Cette distance, dite *de Mahalanobis*, associée à la métrique T^{-1} dans l'espace de départ, équivaut à la distance euclidienne (métrique I) dans l'espace d'arrivée, ce qui donne une valeur informante aux proximités dans cette nouvelle représentation, comme nous le montrons sur les figures du paragraphe suivant.

2. Situations de calcul

Pour illustrer les méthodes d'analyse rappelées ci-dessus, nous nous plaçons dans les quatre situations de calcul suivantes :

Analyse / Données	en composantes principales	factorielle discriminante
.centrées	cas n° 1	cas n° 3
.normalisées à énergie constante .centrées	cas n° 2	cas n° 4

La normalisation des données-spectres à énergie totale constante des cas 2 et 4 est faite pour tenter de s'affranchir des variations d'intensité : réglage de volume, distance au microphone, niveau de la voix. Il faut remarquer que l'on réduit alors la dimension de l'espace des paramètres de départ d'une unité.

Nous donnons dans la suite des illustrations comparatives pour ces quatre situations, après avoir décrit le logiciel d'apprentissage, d'analyse et de visualisation que nous avons mis au point.

III - LOGICIEL D'APPRENTISSAGE, D'ANALYSE ET DE VISUALISATION - EXEMPLES [HATO - 79]

Le logiciel mis au point comporte trois modules :

1. acquisition supervisée des données vocales : seuil d'énergie contrôlé, attaques et queues de phonèmes supprimées, etc. Les classes retenues correspondent la plupart du temps à des sons tenus (ce qui exclut les consonnes plosives) ou des ensembles de tels phonèmes. Des vecteurs moyens sont calculés à partir de zones stables de quelques centièmes de seconde. La saisie "quasi immatérielle" des données peut être considérée comme une originalité dans le domaine de l'analyse d'importants corpus,

2. réduction des données en mode conversationnel avec contrôle point par point des calculs intermédiaires et des rangs des matrices pour chacun des quatre cas retenus,

3. acquisition de données qui n'ont pas participé à l'apprentissage, évaluation de proximité aux données d'apprentissage et visualisation en temps réel. La visualisation se fait dans le plan des deux vecteurs propres correspondant aux plus grandes valeurs propres, soit de T, soit de $T^{-1}E$ suivant le cas.

Le schéma B-27 indique les liens entre ces trois modules.

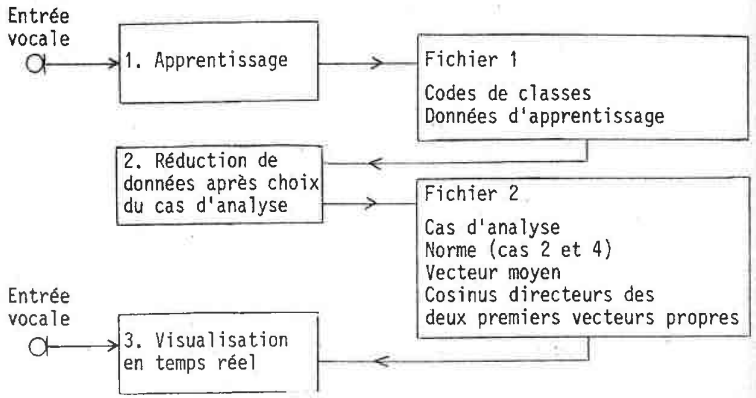


Fig. B-27 : Logiciel d'apprentissage, de réduction et de visualisation de données spectrales.

Les figures qui suivent donnent les résultats de l'apprentissage effectué par S. (11 ans, normalement entendant) pour quelques essais :

- distinction /f/-/s/-/ʃ/ . La figure B-28 représente les nuages des données-spectres d'apprentissage dans le plan des deux premières composantes principales ou des deux premiers facteurs discriminants suivant la situation d'analyse choisie. Les cas 1 et 2 donnent une idée de la dispersion des données d'origine. Les cas 3 et 4 montrent l'avantage du cas n° 3 (analyse factorielle discriminante sur les données centrées non normées) où le rôle décisionnel de la représentation est optimal.

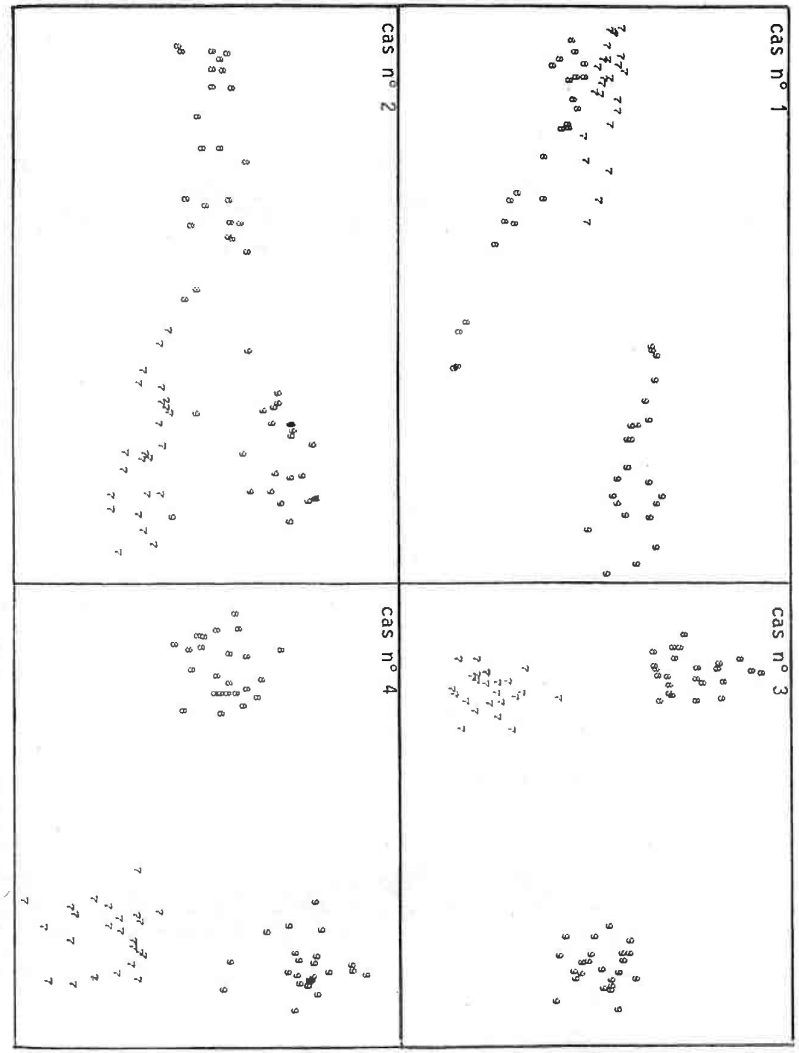


Fig. B-28 : /f/, /s/ et /ʃ/ notés respectivement 7, 8 et 9
Représentation des données d'apprentissage.

La figure B-29 représente les spectres moyennes de classes ainsi que les quantités $\{\frac{E(i,i)}{T(i,i)}$, i = 1, ..., $n_p\}$ qui donnent le pouvoir discriminant associé à chacun des filtres de l'analyseur spectral,

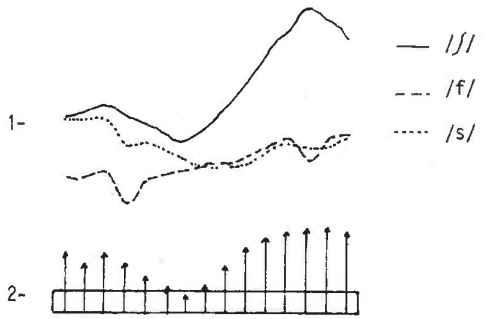


Fig. B-29 : Distinction /f/-/s/-/ʃ/
1- Spectres moyens
2- Pouvoir discriminant des filtres

- distinction /y/-/u/ et /o/-/õ/. De la même façon, la figure B-30 illustre les résultats obtenus pour des séries de deux classes. Le pouvoir discriminant des filtres est ici très directement lié à l'écart entre les deux courbes moyennes de classes. Rappelons qu'il n'existe ici qu'un axe factoriel discriminant et que les nuages de données se répartissent sur une droite en principe de part et d'autre du point moyen,

- distinction /a/-/i/-/u/ (figure B-31). On a représenté ici complètement les spectres d'apprentissage par leurs histogrammes sur l'ensemble {0., 0.5, 1., ..., 6.5, 7.} avec en pointillé le spectre moyen pour chacun des trois phonèmes. Le schéma d) indique la répartition fréquentielle des 15 filtres utilisés. En e), on donne leur pouvoir discriminant.

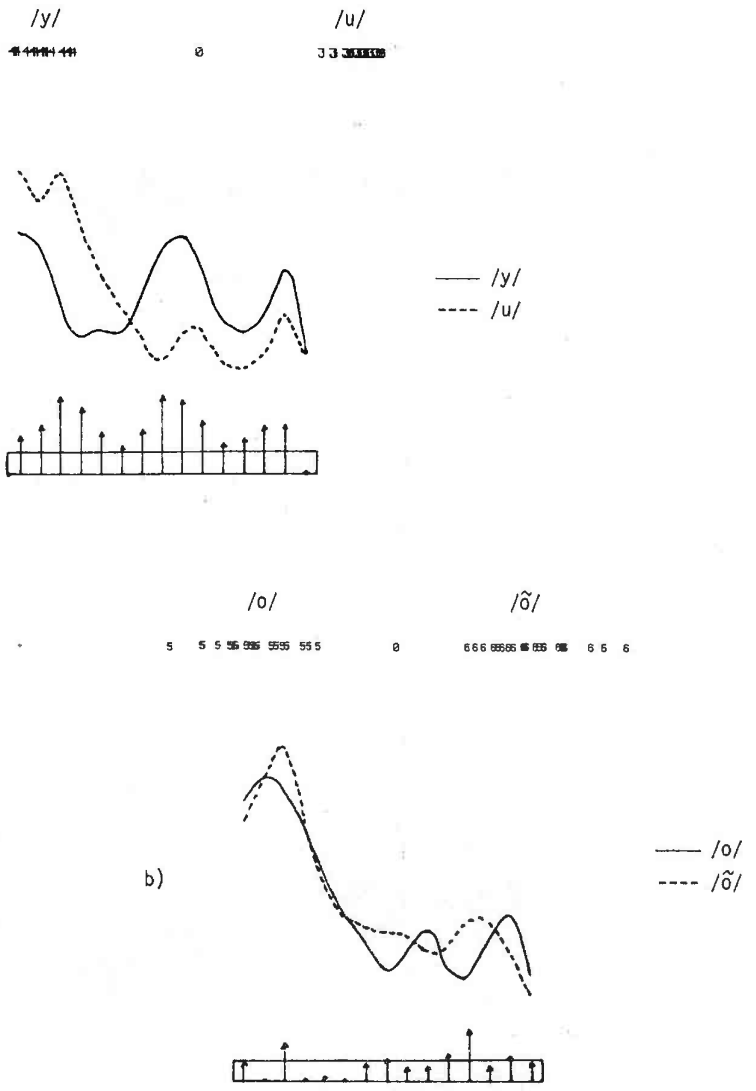


Fig. B-30
a- Distinction /y/-/u/
b- Distinction /o/-/õ/

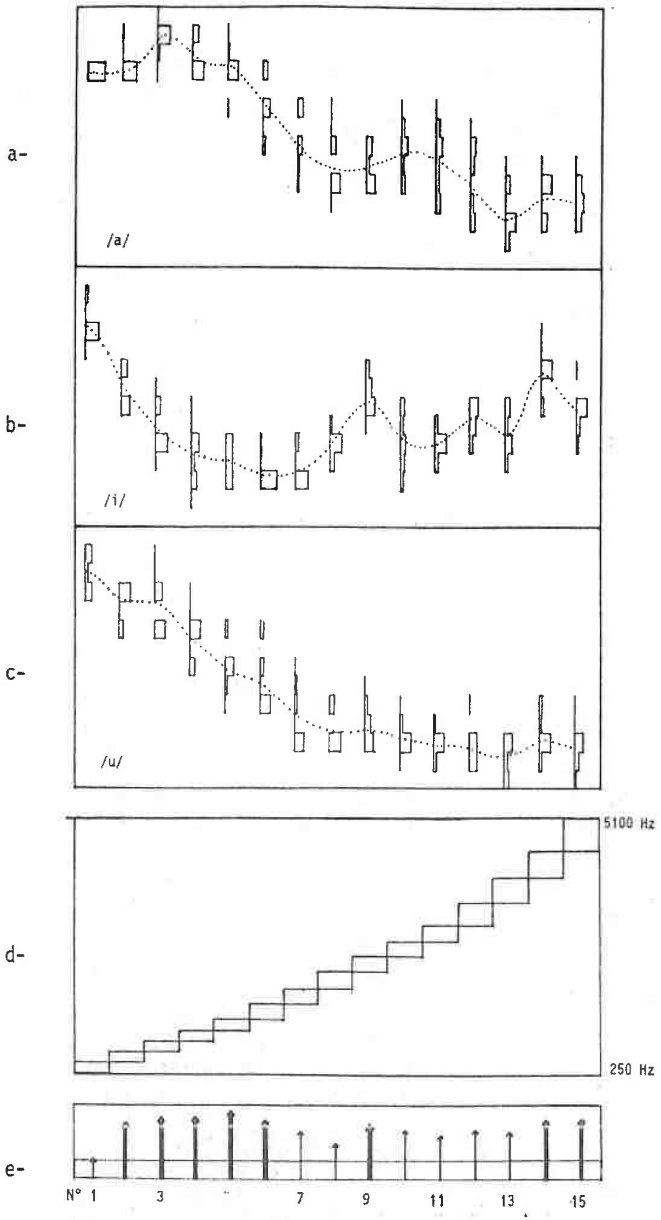


Fig. B-31 : Distinction /a/-/i/-/u/

a, b et c : histogrammes des spectres d'apprentissage

d : répartition fréquentielle des filtres de l'analyseur

e : pouvoir discriminant de chacun des filtres.

La figure B-32, pour terminer, représente en a) les trois nuages d'apprentissage /a/, /ɔ/ et /u/ après analyse factorielle discriminante et en b) un exemple de projection, obtenue en temps réel, de données vocales prononcées par le même enfant.

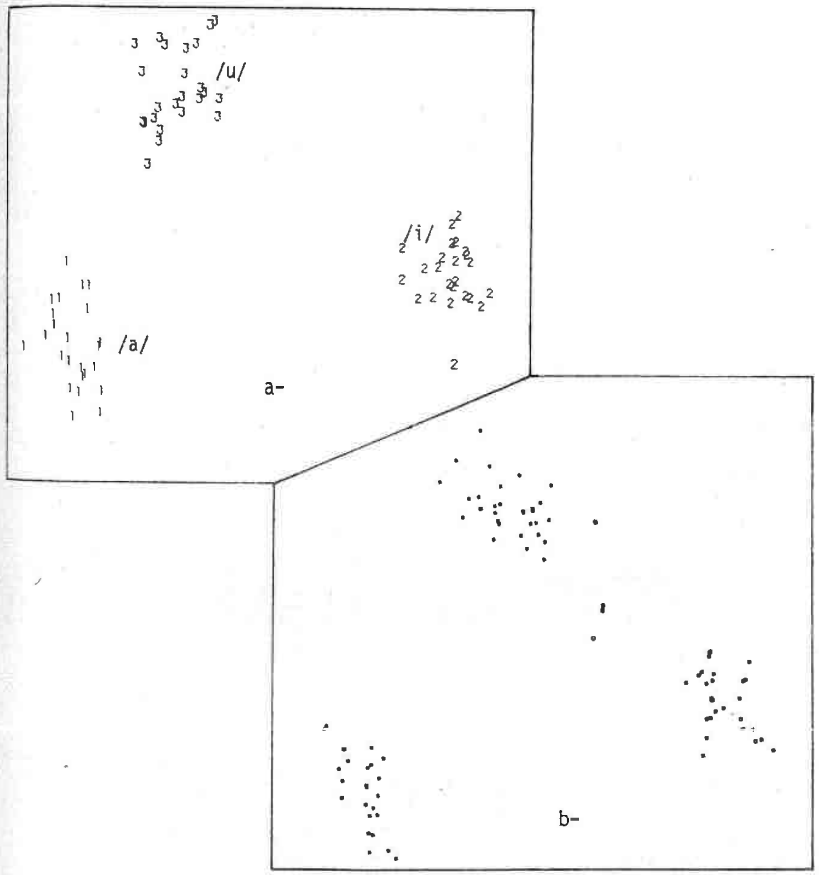


Fig. B-32 : Projection de données vocales

a : données d'apprentissage

b : données projetées en temps réel

IV - CONCLUSION

Les méthodes d'analyse envisagées dans ce chapitre trouvent deux terrains d'application dans notre travail :

• recherche des paramètres (donc des filtres) en nombre restreint permettant de caractériser au mieux un ensemble de classes de phonèmes. Sur les figures B-29 et B-31, les flèches indiquent le pouvoir discriminant de chacun des filtres. L'obtention de tels paramètres présente plusieurs intérêts , en particulier pour la définition de filtres optimaux au sens de la meilleure discrimination à l'intérieur d'un ensemble donné de classes (dans le but, par exemple, de commander par les sorties de filtres des composantes chromatiques pour une association son-couleur, cf. partie D.1.III),

• prise de décision visuelle sur la qualité d'articulation d'un phonème grâce à sa projection dans le plan des deux premiers facteurs obtenus par apprentissage. Il est à remarquer que le centrage avant projection est inutile puisqu'il ne joue que sur une translation identique pour toutes les données. L'interprétation se fait grâce à l'appréciation visuelle de la distance euclidienne entre points représentatifs dans le plan de l'écran de visualisation. Nous précisons et illustrons ce propos en partie D.

PARTIE C

ETUDE DES VOIX

INTRODUCTION A LA PARTIE C

Avant de concevoir des systèmes d'aide à la production vocale, tout comme pour la mise au point de systèmes de reconnaissance ou de compréhension du discours parlé adaptés à une population donnée, il est important de se munir d'outils logiciels d'étude des voix (pathologiques ou non). Ces outils, faisant appel aux techniques décrites en partie B, doivent faciliter la saisie à partir de l'énoncé direct ou d'enregistrements sur support analogique, la segmentation et l'analyse des données vocales.

Le premier chapitre de cette troisième partie concerne ces questions. Nous décrivons pour commencer les outils logiciels que nous avons mis au point en indiquant la cohérence entre les différents modules. Dans l'optique de la caractérisation de contours variant dans le temps (comme le contour mélodique), nous envisageons ensuite le problème de l'approximation et de l'interprétation de contours.

Le chapitre 2 est consacré à la comparaison et au traitement de formes sonores matricielles (essentiellement spectrogrammes en ce qui nous concerne), en vue principalement de l'évaluation des performances et des capacités de commande vocale du sujet locuteur.

Au chapitre 3, nous détaillons dans un premier temps le plan que nous proposons pour l'analyse des voix pathologiques. Nous donnons ensuite un aperçu des perspectives offertes par les travaux en modélisation (production et perception de la parole) et en synthèse vocale.

Nous concluons sur l'intérêt du "bilan vocal" pour l'aide à l'orientation à donner à la rééducation et l'aide au diagnostic médical.

CHAPITRE 1

OUTILS LOGICIELS

I - LOGICIELS D'ACQUISITION, D'ETIQUETAGE ET DE TRAITEMENT DE SEGMENTS DE PAROLE

Dans la nécessité de systématiser et de simplifier les opérations préliminaires à l'extraction de paramètres en temps différé d'éléments de parole, nous avons contribué à la mise au point d'un certain nombre de logiciels entièrement compatibles, fonctionnant en mode conversationnel et faisant appel, si nécessaire, à des modules écrits par d'autres chercheurs du laboratoire.

Nous faisons mention, dans ce paragraphe, des logiciels d'acquisition de parole numérisée simplement ou prétraitée, d'observation et d'étiquetage de segments et de traitement de ces segments.

1. ACQOBS, logiciel d'acquisition numérique de parole avec contrôle auditif éventuel

a) Généralités

Ce logiciel permet la constitution en temps réel, à partir d'une élocution au microphone ou d'un enregistrement analogique sur bande magnétique, de constituer un ou plusieurs des fichiers suivants :

- un fichier dit "*de données convertisseur*" contenant le signal d'origine filtré et numérisé sur 10 bits à une fréquence pouvant varier de 7,5 à 20 kHz (ou plus grâce à une horloge externe). Les filtres d'entrée permettent de se limiter à la zone informante de la parole par élimination des fréquences inférieures à 80 Hz et des fréquences supérieures à 4 500, 5 500 ou 7 500 Hz ,

- un fichier dit *"de données vocodeur"* renfermant les valeurs fournies par l'analyseur spectral, c'est-à-dire (cf. B.1.IV.1) :

. les données spectrales,

. la valeur de la fréquence fondamentale, s'il y a lieu.

Ces données sont gardées sous forme d'une succession de "prélèvements vocodeurs" ; chacun de ces prélèvements étant constitué d'un message de 60 bits il occupe sous forme tassée 4 mots-mémoire de 16 bits. Le prélèvement est décodé à la recherche dans le fichier pour observation, écoute ou traitement éventuel,

- un fichier dit *"de données mélographe"* qui, sur un mot, contient le nombre de fois qu'il y a 64 µs dans la période fondamentale détectée.

Ces trois fichiers peuvent être constitués séparément ou en simultanéité. Dans ce deuxième cas, ils sont synchronisés comme indiqué plus loin au paragraphe I.2., de sorte qu'un segment de parole repéré dans l'un des fichiers peut être aisément retrouvé dans un autre.

Les caractéristiques de l'acquisition sont les suivantes, après observation et réglage du niveau d'enregistrement :

- déclenchement au-dessus d'un seuil d'énergie choisi en fonction du capteur et du bruit ambiant et prise en compte uniquement des séquences de durée supérieure à un "seuil de bruit" préfixé,

- suppression des zones de silence de durée supérieure à un "seuil de silence" au-dessous duquel l'absence d'énergie peut correspondre à l'occlusion d'une consonne plosive,

- calcul en temps réel du nombre de valeurs du signal numérisé qui atteignent -512 ou +511 (donc où il y a saturation) et arrêt de l'acquisition si un taux de dépassement prédéfini est atteint. Il convient, dans ce cas, de réétudier le réglage du volume d'entrée et la distance au microphone s'il y a lieu,

- calcul du nombre des secteurs (suite de 128 valeurs) de signal où l'une au moins des valeurs dépasse un certain niveau d'intensité. Une fois l'acquisition terminée, le rapport de ce nombre au nombre total de secteurs renseigne sur la qualité de l'enregistrement.

L'écoute sélective de portions de fichiers après conversion numérique-analogique, dans un cas, ou synthèse par le vocodeur, dans l'autre, permet de juger du succès et de la qualité de l'acquisition.

b) Diagramme

Le schéma C-1 suivant indique l'organisation du programme conversationnel d'acquisition et d'écoute des fichiers de parole.

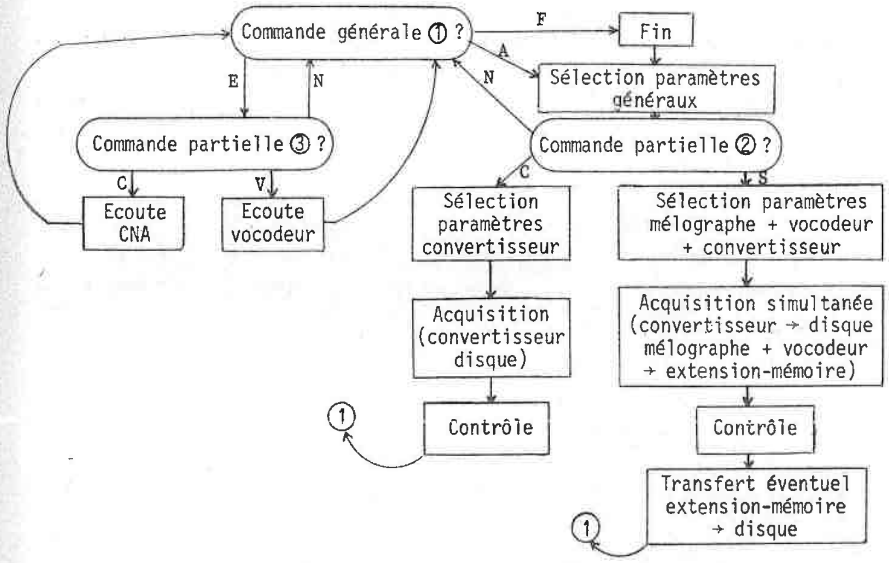


Fig. C-1 : Acquisition et écoute des fichiers de parole

c) Exploitation des fichiers de données de parole

Les fichiers ainsi constitués sont ensuite exploités grâce à des logiciels d'observation et de traitement tel *"l'observateur"* élaboré par C. SANCHEZ [SANC - 77]. Nous mentionnons, dans le sous-paragraphe 2 suivant, un logiciel que nous avons développé à l'origine pour le cas particulier des données fournies par l'analyseur spectral puis étendu aux fichiers obtenus par acquisition simultanée.

2. SONOBS, logiciel d'observation et d'étiquetage d'éléments de parole

a) Généralités

• But du programme -

Le but de ce logiciel est l'observation et l'étude de données vocales acquises en temps réel et stockées en fichiers sur disque. Les traitements peuvent être lancés à partir de l'observation directe ou par recherche d'un segment déjà sélectionné antérieurement et dont les caractéristiques sont rangées dans un fichier dit *"fichier de marques de segmentation"*. Nous ne donnerons ici que les grandes lignes de ce programme.

• Nature des données -

Les "données vocodeur" sont constituées de suites de prélèvements de sortie de l'analyseur spectral à quinze canaux CIT-Alcatel avec lequel nous travaillons. Chacun de ces prélèvements comprend :

. les quinze valeurs de l'intensité du signal capté au microphone dans quinze bandes de fréquence contiguës s'étendant de 250 à 5 000 Hz ,

. la valeur de la fréquence fondamentale du signal.

Ces seize valeurs entières sont codées sur quatre mots et rangées en fichier sous cette forme "tassée" au moment de l'acquisition. Elles sont lues et détassées au moment de leur utilisation.

• Acquisition des données -

Les fichiers sur disque de données vocodeur sont constitués à raison de 50 ou 100 prélèvements par seconde. Il est possible grâce au programme d'acquisition simultanée (cf. C.1.I) d'acquérir en synchronisme le signal de parole filtré et discrétisé par l'intermédiaire du convertisseur analogique-numérique. Dans ce deuxième cas, le "prélèvement vocodeur" en cours est sauvegardé chaque fois que l'on a rempli un tampon de 128 valeurs de "données convertisseur". La fréquence d'échantillonnage du signal qui a été retenue est de 12 kHz de sorte que, par rapport à l'acquisition centiseconde, 8 prélèvements sur 128 sont ignorés. Ce détail ne présente pas d'inconvénients majeurs même au niveau de la réécoute, après synthèse au vocodeur.

• Visualisation des "données vocodeur" -

Le sonagramme est présenté sur l'écran d'une console alphanumérique sous forme de quatre-vingts colonnes de quinze caractères. Chaque colonne correspond à un prélèvement vocodeur. Le temps se déroule de gauche à droite. Suivant les conditions d'acquisition (fréquence 50 Hz, 100 Hz ou acquisition simultanée), un écran complet correspond à une durée de 1.6 , 0.8 ou 0.85 seconde.

• Visualisation des "données convertisseur" et "données mélographe" synchrones -

Le signal de parole synchrone peut à la demande être présenté sur écran graphique. Le tracé correspond ainsi à quatre-vingts secteurs de 128 valeurs. Le contour mélodique peut être superposé au signal d'origine.

La figure C-2 suivante correspond à la visualisation des trois types de données mentionnées précédemment pour la séquence de parole. Le spectrogramme numérique est codé sur 8 niveaux représentés ici par des symboles plus ou moins noirs.

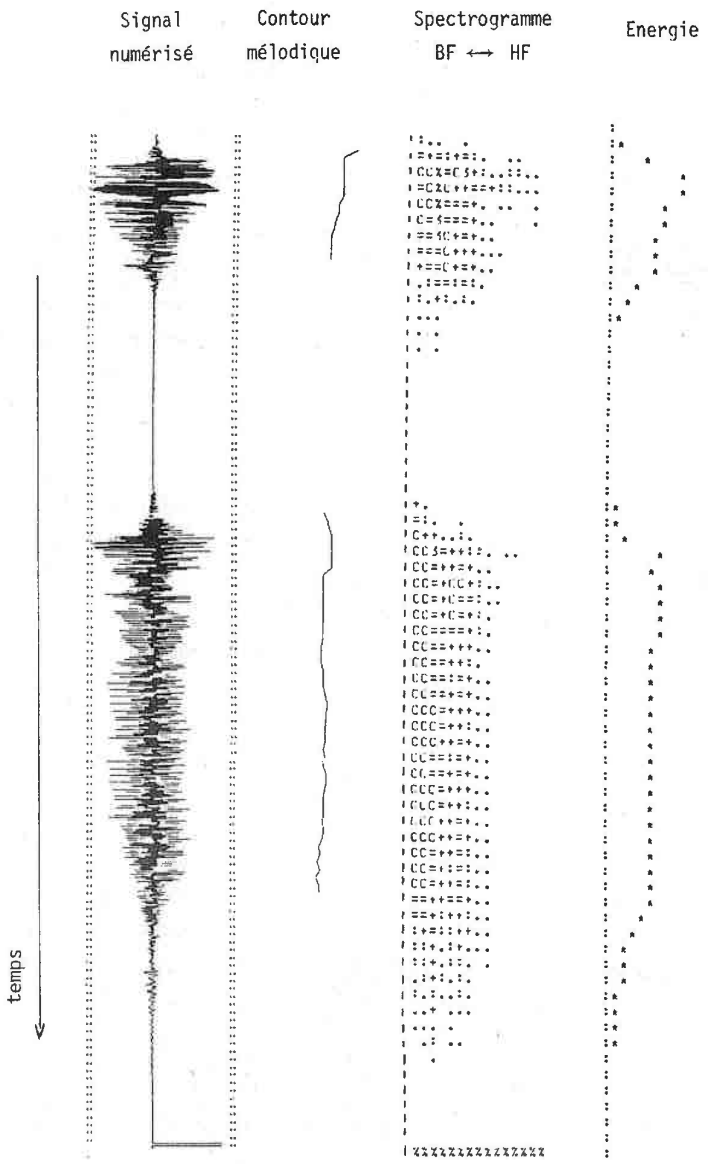


Fig. C-2 : Résultat d'acquisition simultanée
Segment /p a p a/ .

b) Diagramme d'ensemble

Le diagramme C-3 suivant indique l'organisation du programme conversationnel SONOBS.

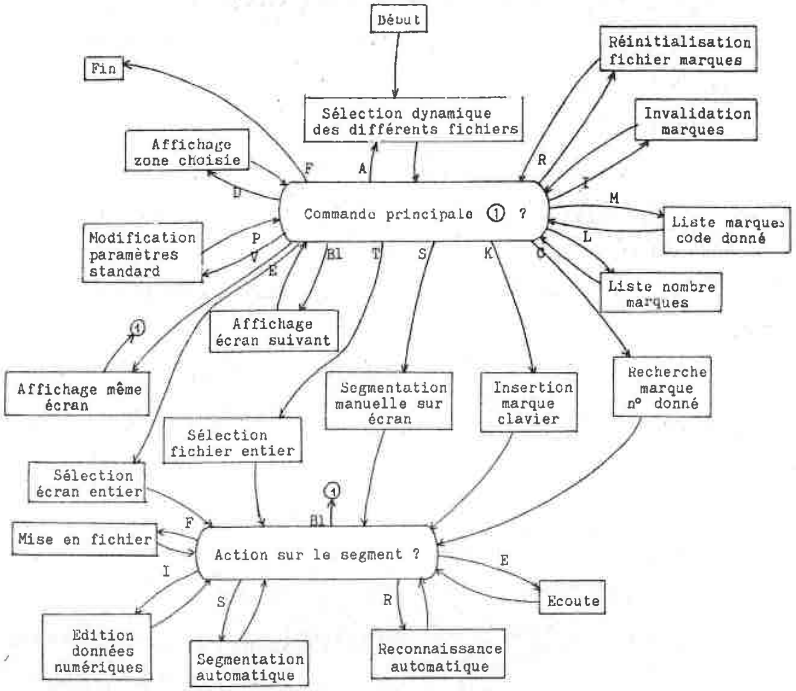


Fig. C-3 : Organisation de SONOBS.

c) Les actions principales

Les actions principales comprennent :

- la modification éventuelle des paramètres standard,

- la sélection au clavier du point à partir duquel on désire l'affichage du sonagramme,

- l'affichage du sonagramme suivant dans l'ordre chronologique,

- la réassignation des fichiers de travail utilisés,

- la sélection comme segment d'étude de l'écran en cours ou du fichier tout entier,

- la segmentation manuelle à partir de l'observation directe sur l'écran,

- l'introduction directe au clavier des paramètres d'un segment.

Dans le cas où l'on met en jeu les échanges avec un fichier de marques de segmentation, les mouvements suivants sont possibles :

- la réinitialisation du fichier de marques,

- l'invalidation d'une marque contenue dans le fichier,

- l'insertion d'une marque à partir du clavier,

- l'édition du contenu du fichier pour chacun des codes de segment,

- la recherche d'une marque repérée par son numéro d'ordre dans le fichier.

d) Les actions sur un segment

L'utilisateur dispose ainsi de trois façons de sélectionner le segment de données de parole sur lequel il veut lancer une action :

1 : la segmentation manuelle à partir de l'observation directe d'une suite d'écrans de 80 prélèvements chacun,

2 : l'introduction manuelle, au clavier, des paramètres permettant de récupérer le segment,

3 : la recherche des paramètres du segment sélectionné antérieurement, gardés dans un fichier de marques de segmentation.

Une fois sélectionné, le segment peut subir les traitements suivants :

- rangement en fichier primaire de marques - si cela n'a jamais été fait - des paramètres repérant le segment. A l'intérieur d'un code donné, les segments sont chaînés par ordre chronologique de début dans le fichier initial donné,

- édition du segment de sonagramme accompagné du contour mélodique et du contour d'intensité (l'intensité étant définie comme la somme des valeurs de sortie des quinze canaux de l'analyseur spectral). A ces tracés peuvent être adjoints différents contours donnant l'évolution de paramètres utiles en segmentation automatique de la parole,

- écoute du segment une ou plusieurs fois de suite,
- reconnaissance automatique, appliquée au segment,
- ajustement du segment par suppression des silences et des bruits,
- segmentation automatique en sous-segments dont les caractéristiques peuvent être rangées dans le fichier secondaire de marques de segmentation.

e) Les fichiers de marques

Comme il a été dit, les marques de segmentation sont chaînées par ordre chronologique pour chaque code de segment possible. Pour des raisons d'organisation du fichier de marques, nous avons retenu 127 codes possibles (notés de 0 à 126). Ce nombre est suffisant pour inclure la trentaine de phonèmes constitutifs du français parlé et des segments tels que éléments de phrases ou syllabes retenus dans nos études.

Une partie catalogue renseigne sur la fonction tête permettant d'accéder à la première marque d'un code donné et renferme, avec l'indication du nombre de marques, une chaîne de caractères permettant d'identifier chacun des codes.

Les paramètres de chacune des marques, auxquels sont joints les liens de chaînage avec la marque prédécesseur et la marque successeur, constituent des enregistrements placés séquentiellement dans l'ordre d'insertion dans le fichier.

L'intérêt de ce chaînage par code réside dans la possibilité de considérer pour un même traitement ultérieur tous les segments répertoriés sous ce numéro de code.

3. ETUDE, logiciel d'extraction de paramètres

Ce logiciel, qui fait appel aux techniques de traitement décrites dans la partie B, fait suite aux modules décrits dans les deux sous-paragraphes précédents.

La figure C-4 traduit l'enchaînement et la compatibilité entre ces différents modules.

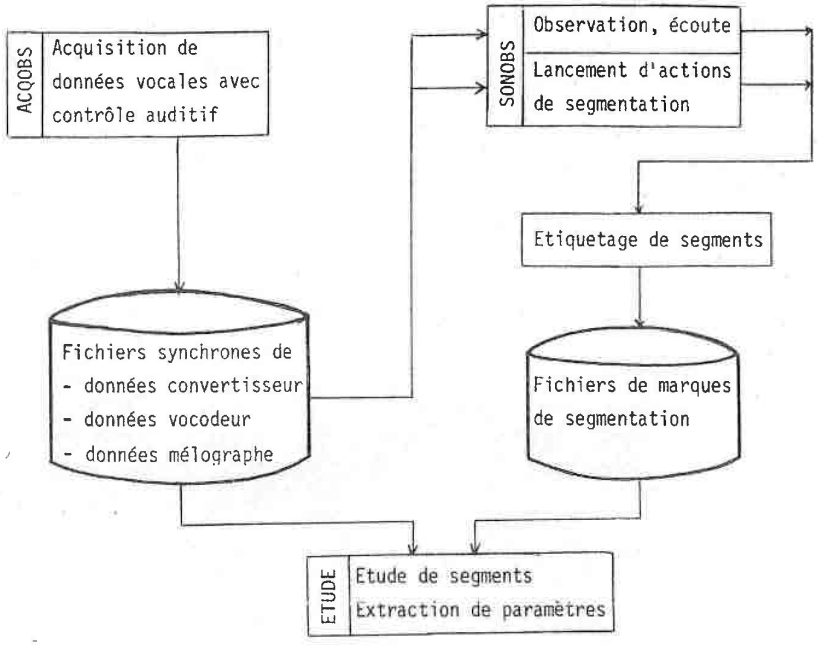


Fig. C-4 : Compatibilité entre les logiciels d'acquisition, d'étiquetage et d'étude de segments de parole.

Le programme d'étude permet, à partir de segments étiquetés et suivant les techniques développées en partie B, de lancer sur les segments d'un code ou d'une suite de codes donnés (par exemple, tel phonème ou tel mot) tout ou partie des traitements suivants :

a) Sur les "données convertisseur" (signal temporel) :

- calcul de l'énergie, du taux de passages par zéro, des aires délimitées par la courbe et l'axe d'amplitude zéro, du nombre d'extrema...,

- étude de la "structure fine" des zones voisées du signal à partir des cycles "minmin" déjà évoqués : taux de variations relatives de l'amplitude, écart des amplitudes et des variations d'amplitude par rapport à la sinusoïde non amortie,

- transformation rapide de Fourier, estimation du fondamental, détection des maxima maximorum,

- traitement de prédiction linéaire, calcul du spectre "lissé", des maxima du spectre, des pôles de la fonction de transfert, étude de la fonction d'aire, calcul du signal résiduel,

- autres traitements numériques : filtrage linéaire, détection de Fo , etc.,

b) Sur les "données vocodeur" :

- calcul de vecteurs moyen ou horizon, lissages,

- recherche de zones de concentration de l'énergie, appréciation du timbre, projections dans des plans de discrimination optimale,

c) Sur les "données mélographe" liées à l'intensité globale :

- étude statistique de Fo ,

- étude du contour mélodique : analyse polynomiale, corrélation fondamental-intensité, etc.

Nous reviendrons sur ces différents points à propos de l'étude des voix pathologiques.

II - LISSAGE ET INTERPOLATION DE CONTOURS

1. Cas d'application

L'utilisation de bancs de filtres ou de techniques de traitement numérique du signal fournit couramment des suites de valeurs pertinentes en nombre limité,

- soit qu'elles correspondent aux paramètres caractérisant un événement à un instant donné (spectre à court terme par exemple),

- soit qu'elles traduisent l'évolution sur un petit laps de temps d'un paramètre unique.

Il est souvent intéressant de donner à ces suites de valeurs numériques une représentation graphique sous forme de contours.

Dans certains cas, la résolution du tracé est telle que le contour obtenu est suffisamment régulier. C'est par exemple le cas des spectres obtenus après traitement de prédiction linéaire où la présentation graphique est directement réalisable. Dans une situation intermédiaire, comme pour un tracé de contour mélodique où l'on ne veut pas faire apparaître de variations locales, une correction de chaque point à partir des k points voisins peut être suffisante.

Au contraire, on peut se trouver en présence de séries de points sans lien très serré entre eux. C'est le cas par exemple des approximations de spectre fournies par un banc de filtres ou des paliers décrivant la configuration instantanée du canal vocal.

Dans cette dernière situation et dans l'esprit de représenter les contours autrement que par des courbes anguleuses ou en paliers et d'en dériver des paramètres exploitables, nous avons recherché une méthode de

calcul susceptible d'effectuer un lissage des données de départ par un procédé mixte en quelque sorte d'approximation et d'interpolation : l'idée est d'augmenter le nombre de points d'arrivée tout en limitant l'ordre des oscillations.

A partir de l'idée d'interpolation trigonométrique dont un des avantages est de limiter les effets de bord aux extrêmités du contour, nous proposons un développement mathématique répondant à nos spécifications. Nous adoptons des notations matricielles qui simplifient considérablement l'écriture et nous montrons qu'un simple produit matriciel permet de passer du vecteur des valeurs de départ à celui des valeurs d'arrivée.

2. Notations et développement théorique

Les n abscisses de départ sont supposées régulièrement réparties sur l'intervalle $]0,\pi[$ et données par :

$$x_i = \frac{2i - 1}{2n}\,\pi \quad , \quad i = 1, 2, \ldots, n \quad .$$

La fonction discrète de départ est la suite des $f(x_i)$.

Les N abscisses d'arrivée, également régulièrement réparties, vérifient :

$$y_j = \frac{2j - 1}{2N}\,\pi \quad , \quad j = 1, 2, \ldots, N \quad .$$

La base de fonctions orthogonales retenue est constituée des m fonctions suivantes :

$$1/\sqrt{2} \ , \cos x \ , \cos 2x, \ \ldots, \cos (m - 1)\, x,$$

m désignant l'ordre de l'approximation.

Approcher la fonction initiale $\{f(x_i)\}$ au sens des moindres carrés par une combinaison linéaire des fonctions de base (avec une fonction de pondération identiquement égale à 1), c'est rechercher l'ensemble de coefficients $\{c_k, k = 1, 2, \ldots, m\}$ qui vérifient :

$$\min_{\{c_k\}} e_m \quad \text{où} \quad e_m = \sum_{i=1}^{n} \left[f(x_i) - \frac{c_1}{\sqrt{2}} - \sum_{k=2}^{m} c_k \cdot \cos(k - 1)\, x_i\right]^2$$

(condition 1) .

Posons $g(x_i) = \dfrac{c_1}{\sqrt{2}} + \sum_{k=2}^{m} c_k \cdot \cos(k - 1)\, x_i$.

La quantité e_m qui caractérise l'erreur d'approximation s'écrit alors :

$$e_m = \sum_{i=1}^{n} \left[f(x_i) - g(x_i)\right]^2 .$$

La condition 1 équivaut aux m conditions suivantes :

$$\frac{\partial e_m}{\partial c_k} = 0 \quad , \quad k = 1, \ldots, m \quad \text{(conditions 2)} .$$

Soient les écritures matricielles suivantes :

F = matrice-colonne des $f(x_i)$, $i = 1, \ldots, n$

G_a = matrice colonne des $g(y_j)$, $j = 1, \ldots, N$

C = matrice-colonne des c_k , $k = 1, \ldots, m$

$$V_d = \begin{bmatrix} \frac{1}{\sqrt{2}} \cos x_1 \ \ldots \ \cos (m - 1)\, x_1 \\ \vdots \\ \frac{1}{\sqrt{2}} \cos x_i \ \ldots \ \cos (m - 1)\, x_i \\ \vdots \\ \frac{1}{\sqrt{2}} \cos x_n \ \ldots \ \cos (m - 1)\, x_n \end{bmatrix} \quad \text{et} \quad V_a = \begin{bmatrix} \frac{1}{\sqrt{2}} \cos y_1 \ \ldots \ \cos (m - 1)\, y_1 \\ \vdots \\ \frac{1}{\sqrt{2}} \cos y_j \ \ldots \ \cos (m - 1)\, y_j \\ \vdots \\ \frac{1}{\sqrt{2}} \cos y_N \ \ldots \ \cos (m - 1)\, y_N \end{bmatrix}$$

Avec ces conventions, e_m s'écrit :

$$e_m = (F^T - C^T V_d^T)(F - V_d C)$$

et les conditions 2) conduisent à :

$$V_d^T F - V_d^T V_d C = 0 \qquad (3) ,$$

soit encore, du fait de la propriété d'orthogonalité des fonctions de base choisies :

$$C = \frac{2}{n} V_d^T F \qquad (4) .$$

(Il est connu en effet que les coefficients c_k qui minimisent l'erreur d'approximation au sens des moindres carrés par un ensemble de fonctions orthonormales v_k sont les produits internes (v_k , f)) .

Nous avons d'autre part :

$$G_a = V_a C \qquad (5) .$$

Soit LISS la matrice de la transformation permettant de passer de $\{f(x_i) , i = 1, n\}$ à $\{g(y_j) , j = 1, N\}$. Cette matrice vérifie :

$$G_a = \text{LISS} \times F \qquad (6)$$

Les relations (4) à (6) entraînent :

$$\text{LISS} = \frac{2}{n} V_a V_d^T \qquad (7)$$

Un élément de la matrice LISS se met alors sous la forme :

$$\text{LISS}(j,i) = \frac{2}{n} \left[\frac{1}{2} + \sum_{k=1}^{m-1} \cos kx_i \, . \cos ky_j\right]$$

$$= \frac{1}{n} \left[\frac{1}{2} + \sum_{k=1}^{m-1} \cos k\alpha + \frac{1}{2} + \sum_{k=1}^{m-1} \cos k\beta\right]$$

où $\alpha = x_i + y_j$ et $\beta = x_i - x_j$.

Un calcul classique conduit au résultat suivant :

$$\text{LISS}(j,i) = \frac{1}{2n} \left[\frac{\sin(2m - 1)\frac{\alpha}{2}}{\sin \frac{\alpha}{2}} + \frac{\sin(2m - 1)\frac{\beta}{2}}{\sin \frac{\beta}{2}}\right] \qquad (8)$$

où $\alpha = \frac{\pi}{2} \left[\frac{2i - 1}{n} + \frac{2j - 1}{N}\right]$ et $\beta = \frac{\pi}{2} \left[\frac{2i - 1}{n} - \frac{2j - 1}{N}\right]$.

Pour un nombre de points de départ n donné, N doit être choisi de façon que $\sin \frac{\beta}{2}$ ne puisse s'annuler. La condition nécessaire et suffisante en est que la fraction $\frac{N}{n}$ ne se réduise pas au rapport de deux nombres entiers impairs (c'est-à-dire au cas où un x_i et un y_j coïncident).

Deux cas peuvent être ainsi distingués :

i) n est impair : il faut et il suffit de prendre pour N n'importe quel entier pair,

ii) n est pair : il suffit de prendre N impair mais bien d'autres valeurs conviennent (si l'on raisonne avec $N > n$, toutes celles telles que l'écart $N - n$ n'est pas un multiple pair de la plus grande puissance de 2 divisant n).

Quelle que soit la parité de n , les expressions suivantes conviennent pour N :

. $n + 2p + 1$ (p entier)

. $2n$

. toute puissance de 2 supérieure à n .

Les éléments de la matrice LISS donnés par (8) ne dépendant que des valeurs de n, N et du choix de m peuvent être calculés en dehors de leur exploitation immédiate et mémorisés. Il est à noter que le temps de calcul nécessité par la transformation (6) est tout à fait indépendant de m et que le choix de cet ordre d'approximation ne dépend que de l'appréciation des oscillations à conserver dans le contour d'arrivée.

3. Applications

Une illustration du lissage obtenu par la méthode exposée dans ce paragraphe a été donnée sur la figure B-12, sur un exemple de fonction d'aire. Nous reprenons cet exemple sur la figure C-5 où l'on voit en superposition :

. la fonction d'aire en paliers obtenue après traitement de prédiction linéaire,

. la même fonction lissée ($n = 12$, $N = 24$) pour deux ordres de lissage différents : $m = 7$ et $m = 11$.

Le choix de m dépend de n (qui représente ici l'ordre de prédiction linéaire) et de l'importance de l'oscillation autorisée. Dans le cas présent, une valeur de l'ordre de n DIV 2 + 2 nous semble convenable. Nous utilisons cette technique de lissage (cf. partie D) pour la présentation en ligne de la fonction d'aire, d'une part, et de contours spectraux obtenus à partir de l'analyse par vocodeur, d'autre part.

a

b

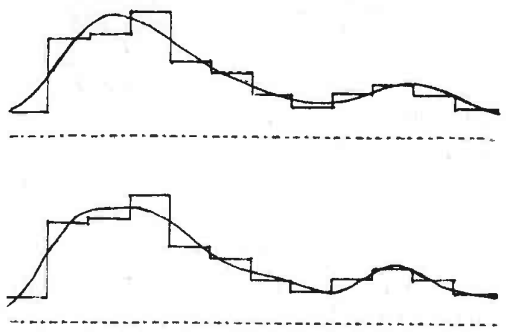


Fig. C-5 : Fonction d'aire en paliers et lissée avec $m = 7$ (a) ou $m = 11$ (b) .

4. Annexe

A l'occasion de ce paragraphe traitant de lissage et d'interpolation de contours, nous précisons la façon dont nous détectons la position d'un extremum à partir des valeurs prises par une fonction discrète f pour des valeurs de la variable régulièrement espacées et ceci avec le minimum de calcul (cas d'un maximum dans un spectre lissé obtenu à partir des sorties discrètes de l'analyseur spectral par exemple).

La figure C-6 représente un pic caractérisé par trois points successifs : celui qui correspond au maximum d'une fonction (abscisse i_M) et ses deux voisins (abscisses $i_M - 1$ et $i_M + 1$).

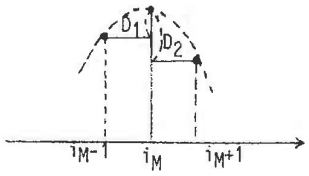


Fig. C-6 : Maximum local d'une fonction discrète.

Soit $D1 = f(iM) - f(i_M - 1)$, quantité positive

et $D2 = f(i_M + 1) - f(i_M)$, quantité négative.

Sur le schéma ci-dessus où $|D2| > D1$, on considère que le maximum correspond à une abscisse comprise entre $i_M - 1$ et i_M . Nous avons choisi d'estimer l'abscisse x_M du maximum en effectuant la correction :

$$x_M \text{ estimé} = \tilde{x_M} = i_M + \frac{D1 + D2}{2(D1 - D2)}$$

qui vérifie les propriétés suivantes :

si $D1 = |D2|$ alors $\tilde{x_M} = i_M$,

si $D1 \to 0$ alors $\tilde{x_M} \overset{>}{\to} i_M - \frac{1}{2}$,

si $D2 \to 0$ alors $\tilde{x_M} \overset{<}{\to} i_M + \frac{1}{2}$.

La même formule permet le repérage de la position des minima.

D'après les remarques formulées en partie B concernant le déplacement des maxima de spectres de résonance par rapport aux pics réels, il est clair que cette méthode n'est pas valide pour de tels spectres (obtenus par FFT après LPC ou non en particulier), sauf si le but est une comparaison pure des maxima à ceux de courbes-témoins.

III - ANALYSE DE PORTIONS DE CONTOURS

1. Introduction et développement théorique

Nous avons déjà évoqué que toute tentative de rééducation vocale pourrait être avantageusement précédée, dans le cadre du bilan orthophonique par exemple, d'une analyse acoustique assez fine de la parole normale.

Nous nous sommes intéressé en particulier aux variations temporelles rapides de la fréquence fondamentale pour tenter de caractériser les défauts de commande au niveau du larynx [HATO - 75].

Pour cela, nous avons choisi de traduire le taux d'oscillation de portions de contour mélodique par leurs coefficients de projection sur une base de polynômes orthogonaux. La base retenue est celle des polynômes de Tchebycheff qui, parmi d'autres, résulte de l'orthogonalisation des

fonctions monômes x^k , $k \geqslant 0$ (mais le raisonnement qui suit est valable pour tout ensemble orthogonal de combinaisons linéaires de monômes à condition de respecter l'intervalle de variation de x).

Soit n le nombre de points de la portion de contour étudiée, les abscisses équiréparties dans l'intervalle $[-1,1]$ vérifiant :

$$x_i = 2 . \frac{i - 1}{n + 1} - 1 \quad , \quad i = 1, \ldots, n$$

et soit $\{f(x_i)\}$ la suite des valeurs du contour.

Le problème envisagé ici est, comme au paragraphe précédent, celui d'une approximation d'ordre m qui peut être égal à n (c'est alors le cas de l'interpolation vraie). Nous avons cherché à le résoudre simplement, sans faire appel aux techniques d'ajustement habituelles (méthode de Cholesky, par exemple) pour les raisons qui suivent.

D'après la propriété de "permanence", l'approximation au sens des moindres carrés fournit des paramètres c_k (cf. condition 1) dont la valeur n'est pas affectée par l'accroissement de l'ordre m . D'autre part, le petit nombre de points du contour ne pose pas de problèmes de temps de calcul ni de place mémoire. Pour ces deux raisons, nous avons calculé simplement la matrice permettant de passer directement des valeurs $\{f(x_i)$, $i = 1, \ldots, n\}$ de départ aux coefficients c_k .

La relation (3) du paragraphe précédent :

$$C = \frac{2}{n} V_d^T . F$$

est toujours valable avec, d'après le choix que nous avons fait de la base des polynômes de Tchebycheff :

$$V_d(i,j) = T_{j-1}(x_i) = \cos[(j-1) \text{ Arc cos} x_i] \quad ,$$
$$i = 1, \ldots, n \quad , \quad j = 1, \ldots, n \quad .$$

On peut calculer facilement la matrice de passage de F à C par l'intermédiaire de la base des monômes x^k .

En effet, les paramètres recherchés sont les coefficients c_k de la combinaison linéaire :

$$g(x) = \sum_{k=1}^{n} c_k \, T_{k-1}(x)$$

telle que $g(x) = f(x)$ pour $x = x_1, x_2, \ldots, x_n$.

Soit T la matrice dont le terme général $T(i,j)$ est le coefficient de x^{i-1} dans le polynôme $T_{j-1}(x)$. Il vient :

$$g(x) = \left[1 \; x \; \ldots \; x^{n-1}\right] T \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = X^T.T.C \qquad (1) .$$

Le vecteur-colonne $T.C$ peut se déduire du vecteur-colonne des $f(x_i)$ grâce à l'introduction d'une matrice L , du type des matrices de Lagrange et calculée pour l'intervalle $[-1,1]$, suivant :

$$T.C = L.F .$$

La matrice T étant inversible, on obtient :

$$C = T^{-1}.L.F \qquad (2) .$$

La relation (2) ci-dessus nous permet d'atteindre directement les coefficients de l'interpolation de $\{f(x_i)\}$ de façon peu coûteuse, quelle que soit l'importance des variations rapides du contour.

La matrice carrée $T^{-1}.L$ (nxn) est calculée pour la valeur particulière choisie pour n. Pour $n = 5$, par exemple, on aurait :

$$L = \frac{1}{6} \times \begin{bmatrix} 0 & 0 & 6 & 0 & 0 \\ 1 & -8 & 0 & 8 & -1 \\ -1 & 16 & -30 & 16 & -1 \\ -4 & 8 & 0 & -8 & 4 \\ 4 & -16 & 24 & -16 & 4 \end{bmatrix} \quad \text{et} \quad T^{-1}.L = \frac{1}{12} \begin{bmatrix} 2 & 4 & 0 & 4 & 2 \\ -4 & -4 & 0 & 4 & 4 \\ 3 & 0 & -6 & 0 & 3 \\ -2 & 4 & 0 & -4 & 2 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix}$$

Après obtention des coefficients c_k grâce à la relation (2), l'erreur d'approximation faite en se limitant à l'ordre m (inférieur ou égal à n) s'évalue facilement. En effet, soit $\mathcal{F}_{i,j}$ la quantité :

$$\mathcal{F}_{i,j} = \sum_k T_i(x_k) \cdot T_j(x_k) .$$

L'erreur d'approximation commise en s'arrêtant à l'ordre m (polynôme $T_{m-1}(x)$ compris) s'écrit :

$$e_m = \sum_{x_k} [f(x_k) - \sum_{i=1}^{m} c_i \, T_{i-1}(x_k)]^2 \quad \text{pour} \quad 0 < m \leqslant n .$$

On a bien sûr $e_n = 0$.

Soit $e_o = \sum_{x_k} [f(x_k)]^2 = F^T . F$.

On montre alors que les quantités e_m se calculent par récurrence suivant la relation :

$$e_m = e_{m+1} + c_{m+1}^2 \, \mathcal{F}_{m,m} + 2c_{m+1} \sum_{\substack{p=m+2 \\ \text{et } p \leqslant n}}^{n} c_p \mathcal{F}_{m,p-1}$$

avec comme condition initiale $e_n = 0$.

e_o s'écrit simplement :

$$e_o = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, \mathcal{F}_{i-1,j-1} .$$

On peut remarquer que, les fonctions $T_{2p}(x)$ et $T_{2q+1}(x)$ étant respectivement paires et impaires, on a toujours :

$$\mathcal{F}_{2p,2q+1} = 0 ,$$

ce qui limite le nombre des multiplications à effectuer dans le calcul de l'erreur.

L'ordre m à partir duquel l'erreur devient inférieure à un seuil préfixé permet de chiffrer la stabilité ou au contraire le caractère erratique du contour étudié. Le critère retenu est souvent l'importance de la différence $e_o - e_m$.

Il est intéressant ici de noter les expressions de la moyenne μ et de la variance σ^2 de l'ensemble des $f(x_k)$. En effet, on peut montrer que l'on a :

$$\mu = \frac{1}{n} \sum_{i=1}^{n} c_i \, \mathcal{F}_{0,i-1}$$

et $$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, [\mathcal{F}_{i-1,j-1} - \frac{1}{n} \mathcal{F}_{0,i-1} \times \mathcal{F}_{0,j-1}] .$$

On a bien sûr :

$$\sigma^2 = \frac{e_o}{n} - \mu^2 .$$

2. Exemple d'application

Nous étudions des portions de contour mélodique échantillonnées à raison de 100 valeurs par seconde. Pour mettre en évidence les petites variations de la fréquence fondamentale tout en limitant l'importance des calculs et des tableaux de données, nous isolons de courtes portions du contour grâce à une fenêtre temporelle de 80 ms , ce qui correspond à la durée d'une phonème court.

Chaque fenêtre contient une suite de neuf points et balaie le contour dans le temps par déplacement de quatre intervalles. A partir de N valeurs consécutives de F_0 , l'analyse fournit ainsi neuf paramètres pour (N - 5) DIV 4 fenêtres .

Un exemple de résultats est donné au paragraphe C.3.II ou dans [HATO - 75]. Nous nous bornons ici à donner les courbes approchant une portion de contour pour des valeurs croissantes de l'ordre m (figure C-7). L'erreur d'approximation est mise en évidence sur la figure C-8 . En ce qui concerne le contour mélodique, la conclusion sera la suivante : plus il faut de coefficients c_i pour caractériser le contour, plus il y a de variations locales mal contrôlées. A l'inverse, de très faibles coefficients c_i , mis à part c_1 , évoqueront une voix sans expression aucune.

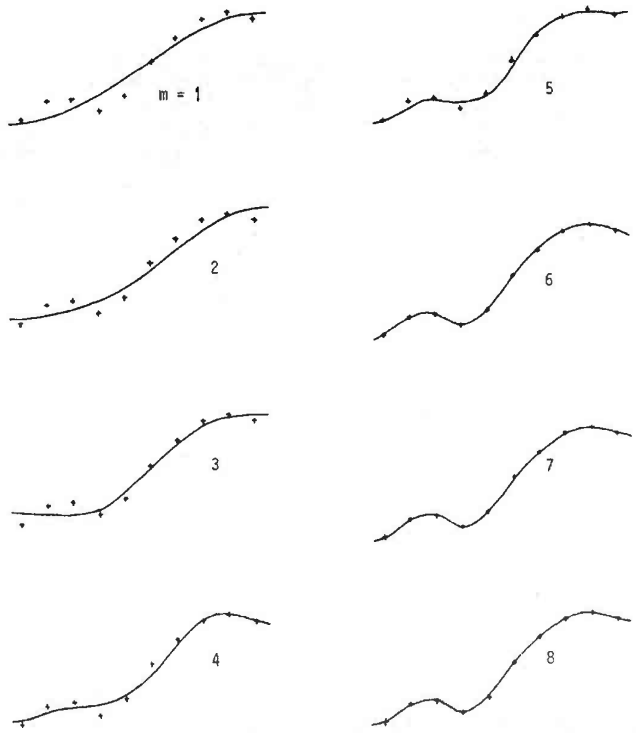


Fig. C-7 : Approximation de portion de contour mélodique pour m variant de 1 à 8.

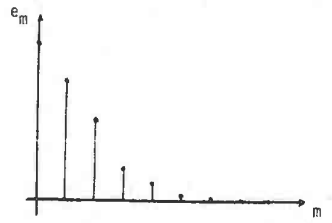


Fig. C-8 : Erreur d'approximation e_m en fonction de m .

CHAPITRE 2

COMPARAISON ET TRAITEMENT DE FORMES SONORES MATRICIELLES

I - INTRODUCTION

Dans l'optique de l'apprentissage oral par un élève d'un vocabulaire de mots et de l'étude des possibilités de commande orale de sujets I.M.C. (Infirmes Cérébraux Moteurs), nous avons été amené à réexaminer les méthodes de comparaison dite "globale" entre deux formes, ceci dans un double but :

- . la prise d'une décision à partir du taux de comparaison (reconnaissance, appréciation de l'élocution),
- . l'élaboration de formes moyennes à partir d'élocutions multiples du même mot.

Après avoir rappelé la structure des formes envisagées, nous présentons dans ce chapitre différents modes de comparaison de formes matricielles, leurs caractéristiques et leurs performances dans le contexte d'un logiciel général de test et d'étude de ces méthodes.

Nous décrivons ensuite une méthode d'élaboration de formes moyennes pour la constitution de vocabulaires de référence qui tiennent compte de différentes populations de locuteurs.

Nous terminons ce chapitre par les conclusions tirées de l'étude de mots enregistrés, prononcés par des sujets I.M.C., en ce qui concerne leurs possibilités d'action par la voix sur leur environnement.

II - TYPE DES DONNEES

Afin de diminuer au maximum les temps de traitement, nous nous plaçons dans les conditions d'acquisition numérique et de stockage de données spectrales en temps réel, soit en mémoire centrale, soit sur disque, comme l'indique la figure C-9 :

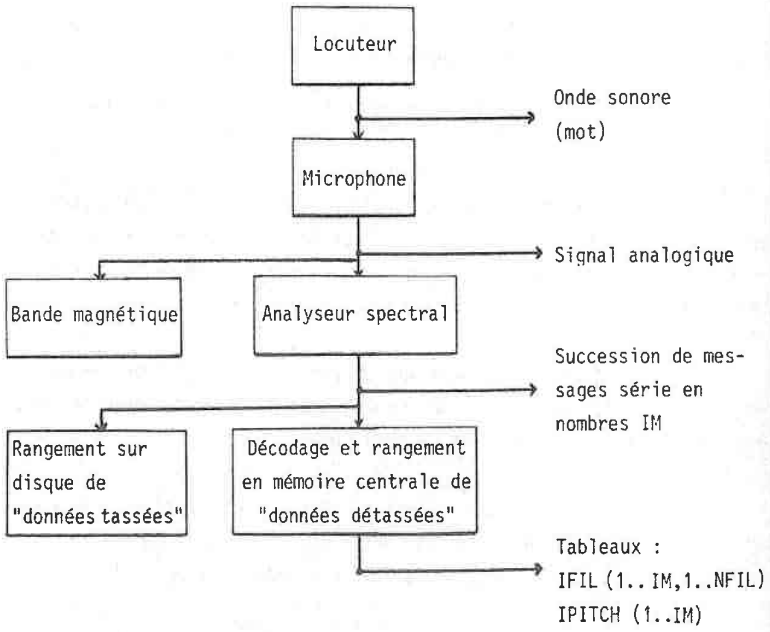


Fig. C-9 : Acquisition de données spectrales

Après décodage des messages fournis par l'analyseur spectral, on dispose de deux tableaux de données :

- **IFIL** qui renferme le spectrogramme du mot prononcé, succession dans le temps de IM "prélèvements". Chaque prélèvement est constitué des NFILT valeurs de sortie des filtres de l'analyseur (15 généralement),
- **IPITCH** qui renferme la suite des valeurs de la fréquence fondamentale.

Nous évoquons, aux paragraphes V et VI, l'élaboration de formes d'autres types mais respectant la même structure.

III - CONDITIONS DE COMPARAISON

Par la suite, les termes "*forme*" ou "*mot*" désignent, suivant le cas :

- soit un élément de phrase,
- soit sa représentation sous la forme matricielle d'un spectrogramme avec indication éventuelle du voisement.

Comparer deux mots revient à rechercher et à chiffrer les similitudes ou les dissemblances entre deux tableaux à deux dimensions dont l'une est variable, dans lesquels le temps se déroule de façon monotone. On cherche habituellement pour cela à faire coincider de façon optimale une partie ou l'ensemble des prélèvements de l'une des formes avec une partie ou l'ensemble des prélèvements de l'autre.

Pour ce faire, on procède par cheminement dans les deux formes suivant des indices croissants. Le résultat se présente comme un taux de dissemblance dont la valeur dépend du mode de progression choisi.

Dans le plan des indices des prélèvements des deux formes à comparer (de longueurs respectives IM et JM), le score final dépendra ainsi du chemin $\mathscr{C}_L$ suivi dans la progression, comme schématisé sur la figure C-10 ci-dessous.(Dans les schémas de la suite, suivant le cas, l'axe des indices de la forme n° 2 sera orienté vers le haut ou vers le bas, pour des raisons de commodité).

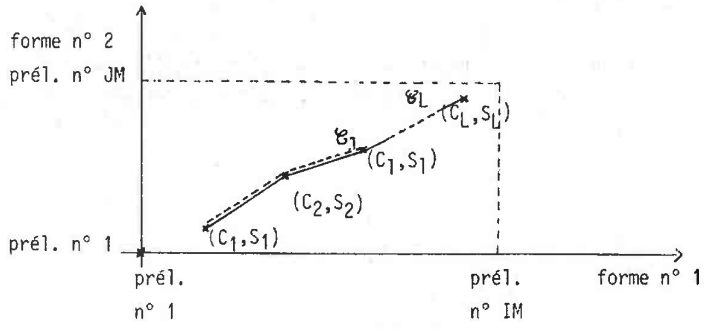


Fig. C-10 : Chemins de comparaison entre deux formes, vues l'une horizontalement, l'autre verticalement.
En pointillé : chemin partiel $\mathscr{C}_1$
En trait plein : chemin complet $\mathscr{C}_L$

La quantité S_1 désigne le score de la comparaison des deux formes, une fois parvenu à la coincidence C_1 des prélèvements d'indices i_1 et j_1 en suivant le chemin partiel $\mathscr{C}_1$.

On conserve pour score final la quantité S_L obtenue lors de la dernière mise en coincidence, le plus souvent sous la forme d'une quantité $\mathscr{S}$ "normalisée" de façon à tenir compte des longueurs IM et JM des formes ou de la longueur L du chemin $\mathscr{C}_L$.

Pour tenir compte de la réalité temporelle des deux formes, on impose une relation d'ordre croissant sur les valeurs des indices $i_1, i_2 \ldots, i_L$, d'une part, et des indices $j_1, j_2 \ldots, j_L$, d'autre part, ce qui interdit toute "remontée dans le temps".

Nous écrirons S_L comme une fonction des étapes antérieures et de la coincidence C_L, donc du chemin $\mathscr{C}_L$, soit encore, une fois les deux formes posées, comme une fonction du critère de cheminement Cr_C :

$$S_L = f\,(Cr_C)$$

La valeur finale retenue $\mathcal{S}$ dépend de plus d'une fonction de normalisation f_N :

$$\mathcal{S} = f\,(Cr_C,\ f_N)$$

Nous distinguerons au paragraphe III les méthodes rapides équivalant au "coup d'oeil" porté sur les deux formes et les méthodes dites "dynamiques" permettant de minimiser le score S_L au sens de critères Cr_C plus ou moins élaborés, comme indiqué ci-dessous (fig. C-11) :

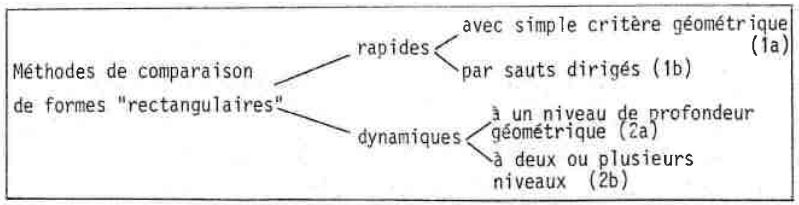


Fig. C-11 : Méthodes de comparaison décrites dans la suite.

Quelle que soit la méthode employée, il est astucieux pour réduire le nombre des calculs de limiter la zone d'étude à la région des coïncidences optimales les plus probables, ce qui revient en plus des relations d'ordre sur les i_1 et j_1 à imposer des conditions sur leurs valeurs. Nous définissons ainsi des "fenêtres-limite" de trois types, illustrées sur la figure C-12 :

a) F_B, fenêtre suivant la première bissectrice du plan des indices, qui présente l'avantage d'imposer une relation simple entre les indices i_1 et j_1 indépendamment des durées réelles des deux formes mais nécessite la définition d'une zone de prolongement si l'on veut poursuivre la comparaison jusqu'à la fin de la forme la plus longue,

b) F_D, fenêtre diagonale qui favoriserait les faibles distorsions entre les deux formes dont l'une serait le simple étirement de l'autre,

c) F_F, fenêtre en fuseau qui forcerait, par exemple, les coïncidences aux deux extrémités tout en laissant une plus grande latitude dans la partie moyenne. F_D et F_F peuvent être imposées simultanément (comme sur d).

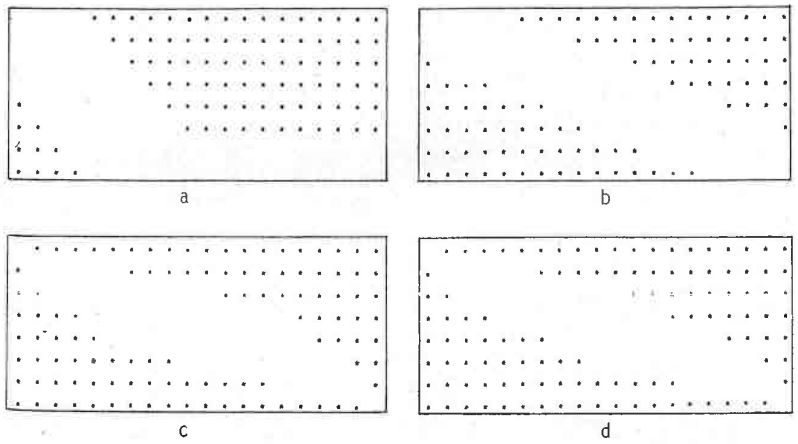


Fig. C-12 : Fenêtres-limite.

IV - METHODES DE COMPARAISON

Nous précisons, dans ce paragraphe, les quatre types de méthodes mentionnées plus haut en indiquant leurs caractéristiques propres.

1. Comparaison rapide

a) Avec critère de progression "géométrique"

Le chemin est prédéterminé dans le cadre $[1..IM, 1..JM]$ et le score se calcule par simple cumul des coûts des coïncidences. Ce qui s'écrit :

$$\left.\begin{aligned} C_{l+1} &= \text{fonction géométrique } (C_l) \\ S_{l+1} &= S_l + \text{coût } (C_l) \\ \varphi &= \frac{S_L}{L} \end{aligned}\right\} \quad (1a)$$

En choisissant pour chemin la ligne diagonale telle que d'une coïncidence à la suivante les indices i et j progressent d'une unité au plus, on se trouve dans la situation générale qui ne dépend que des valeurs respectives de IM et JM et qui équivaut à une normalisation temporelle linéaire des formes suivie d'une comparaison prélèvement par prélèvement.

b) Localement dirigée

La comparaison se fera par une suite de décisions locales lorsqu'une coïncidence de moindre coût sera détectée selon l'algorithme suivant :

Nombre de coïncidences n = 0
Coïncidence arbitraire de départ = C_o ($i_o = 1$ et $j_o = 1$ par exemple)
S = 0
Fin = faux
Tant que non fin faire
  Définir une zone d'étude Z_n à partir de C_n
  n = n + 1
  Rechercher la coïncidence C_{n+1} de moindre coût dans Z_n
  S = S + coût (C_{n+1})
  Fin = (Z_n contient le point-limite extrémité)
Fin tant que
L = n
φ = S/L

Il se trouve ainsi que la décision au rang n+1 n'est conditionnée que par la coïncidence C_n et les coûts dans Z_n, ce qui s'écrit :

$$\left.\begin{aligned} C_{n+1} &= f(Z_n) \\ S_{n+1} &= S_n + \text{coût } (C_{n+1}) \\ \varphi &= \frac{S_L}{L} \end{aligned}\right\} \quad (1b)$$

De même que dans le cas (1a), la simplicité de l'algorithme entraîne une mise en oeuvre rapide de la méthode. L'inconvénient majeur en est bien sûr le côté approximatif qui la rend sensible au choix de la taille de la zone d'étude.

2. Comparaison dynamique

La comparaison dynamique revient au calcul des coûts de tous les chemins respectant certaines contraintes géométriques et, par suite, la détermination du score de comparaison (coût total "normalisé") minimal. Seule la décision finale peut permettre par "remontée" de déduire la suite de coïncidences donc le chemin $\mathscr{C}$ affecté de ce coût minimal.

Nous désignerons par $C_p(i,j)$ une mise en coïncidence provisoire des prélèvements i d'une forme et j de l'autre, provisoire en ce sens que localement rien ne permet de conclure que la coïncidence appartiendra ou non au chemin optimal $\mathscr{C}$.
$c(i,j) = \text{coût}\,(C_p(i,j))$ désignera le coût de cette mise en coïncidence et $S(i,j)$ le score provisoire associé au chemin optimal passant par le point (i,j).

Parmi les contraintes géométriques en dehors de la fenêtre-limite envisagée plus haut en II, on peut retenir deux schémas de base tels que, d'une coïncidence provisoire à la suivante :

i) ou bien les indices i ou j progressent au plus d'une unité, ce qui conduit à un algorithme que nous dirons *"à un niveau de profondeur géométrique"*,

ii) ou bien l'un au moins des indices peut progresser de deux ou plusieurs unités, ce qui conduit à un algorithme *"à deux ou plusieurs niveaux"*.

Nous envisageons successivement chacun de ces deux points.

a) Comparaison dynamique à un niveau

La décision locale pour $C_p(i,j)$ se fait en fonction des trois plus proches voisins respectant la relation d'ordre sur les indices, ce qui peut s'écrire :

$$S(i,j) = f\big(S(i-1,\,j),\ S(i-1,\,j-1),\ S(i,\,j-1),\ c(i,j)\big)$$

On adopte souvent la relation [HATO - 74a] :

$$S(i,j) = \min\left\{\begin{matrix} S(i-1,j) \\ S(i-1,j-1) \\ S(i,j-1) \end{matrix}\right\} + c(i,j) \qquad (2a)$$

La figure C-13 indique les chemins optimaux provisoires aboutissant aux trois voisins de (i,j).

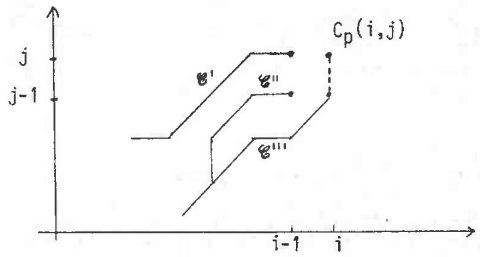


Fig. C-13 : en plein = chemins optimaux provisoires précédents
en pointillé = prolongement du chemin $\mathscr{C}'''$ vers le point (i,j) par suite d'une décision locale à partir des scores S', S'' et S''' ($S'>S'''$ et $S''>S'''$)

Suivant la relation (2a) ci-dessus, seul l'un des trois chemins provisoires se prolonge jusqu'au point (i,j), comme indiqué en pointillé sur la figure C-13. Sur cet exemple, le chemin "↗" est définitivement abandonné.

Il en sera de même pour la coïncidence C(IM,JM) en admettant que l'on impose que le chemin définitif atteigne l'extrémité des deux formes. Le score S(IM,JM) représentera ainsi le coût du chemin optimal.

Pour éviter de favoriser les formes courtes par rapport à d'autres formes plus longues mais plus proches du modèle, on peut choisir comme formule de normalisation :

$$\mathcal{S} = S(IM,JM) \;/\; (IM + JM)$$

ou bien :

$$\mathcal{S} = S(IM,JM) \;/\; L$$

où L désigne le nombre d'arcs entre les coïncidences effectives du chemin optimal.

L'algorithme ainsi décrit revient donc à amorcer l'exploration des chemins composés d'arcs :

→ ou ↑ ou ↗

intérieurs à la fenêtre-limite. Ces chemins sont par décision locale ou prolongés ou abandonnés. Un seul d'entre eux fera la liaison entre les coïncidences forcées de l'origine et de l'extrémité. Le résultat final

correspond à une sorte de mise en correspondance des deux formes par "*pliage accordéon*" de façon à les rapporter à durées égales.

Le schéma de base permettant de caractériser cet algorithme peut se figurer ainsi :

schéma méthode (2a)

L'algorithme (2a) peut s'écrire sous forme de deux itérations imbriquées sur les indices i et j sans perdre sa symétrie de la façon suivante :

pour une valeur de j donnée, on calcule tous les S(i,j) correspondant à des coïncidences provisoires intérieures à la fenêtre-limite. Il suffit, au moment du calcul de S(i,j), d'évaluer c(i,j) et d'avoir conservé en memoire les scores provisoires indiqués sur la figure C-14 :

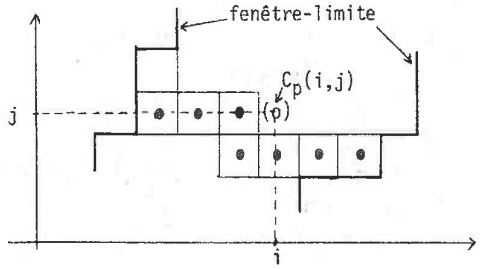


Fig. C-14 : Illustration de la programmation de l'algorithme 2a
(o) : coût ($C_p(i,j)$) = c(i,j)
• : scores provisoires calculés avant d'envisager $C_p(i,j)$

On attribue une valeur arbitrairement très grande aux scores provisoires des coïncidences extérieures à la fenêtre-limite.

b) Comparaison dynamique à deux ou plusieurs niveaux

Le schéma de base précédent autorise à mettre en coïncidence un prélèvement unique d'une forme avec une succession de plusieurs prélèvements de l'autre, ce qui peut se justifier sur des zones stables de durées nettement différentes. On s'éloigne pourtant de cette façon de la zone diagonale, ce que l'on peut empêcher par exemple en interdisant trois coïncidences successives sans variation de l'un des indices.

Ce parti-pris conduit au schéma de base suivant [SAKO - 78]

dans lequel, quelle que soit la façon dont se terminent les chemins provisoires précédents, le prolongement commence toujours par un arc incliné à 45°.

Le score associé à $C_p(i,j)$ peut s'exprimer suivant la relation (2b) suivante pour i et $j \geqslant 3$:

$$S(i,j) = \min \begin{Bmatrix} g[S(i-2,j-1),\ c(i-1,j),\ c(i,j)] \\ h[S(i-1,j-1),\ c(i,j)] \\ g[S(i-1,j-2),\ c(i,j-1),\ c(i,j)] \end{Bmatrix} \quad (2b)$$

L'algorithme est, là aussi, symétrique en i et j . Au moment du calcul de $S(i,j)$, il suffit d'évaluer $c(i,j)$ et d'avoir conservé en mémoire les quantités indiquées sur la figure C-15 ci-dessous :

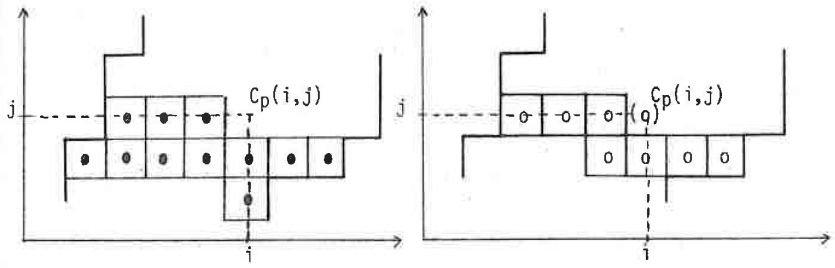


Fig. C-15 : Illustration de la programmation de l'algorithme 2b.

o , marqué en (i_k,j_k) , désigne $c(i_k,j_k)$

• , - - - - $S(i_k,j_k)$

(o) coût du nouveau $C_p(i,j) = c(i,j)$

La figure C-16 illustre des expressions simples des fonctions g et h :

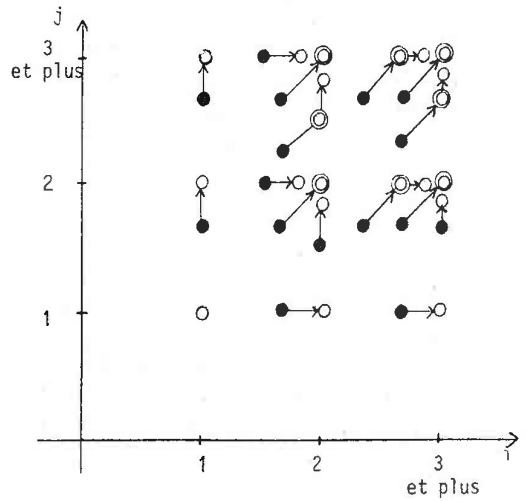


Fig. C-16 : Exemples de schémas de base pour une comparaison dynamique à deux niveaux de profondeur géométrique

Le signe ○ désigne le coût local de la coïncidence (à un coefficient α multiplicatif près)

Le signe ● désigne le score cumulé (c'est-à-dire le coût du chemin optimal qui conduit de l'origine au point considéré)

Enfin ◎ désigne le coût local affecté d'un coefficient $\beta > \alpha$

Les coûts situés sur le même arc se cumulent.

Par exemple, pour $j = 2$ et $i = 3$, on a :

$$S(3,2) = \min \begin{cases} S(1,1) + \beta\, c(2,2) + \alpha\, c(3,2) \\ S(2,1) + \beta\, c(3,2) \\ S(3,1) + \alpha\, c(3,2) \end{cases}$$

3. Illustrations des méthodes décrites en 1. et 2.

Comme indiqué au paragraphe I de ce chapitre, nous traitons communément des spectrogrammes acquis en temps réel à partir d'un analyseur spectral (constitué de filtres analogiques ou numériques indifféremment). Du fait de la compression de dynamique et du codage numérique des intensités de sortie des filtres, on ne peut leur accorder de valeurs en absolu. Nous avons cependant considéré que l'écart des valeurs dans le temps pour un même canal caractérise assez bien la variation d'énergie dans ce canal ; c'est pourquoi nous avons, dans le cas de ces spectrogrammes, conservé comme coût associé à une mise en coïncidence la fonction :

$$c(i,j) = \sum_{k=NF1}^{NF2} \left| e_i^k - e'_j{}^k \right| \quad (3)$$

expression dans laquelle :

- NF1 et NF2 désignent les numéros de filtres limitant la zone fréquentielle retenue, généralement, sauf exception, pris égaux à 1 et NFILT (nombre total) respectivement,

- e_i^k désigne la $k^{ième}$ composante du vecteur e_i représentant le $i^{ème}$ prélèvement de la première forme,
- $e'_j{}^k$ désigne la $k^{ième}$ composante du vecteur e'_j représentant le $j^{ième}$ prélèvement de la deuxième forme.

Ce choix de $c(i,j)$ trouve une justification complémentaire dans la simplicité et la rapidité du calcul. Il présente le désavantage d'être sensible au niveau d'acquisition qui se répercute sur les valeurs quantifiées donc sur leurs différences. La plupart du temps cependant, le fait de rester dans les mêmes conditions de segmentation parole - non parole à l'apprentissage et à l'utilisation minimise cet inconvénient.

Une fois posée la fonction $c(i,j)$, on peut caractériser le lien entre les deux formes étudiées par la matrice des coûts, chaque prélèvement de l'une étant mis en coïncidence provisoire avec chaque prélèvement de l'autre. Des exemples de matrices des coûts seront donnés plus loin.

La figure C-17 représente les spectrogrammes accompagnés de la valeur du fondamental, du contour mélodique et du contour d'intensité pour des occurences des mots suivants :

- "début", servant de référence dans la suite,
- "début" et "plus vite", jouant le rôle de mots inconnus.

Nous donnons sur les figures suivantes des illustrations des quatre méthodes définies sur le schéma C-11 :

- méthode 1a (fig. C-18a) où l'on impose un chemin diagonal. On représente sur les schémas deux types de renseignements obtenus en cours de comparaison :

1 : l'évolution des scores dans la comparaison du mot "début" à un vocabulaire de neuf mots (commande de fauteuil roulant). Une barre-limite rejette les mots de référence de score trop élevé avant que la comparaison soit terminée. Le score minimal est obtenu pour le mot de référence numéro 1 ("début"),

2 : les chemins de progression dans le plan des deux formes, ici rectilignes.

En comparant les deux situations de la figure, on voit le danger de la contrainte géométrique dans le cas de distorsion temporelle dans l'une des formes comparées,

- méthode 1b où la comparaison est localement dirigée. La figure C-18b reprend l'exemple précédent avec le même risque de divergence en cas de distorsion mais avec un nombre limité de comparaisons,

- méthode 2a de comparaison dynamique à un niveau de profondeur géométrique. Sur la figure C-19a, on a successivement les chemins provisoires (avec en gras le chemin optimal) dans la comparaison :

1 : de deux occurences du mot "début" avec fenêtres diagonales,

2 : de "début" et "plus vite" : on remarque le recadrage entre les deux mots sur la succession consonne plosive-liquide et voyelle-consonne plosive-voyelle,

3 et 4 : de deux occurences du mot "début" sans fenêtre limite. En 4 le /e/ du "mot inconnu", présenté horizontalement, est allongé,

- méthode 2b de comparaison dynamique. La possibilité de calcul du coût de mise en coïncidence de deux prélèvements suivant le schéma C-16 apparaît sur les tracés locaux de la figure C-19b. Comme dans les autres cas, le chemin optimal, lié au score associé à la comparaison, est obtenu par retour en arrière à partir du point d'arrivée. Notons que,

dans tous nos exemples, les deux formes considérées sont explorées du début à la fin étant données les conditions de travail (mots isolés par segmentation automatique parole-non parole).

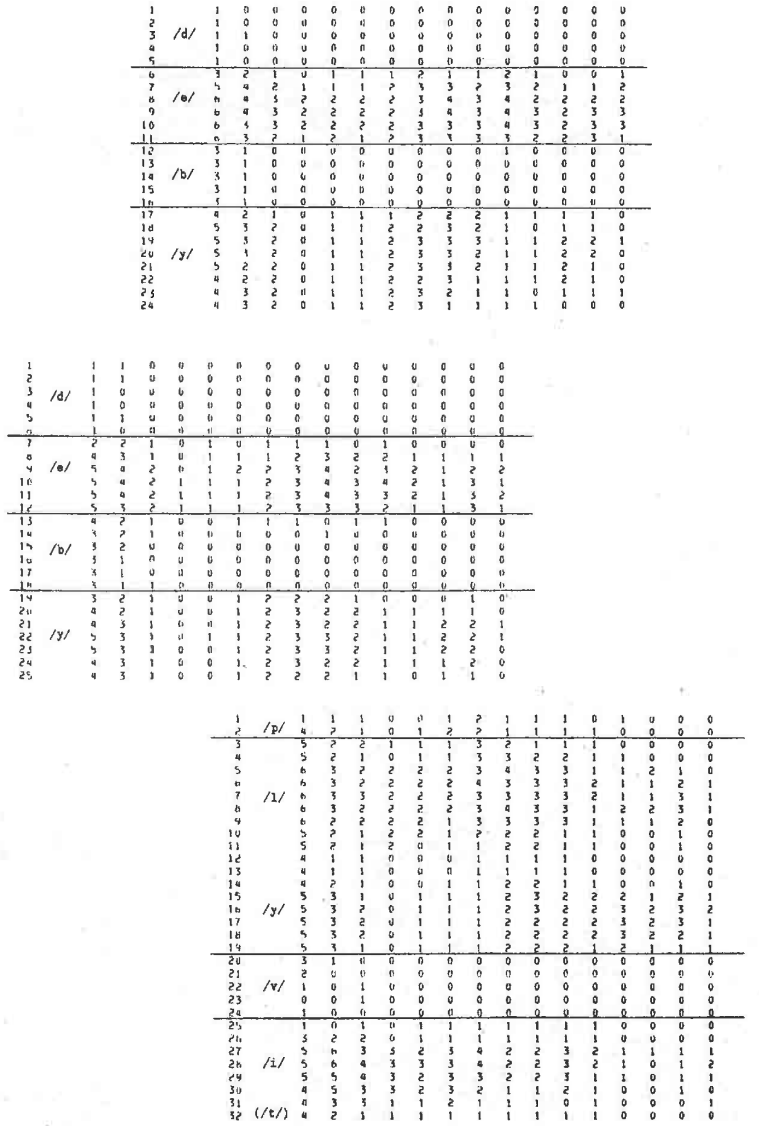
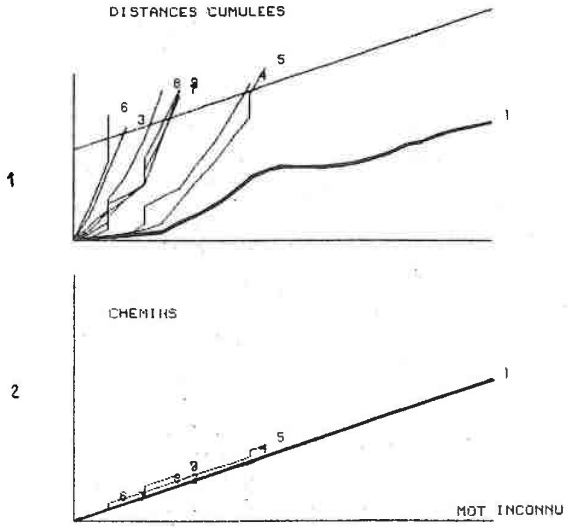


Fig. C-17 : Spectrogrammes a- "début" (référence)
b- "début"



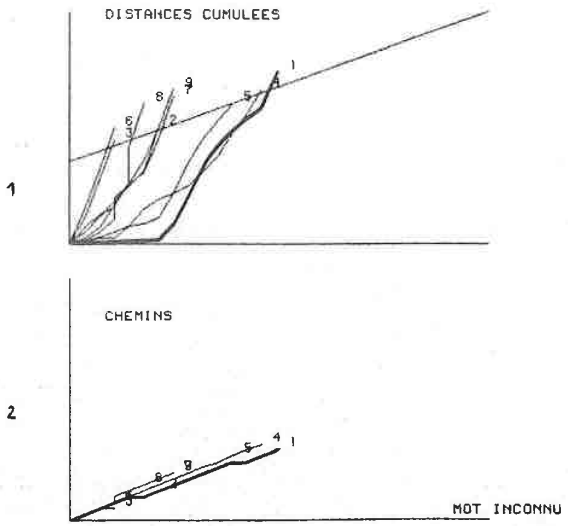


Fig. C-18a : Comparaison de type 1a avec pour mot inconnu :
. en haut : /deby/

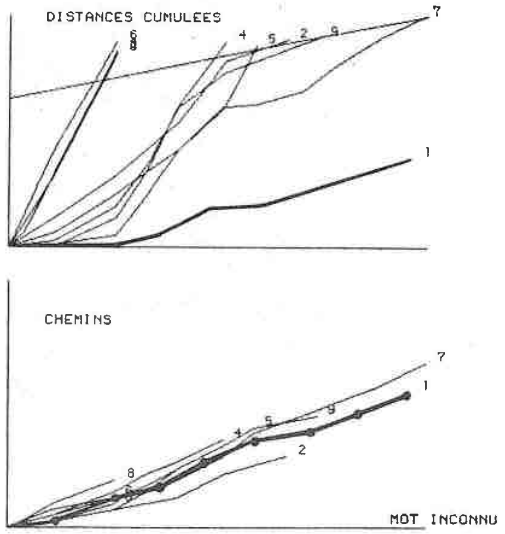


Fig. C-18b : Comparaison de type 1b

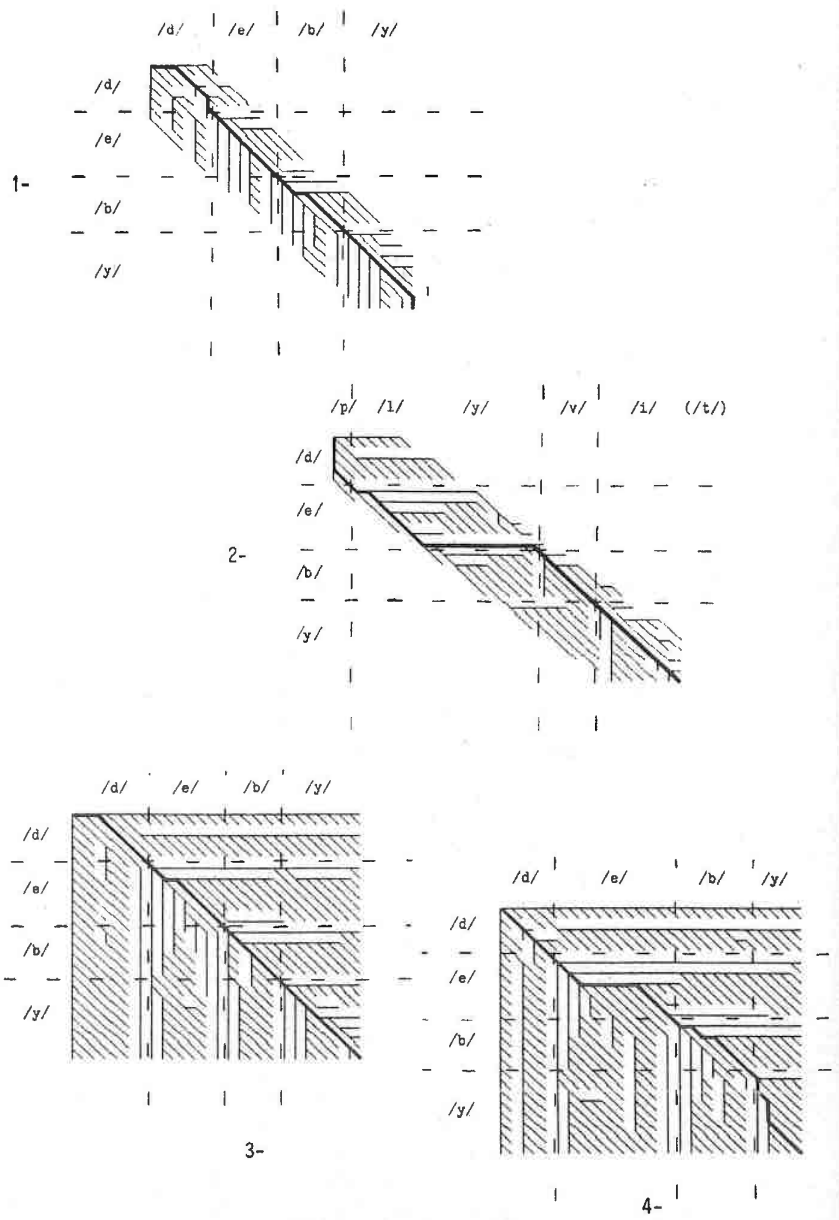


Fig. C-19a : Comparaison de type 2a

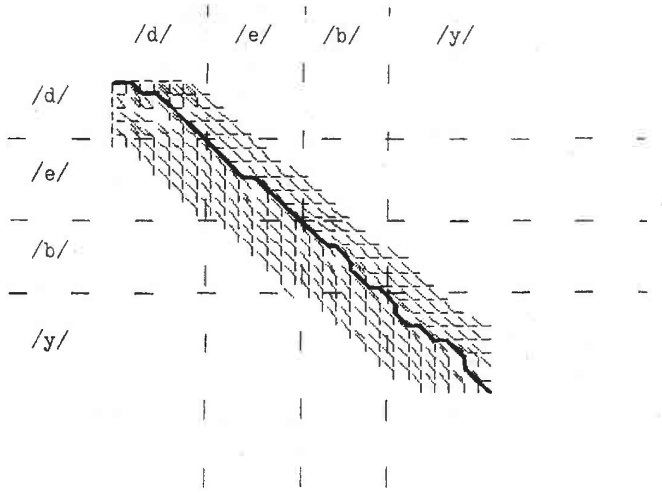


Fig. C-19b : Comparaison de type 2b (deux niveaux de profondeur géométrique).

V - PRETRAITEMENT DES FORMES AVANT COMPARAISON

Les variations de durée des formes sonores, même pour l'élocution d'un même mot par une même personne, peuvent être considérées comme un facteur de complexité des algorithmes de comparaison.

D'un autre côté, pour des raisons de place mémoire et de temps de reconnaissance, nous avons habituellement travaillé sur des formes acquises à la fréquence de 50 prélèvements par seconde. Il se trouve encore qu'avec cette cadence d'acquisition, un mot bien articulé de durée moyenne comprend plusieurs dizaines de prélèvements ; or on peut remarquer sur les zones stables des répétitions de prélèvements peu différents qui ne renseignent que sur les durées de ces zones mais qui apportent peu au niveau de la comparaison.

C'est pourquoi il vient à l'esprit d'effectuer un prétraitement des formes à comparer qui pourrait être de deux types :

- une normalisation en durée,
- une sorte de "squelettisation" de la forme.

1. Normalisation temporelle

La normalisation en durée des formes matricielles présente deux avantages fondamentaux, qui sont :

- la simplification de l'algorithme de comparaison choisi par suite d'une parfaite symétrie du rôle des indices i et j des prélèvements, notamment au niveau de la fenêtre-limite,
- par contrecoup, une amélioration du temps de comparaison d'une forme à un ensemble de formes de référence grâce à un accès facilité à ces formes et une simplification notable en particulier dans le cas d'une comparaison effectuée en parallèle sur l'ensemble des références.

Cependant, il est clair que la normalisation en durée ne peut se faire en temps réel au moment de l'élocution du mot. Si le traitement en temps différé ne se fait pas sentir sur les références au moment de l'apprentissage, il peut être pénalisant en cours d'identification même si ce traitement peut ne résider qu'en un repérage des prélèvements à supprimer ou doubler.

Nous avons en effet choisi d'effectuer la normalisation temporelle sans tenir compte de la structure fine du mot mais par simple extension ou comparaison homothétique pour atteindre la durée préfixée.

2. Squelettisation

Un autre prétraitement consiste non plus à simplifier l'algorithme de comparaison ou l'accès aux formes de référence, mais à réduire la taille des formes en jeu en ne conservant qu'un squelette dans lequel la succession des prélèvements est caractéristique des variations spectrales mais où la durée des parties stables n'est plus respectée.

Un algorithme possible revient, à partir de l'ensemble ordonné de départ des prélèvements d'une forme, à ne conserver qu'un sous-ensemble ordonné tel que le coût de mise en coïncidence d'un de ses éléments avec ses voisins immédiats soit inférieur à un certain seuil. Ce seuil doit établir un compromis entre le gain au niveau de la longueur des formes, donc également au niveau de l'espace de stockage et des temps de comparaison, et la richesse de l'information conservée. Le prétraitement peut se faire en temps réel au moment de l'acquisition mais alors il n'est plus question d'agir sur le seuil de squelettisation.

3. Conclusion

Les tests comparatifs ne nous ont pas permis de donner un avantage remarquable aux formes prétraitées suivant les algorithmes ci-dessus, étant donné les contraintes qu'elles imposent ; aussi pensons-nous que leur seul intérêt pourrait résider dans la compression d'information, à condition de conserver l'information sur les durées locales.

VI - LOGICIEL DE PRISE EN COMPTE ET DE COMPARAISON DE FORMES MATRICIELLES

Dans ce paragraphe, nous donnons une vue synthétique du logiciel que nous avons mis au point pour :

1. le test des méthodes de comparaison,
2. la constitution de fichiers de données de référence, l'apprentissage d'une forme dite "inconnue" et la reconnaissance globale en ligne,
3. l'apprentissage et la comparaison de successions d'élocutions d'un même mot pour études statistiques en différé.

Il sera fait référence implicitement aux sujets traités dans les paragraphes précédents de ce chapitre. Les paragraphes VI à VIII suivants décriront des utilisations de ce logiciel dans trois cas :

- l'élaboration de formes moyennes,
- la prise en compte d'autres types de formes matricielles que les spectrogrammes,
- le résultat des tests sur les voix de sujets I.M.C.

1. COMTES, Test des méthodes de comparaison

Ce test est effectué grâce à un logiciel simple fonctionnant sur un jeu d'essai fixe. L'introduction d'une nouvelle méthode, en plus de la procédure correspondante, exige simplement l'adjonction :

- d'une zone de données communes où apparaissent ses paramètres spécifiques,

- de modules de modification en ligne et d'édition de ces paramètres.

Dans chaque essai, il est possible d'agir sur :

. les paramètres généraux de comparaison : indicateurs de prétraitement, de rejet sur une trop grande disparité des longueurs, etc,

. les paramètres propres de la méthode testée : critère de cheminement, zone d'études, indicateur de rejet sur un accroissement trop rapide du coût partiel S , etc,

. les paramètres de la fenêtre-limite.

2. COMPAR, Apprentissage et reconnaissance de formes sonores

a) Organisation

Ce logiciel permet de constituer en ligne des vocabulaires de formes de référence pour différents locuteurs repérés eux-mêmes par leur signature vocale, l'accès à l'un de ces vocabulaires (formes dites de référence), la présentation en ligne d'une forme dite inconnue et la reconnaissance à partir de la comparaison aux références.

La figure C-20 illustre l'organisation de ce logiciel qui utilise deux fichiers de données :

- le fichier des signatures vocales des locuteurs, traité comme un fichier de un vocabulaire,
- le fichier des vocabulaires classés dans l'ordre des locuteurs.

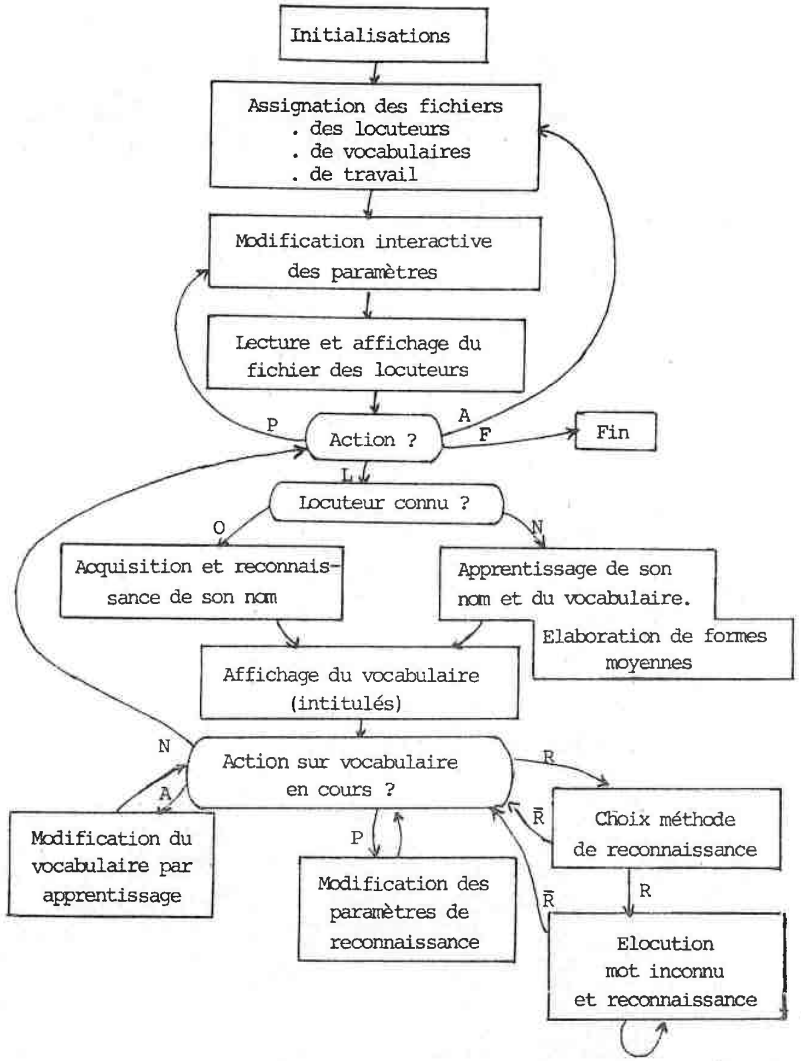


Fig. C-20 : Organisation du logiciel COMPAR

L'accès à l'un des vocabulaires de référénce se fait par identification de la signature vocale ou, à défaut, par entrée de son numéro d'ordre.

L'acquisition des mots prononcés se fait grâce à un algorithme de détection parole-non parole avec comme paramètres d'acquisition :

- un seuil global d'intensité au-dessus duquel on estime que le son capté par le microphone est à prendre en compte,
- un indicateur de voisement qui permet de déclencher une acquisition sur une consonne plosive ou fricative voisée par exemple,
- un seuil d'intensité en haute fréquence qui permet de prendre en compte les sons fricatifs,
- un seuil de durée de bruit tel qu'un son de durée inférieure est à éliminer,
- un seuil de durée de silence au-delà duquel l'acquisition s'arrête mais qui doit être ajusté de façon à ne pas s'arrêter sur une explosion de consonne plosive en particulier.

Il n'est pratiquement pas possible de garantir l'acquisition correcte en début de mot d'une consonne plosive non voisée. Cela ne présente pas d'inconvénient majeur si le phénomène est répétitif et si cette consonne n'intervient pas dans la différenciation des mots du vocabulaire, ce à quoi il faut veiller.

Afin de garantir à l'apprentissage que la forme retenue pour un mot du vocabulaire est "représentative" et éviter la sauvegarde de formes non convenables, on visualise immédiatement le résultat de l'acquisition et l'on choisit de prendre en compte ou non la forme acquise. D'autre part, la possibilité est donnée au locuteur de répéter le mot courant plusieurs fois. On retient alors :

- ou bien la forme plus proche voisine de toutes les autres (au sens d'un coût moyen de comparaison minimum) ; la forme appartient alors au corpus d'origine,
- ou bien une forme artificielle "moyenne" qui intègre les particularités de toutes les formes prononcées ; elle est alors étrangère au corpus d'origine. Cette élaboration de formes moyennes est traitée au paragraphe VII suivant.

Au moment de la reconnaissance, la forme prononcée est comparée aux formes du vocabulaire de référence. La comparaison n'est effective que si les longueurs respectives ne sont pas trop éloignées ; elle peut s'arrêter avant exploration totale si le coût devient trop élevé. Enfin, la décision de reconnaissance du mot "inconnu" ne se fait que si l'avantage du meilleur score est très net. On s'attachera, dans le cas de commande vocale de l'environnement, à ne lancer une action que sur une reconnaissance sûre de la commande.

b) Exemples de résultats

Nous citerons en exemple des tests en ligne réalisés dans les conditions ci-dessous :

- locuteur féminin unique,
- vocabulaire des 10 mots suivants :

caveau	cabot	calot
cageot	cadeau	carreau
		caillot
cachot	capot	canot

- une forme de référence par mot,
- 10 élocutions des 10 mots dans un ordre quelconque,
- méthode de comparaison (2a) avec fenêtre-limite diagonale de largeur 6 unités,
- aucun prétraitement des spectrogrammes.

La figure C-21 donne, dans une matrice de confusion, des indications du type n/m où n (resp. m) désigne le nombre de fois que le mot de référence est reconnu en 1ère (resp. 2ème) position. Les cas d'ambiguité à cause de scores très proches sont pris en compte.

La figure C-22 correspond à une autre série d'essais dans lesquels l'indication de voisement permet sans ambiguité de scinder les données en deux catégories suivant la consonne médiane.

Enfin, la figure C-23 illustre les résultats de tests effectués dans les mêmes conditions sur un vocabulaire de commande de fauteuil roulant par exemple. La commande de confirmation (Oui/Non) est toujours reconnue sans ambiguïté. On indique en plus pour chaque mot prononcé (et dans le cas où il est toujours bien reconnu) une valeur numérique RM donnant le rapport moyen du deuxième au premier meilleur score, ce dernier étant le plus faible.

prononcé →

reconnu ↓	v	ʒ	b	d	l	R	j	n	ʃ	p
v	8/2		/1			/8				
ʒ		10/			/2		/4		/10	
b	2/8		9/1	1/6						3/4
d			1/7	9/1		/1		/5		/2
l					10/		/5	/5		
R				/1		10/				/1
j		/2					10/			
n					/8		/1	10/		
ʃ		/8							10/	
p			/1	/2		/1				7/3

Fig. C-21 : Matrices de confusion pour une série de mots se distinguant par la consonne médiane (indiquée sur le schéma).

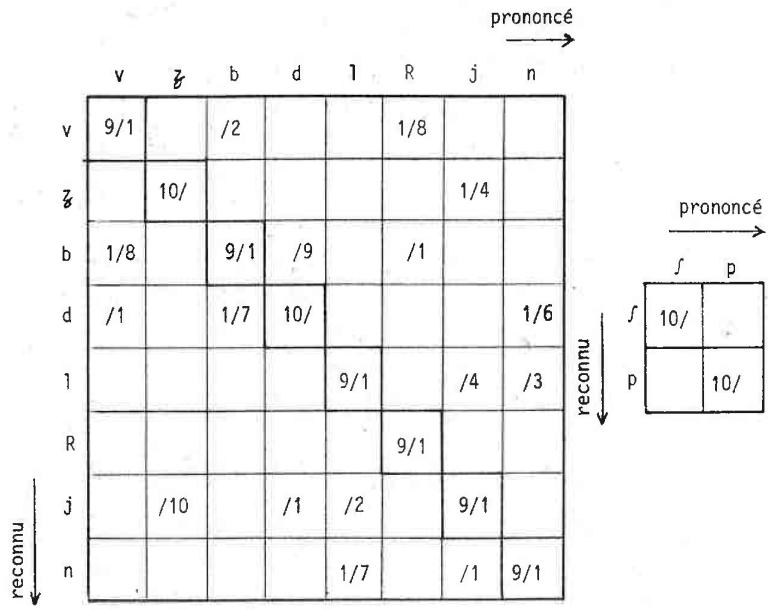
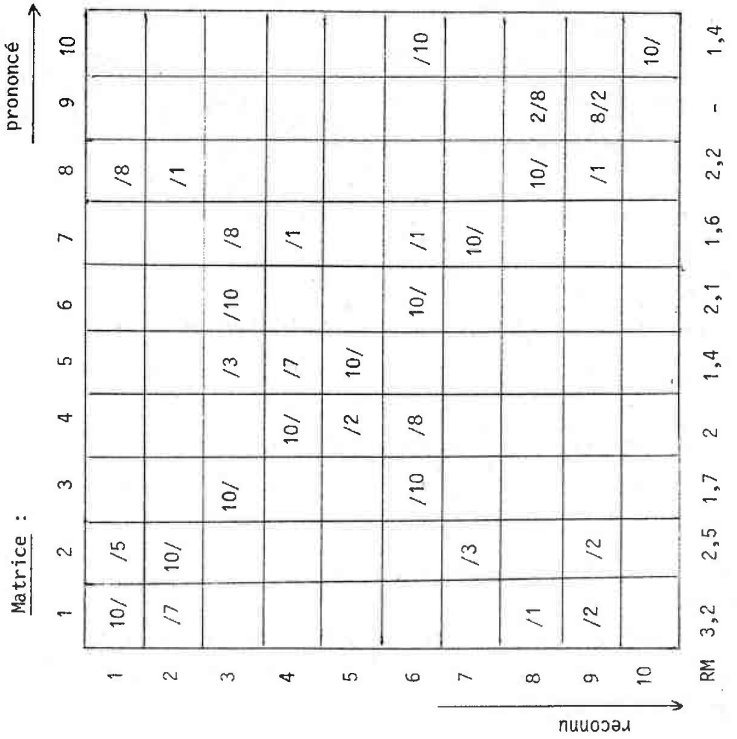


Fig. C-22 : Identification avec distinction voisé - non voisé.



Vocabulaire :

1 Début
2 Marche
3 Arrêt
4 A droite
5 A gauche
6 Avant
7 Arrière
8 Plus vite
9 Moins vite
10 Encore

Fig. C-23 : Matrice de confusion pour un vocabulaire de commande.

c) Remarque

Les matrices présentées ci-dessus donnent la vision la plus optimiste des résultats de reconnaissance ; pour les tempérer, il faut rappeler que :

- le locuteur est unique et possède en plus, grâce à l'habitude, une aptitude à prononcer les mots au microphone dans les conditions de l'expérience,

- on ne fait pas apparaître dans la figure les cas où, dans une situation réelle, il aurait fallu douter de la reconnaissance par précaution (scores trop voisins) et demander confirmation.

Il est clair que dans une phase d'apprentissage, quel que soit le vocabulaire retenu, on doit s'intéresser à des facteurs tels que le rapport moyen RM défini ci-dessus (avec éventuellement son écart-type), les risques de confusion intravocabulaire..., pour aider à la décision de reconnaissance.

3. COMSTA, études statistiques liées à la reconnaissance globale de mots

Ce logiciel permet

- d'une part, la constitution de vocabulaires de mots-tests à partir du microphone ou de bandes magnétiques, étiquetés par un intitulé en clair et classés par locuteur,

- d'autre part, le déroulement systématique de la comparaison en vue de la reconnaissance avec sélection préalable :

. du vocabulaire des mots-tests,

. du vocabulaire des mots de référence,

. de la méthode et des conditions de comparaison.

Il est ainsi possible à partir de la constitution préalable de vocabulaires de base (sous forme de spectrogrammes ou autres) d'envisager l'influence de facteurs tels que la technique utilisée ou la "complexité" du vocabulaire grâce à la comparaison des performances "intra" et "inter" locuteurs.

Des résultats complets sont donnés dans [DAME - 83] dans lesquels sont envisagées également des formes autres que les spectrogrammes.

En effet, les études sur le codage de la parole ont conduit les chercheurs à envisager d'autres représentations que les spectrogrammes [DAVIS - 82], pour certaines plus conformes à ce que l'on connaît du mécanisme d'analyse de l'oreille et pas forcément limitées au domaine fréquentiel. Dans ces représentations, on trouve en particulier, en plus des spectrogrammes (type 1) :

- les "*cepstrogrammes*" suivant l'échelle Mel de fréquence ($f_{Mel} = 1\,000 \log_2 (1 + \frac{f}{1\,000})$) et dits MFCC (type 2),

- les "*cepstrogrammes*" obtenus par transformation directe des sorties du "vocodeur" que l'on suppose comprimées suivant une loi logarithmique et dits LFCC (type 3).

Un système de référence fondé sur ces représentations a été proposé [CHOL - 82] afin d'étudier les performances des systèmes de reconnaissance globale de mots.

A partir de cette idée, nous avons étendu systématiquement ce qui vient d'être décrit dans ce chapitre aux modes de codage de types 2 et 3 avec comparaison aux performances obtenues dans la représentation classique par spectrogrammes en suivant le protocole ci-dessous :

- acquisition des vocabulaires de référence et des mots-tests à partir de l'algorithme *"d'acquisition simultanée"* décrit sur la figure C-1, avec contrôle visuel et auditif systématique, ce qui permet tout d'abord de retenir une élocution sous deux formes : son spectrogramme (type 1) et sa représentation temporelle échantillonnée et numérisée,

- traitements en différé fournissant les types 2 et 3,

- étude statistique qui traite systématiquement les trois modes de représentation des mots.

On trouve dans [DAME - 83] de nombreux résultats à ce sujet. Nous citons seulement les principales conclusions appelées par ces résultats qui mettent en particulier en lumière la supériorité très nette des cepstrogrammes en ce qui concerne :

- la réduction du nombre des ambiguïtés,

- la discrimination dans l'ensemble des voyelles / i y ɛ œ / ,

- la neutralité vis-à-vis du sexe du locuteur.

L'inconvénient majeur de ces formes est le coût en temps de leur élaboration qui limite la portée de leur utilisation. Jusqu'à présent, nous continuons à utiliser les spectrogrammes pour la reconnaissance globale.

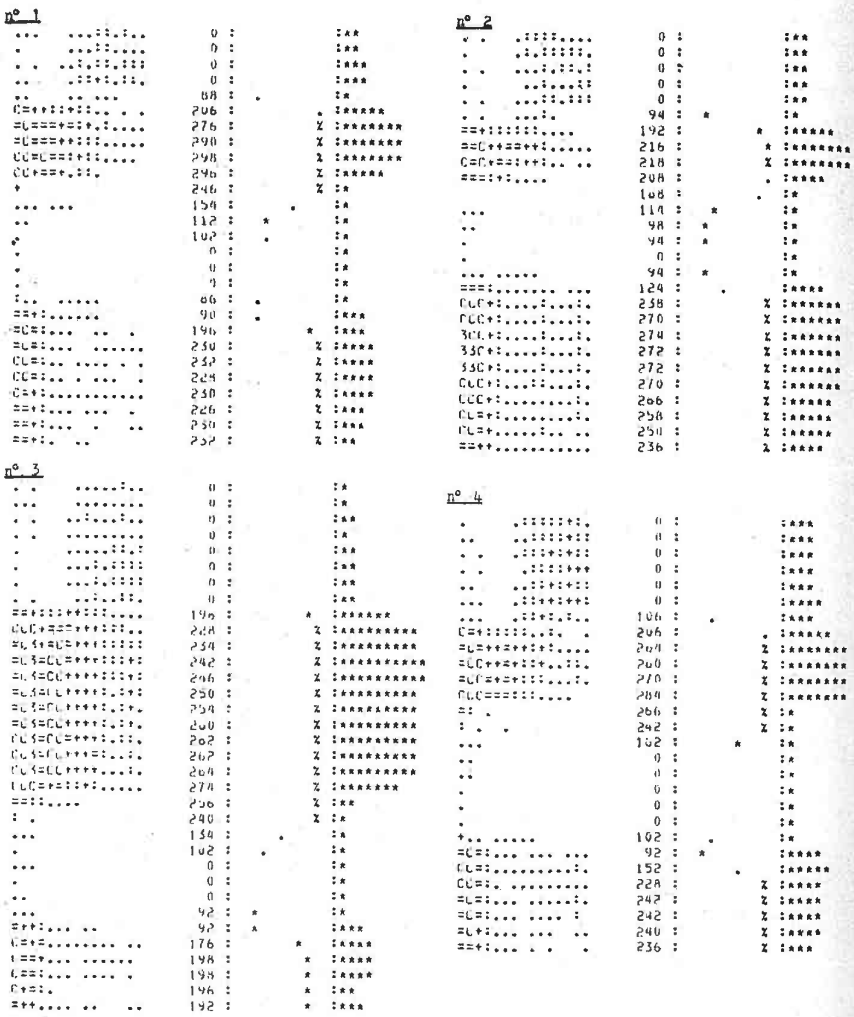
VII - ELABORATION DE FORMES-TYPES A PARTIR DE DIFFERENTES ELOCUTIONS D'UN MEME MOT

Dans le schéma C-20 donnant l'organisation du logiciel COMPAR, nous avons fait mention de l'élaboration de formes représentatives lors de la constitution de vocabulaires de référence. nous décrivons dans ce paragraphe deux façons que nous avons envisagées pour obtenir une forme "représentative" à partir de plusieurs élocutions du même mot.

1. Sélection à partir d'un critère de proximité

La proximité de deux formes est chiffrée par le score résultat de la comparaison dynamique de type 2a. Ce score ne respecte pas la définition d'une distance mais traduit correctement l'idée de proximité. D'une série d'élocutions différentes, on retient la forme dont la proximité à l'ensemble de toutes les autres est la plus grande. La forme retenue fait donc partie du corpus de départ.

La figure C-24 illustre cette sélection dans le cas de quatre élocutions du mot /ʃ a p o/ avec l'indication des scores croisés de proximité entre ces formes.



Forme	n° 1	n° 2	n° 3	n° 4
Longueur	27	27	34	28
Score avec				
n° 1	0	31	41	24
n° 2	31	0	43	28
n° 3	41	43	0	37
n° 4	24	28	37	0

Fig. C-24 : Sélection d'une forme-type par proximité : la forme n° 4

2. Elaboration d'une forme moyenne

La comparaison dynamique permettant au sens d'un critère donné de mettre deux formes en coïncidence de façon optimale, nous avons proposé de tenir compte de la suite des coïncidences optimales fournies par la méthode 2a pour fabriquer une forme dérivée intégrant des caractéristiques de l'une et l'autre formes de départ. La forme finale n'appartient pas au corpus d'origine.

L'exemple suivant (figure C-25) relatif à des mots de durées respectives 4 et 7 illustre l'algorithme retenu : à chaque ensemble de points entourés est associé un prélèvement artificiellement fabriqué.

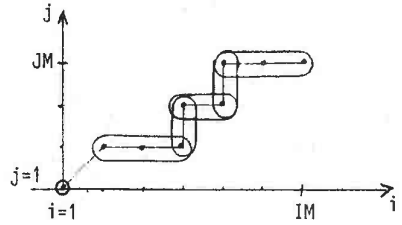


Fig. C-25 : Illustration de l'algorithme d'élaboration de forme moyenne.

Dans cet exemple, le forme finale comporte 6 prélèvements. On rencontre deux difficultés principales dans l'application de l'algorithme choisi :

• tout d'abord, la durée KM de la forme obtenue peut différer notablement des durées IM et JM de départ. Le plus souvent

dans les cas concrets, pour des durées de départ voisines, on a :

$$KM < \min (IM, JM).$$

Nous avons vérifié par le calcul cette tendance au raccourcissement en faisant l'hypothèse (approximative) de l'équiprobabilité de tous les chemins permettant de faire coïncider origines et extrêmités des deux formes et respectant la contrainte géométrique de la méthode de comparaison retenue (les indices progressent toujours dans le même sens et de 0 ou 1 unité seulement sur chacun des arcs).

La figure C-26 donne la longueur résultante moyenne (trait plein) et la longueur résultante la plus probable (croix) en fonction de la durée IM de l'une des deux formes de départ variant de 1 à 15 dans deux cas :

a) IM = JM

Les courbes sont toujours situées sous la ligne bissectrice avec une pente moyenne de l'ordre de 90 % ,

b) JM est fixé à 15

Les longueurs résultantes sont toujours beaucoup plus proches de la plus faible des longueurs avec même, à partir de IM = 12 , une tendance à être inférieures à la plus petite des deux longueurs de départ.

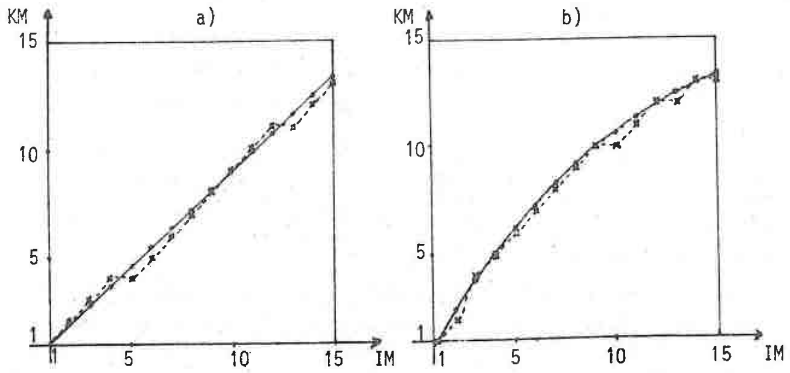


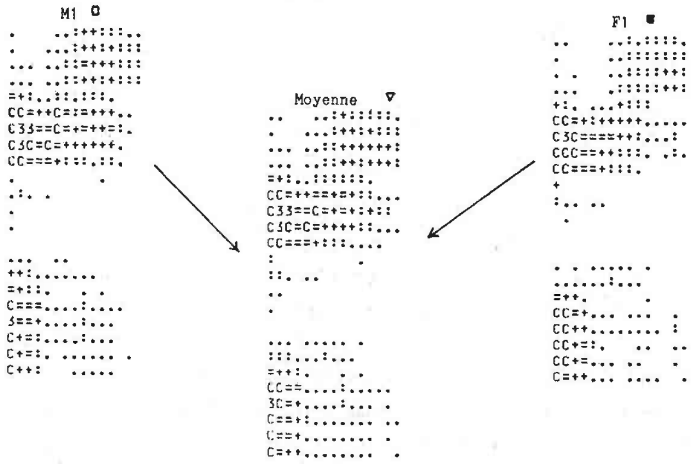
Fig. C-26 : Longueurs résultante moyenne (•—•—•)
et la plus probable (×···×···×)

Du fait de cette tendance, nous avons été amené à modifier l'algorithme illustré en C-25 essentiellement sur les plages horizontales et verticales du chemin de comparaison,

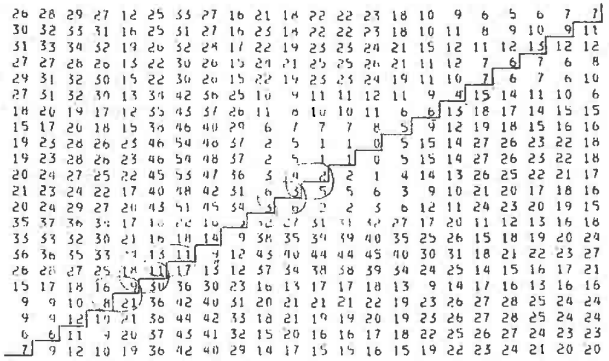
• ensuite, le type entier des données constituant les spectrogrammes pour des raisons de quantification numérique, d'une part, et d'occupation d'espace mémoire et de calcul, d'autre part, nous oblige à arrondir les variables caractérisant la forme moyenne. Suivant le parti choisi, on risque d'introduire artificiellement du bruit noyant la forme ou du "blanc" équivalant à un réglage trop faible du niveau d'entrée de l'analyseur. Ce défaut ne peut se corriger qu'en agissant sur les conditions de travail mentionnées dans ce sous-paragraphe.

Nous montrons sur la figure C-27 la forme moyenne obtenue à partir de deux élocutions du même mot //apo/ par deux locuteurs différents (M1 masculin et F1 féminin). Il est clair que l'algorithme peut être appliqué à un nombre d'élocutions plus important, la forme finale se construisant par étapes successives.

a-



b-



c-

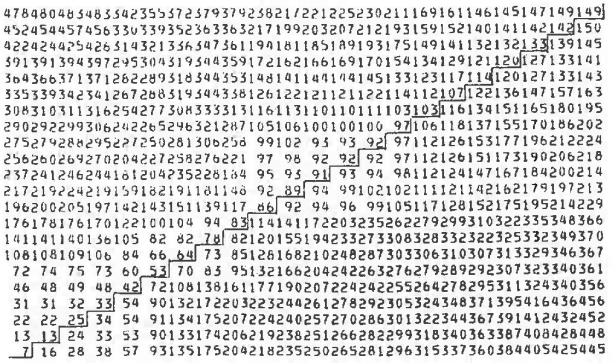


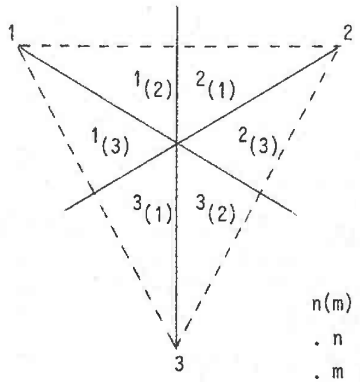
Fig. C-27 : Elaboration de forme moyenne

3. Application

Dans le cas n° 1, il est évident que le critère de plus grande proximité contribue pour la reconnaissance :

- à améliorer le pourcentage global de mots correctement reconnus,
- à favoriser le locuteur dont l'élocution est retenue.

Il est plus intéressant de voir l'effet de l'algorithme du cas n° 2 sur la reconnaissance. Nous en donnons successivement trois exemples dans lesquels trois formes de référence ont été retenues : deux formes de base et la forme moyenne. Ces trois références sont figurées aux sommets d'un triangle. La position des formes "inconnues" comparées aux références est figurée par un point distant des longueurs d_1 , d_2 et d_3 aux trois sommets, proportionnelles aux scores de comparaison S_1 , S_2 et S_3 aux formes correspondantes. Le plan du triangle peut être ainsi divisé en six zones (secteurs angulaires centrés sur le barycentre du triangle) qui permettent de repérer les premiers et seconds meilleurs scores comme suit (figure C-28) :



n(m) indique :
. n 1er meilleur score
. m 2ème meilleur score

Fig. C-28 : Schéma de proximité d'une forme inconnue à trois formes de référence

a) (figure C-29)

Forme 1 : / ʃ a p o / locuteur féminin F1
Forme 2 : / ʃ a p o / même locuteur féminin F1
Forme 3 : moyenne des deux précédentes
Formes présentées : / ʃ a p o / même locuteur F1

Les formes présentées sont toujours plus proches des formes 1 et 2 que de la moyenne. Les rapports $\frac{S_3}{S_1}$ ou $\frac{S_2}{S_1}$ varient entre 0.5 et 0.85 .

b) (figure C-30)

Forme 1 : / ʃ a p o / locuteur féminin F1
Forme 2 : / ʃ a p o / locuteur masculin M1
Forme 3 : moyenne des deux précédentes
Formes présentées : / ʃ a p o / locuteur F1 (●)
/ ʃ a p o / locuteur M1 (o)

Dans nos essais, la forme moyenne n'est jamais reconnue en première ligne mais elle apparaît en deuxième ligne dans la proportion de neuf fois sur dix pour M1 et deux fois sur dix pour F1 . Cette tendance à la prédominance du locuteur masculin vient certainement du choix de l'arrondi à l'entier supérieur dans la construction de la forme moyenne.

c) (figure C-31)

Forme 1 : / o / locuteur F1
Forme 2 : / ɔ / même locuteur
Forme 3 : moyenne des deux précédentes
Formes présentées : - figure C-31a : / o / locuteur F1
- figure C-31b : / ɔ / locuteur F1

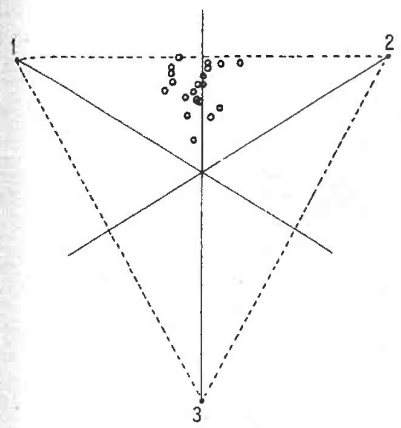


Fig. C-29

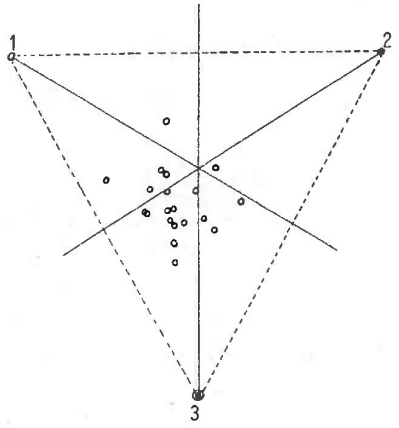


Fig. C-31a

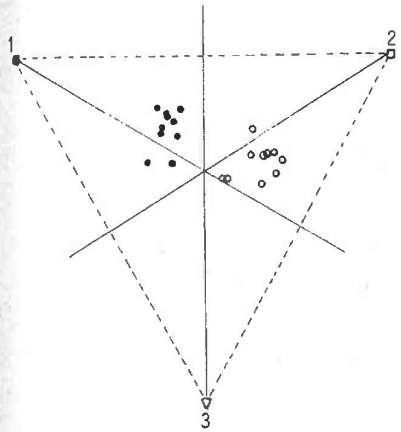


Fig. C-30

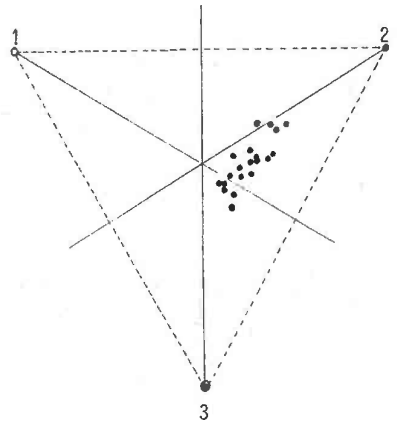


Fig. C-31b

Proximités de formes prononcées à trois formes de référence.

4. Utilisation en segmentation assistée

La mise en coïncidence, prélèvement par prélèvement, permise par la méthode 2a retenue ici peut trouver une application intéressante en segmentation automatique à partir d'une forme de référence accompagnée de repères prédéterminés, ainsi que l'indique la figure C-32.

A cause du schéma retenu (double incrémentation pour les schémas $\uparrow^{\rightarrow}$ ou $\rightarrow^{\uparrow}$ par rapport au schéma $\nearrow$), la coupure se fera dans le mot inconnu à une unité d'indice près.

Ce principe peut être exploité dans un certain nombre d'applications, par exemple :

- normalisation temporelle d'une forme à une durée fixe par calage imposé sur une forme de référence "préfabriquée",
- mises en coïncidence locales de phonèmes et détection d'anomalie d'élocution. Ce point est illustré en partie D (figures D-18 et D-19),
- étude de la déviation du rythme et des durées par rapport à la normale.

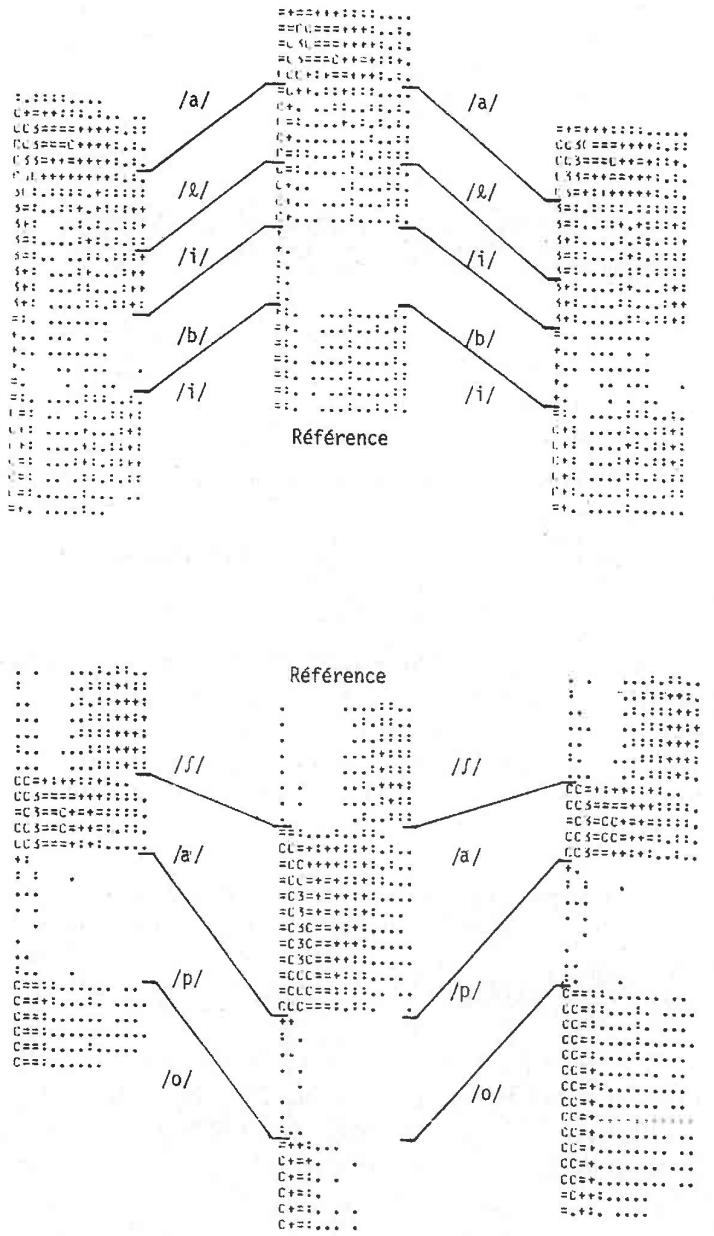


Fig. C-32 : Segmentation automatique de deux formes à partir d'une forme de référence.

VIII - UN EXEMPLE D'APPLICATION : L'ETUDE DES POSSIBILITES DE COMMANDE VOCALE DE SUJETS I.M.C.

Afin de déterminer l'aptitude de sujets infirmes (en particulier Infirmes Moteurs Cérébraux) à la commande orale de dispositifs d'assistance, nous avons appliqué les techniques décrites dans ce chapitre pour successivement :

- sélectionner les commandes les "meilleures" parmi un ensemble d'élocutions,
- étudier le pouvoir de discrimination du système de reconnaissance.

Nous avons fait une étude à partir d'enregistrements analogiques sur cassettes de qualité moyenne de séries de mots prononcés par des enfants I.M.C., effectués d'une part à Berne - Suisse (Institut Orthopédique Expérimental) par J.C. GABUS [GABU - 82], d'autre part à Flavigny - Meurthe et Moselle (Centre de Rééducation Fonctionnelle dirigé par le Professeur J.P. ANDRE).

Une première écoute des bandes magnétiques nous a conduit à éliminer de l'étude la majorité des locuteurs pour qui l'absence de répétitivité des élocutions ou l'impossibilité d'en faire l'acquisition globale convenable étaient évidentes.

Pour les sujets retenus, nous avons distingué trois listes de mots que nous avions définies avant les enregistrements : des commandes chargées de sens (comme "*sortir*", "*mal*", "*toilette*"), l'opposition "*oui/non*" et des syllabes non signifiantes telles que "*sou*", "*ma*" ou "*di*".

Les tableaux numérotés C-34 concernent Véronique (10 ans) dont l'intelligibilité, parmi les enfants retenus, était la meilleure et qui avait pu répéter dix fois chacun des mots. L'élocution reconnue comme la plus proche de toutes les autres (au sens du paragraphe IV.1. précédent) a été choisie comme référence, les autres représentants ont constitué l'ensemble des mots-tests.

Les commandes "*oui*" et "*non*" ont été traitées séparément dans l'idée de les utiliser pour valider la décision de reconnaissance automatique dans une situation réelle de commande vocale. Le schéma C-33 indique les proximités des formes présentées aux formes-témoins "*oui*" et "*non*". Les points représentatifs sont placés de façon qu'en abscisse le rapport des distances aux pôles soit égal au rapport des scores de comparaison (valeur de RM déjà rencontrée). L'ordonnée est proportionnelle au score de comparaison avec la forme attendue. On peut remarquer que deux paramètres de rejet sont ajustables :

- la barre-limite en absolu,
- et la largeur de la zone d'ambiguïté

qui, dans l'exemple donné, permettraient d'éliminer les "*oui*" mal reconnus. Cette sévérité risque bien sûr d'éliminer des commandes "bien reconnues" mais accroît la fiabilité de la commande vocale. Sur tout cas particulier de tels paramètres doivent être étudiés soigneusement.

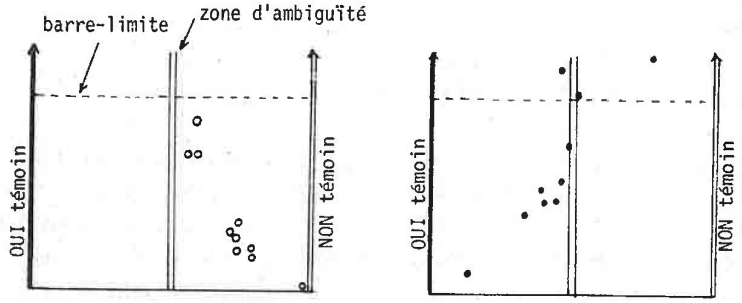


Fig. C-33 : Proximité des formes présentées
(o = NON, ● = OUI) aux
formes-témoins de Véronique.

MOT PRONONCÉ — TOTAL — NBRES AMBIGUITÉ — NUMÉRO DU MOT RECONNU

		T	A	1	2	3	4	5	6	7	8	9
1 PAPA	:	10	0	0	0	5	0	0	0	0	4	1
2 MAMAN	:	10	0	0	6	2	0	0	0	1	1	0
3 MANGER	:	10	0	0	0	2	0	0	0	8	0	0
4 TOILETTE	:	10	0	0	0	0	6	3	0	1	0	0
5 MAL	:	10	0	0	0	0	0	5	0	4	1	0
6 OUI	:	10	0	1	0	2	0	0	6	1	0	0
7 NON	:	10	0	0	0	0	0	1	0	9	0	0
8 DORMIR	:	10	0	1	0	3	0	0	0	2	4	0
9 SORTIR	:	10	0	5	0	1	0	0	0	2	0	2

	% BIEN RECONNUS	% AMBIGUITÉ	% MAL RECONNUS
1 PAPA	0.00	0.00	100.00
2 MAMAN	60.00	0.00	40.00
3 MANGER	20.00	0.00	80.00
4 TOILETTE	60.00	0.00	40.00
5 MAL	50.00	0.00	50.00
6 OUI	60.00	0.00	40.00
7 NON	90.00	0.00	10.00
8 DORMIR	40.00	0.00	60.00
9 SORTIR	20.00	0.00	80.00
*	44.44 *	0.00 *	55.56 *

		T	A	1	2
1 OUI	:	10	0	8	2
2 NON	:	10	0	0	10

Fig. C-34 : Commandes vocales de Véronique :
performances.

Pour les autres résultats, les faibles taux de réussite s'expliquent par différents facteurs dont le principal est que nous avons respecté, à partir des cassettes magnétiques, les conditions d'acquisition automatique (critères de segmentation parole - non parole, seuils de silence, ...) pour nous placer dans les conditions réelles où aucune intervention extérieure n'indique la fin du message. La parole hachée, les pauses anormalement longues, la perte de souffle, la désonorisation, etc. sont autant de facteurs défavorables à l'utilisation fiable des entrées vocales.

CHAPITRE 3

ANALYSE DES VOIX

I - INTRODUCTION

Dans ce chapitre, nous envisageons la question de la paramétrisation de la parole dans une optique autre que celle de la reconnaissance ou de la compréhension automatique. En dehors de ces aspects, il nous semble que l'analyse des voix, et particulièrement des voix pathologiques, revêt un intérêt particulier dans les applications suivantes :

- perfectionnement de la connaissance que l'on a des mécanismes de production de la parole, définition de ce que l'on peut nommer "*parole normale*", classification des locuteurs dits normaux, tolérance vis-à-vis de la norme, caractéristiques du discours parent - enfant comme étudié dans [STEV - 80], etc.,

- pathologie et aide au diagnostic, détermination de l'importance de la déviation par rapport à la parole normale, classification des voix pathologiques, recherche de corrélation entre pathologie et manifestations acoustiques de la parole,

- détermination de l'orientation à donner à la rééducation vocale, le cas échéant : recherche des défauts majeurs qui affectent la qualité et l'intelligibilité de la parole,

- évaluation en vue de l'appréciation des progrès et des performances vocales,

- recherche de formes vocales répétitives pour servir de commandes orales de l'environnement par le sujet.

A la suite de la présentation, aux chapitres précédents, des méthodes et des outils mis en place pour l'analyse automatique de la parole, nous proposons dans ce chapitre un plan pour l'étude des voix et la constitution

de fichiers de paramètres objectifs. Nous terminons sur les perspectives offertes par les travaux en modélisation des appareils de production et de perception et en synthèse de la parole et sur l'extension de la paramétrisation du signal aux problèmes de rééducation vocale et d'aides au diagnostic.

II - PARAMETRISATION DU SIGNAL DE PAROLE

Aux chapitres 1 et 2 de cette partie, nous avons décrit les logiciels mis en place pour faciliter les opérations d'acquisition et de dépouillement des données vocales. Ceci nous a conduit à faire le choix d'une paramétrisation adaptée à l'objectif de l'étude et en relation directe avec les possibilités de rééducation vocale lorsque celle-ci est envisagée (cf. partie D).

1. Etude subjective

Préalablement à toute analyse automatique, nous jugeons indispensable de faire une analyse subjective de la voix à partir, si possible, de l'écoute directe, ce qui permet de porter également un jugement sur le comportement du sujet (signes d'agitation, hochements de tête, mouvements articulatoires), et à partir de l'écoute d'enregistrements qui permettent de faire appel au jugement de différents auditeurs, spécialistes ou non. Dans la mesure du possible, les enregistrements doivent comporter des phrases prononcées spontanément et des mots ou phrases imposés.

L'appréciation demandée aux auditeurs doit concerner :

- la qualité de la voix : attribution d'un qualificatif donnant l'impression générale, choix de qualificatifs dans une liste donnée,

appréciation chiffrée sur une échelle quantitative de facteurs divers comme la richesse intonative, le timbre, la rapidité d'élocution, la raucité, etc.,

- l'intelligibilité : taux de compréhension de mots ou d'éléments de phrase par l'auditeur entraîné ou non (la distinction entre ces deux types de jugement étant fondamentale),

- le degré d'évolution (ou de dégradation) du système phonétique grâce à un bilan complet qui peut aller, dans certains cas pathologiques, jusqu'à la question de savoir si les sons produits s'organisent dans un système cohérent [CHRI - 77].

Cette première étape permet la recherche de corrélations entre caractéristiques physiques, physiologiques et acoustiques subjectives correspondant au "coup d'oreille" (plutôt qu'au coup d'oeil) du spécialiste. Ces corrélations apparaissent assez bien dans le cas d'enfants sourds dont les voix présentent un certain nombre de caractéristiques voisines, du moins lorsqu'on distingue surdité congénitale ou de la prime enfance et surdité acquise.

Pour notre part, nous avons travaillé pour différentes raisons sur des enregistrements provenant :

- des enfants malentendants ou étrangers ayant participé à nos sessions de rééducation vocale [DUTE - 79] (cf. chapitre D.2.),

- des enfants infirmes moteurs cérébraux pour lesquels nous avons étudié les possibilités de commande vocale (cf. chapitre C.2.).

Nous nous contenterons ici de résumer les défauts observés à l'écoute des voix d'enfants sourds. Ils ne sont pas représentatifs de l'ensemble des voix pathologiques (neuro ou psychopathologies, trachéotomies, etc.) mais, étant donnée leur cause, ils englobent en particulier nombre de défauts d'origine motrice.

Les défauts observés à l'oreille peuvent être classés grossièrement en deux catégories :

1. les défauts qui affectent l'intelligibilité de la parole dans lesquels on peut inclure la mauvaise articulation des phonèmes, les défauts de liaisons entre les différents sons et le défaut de commande de la respiration qui réagit sur le rythme de la phrase [OSBE - 79] et les durées locales et globales,

2. les défauts qui affectent la qualité : mauvaise distinction voisé - non voisé, défauts de la fréquence fondamentale et du niveau de l'intensité.

Cette classification doit être modulée car chacun des facteurs cités intervient dans la qualité et l'intelligibilité de l'élocution. Il appartient en particulier aux tests de perception de parole synthétique de déterminer les influences relatives de ces facteurs, influences en réalité très complexes.

On retrouve souvent, à l'écoute des voix des enfants sourds, les caractéristiques suivantes :

- la voix a une couleur particulière, elle manque d'expression, la hauteur n'est pas esthétiquement correcte. L'étude des cris et pleurs de petits enfants [JONE - 67] montre que la fréquence fondamentale avant l'âge de six mois est indépendante du degré de surdité, mais qu'ensuite elle est nettement plus élevée chez les enfants déficients auditifs. Souvent, les variations du fondamental suivent les variations d'intensité et inversement, ce qui est sensible dans les exercices de rééducation vocale,

- l'intensité varie par à-coups suivant les efforts articulatoires,

- les erreurs de rythme sont fréquentes : la phrase est attaquée rapidement. Les pauses deviennent ensuite plus nombreuses et plus longues ou bien la parole est prolongée jusqu'à épuisement du souffle. Durant les sons voisés, les cordes vocales laissent passer un flot d'air trop important qui réclame jusqu'à deux fois plus d'inspirations que la normale,

- la durée des sons individuels est incorrecte : les voyelles et sons continus sont étirés, jusqu'à un rapport de deux à quatre pour les sons fricatifs. Il s'ensuit que les phrases elles-mêmes sont nettement plus longues que la normale,

- l'attaque des sons peut être aspirée ou en coup de glotte à cause d'une pression sous-glottique exagérée,

- les troubles de la nasalité, par excès ou par défaut, se remarquent ouvent. Localement, on peut rencontrer une tendance à la nasalisation des voyelles orales ou des consonnes plosives. Mais la tendance à la nasalité peut affecter plus généralement les productions vocales et intervient alors comme un trait suprasegmental,

- le positionnement imparfait des articulateurs est responsable de substitutions entre phonèmes. Les voyelles tendent vers la voyelle neutre, celle que l'on émet en cherchant ses mots dans une conversation. Pour les consonnes, on observe généralement des erreurs de mode d'articulation plutôt que de lieu d'articulation (désonorisation des consonnes, nasalisation comme indiqué plus haut, etc.). Certains sons sont souvent omis, comme les phonèmes médians ou d'arrière (l'exemple du /R/ est caractéristique)...

Les erreurs et leur fréquence d'apparition peuvent varier considérablement d'un sujet à un autre. C'est pourquoi tout programme de rééducation vocale doit être adopté à chaque cas particulier, qu'il existe un support matériel techniquement évolué ou pas. Il est ainsi particulièrement

intéressant de pouvoir mettre en oeuvre des techniques d'analyse des voix au niveau du bilan orthophonique puis aux différentes étapes de la rééducation.

Le bilan sommaire suivant a été établi pour Edith (17 ans) :

. scolarisée au C.E.S., en 6ème,
. surdité de perception sévère - appareillée,
. F_0 moyen = 250 Hz - modulation faible,
. phonation commençant souvent à bout de souffle,
. intensité faible,
. timbre sourd, étouffé, nasonné,
. articulation : tendance à la désonorisation,
élision des consonnes finales,
sigmatisme occlusif,
. rythme très altéré.

2. Fréquence fondamentale (Fo) et mélodie

En l'absence de problème organique, toutes les difficultés rencontrées au niveau de la production du son glottal peuvent être attribuées à un défaut de commande des muscles du larynx et/ou à un défaut de coordination du système respiratoire. Les points suivants peuvent être étudiés :

- l'importance relative dans la voix normale de l'information linguistique et de l'information non-linguistique (relative à l'idiosynchrasie du locuteur) [FUJI - 81],

- la répartition des valeurs de Fo : histogrammes du 1er ordre et d'ordres supérieurs (pour l'étude de la régularité des vibrations) et leurs approximations triangulaires [ASKE - 78], calcul de la moyenne et de l'écart-type sur des sons tenus.

- distribution des valeurs de Fo en début et fin de phrase, valeur moyenne et valeur maximale [CONT - 73], calcul d'un indice d'expression,

- taux de voisement dans la phrase,

- appréciation des variations locales aléatoires par rapport à la ligne mélodique moyenne : statistique de VON NEUMANN [CHEV - 77], analyse polynomiale du contour mélodique [HATO - 77],

- corrélations entre :
. qualificatifs attribués à la voix et variations de Fo ,
. valeur moyenne de Fo et âge du sujet,
. valeur moyenne et appréciation subjective de la hauteur,
. variabilité locale et degré de perte auditive,
. écart-type et appréciation subjective de la richesse intonative.

Les figures C-36 et C-37 donnent un exemple de résultats obtenus dans l'analyse de la phrase *"tu viens jouer"* prononcée par deux enfants avant et après rééducation de la mélodie dans un travail d'une année [SALL - 80] au C.T.O.P. de Fougères. L'analyse est effectuée suivant le schéma ci-dessous (figure C-35) [HATO - 77] :

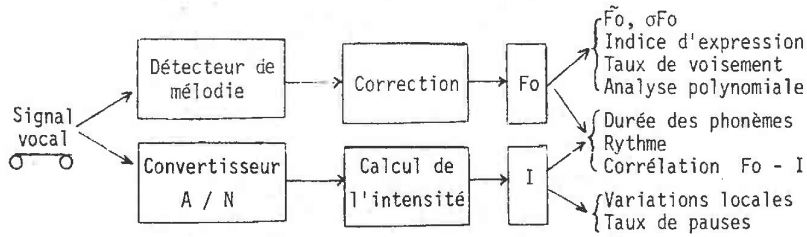


Fig. C-35 : Procédure d'analyse des paramètres prosodiques.

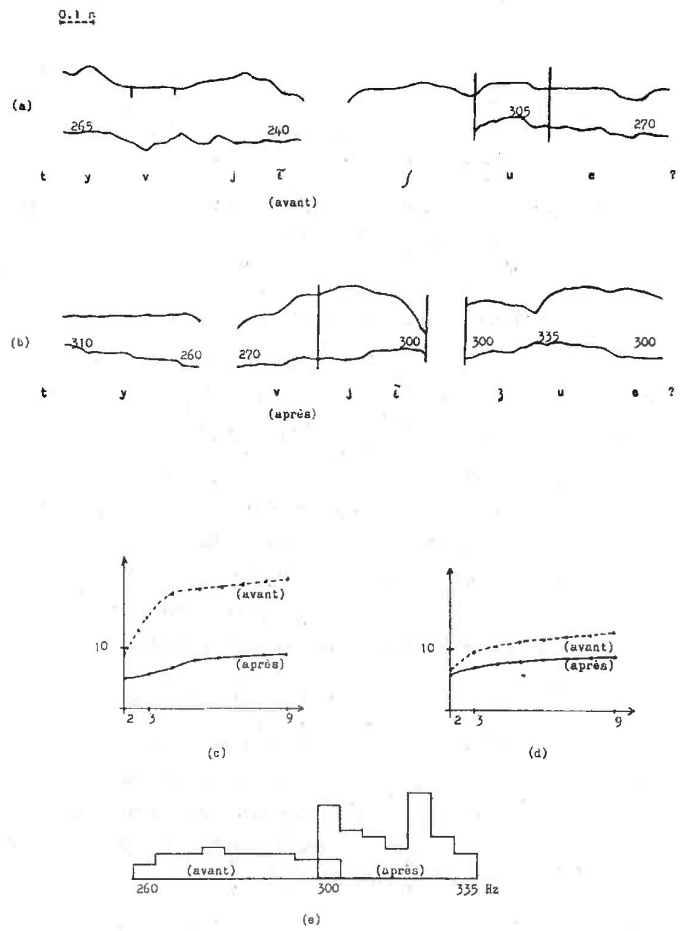


Fig. C-36 : Analyse de la phrase :
"tu viens jouer ?"
(Philippe).

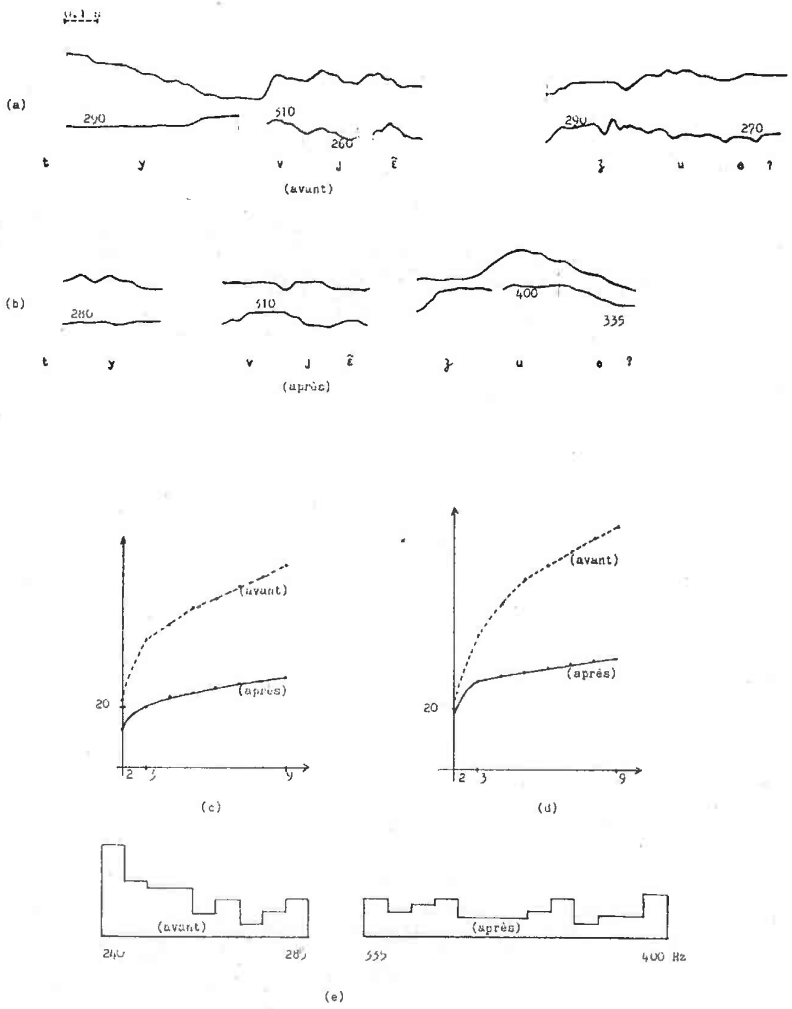


Fig. C-37 : Analyse de la phrase :
"tu viens jouer ?"
(Valérie).

Les deux enfants, au début de la rééducation, ont les caractéristiques suivantes :

	Philippe	Valérie
Degré de déficience auditive	*profond 2ème degré*	*profond 3ème degré*
Surdité	*héréditaire*	*congénitale*
Age	*6 ans 10 mois*	*6 ans 1 mois*
Nombre d'années de rééducation au centre	*3 ans*	*3 ans*

Le travail s'est déroulé en trois phases : tenue du fondamental, variations volontaires, essai d'imitation de contours intonatifs imposés dans l'élocution de phrases.

Nous donnons successivement pour chacun des enfants :

- le contour d'intensité et le contour mélodique avant (a) et après (b) rééducation,

- les valeurs cumulées $\sum_{i=1}^{m} c_i$ pour m variant de 1 à 9, où c_i désigne la valeur moyenne du coefficient d'ordre i de l'approximation polynomiale du contour mélodique, pour l'ensemble de la phrase (c) et pour la partie interrogative (mot "*jouer*") seule (d), avant et après rééducation,

- les histogrammes du premier ordre de la partie interrogative avant et après rééducation (e).

Le tableau C-38 ci-dessous rassemble quelques résultats numériques. La dernière ligne concerne un paramètre dont il est fait mention au paragraphe 3 suivant.

	E. (voix de référence)	Philippe		Valérie	
		Avant	Après	Avant	Après
Phrase entière					
. Durée (s)	*1.1*	*2.2*	*2.2*	*2.8*	*2.3*
. Taux de pauses (%)	*0*	*8*	*12*	*17*	*18*
. Taux de voisement (%)	*100*	*78*	*100*	*93*	*97*
. Ordre moyen retenu pour l'approximation	*3*	*5*	*4*	*9*	*8*
Partie interrogative					
. Durée (s)	*0.6*	*1.2*	*0.8*	*0.9*	*0.8*
. Taux de voisement (%)	*100*	*60*	*100*	*100*	*96*
. Ordre moyen retenu pour l'approximation	*3*	*5*	*3*	*9*	*8*
. Indice d'expression (demi tons)	*4*	*2.5*	*2*	*3*	*3*
. Déviation Fo type / Fo (Hz)	*0*	*30*	*37*	*32*	*58*
. Corrélation des pentes Fo type / Fo	*1*	*0.51*	*0.65*	*0.48*	*0.52*
. Corrélation des pentes Fo / I	*0.50*	*0.82*	*0.50*	*0.54*	*0.54*

Fig. C-38 : Données numériques sur la phrase *"Tu viens jouer ?"*.

On peut, dans l'analyse des résultats de Philippe par exemple, faire les remarques suivantes : après rééducation du fondamental dans une phrase interrogative à intonation montante, on constate une amélioration sur le plan de la durée et du taux de voisement et une plus grande stabilité du fondamental. Les efforts pour élever la hauteur de la voix se font au détriment de la plage de variation du fondamental. La montée de l'intonation est correcte mais chute trop rapidement, l'enfant maintenant alors son effort au niveau de l'intensité totale. Cette dernière tendance assez fréquente apparaît dans les résultats qui montrent que la corrélation Fo-I a une valeur raisonnable mais que la courbe mélodique n'a pas, jusqu'à la fin du dernier mot, l'allure attendue (corrélation Fo type / Fo).

3. Intensité

L'étude de l'intensité du signal permet de chiffrer les facteurs suivants :

- la durée totale du message,
- le taux de dépassement des seuils extrêmes associés aux références trop faible - trop fort,
- les chutes brutales dues à des défauts d'expiration (pertes d'air exagérées),
- le taux de pauses associées aux problèmes de respiration et aux efforts articulatoires.

Une segmentation en noyaux syllabiques, grâce à l'observation simultanée des contours mélodique et d'intensité, et la comparaison à des courbes-modèles permettent de mettre en évidence :

- les défauts de rythme et de durée,
- les syllabes en trop ou manquantes, s'il y a lieu,
- la place des coupures entre mots et syllabes.

Déterminé à partir des mêmes données, le temps d'établissement du voisement (par rapport au début de l'élocution d'un phonème sonore) est intéressant pour renseigner sur le degré de maturité vocale [KENT - 76]. C'est en effet un indice de coordination des activités des systèmes phonatoire et articulatoire. Il sera particulièrement aisé de l'étudier dans l'élocution de syllabes composées d'une consonne plosive et d'une voyelle.

Enfin, pour mesurer le degré d'indépendance des efforts de tension des cordes vocales et des efforts d'expiration, on peut rechercher la corrélation entre contours de fondamental et d'intensité. Il est ainsi possible :

- de représenter l'un des facteurs en fonction de l'autre et de rechercher la tendance de la régression [GRAI - 77],
- après lissage des contours d'étudier leurs pentes respectives dans des exercices de tenue ou de montée mélodique (cf. tableau C-38),
- ou de calculer un coefficient de corrélation sur une fenêtre, par exemple un "*facteur cosinus*" à partir de la formule

$$\frac{\Sigma x_i \cdot y_i}{\sqrt{\Sigma x_i^2} \cdot \sqrt{\Sigma y_i^2}}$$

où x_i et y_i représentent les $i^{èmes}$ occurences des deux paramètres mis en correspondance. Dans un exercice de montée mélodique, la valeur de ce facteur est de 0.88 pour Edith (surdité sévère) pour une valeur normale de 0.69 correspondant à l'exercice, ce qui indique un effort d'expiration exagéré au lieu d'un effort de tension des cordes vocales.

4. Nasalité

Ce trait se détermine avec difficulté si l'on ne met pas en place des dispositifs spéciaux à l'enregistrement, d'autant plus qu'un défaut de nasalisation est dû à un couplage incorrect du conduit nasal au conduit oral, ce qui ne permet pas une modélisation fiable du système de transmission. Il est plus simple ici de s'en remettre à l'appréciation subjective d'un auditeur entraîné.

5. Paramètres fréquentiels

Un certain nombre de facteurs sont intéressants pour caractériser le timbre et l'articulation. On peut retenir en particulier pour des sons tenus :

- les fréquences et largeurs de bande des premiers formants. La figure C-39 montre les valeurs obtenues par traitement de prédiction linéaire pour les voyelles extrêmes /a/, /i/ et /u/ articulées par Véronique (10 ans et demi, enfant I.M.C.) et Hélène (9 ans, ouïe normale, voix tendant vers l'aigu) dans le plan $F_1 - F_2$. Les segments visualisent la largeur de bande (sur la même échelle). On remarque dans la parole pathologique un resserrement accentué du triangle des voyelles et un amortissement supérieur à la normale,

- le barycentre de l'énergie dans les zones 100 - 800 Hz, 700 - 3000 Hz, et au delà, par exemple,

- un indice de la *"couleur de la voix"*, rapport du taux de hautes fréquences au taux de basses fréquences. Ainsi pour le son /a/, en faisant une coupure à 1 000 Hz, on obtient 60 % pour Véronique et 72 % pour Hélène (cf. plus haut),

- l'évolution de tous ces paramètres dans le temps,

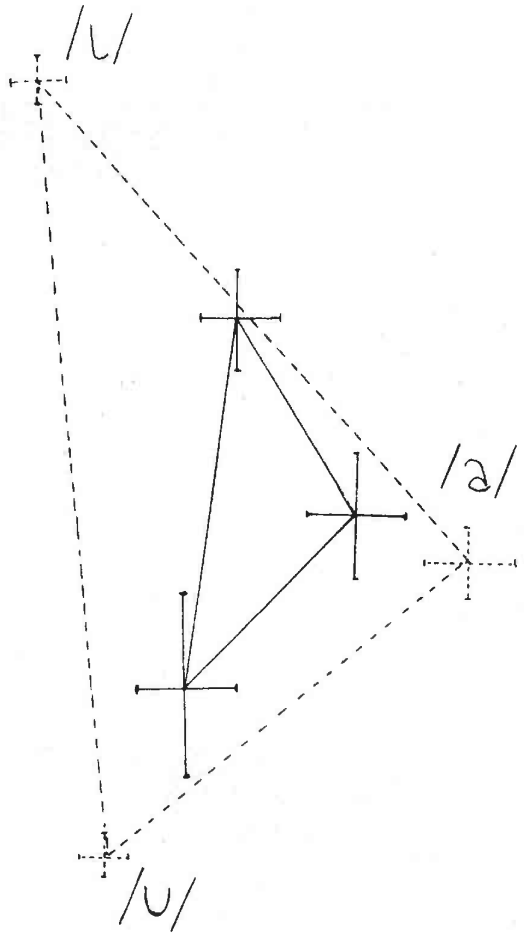


Fig. C-39 : Triangle des voyelles /a/, /i/ et /u/
Hélène (trait pointillé)
Véronique (trait plein).

- les spectres moyen et horizon pour des élocutions répétées de sons individuels ou pour des mots.

La figure C-41 illustre ces deux derniers points pour le mot / m a m ã / prononcé par Véronique et Hélène. Les spectrogrammes numériques et le contour mélodique sont donnés sur la figure C-40.

Le schéma suivant, C-42, adapté de [KIM - 74] et [HIKI - 76], qui donne la relation entre les mouvements des articulateurs et le déplacement du point représentatif du son produit dans le plan $F_1 - F_2$, permet d'apprécier les défauts majeurs du sujet au niveau articulatoire.

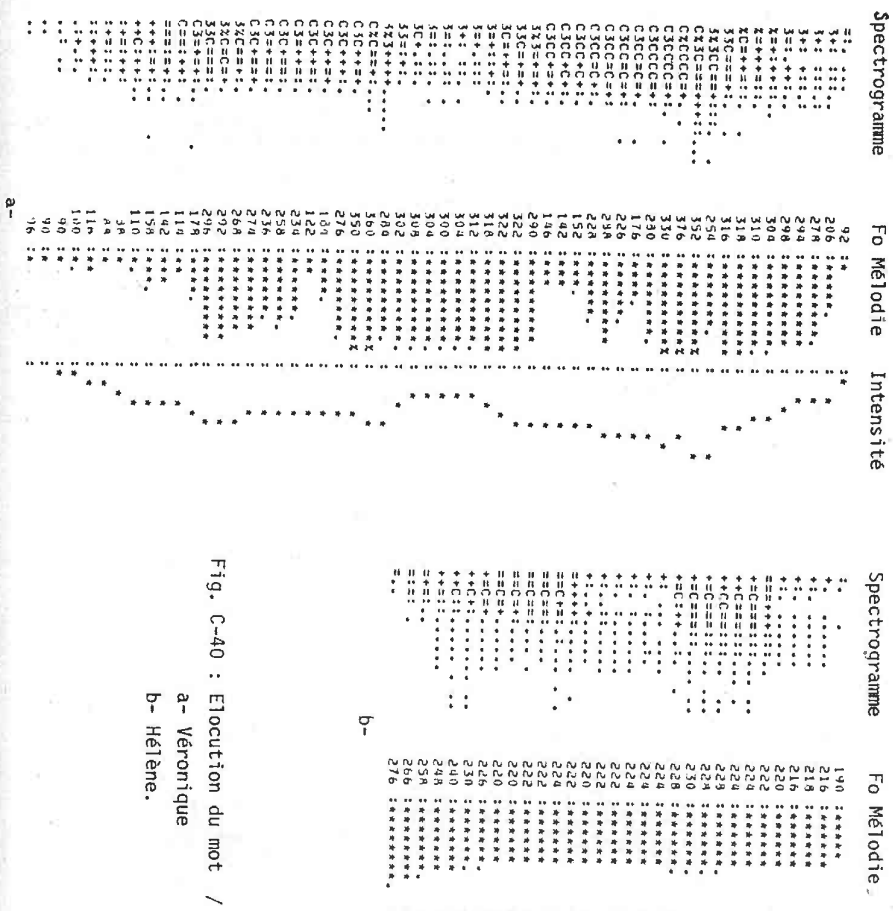
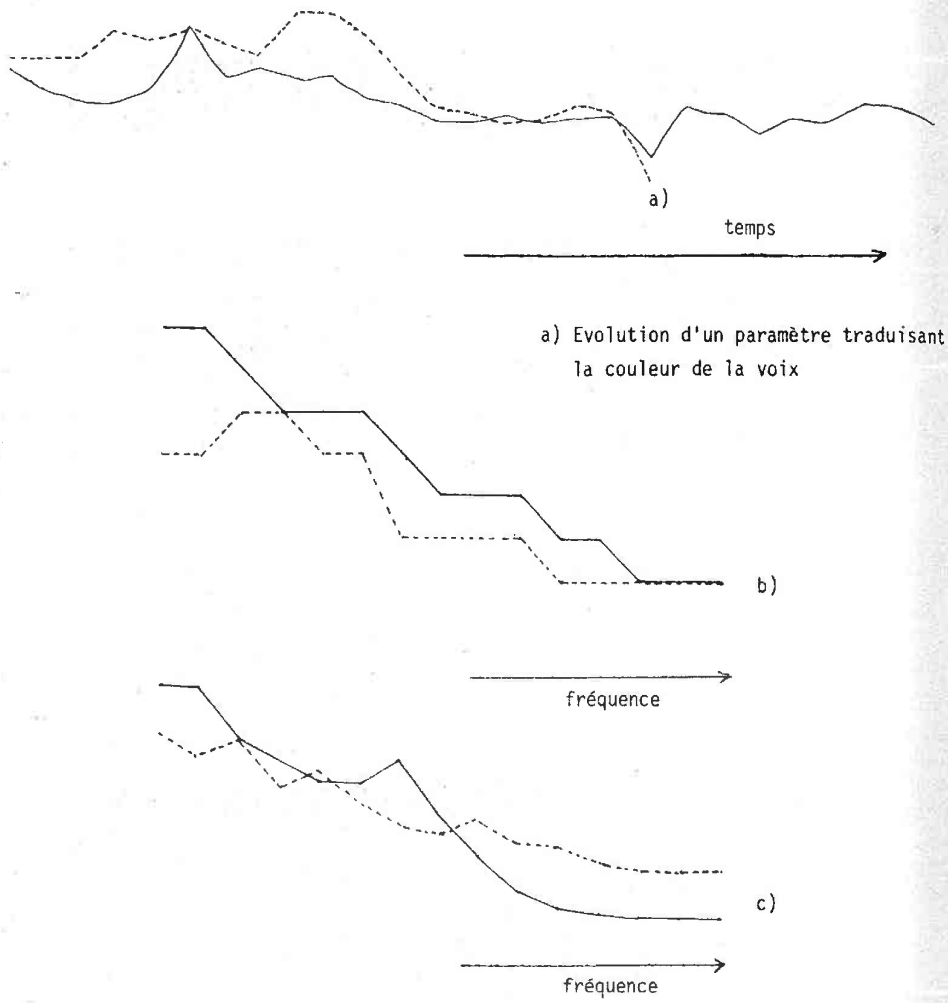


Fig. C-40 : Elocution du mot / m a m ã /
a- Véronique
b- Hélène.



a) Evolution d'un paramètre traduisant la couleur de la voix

b) Spectre horizon supérieur

c) Spectre moyen

Fig. C-41 : Elocution de / m a m ã / par Hélène (trait pointillé) et Véronique (trait plein).

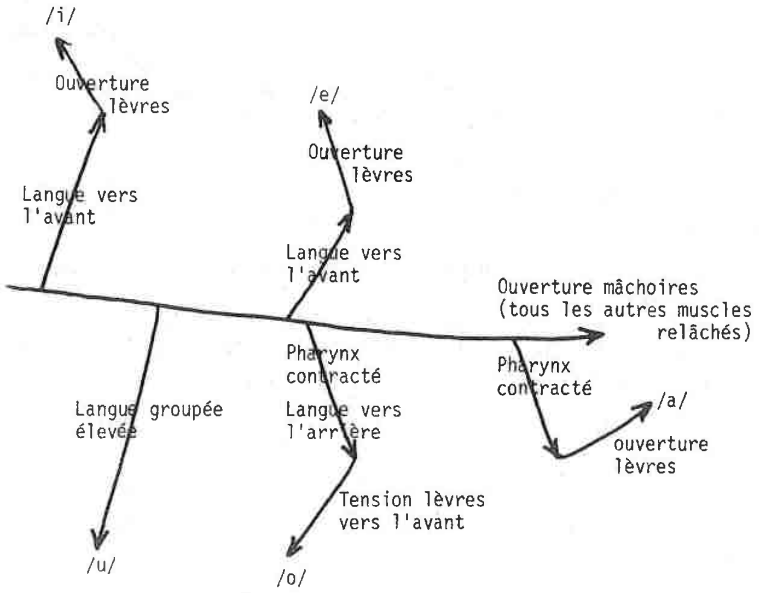


Fig. C-42 : Evolution des deux premiers formants en fonction de l'articulation.

6. Etude des segments

Le bilan des erreurs phonétiques, établi à partir de tests subjectifs et de mesures après segmentation du signal acoustique, permet de faire la part :

- des erreurs articulatoires pures : défaut de mode d'articulation le plus souvent ou de lieu d'articulation dans l'opposition s/ʃ ,

- des erreurs dues à la concaténation des mots qui n'apparaissent pas sur les mots isolés,

- des défauts dus à l'apparition de bruits parasites (friction du phonème suivant une consonne fricative, par exemple),

- des défauts de commande du voisement, du mouvement du velum et défauts de coordination.

La comparaison des fréquences d'occurence des phonèmes à la norme théorique et l'étude de la matrice de confusion, établie à partir des substitutions de phonèmes, apportent des renseignements de base sur :

- la maturité du système phonétique du sujet,

- les erreurs segmentales qui affectent de façon majeure l'intelligibilité de la parole.

Une mesure de l'information contenue dans la parole étudiée et de la redondance peut se faire, si l'on dispose de données statistiques suffisamment vastes, en considérant l'apparition d'un phonème x_i comme une manifestation d'une variable aléatoire X et en assimilant de ce fait les fréquences d'occurence $\{p(x_i), i = 1, \ldots, n\}$ à des probabilités [HATO - 74b]. L'auto-information du corpus se calcule alors suivant la relation :

$$H(X) = - \sum_{i=1}^{n} p(x_i) \log p(x_i) \ .$$

La redondance en ce qui concerne les sons isolés peut être chiffrée par :

$$R(X) = 1 - \frac{H(X)}{\log n}$$

où $\log n$ correspond à l'auto-information continue dans un corpus de phonèmes qui seraient équiprobables.

L'étude des fréquences d'occurence des couples ou des triplets de phonèmes (diphones et triphones) permet de la même façon d'atteindre des renseignements aux niveaux supérieurs. Il nous semble qu'il serait intéressant d'étudier la corrélation entre tous ces résultats et les taux d'intelligibilité de la parole étudiée.

III - TRAVAUX EN MODELISATION ET EN SYNTHESE

1. Modélisation

Les travaux en modélisation sont de deux types : ils peuvent concerner le système auditif, périphérique et central, ou le système phonatoire et articulatoire. Nous nous contentons de faire référence à certains des travaux concernant la production vocale et à conclure sur leur intérêt :

a) modèles de conduit vocal, destinés à faire le lien entre conduits et formants [JOSP - 77] ou [STEV - 80],

b) modèles du larynx :

. modèle à deux masses des cordes vocales [FLAN - 72],

. incidence de l'onde de débit, de la pression sous-glottique et de la tension des cordes vocales sur la fréquence fondamentale et l'intensité [GUER - 77] ou [MAED - 77],

. modèle faisant intervenir les résultats d'exploration par électromyogramme et mesure du mouvement du larynx [KAKI - 76],

. étude de la morphologie, de l'élasticité des tissus des cordes vocales et de leur comportement biomécanique pour la mise au point d'un modèle de commande musculaire du larynx [TITZ - 79].

La modélisation devient d'un grand intérêt lorsque les écarts entre modèle et appareil humain sont du même ordre de grandeur que les différences entre locuteurs. Elle permet alors :

. de relier les variations des paramètres physiologiques à celles des quantités acoustiques mesurables,

. de progresser ainsi dans la connaissance du processus de production vocale,

. de tirer enfin des enseignements sur l'origine des perturbations rencontrées dans la parole pathologique.

2. Synthèse vocale

La synthèse peut être considérée comme résultant d'une paramétrisation du signal suivant un modèle de production vocale. Ses avantages pour notre propos sont ceux de la modélisation auxquels, dans le cas des voix de déficients auditifs en particulier, il convient d'ajouter l'aide à la détermination de défauts majeurs affectant l'intelligibilité de la parole.

Parmi les travaux allant dans ce sens, nous retenons les suivants :

. synthèse par règles à partir de paramètres neurophysiologiques [HIKI - 74],

. recherches en synthèse articulatoire : par exemple, étude du mouvement des lèvres et de leur modélisation prédictive,

. détermination de l'effet de différents paramètres sur l'intelligibilité par modification de synthèse par règles de haute qualité (celle de D. KLATT) de façon à se rapprocher de la voix pathologique d'un sujet donné (JARED et BERNSTEIN). Deux paramètres composants de la parole sont considérés sous l'hypothèse de leur indépendance linguistique : les erreurs consonantales - omissions ou substitutions - et la distorsion temporelle suprasegmentale,

. remodelage du signal : correction des erreurs de rythme et de durée pour l'étude de leur importance dans l'intelligibilité par rapport à la parole normale [OSBE - 79].

IV - CONCLUSION

Les travaux en paramétrisation de la parole peuvent, dans une certaine mesure, répondre au besoin du spécialiste médecin de mettre en évidence les liens entre la déviation vocale et la pathologie, en particulier la pathologie de la voix.

En effet, les techniques d'exploration mises à la disposition du phoniatre présentent des inconvénients de subjectivité (tests à l'écoute sans critère quantitatif), d'inconfort et de difficulté de mise en place (laryngoscopie, ces défauts étant à nuancer avec l'utilisation des fibres optiques), de manque de détail local sur le comportement des cordes vocales (stroboscopie), de danger an cas de répétitions (radiographie) ou bien posent d'autres problèmes tels que ceux de l'analyse d'image (cinématographie ultra-rapide).

Par ailleurs, il n'existe pas de technique équivalant aux enregistrements électrocardiographiques ou audiométriques pour l'oreille.

Il paraît alors intéressant de proposer les mesures acoustiques comme outil clinique, ce qui aurait pour conséquence supplémentaire de fournir une terminologie unifiée pour décrire les désordres vocaux autrement qu'en termes qualitatifs.

Ce chapitre consacré à l'analyse des voix a montré l'intérêt présenté par les travaux en paramétrisation du signal, en modélisation et en synthèse. Deux domaines sont particulièrement concernés :

1. L'aide au diagnostic médical

Nous citerons quelques exemples récents illustrant cet aspect des études de voix où, d'ailleurs, les problèmes phonatoires (niveau laryngé) tiennent une place prépondérante par rapport aux problèmes articulatoires.

Le tableau de la figure C-43 rassemble ces exemples.

2. L'aide à l'orientation à donner à la rééducation

Grâce à la détermination préalable de l'acquis et des fautes affectant surtout l'intelligibilité de la parole, de façon à exploiter au mieux la phase d'éducation vocale. L'aspect articulatoire joue ici son rôle ainsi que la coordination des différentes activités intervenant dans le mécanisme de production.

Un exemple permet d'illustrer ce point. Il est en particulier important de définir si une valeur trop élevée du fondamental est due à un défaut purement phonatoire ou la conséquence d'efforts articulatoires exagérés. La déviation de F_0 par rapport à la normale peut parfois être corrélée avec une utilisation de stratégies articulatoires extrêmes (aspiration excessive, mouvements des mâchoires ou de la langue exagérés, etc.).

Observations - Diagnostic médical	Corrélats acoustiques	
Personnalité + considérations linguistiques	intensité I	[FLET - 76]
Désordres neurologiques	intensité : difficulté de maintenir I constant	
Etat d'esprit - état émotionnel	intonation et timbre	
peur	petite modulation périodique de la mélodie	
emphase - émotion	variations importantes de F_0	
âge	contrôle de F_0 plus difficile - F_0 plus haut - réduction du débit	
ennui	syllabes étirées - montée et chute lentes	[ASKE - 80]
	F_0 plus bas - ΔF_0 faible - plus de pauses	
Dépression	chute de F_0 et I - manque d'emphase	
	taux de $\frac{\Delta F_0}{F_0}$ locaux faibles	
	souffle et articulation pouvant être affectés par inhibition psychomotrice	
Schizophrénie	variabilité de F_0 (logorrhée et maniérisme) ou monotonie reflétant le comportement général	[CHEV - 77]
Maladie de Parkinson (discrimination cas graves / moins graves)	variance de F_0 basse / élevée débit affecté (tachylalie parkinsonienne)	
Dyslalie (sans anomalie physique)	démarche : recherche d'éléments symbolisables et essai de voir s'ils s'organisent dans un système phonétique	[CHRI - 77]
Hypernasalité	rapport du taux flux nasal / flux oral exagéré	[FLET - 76]
Anomalies des cris d'enfants	problèmes conjugués de respiration, de défaut de contrôle laryngé - conduit vocal contracté	[STEV - 82]
Pathologie du larynx	bruit additif influençant les formants perturbation de l'onde glottale	
voix rauque, dysphonie spastique, nodules...	taux de "jitter" (défaut de régularité temporelle) et "shimmer" (défaut de régularité de la puissance) élevés	[DAVI - 76]
paralysie des cordes vocales	dispersion de la répartition statistique de F_0 - pauses accrues	[GUBR - 77]
sévérité de l'atteinte (cancers, nodules, polypes)	facteurs ci-dessus pour discrimination (+ cas récalcitrants)	[DELL - 82]
diplophonie (déséquilibre des cordes vocales)	anomalie de période à période donnant l'impression de deux fréquences fondamentales différentes simultanées	[SCHO - 77]

Fig. C-43 : Aide au diagnostic médical en phoniatrie - Exemples

PARTIE D

LE SYSTEME SIRENE

INTRODUCTION A LA PARTIE D

Après avoir, dans les parties précédentes, placé notre travail dans le contexte de la communication parlée, développé les techniques mathématiques permettant d'accéder à différents paramètres pertinents ou à différents modes de présentation de la parole et envisagé la question des études de voix en temps différé, nous présentons dans cette dernière partie le système SIRENE, "Système Interactif pour la Rééducation vocale des Enfants Non-Entendants".

Dans le premier chapitre, nous verrons successivement les idées de base qui ont conduit à la conception et à la réalisation du système, comment y sont intégrées les techniques d'E.I.A.O. (Enseignement Intelligemment Assisté par Ordinateur), quels sont les apports de l'informatique et des méthodes de la reconnaissance des formes et de l'intelligence artificielle (comparaison dynamique, analyse de réponse, ...).

Dans le deuxième chapitre, nous décrivons les conditions d'expérimentation du système SIRENE et les réflexions que les essais nous ont inspirées au niveau psychopédagogique et technique.

Nous terminons, dans un troisième chapitre, sur la définition d'une version de SIRENE sur micro-ordinateur et sur les extensions et adaptations possibles du système à des domaines autres que l'aide aux malentendants.

CHAPITRE 1

STRUCTURE DE SIRENE

I - IDEES DE BASE

1. Education de la parole

S'il est certain que parler est quotidiennement la façon la plus naturelle de communiquer, il est évident également que la parole, dans certains cas comme celui des sourds de naissance, n'est pas une forme naturelle d'expression.

Or, un certain nombre de raisons font qu'il est difficile de lui substituer d'autres moyens de communication pour des considérations techniques d'abord mais surtout du fait que le bain auditif permanent dans lequel est plongé l'enfant sans handicap l'amène à acquérir la parole et le langage sans effort et favorise du même coup son développement social.

Pour une meilleure insertion de l'enfant déficient de l'ouïe et de la parole, il est alors nécessaire, dans une collaboration entre scientifiques et rééducateurs, de mettre en oeuvre des techniques d'enseignement modernes dans les conditions les meilleures possible (diagnostic précoce, suivi de la part de l'entourage, ...) pour favoriser l'éducation auditive et la communication par la parole.

2. SIRENE, aide visuelle

La conception et la réalisation de systèmes d'aide à la production ou la compréhension de la parole a connu un nouvel élan depuis 1972 grâce à l'introduction de la mini- puis micro-informatique. Notre système fut le premier en France à utiliser largement ses possibilités, après principalement les expériences américaines [KALI - 72], [NICK - 73] et à la suite de la définition dans différents pays (Etats-Unis, Suède, Japon, ...) d'aides de nature essentiellement analogique.

Le système SIRENE qui s'intéresse à la production vocale en situation d'apprentissage est conçu comme une aide au rééducateur ou au professeur (que nous désignerons par "*le maître*" dans la suite). Il est destiné à seconder celui-ci dans sa tâche grâce à des exercices faisant appel à la présentation visuelle de différents paramètres de la parole ou des jeux dans lesquels les performances vocales sont déterminantes. L'appel au sens de la vue pour compenser l'ouïe déficiente est à la base même du système comme de bien d'autres (cf. A.2). Au jugement du maître et à la sanction brute d'un module d'évaluation "*non psychologue*" s'ajoute la possibilité pour l'élève de faire une estimation de sa performance, de relier l'effort au résultat.

3. Vision et audition

Pour le malentendant, le sens de la vue intervient comme un palliatif en situation d'apprentissage. Il est par ailleurs largement sollicité -éventuellement avec le toucher et l'ouïe résiduelle- en situation de communication, ne serait-ce que dans le cas du langage signé (ou codé) lorsqu'il est utilisé dans le milieu familial et à l'école ou dans le cas de la lecture faciale.

Quelques considérations physiologiques permettent de dégager les grandes caractéristiques de la vue par rapport à l'ouïe et de comparer leurs champs d'application, notamment en ce qui concerne :

- la quantité d'information délivrée au cerveau. La rétine sensorielle de l'oeil est composée d'environ 130 millions de cellules réceptrices auxquelles correspondent un million de fibres dans le nerf optique. De son côté, l'organe de Corti, qui est l'organe d'analyse de l'oreille, est formé de 30 000 cellules environ transformant l'énergie acoustique en potentiels neurosensoriels et reliées à 30 000 cellules-ganglions dans le nerf auditif. La neuro-anatomie de l'oeil fait alors qu'il délivre par unité de temps plusieurs fois la quantité d'information délivrée par l'oreille. Il est à noter que l'information visuelle est de nature variée : vision des formes, du fond, des contours, perception des couleurs, ...,

- la dynamique définie comme le rapport de l'intensité d'un stimulus au seuil de douleur à celle d'un stimulus au seuil de sensibilité. La dynamique est tout à fait comparable dans les deux cas (15 pour l'oeil - 12 pour l'oreille). Il faut cependant considérer que la vision fait appel à deux systèmes sensoriels par l'intermédiaire de deux types de récepteurs : les cônes qui interviennent en vision photopique (couleurs, acuité visuelle) et les bâtonnets qui interviennent en vision scotopique (formes, images de très faible luminance). En réalité, la dynamique disponible à un instant donné est la plus grande pour l'oreille,

- l'adaptation à des variations d'intensité du stimulus. L'oeil a une adaptation lente à des variations d'intensité du stimulus visuel (l'adaptation à l'obscurité prend jusqu'à 30 mn) alors que l'oreille a un pouvoir d'adaptation particulièrement remarquable : un bruit fort, par exemple, n'affecte le seuil d'audibilité que pendant quelques millisecondes,

- la discrimination temporelle. Bien qu'il soit difficile de trouver une définition analogue pour l'oeil et l'oreille, on peut comparer leurs performances par des expériences de détermination de fréquence de fusion critique (fréquence à laquelle une succession de stimuli régulièrement espacés se confondent) ou de mesures de la capacité à déceler l'ordre d'émission de deux ou plusieurs stimuli. On déduit que l'oreille a un pouvoir de résolution temporelle supérieur et une meilleure aptitude à déceler l'ordre d'apparition des stimuli. Alors que l'oreille est mieux adaptée à la perception du successif, on peut conclure au contraire que l'oeil est supérieur dans la perception du simultané.

En ce qui concerne la perception d'événements de grande variabilité et de grande dynamique comme la parole, il est bien évident que les caractéristiques de l'oreille en font un outil tout à fait adapté. L'ordre du décodage, quant à lui, n'est pas absolument nécessaire pour la compréhension. Il existe en effet dans la parole une richesse du message qui permet la levée de certaines ambiguïtés grâce à la redondance, le nombre limité d'associations phonétiques possibles et le contexte.

La transposition du message sonore en message visuel ne peut se faire par une simple traduction des changements d'intensité d'un code dans l'autre du fait des différences citées plus haut. Il faut par conséquent chercher d'autres types de visualisation faisant appel au caractère plus global de la perception visuelle [ROUF - 82] Dans ce cas, l'oeil est tout à fait apte à faire la discrimination entre des formes différentes. On peut par ailleurs considérer que la difficulté de résoudre les différences temporelles peut être levée par la représentation visuelle où se fait la distinction spatiale.

Des événements non détectés ou non différenciés pendant l'élocution peuvent l'être ensuite grâce au retour visuel dans un processus de retour en arrière voisin de celui de la compréhension lors de la lecture.

Un dernier point dont il faut tenir compte est la capacité de perception visuelle de l'enfant sourd. Contrairement à des théories du passé, il semble acquis qu'il n'existe pas de compensation organique, ni de compensation perceptive, si ce n'est la tendance de se fixer des points de repère visuels dans le contact quotidien avec l'environnement ou l'utilisation de matériel didactique imagé. Bien que les résultats expérimentaux apportent des résultats parfois contradictoires, on peut considérer, en général, que l'enfant sourd présente une faiblesse au niveau de l'organisation perceptive. On constaterait un retard dans les capacités visuomotrices traduisant un défaut de stratégie de recherche dans l'appréhension d'une image.

On peut, à partir des réflexions précédentes, considérer que l'oeil peut jouer un rôle en situation de rééducation vocale, à condition de lui fournir des équivalents visuels des événements sonores qu'il puisse directement appréhender : tenir compte des possibilités de saisie globale de l'oeil et de la richesse de l'information fournie ensuite au cerveau, ne pas chercher à transposer directement les sons dans le code visuel, donner une image dans laquelle apparaît clairement la notion de temps lorsque c'est nécessaire, et, de façon générale, réduire au maximum l'organisation *"temporo-séquentielle"* au profit d'une organisation *"spatio-séquentielle"*.

4. Représentation visuelle des paramètres de la parole

La contreréaction visuelle est, chaque fois que cela est possible et jugé nécessaire, proposée en temps réel. Pour le cas où l'analyse requise ne peut se faire en cours d'acquisition de parole ou lorsque l'on ne visualise que le résultat d'une commande vocale, l'image apparaît en temps légèrement différé. Le décalage peut d'ailleurs, dans les deux cas, être accentué volontairement si l'on désire que l'élève prenne du recul par rapport à ses productions vocales et s'affranchisse de la présence de l'image-sanction.

Les possibilités actuelles de gestion d'écran et de manipulation d'images en couleur et de graphismes permettent d'envisager des représentations visuelles de variétés multiples depuis les plus simples (tracés, contours) aux plus sophistiquées (dessins en couleur, images tridimensionnelles, ...). Pour notre part, nous utilisons dans la version de base une console graphique unicolore et nous nous sommes surtout attaché à donner une cohérence sémantique aux représentations sur écran de façon que l'élève puisse faire une correspondance biunivoque entre le tracé et la performance phonatoire ou articulatoire.

Le maître dispose, de son côté, d'une console alphanumérique à clavier et sélectionne le jeu voulu après affichage d'un menu. A côté d'une version standard, la possibilité de modifier les facteurs qui conditionnent la difficulté de l'exercice, les vitesses de tracé, les procédures de segmentation parole/non parole, etc. est laissée à son appréciation.

II - CARACTERISTIQUES GENERALES

Les avantages de l'introduction de l'informatique dans un système tel que SIRENE sont tout à fait classiques. Aussi nous limitons-nous à une énumération organisée des possibilités d'une version *"informatique"* d'un tel équipement par rapport à une version purement *"électronique"* :

[1] réduction des problèmes pratiques d'utilisation (ajustements simplifiés, protection contre les erreurs de manipulation, auto-apprentissage possible quand l'élève est capable de juger ses productions et de travailler seul),

[2] utilisation de processeurs numériques spécialisés pour réaliser des tâches diverses comme :

- la préanalyse du signal, la réduction de données vocales, l'extraction de paramètres de base telles les données spectrales ou la fréquence fondamentale,
- la reconnaissance automatique de la parole,
- la gestion d'écran,

[3] introduction du mode conversationnel pour le choix des procédures de rééducation, l'ajustement des seuils, le lancement d'actions diverses : répétitions d'exercices dans les conditions choisies, stockage des formes sonores pour études ultérieures, ...,

[4] modularité de la structure d'ensemble,

[5] utilisation de méthodes complexes de calcul numérique et d'analyse,

[6] adjonction de modules d'aide à la progression de la rééducation :

- évaluation,
- émission automatique de conseils et d'encouragements,
- appel à des procédures spéciales suivant les performances,
- mémorisation de la progression, etc.,

[7] dans une vue plus globale, réalisation, implantation et test de systèmes à moindre coût.

III - STRUCTURE ET PRINCIPES D'UTILISATION

Le système SIRENE, dans sa version prototype sur minicalculateur, est organisé autour d'un MITRA 125 et fait appel aux organes suivants :

- un microphone avec casque amplificateur,
- un lecteur de cassette ou de bande magnétique,
- un analyseur spectral ("vocodeur"),
- un détecteur de fréquence fondamentale de la voix,
- un convertisseur analogique - numérique,
- un écran - élève et un écran - professeur avec clavier.

L'ensemble des modules (analyse fine, calcul, conversation, évaluation, visualisation, ...) est réalisé sous forme de logiciels écrits en FORTRAN.

La figure D-1 résume la configuration du système.

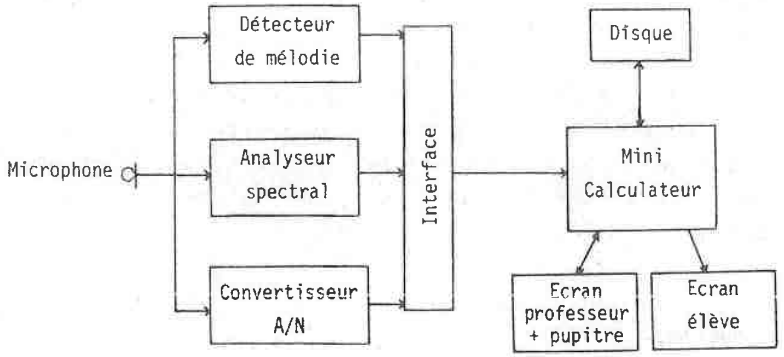


Fig. D-1 : Configuration du système SIRENE sur mini-ordinateur.

Les données qui seront ensuite traitées sont essentiellement :

- le fondamental de la voix,

- la succession dans le temps de spectres à court terme (notés prélèvements dans la suite) constitués des valeurs des intensités de sortie des canaux de l'analyseur spectral (12 à 15 canaux répartis dans la zone informante de la parole soit environ 200 Hz à 5 000 Hz),

- le signal de parole échantillonné et numérisé qui, pour des besoins d'analyse plus fine, peut être conservé en mémoire et traité en temps différé.

Le système est divisé en trois grands modules de jeux (figure D-2) :

. PP - paramètres prosodiques,
. PF - paramètres fréquentiels,
. VM - vocabulaire de mots.

Un menu global est proposé pour chacune de ces rubriques dans lequel le maître sélectionne l'exercice choisi, modifie éventuellement des seuils ou des conditions de travail.

Certains exercices font référence à un modèle affiché en partie haute de l'écran. Le maître peut, dans ce cas, soit l'énoncer directement au microphone, soit le choisir parmi des formes-types gardées en fichier. Pour des essais répétés, après effacement de l'écran, la réapparition du modèle est automatique.

La figure D-3 donne l'automate de base des exercices avec forme-témoin de référence.

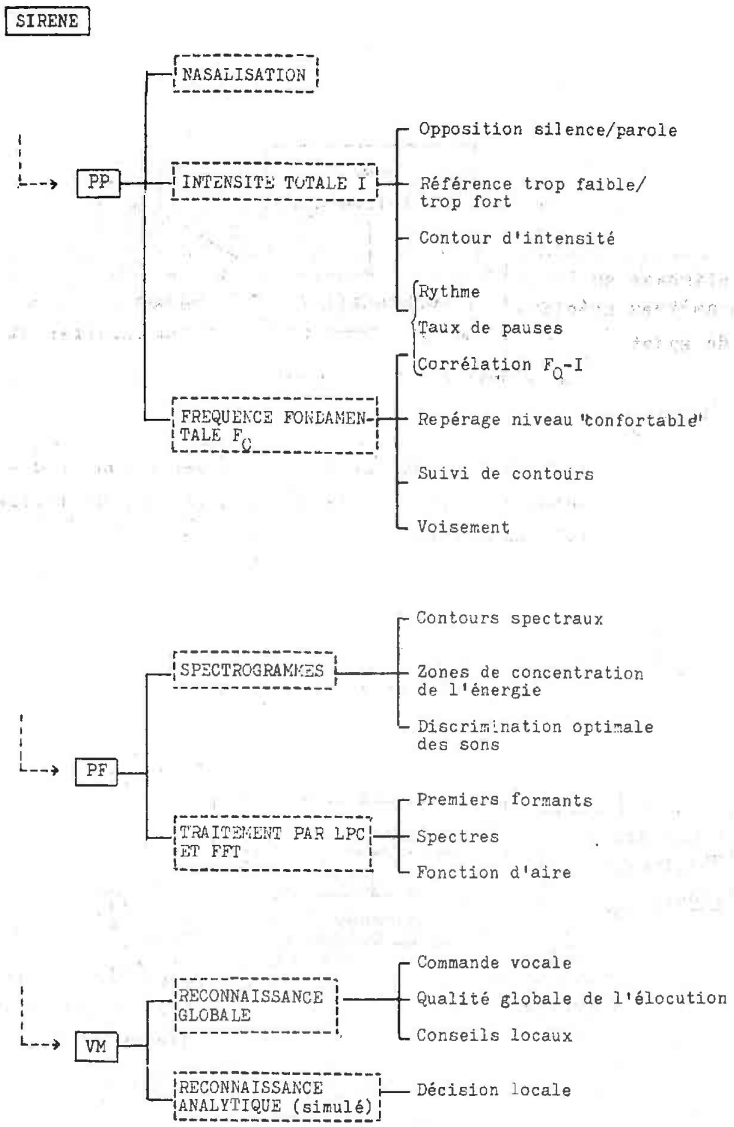


Fig. D-2 : SIRENE : division en trois modules d'exercices.

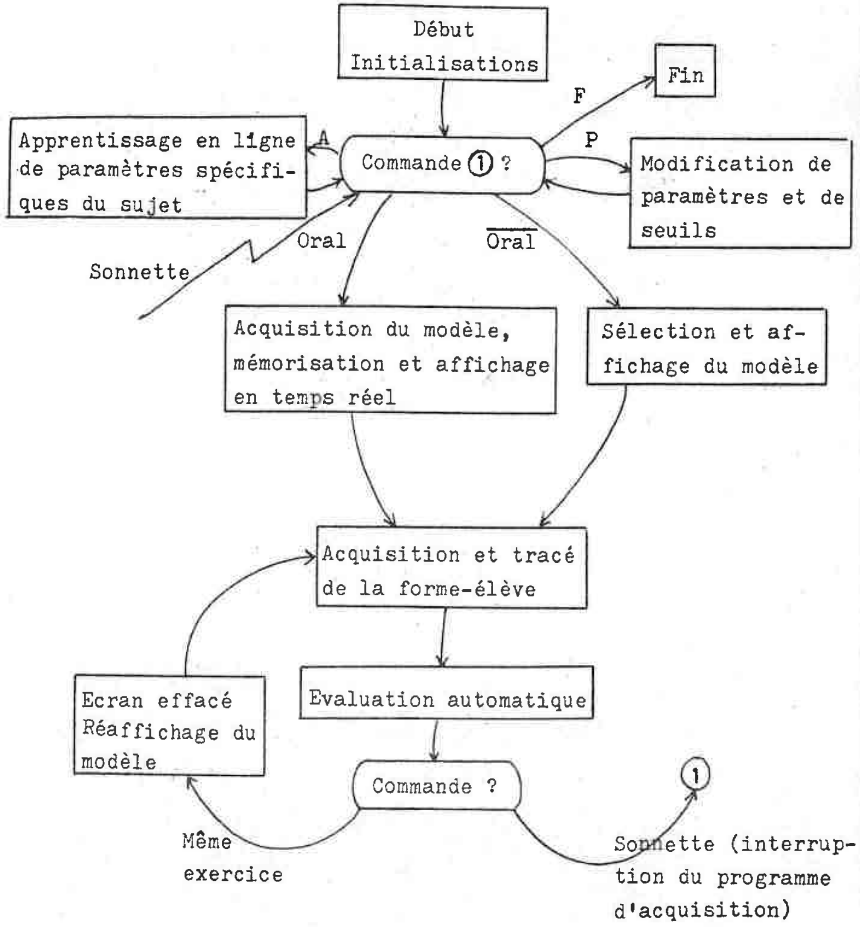


Fig. D-3 : Organisation des exercices avec forme-modèle de référence.

Nous allons donner les caractéristiques principales de ces trois catégories de jeux.

1. PP - Paramètres prosodiques

Les exercices et leur combinaison permettent de s'intéresser aux paramètres suprasegmentaux : hauteur et mélodie, intensité, durée, rythme.

L'acquisition des données vocales se fait en continu avec suppression des zones de silence. La succession des prélèvements utiles est repérée par leur position dans des tampons en bascule, de taille variable suivant la cadence future de présentation visuelle en temps réel, compte-tenu des actions qui suivent l'acquisition.

Au niveau de ces actions, on rencontre :

- la suppression contrôlée des attaques (coups de glotte),
- le calcul de la fonction à visualiser, après corrections et lissages,
- s'il y a lieu, l'évaluation cumulée de la performance,
- le stockage éventuel de la forme,
- la visualisation proprement dite.

Ces traitements sont conditionnés par un certain nombre de paramètres ajustables tels que :

- le délai de prise en compte de la parole à l'attaque,
- la cadence de présentation,
- l'étalement sur écran des images,
- les conditions du jeu et les seuils de satisfaction, etc.

Après avoir, dans une phase préalable, apprécié de façon automatique le fondamental confortable de l'enfant, ses capacités de niveau sonore ou de souffle, le maître peut sélectionner dans un menu l'un des exercices suivants :

- Opposition son - silence. Suivant le capteur et le choix des seuils, peuvent être mis en évidence la présence de parole, de voisement, de nasalité, de souffle... L'action résultante (vitesse de tracé d'un dessin par exemple) est conditionnée par l'énergie du signal fourni par le capteur.

- Evolution du niveau d'intensité dans le temps pour l'évocation du niveau global, des niveaux locaux, des pauses et de la durée (figure D-4a).
On a adjoint un aspect fréquentiel à cet exercice avec la possibilité de se limiter à une zone de fréquence choisie pour des distinctions entre sons tenus tels que /s/ - /ʃ/ ou /i/ - /e/ . La figure B-29 qui montre les réponses de l'analyseur spectral à des phonèmes fricatifs sourds d'enfants, met clairement en évidence que l'énergie en haute fréquence des sons /s/ est rejetée au-delà de la limite du dernier filtre. Le choix de la zone des derniers canaux de l'analyseur permet d'obtenir les formes de la figure D-4 b. Cette limitation de la zone fréquentielle d'étude permet en outre d'interdire à l'élève de contourner l'exercice en "forçant" un phonème pour en simuler un autre. De la même façon, la figure D-4 c montre grâce à l'étude des basses fréquences l'évolution de l'intensité pour des successions de voyelles pour lesquelles le degré d'aperture (aux lèvres) va croissant.

- Opposition binaire voisé - non voisé mise en évidence sur des syllabes ou des énoncés courts (figure D-5a).
Comme exemple choisi parmi d'autres, les lignes suivantes illustrent l'algorithme suivi pour la construction des crêneaux de la fonction binaire :

Lexique :

entiers	I	intensité du prélèvement courant,
	Seuil I	seuil de parole/non parole,
	F_0	fréquence fondamentale courante,
	Seuil F_0	seuil voisé/non voisé,
	N	numéro d'ordre,
	NSilence	seuil de silence (en nombre d'intervalles)
booléen	Ind	indice de voisement

Initialisations : N = NSilence + 1
Ind = faux

Traitement en temps réel en cours d'acquisition continue

	I > Seuil I (1)		I < Seuil I (2)		
	F_0<Seuil F_0	F_0>Seuil F_0	N<NSilence	N = NSilence	N>NSilence
Ind = faux	N = 0 Ind = faux →	N = 0 Ind = vrai ↑→	N = N + 1 Ind = faux →	N = N + 1 déplacement sans tracé	pas de traitement
Ind = vrai	N = 0 Ind = faux ↓→	N = 0 Ind = vrai →	N = 1 Ind = faux ↓→		

Le traitement de la partie (2) du tableau garantit la continuité du tracé dans le cas des phonèmes occlusifs sourds.

Pour des exercices portant également sur les inflexions mélodiques, la valeur de F_0 est figurée en absolu.

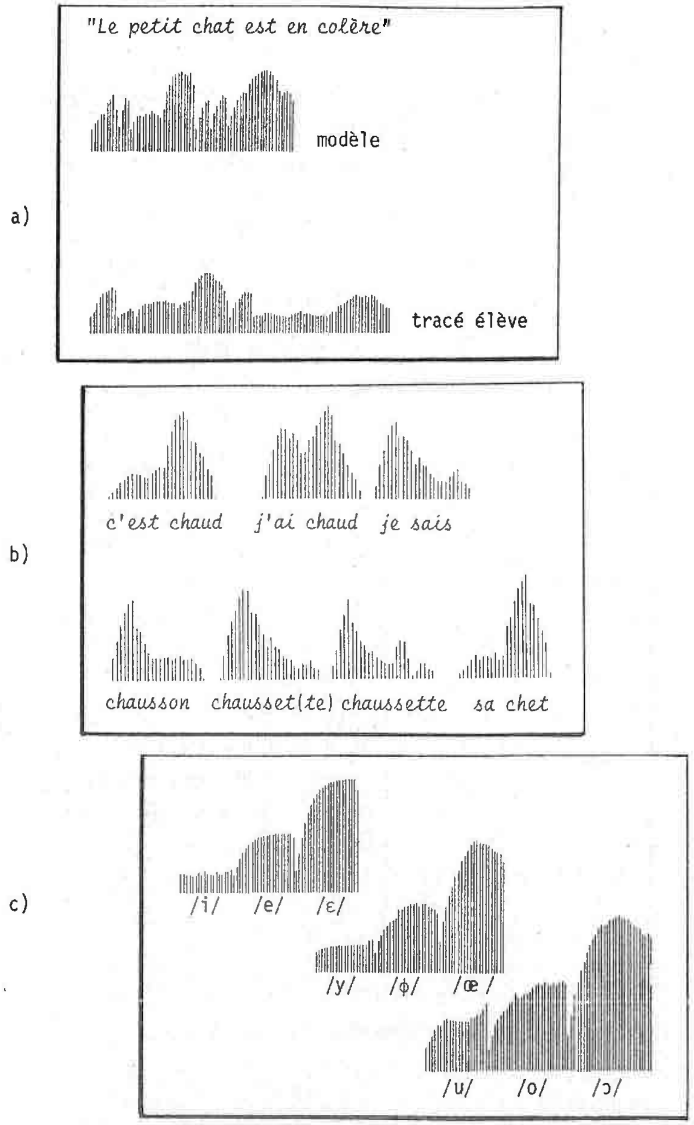


Fig. D-4 : Evolution dans le temps de l'intensité totale (a), en haute fréquence (b) et en basse fréquence (c).

- Contour mélodique suivant une échelle logarithmique pour une comparaison absolue des écarts de fondamental. Les exercices peuvent porter sur les points suivants :

  . tenue d'un niveau jugé esthétiquement convenable pour l'élève, puis à des niveaux "*bas, moyen et haut*",

  . évolution dans un tunnel, niveau final à atteindre (figure D-5b),

  . essai de chant en échelle absolue ou relativement à une note de base déterminée par apprentissage préalable,

  . reproduction d'un contour-modèle énoncé par le maître ou sélectionné dans une base de données sur un son tenu ou des énoncés courts choisis pour l'aspect informant de l'intonation(fig. D-5c).

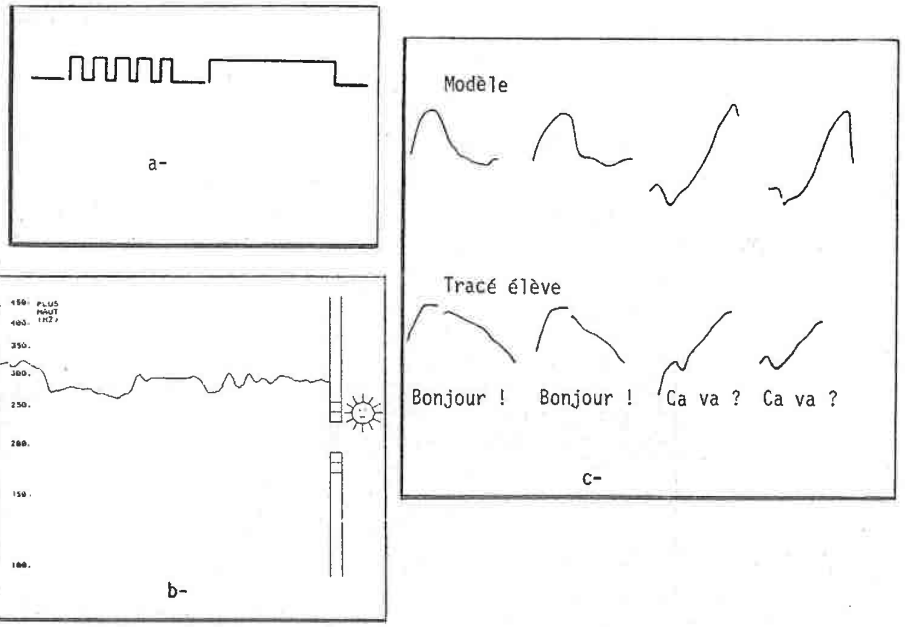


Fig. D-5 : a- Voisement
b- Fondamental : valeur-cible à 200 Hz
c- Reproduction de contours-types

- Rythme mis en évidence à partir d'un algorithme simple de détection des noyaux syllabiques étudié , après avis de spécialiste, pour que des segments tels que "*fleur*" constituent un tout, ce qui exclut des énoncés-tests de plusieurs syllabes renfermant des consonnes liquides ou nasales (figure D-6). Un tel algorithme associé à une détection de sons fricatifs permet d'affiner la mise en évidence des différences entre couples de mots voisins. La visualisation doit alors distinguer les parties voisées ou non, les parties fricatives ou non et plosives ou non.

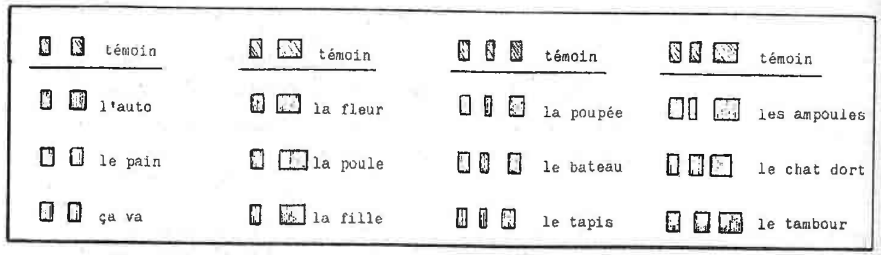


Fig. D-6 : Rythme : reproduction de schémas-témoins

- Décorrélation fondamental - intensité : variations rapides des deux facteurs ou contour mélodique modulé par l'intensité. Les exercices concernent les essais de contrôle des cordes vocales, par exemple montée mélodique sans perte de souffle (figure D-7a) ou, par opposition, tenue de fondamental avec variations d'intensité (figure D-7b. En haut , avec S. normalement entendant, l'on peut remarquer une corrélation naturelle due à l'effort de souffle).

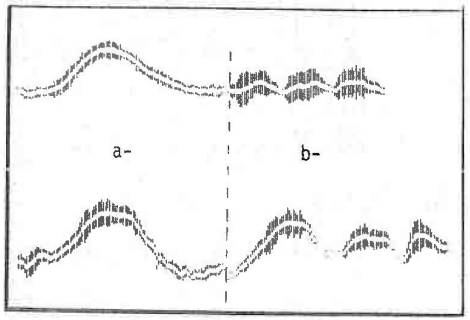


Fig. D-7 : Corrélation F_0-I a- Montée mélodique à I constant
b- Exercice de tenue de F_0

La performance vocale de l'enfant est guidée et sanctionnée par le retour visuel et par des messages d'appréciation : sens de l'effort à fournir, erreurs, référence aux essais antérieurs. A titre d'exemple, nous décrivons une procédure d'évaluation de la performance dans le cas de l'imitation d'un contour-témoin, à partir de la mise en coïncidence optimale de la forme-témoin et de la forme produite par l'élève. A chacun des deux contours est associée la fonction ternaire qui code la pente sur trois niveaux, comme indiqué sur la figure D.8, moyennant un seuil de pente ajustable.

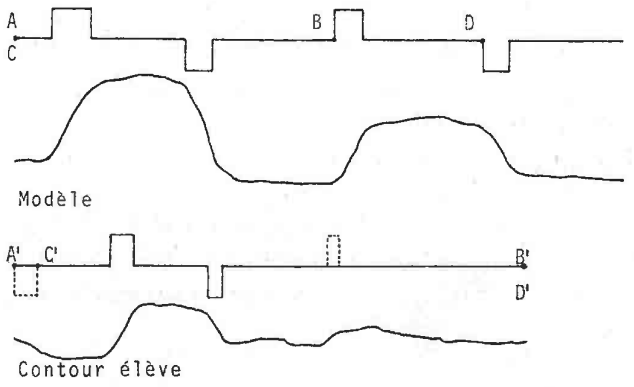


Fig. D-8 : Imitation de contour : évaluation de la performance

1er cas : seuil de pente moyen

L'algorithme de cadrage met en coïncidence les zones [A,B] et [A',B'] et fait les repérages suivants :

a) démarrage trop lent,

b) pour le premier motif : montée faible, durée correcte, descente à un niveau insuffisamment bas,

c) deuxième motif absent malgré durée prolongée.

2ème cas : seuil de pente plus faible (traits pointillés)

Le cadrage met en coïncidence [C,D] et [C',D'].
L'algorithme d'évaluation fait les repérages suivants :

a') et b') identiques à a) et b), mais en plus,

c') chute non prévue au départ. (En réalité, ceci peut être considéré comme un ajustement à ne pas pénaliser),

d') deuxième motif : montée très faible, descente non perçue.

La note finale tient compte de ces différents points. Lorsque le décalage est trop grand ou la zone de coïncidence trop peu importante, aucune évaluation n'est faite.

On peut noter que l'ensemble des modules PP s'adapte directement à l'apprentissage d'une langue étrangère (phénomènes d'accent et d'intonation), faisant de SIRENE un outil d'enseignement assisté des langues.

2. PF - Paramètres fréquentiels

Les principes de base sont identiques à ceux de la catégorie précédente. Les choix possibles sont les suivants :

- Estimation du spectre à court terme : sorties brutes des canaux de l'analyseur spectral ou forme traitée par lissage trigonométrique, ce qui réduit légèrement la cadence de visualisation (figure D-9a).

- Mise en évidence des deux premières zones de concentration de l'énergie : évolution dans le temps des fréquences ou figure paramétrique.

- Projection d'un son dans un plan optimal au sens de la meilleure discrimination entre groupes de phonèmes, comme décrit au chapitre B-2, grâce à un apprentissage préalable et à la mémorisation des facteurs de projection. Dans la suite, nous donnons l'organisation de cet exercice et des exemples de visualisation (figure D-10). La plus ou moins grande proximité des points représentatifs à la cible (imposée ou meilleure performance antérieure de l'élève) permet le calcul d'une note de réussite et la prise de conscience visuelle par l'élève de sa performance.

Différents facteurs favorisent la dispersion des points représentatifs des sons prononcés, en particulier :

[1] le bruit de quantification des données : pour limiter l'influence de ce facteur, un médaillon affiché systématiquement donne l'indication "*trop faible*" ou "*trop fort*" de façon à contrôler le niveau d'émission vocale,

Fig. D-9a : Spectres "vocodeur" bruts (à gauche) ou lissés (à droite) affichés en ligne

Fig. D-9b : Spectres "vocodeur" lissés

Fig. D-9c : Spectres de prédiction linéaires

a-

b-

Fig. D-10 : Exercice de discrimination entre classe de sons (a- organisation, b- illustration).

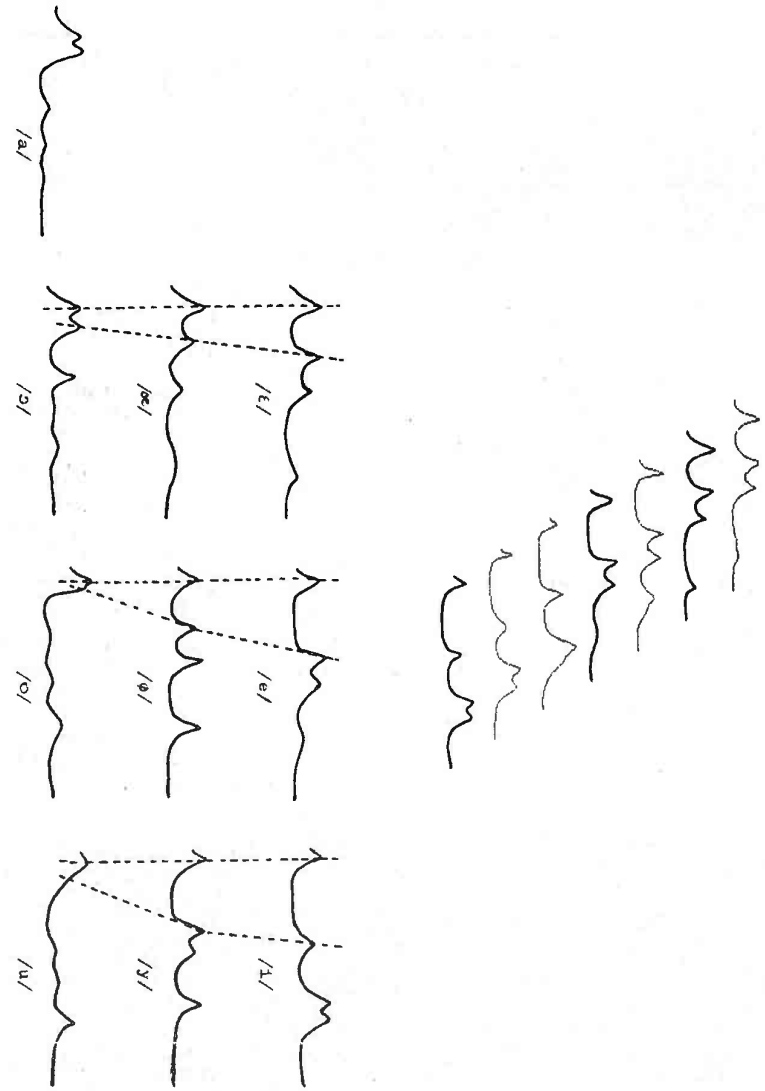


Fig. D-11 : Spectres de prédiction linéaire des voyelles calculés et affichés en ligne (enfant S., 12 ans)

[2] l'attaque et la chute des phonèmes : ce dernier problème est résolu en partie en encadrant le niveau sonore autorisé mais seule l'imposition de seuils variables suivant les exercices pourrait empêcher l'affichage de points s'éloignant de ceux de la partie stable. Le problème de l'attaque est résolu en n'envisageant l'affichage qu'après un laps de temps ajustable suivant la décision de parole.

- Configuration du conduit vocal, spectres et formants grâce à un traitement logiciel de prédiction linéaire. Dans nos conditions de travail, l'acquisition du signal numérisé en mémoire centrale et la sélection de la zone à soumettre au traitement entraîne que la cadence d'affichage est de quelque formes par seconde.

* la figure D-12 montre des fonctions d'aire affichées en cours d'élocution, pour S. (garçon, 11 ans).

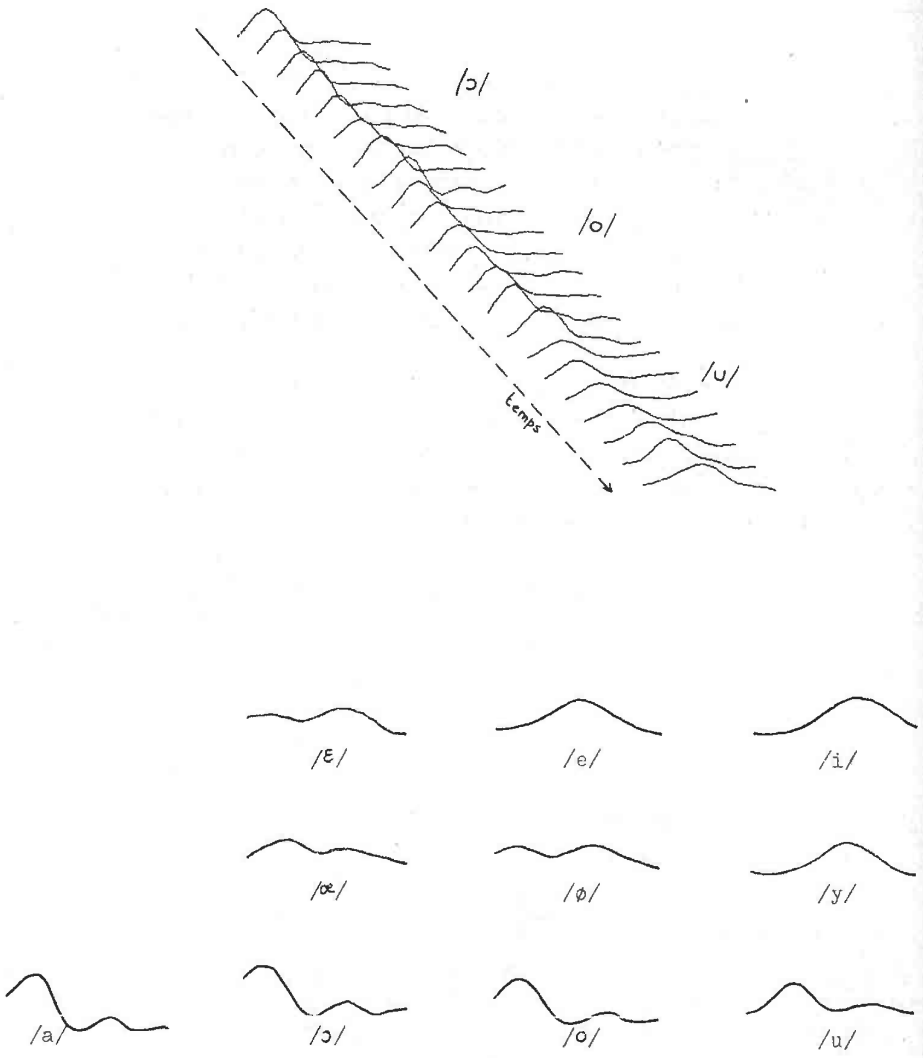


Fig. D-12 : Présentation en ligne des fonctions d'aire (voyelles enfant S.).

* Sur la figure D-13, différents tableaux indiquent des proximité entre fonctions d'aire (différence de logarithmes) dans quatre situations :

  a) voyelles M.C.,
  b) voyelles S. (garçon, 12 ans),
  c) voyelles S. contre voyelles M.C.,
  d) voyelles /a/, /i/ et /u/ de S. comparées à l'ensemble des voyelles de M.C. et aux autres de S. On peut remarquer que le /i/ de S. est plus proche de son propre /y/ que du /i/ de M.C. Dans les autres cas, au contraire, la voyelle de M.C. pourrait être choisie comme référence pour l'exercice.

  Généralement, des références obtenues à partir d'un enfant de même sexe et d'âge voisin conviennent.

* Les figures D-9b et D-9c indiquent les spectres lissés obtenus respectivement à partir de l'analyseur spectral et le traitement numérique de prédiction linéaire pour les voyelles extrêmes. Notons que les limites sur l'échelle des fréquences sont respectivement les plages 250-5000 Hz et 0-6000 Hz. L'approximation des formants sur les formes D-9c ou D-11 (voyelles visualisées en ligne) se fait très rapidement par détection des premiers maxima avec le risque déjà exposé de ne pas repérer le deuxième formant (cas de /o/ et /u/ sur la figure). La recherche des pôles et le calcul des fréquences de formants (cf. paragraphe B.1.V.) permet de résoudre cette difficulté mais au détriment de la cdence de visualisation.

M.J. (femme)

	/a/	/ɔ/	/o/	/u/	/œ/	/ø'/	/y/	/ɛ/	/e/	/i/
/a/	0	398	711	812	692	764	1380	818	1116	1601
/ɔ/		0	107	139	195	318	705	349	629	839
/o/			0	29	129	342	595	369	567	763
/u/				0	77	236	435	248	401	545
/œ/					0	110	437	128	259	430
/ø/						0	427	101	235	210
/y/							0	321	110	346
/ɛ/								0	129	270
/e/									0	239
/i/										0

a-

S. (garçon, 12 ans)

	/a/	/ɔ/	/o/	/u/	/œ/	/ø/	/y/	/ɛ/	/e/	/i/
/a/	0	474	539	916	1199	1396	877	1035	1386	
/ɔ/		0	77	258	232	568	662	448	308	561
/o/			0	129	236	568	607	671	304	527
/u/				0	250	285	251	600	103	208
/œ/					0	383	482	329	279	488
/ø/						0	45	254	116	66
/y/							0	415	95	25
/ɛ/								0	306	396
/e/									0	57
/i/										0

b-

M.C.

	/a/	/ɔ/	/o/	/u/	/œ/	/ø/	/y/	/ɛ/	/e/	/i/
/a/	0	247	226	506	600	596	750	1287	760	1144
/ɔ/		0	594	149	61	78	89	343	599	337
/o/			0	778	106	12	19	122	300	558
/u/				0	1186	316	171	77	161	254
/œ/					0	559	158	246	198	199
/ø/						0	1210	652	627	436
/y/							0	1493	782	669
/ɛ/								0	724	610
/e/									0	1144
/i/										0

c-

	v S.	toutes M.C. sauf v	toutes M.C. avec v	v M.C.	toutes S. sauf v	écart mi avec au S.
/a/		820	763	247	914	399 (/æ/
/ɪ/		452	422	149	413	25 (/ɣ/
/u/		354	326	77	333	103 (/ɐ/

d-

Fig. D-13 : Distances entre fonctions d'aire.

La détection des formants (ou des zones de concentration de l'énergie) et l'introduction de la couleur dans un système d'aides visuelles nous conduisent à proposer l'utilisation de la synthèse additive trichrome pour traduire en couleur les voyelles et les transitions formantiques. Nous choisissons pour l'explication de synthétiser une couleur à partir de ses trois composantes fondamentales : bleu B, rouge R et jaune J, aussi saturées que possible. Dans l'espace de ces composantes, une couleur de point représentatif C sera caractérisée par l'intersection C' de $\overrightarrow{OC}$ (figure D-14a) avec le plan b + r + j = 1, qui décrit la teinte, et la longueur OC, qui traduit la puissance totale.

Une couleur synthétisée C aura donc une teinte conditionnée par les taux de bleu, rouge et jaune et une puissance modulée par OC. Notre idée est de faire correspondre le triangle des voyelles extrêmes u a i au triangle BRJ. L'inconvénient de ce choix est le déplacement fréquentiel du triangle u a i dans le plan F1-F2 des deux premiers formants en fonction de la longueur du conduit vocal. Des mesures relatives à F3 par exemple permettraient de réduire cet inconvénient. C'est le parti pris dans [WATA - 78]. Cependant, la difficulté et le temps de détermination de F3 nous conduisent vers ce choix simple. Un tel choix des couleurs fait qu'une voix plus grave paraîtra plus froide et une voix plus aigue plus chaude.

La figure D-14b indique :

- en plein, le triangle extrême moyen pour une voix masculine,
- en pointillé, le triangle moyen pour des voix d'enfants,
- en gras, le triangle limite que nous associons au triangle BRJ.

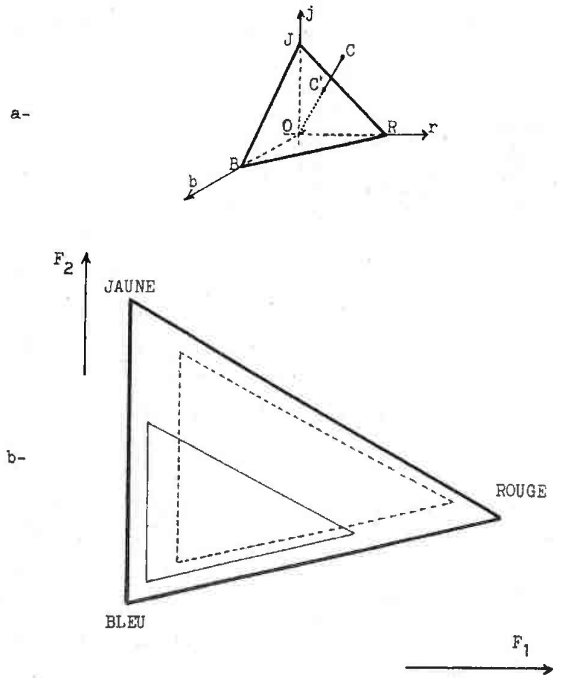


Fig. D-14 : Synthèse additive trichrome
a- Représentation d'une couleur C
b- Triangles extrêmes

A titre d'exemple, les fonctions à paramètres ajustables de F_1 et F_2 peuvent avoir l'allure suivante :

bleu = $-0.8\ F_1 - 0.3\ F_2 + 1.3$
rouge = $1.1\ F_1 - 0.2$
jaune = $-0.3\ F_1 + 0.3\ F_2 - 0.1$,

avec la contrainte que les valeurs doivent rester dans l'intervalle $[0,1]$.

La voyelle neutre a la couleur blanc tirant sur le bleu pour une voix masculine, blanc tirant sur le rouge pour un enfant. Les voyelles "*resserrées*" de l'enfant sourd apparaîtront comme peu colorées et peu contrastées entre elles.

3. VM - Vocabulaire de mots

Les exercices de cette rubrique font appel à des techniques de reconnaissance automatique de la parole. Ils constituent un des apports originaux de SIRENE et sont présentés dans le paragraphe IV suivant.

IV - APPORT DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Le sytème SIRENE fut le premier à utiliser les performances d'un module de reconnaissance automatique de la parole pour encourager la prononciation de mots ou d'éléments de phrase. La difficulté d'une telle utilisation réside dans le fait que les modules actuels fonctionnent pour le locuteur qui a fait l'apprentissage et ne s'adaptent pas facilement à des locuteurs multiples. A ce propos, nous nous bornerons à citer un travail visant à l'adaptation automatique au locuteur [PIST - 84]. Le cas des voix pathologiques posent des problèmes supplémentaires, notamment à cause des pauses trop longues qui faussent le test de segmentation parole - non parole à l'acquisition. Cet ennui a été mis à profit pour obtenir de l'élève que son élocution soit fluide plutôt que heurtée à cause d'un effort articulatoire exagéré.

Dans la suite de ce paragraphe, nous décrivons les jeux qui font appel au jugement de modules de reconnaissance sous deux aspects :

- le premier, dans le but d'encourager l'élève à s'exprimer sans s'arrêter à des détails d'articulation mais avec certaines contraintes globales,

- le second, en vue de l'émission de conseils articulatoires fins portant sur des substitutions de phonèmes en contexte, par exemple, pour lesquels il est nécessaire d'avoir une vue analytique du mot prononcé. Cette dernière étape suppose l'acquisition antérieure par l'enfant d'aptitudes au niveau de la prosodie et de la distinction entre sons élémentaires.

1. Comparaison globale

Nous utilisons l'algorithme de comparaison dynamique (2a) évoqué en C.2.III dans trois situations de jeux décrites dans la suite.

Dans un premier temps, nous nous référons à des formes-types énoncées par des enfants, garçons et filles, de 6 et 12 ans et par des adultes.

Le choix du ou des vocabulaires de référence dépend de l'exercice et de l'élève. Nous proposons ensuite de faire référence à la meilleure production intelligible de l'enfant (selon le jugement du maître), celle-ci pouvant s'affiner au cours du temps.

a) le mot prononcé commande une action -

L'action peut être, par exemple, le déplacement d'un mobile sur l'écran grâce à un vocabulaire (*"marche" "arrêt" "à droite"*...) du type du vocabulaire de commande des mouvements d'un véhicule roulant.

Cet exercice, outre l'aspect vocal, présente l'intérêt de contrôler chez l'enfant la maîtrise de son environnement spatial qui n'est pas toujours parfaitement réalisée.

Chaque commande de l'élève est comparée à un vocabulaire de référence unique où chaque mot peut avoir plusieurs représentants et constitué comme indiqué en C.2.V avec l'aide d'enfants de sexe et d'âge choisis en fonction de l'élève.

Deux cas provoquent l'échec de l'action :

- ou bien la commande fournie est de trop mauvaise qualité (rejet sur la durée, le score de comparaison à la référence ou une ambiguïté),
- ou bien elle ne correspond pas à la commande attendue.

En cas d'échecs successifs, on passe à la situation suivante, pour ne pas décourager l'élève, au bout d'un nombre d'essais prédéterminé.

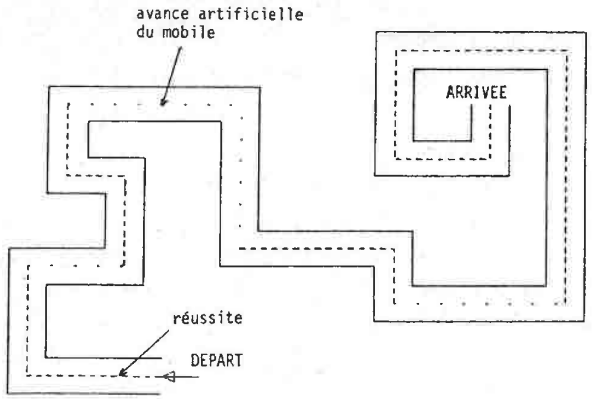
L'algorithme suivant (fig. D-15) donne l'exemple du déplacement d'un mobile dans un labyrinthe (illustré sur la figure D-16a). Le tétraèdre du schéma D-16b traduit les performances de l'élève.

```
Tant que mobile-dans-le-labyrinthe faire
    bon:=faux
    nbessais:=0
    tant que nbessais < seuil et non bon faire
       élocution (en sortie : commande)
       nbessais:=nbessais + 1
       évaluation (en entrée : commande ; en sortie : motreconnu,rejet)
       si non rejet et motreconnu = motattendu
                    alors bon:=vrai
                         déplacement
       fin si
    fin tant que
    si non bon alors déplacementartificiel
    fin si
fin tant que
```

Fig. D-15 : Exercice de commande de déplacement d'un mobile dans un labyrinthe.

FAUTES : **********

a-



b-

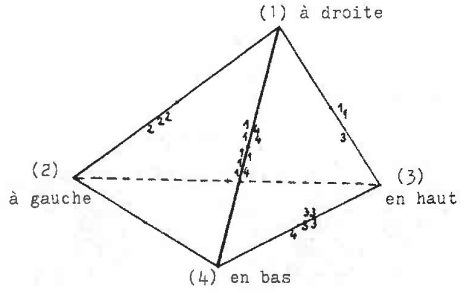


Fig. D-16 : a- Jeu du labyrinthe
b- Tétraèdre des performances (enfant O., 13 ans).

b) la qualité du mot prononcé est déterminante -

Elle provoque une action (élaboration d'un dessin, par exemple, comme sur la figure D-17d) ou le rejet du mot. Suivant le cas, on passe au mot suivant ou on demande un nouvel essai.

Ici, on s'intéresse uniquement aux taux de ressemblance de la forme prononcée à plusieurs formes de référence correspondant au mot en question. De même que dans la situation a), les références sont élaborées dans une phase d'apprentissage où l'on retient cette fois les scores croisés de comparaison des mots entre eux. De cette façon, on associe à chaque mot de la suite à prononcer un score moyen auquel on se réfère pour juger de la qualité globale de l'élocution, moyennant un facteur de sévérité modulé en fonction de l'élève. Cette phase d'ajustement des seuils de satisfaction est nécessairement non optimale dans la mesure où les scores de comparaison subissent l'influence, et dans des proportions très variables, de facteurs tels que les défauts de durée ou de rythme ou les substitutions de phonèmes (figure D-11a).

La figure D-17b donne un exemple de vocabulaire composé de dix mots prononcés par quatre locuteurs, la durée de ces différentes élocutions et le score de base associé à chacun des mots. Les scores croisés de reconnaissance entre les quatre élocutions du même mot sont donnés dans les tableaux de la figure D-17c. A titre d'exemple, précisons que pour l'enfant entendant S. (garçon, 12 ans), le facteur de tolérance doit être ajusté à 1.1 quand il fait partie des locuteurs de référence et à 1.5 quand il est remplacé par un autre garçon du même âge. Moyennant le critère retenu de jugement de la qualité de l'élocution, l'exercice ne peut avoir qu'une valeur approximative. Il permet d'encourager l'enfant à s'exprimer, de le corriger sur des défauts tels que la durée et le rythme mais peut donner un jugement favorable dans des cas où l'exercice serait détourné de son but (mot prononcé autre que le mot attendu).

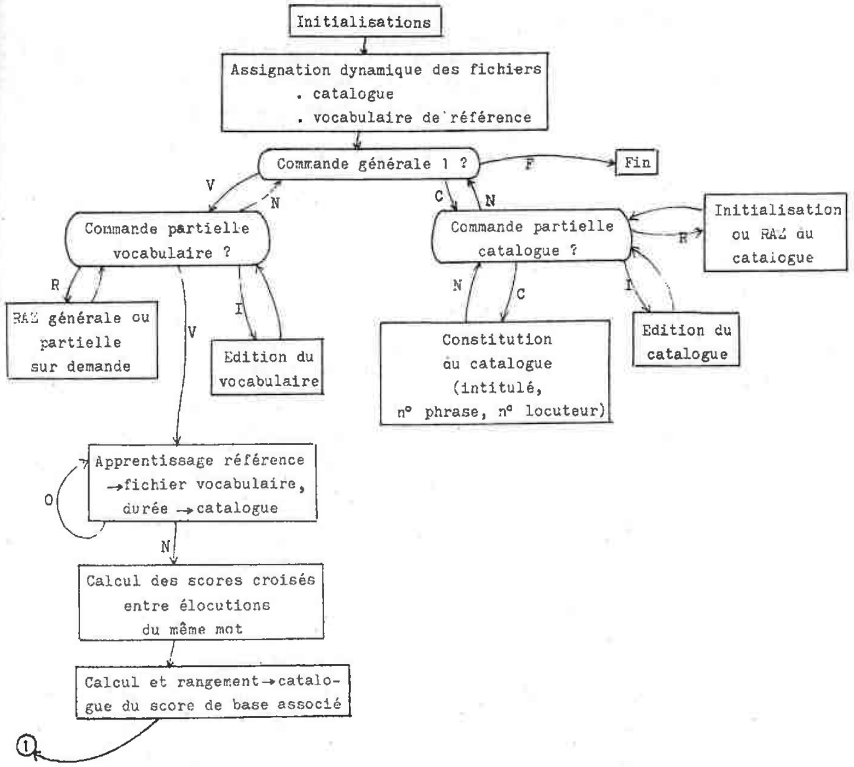


Fig. D-17a : Constitution du catalogue et du vocabulaire de référence.

	LOCUTEUR NO :	1				5				10		
1	BONJOUR	26	25	25	26	0	0	0	0	0	0	38
2	AU REVOIR	27	26	31	32	0	0	0	0	0	0	133
3	CA VA ?	22	21	19	20	0	0	0	0	0	0	37
4	TRES BIEN	21	25	21	27	0	0	0	0	0	0	138
5	MERCI	22	19	21	18	0	0	0	0	0	0	66
6	ET VOUS?	20	20	20	22	0	0	0	0	0	0	38
7	EN VACANCES	28	28	28	29	0	0	0	0	0	0	49
8	A L'ECOLE	33	35	36	34	0	0	0	0	0	0	31
9	J'AI FAIM	28	26	23	24	0	0	0	0	0	0	108
10	J'AI SOMMEIL	36	38	34	32	0	0	0	0	0	0	63

Fig. D-17b : Le vocabulaire, les durées des quatre élocutions (facteur de 20 ms) et les scores associés pour chacun des mots.

BONJOUR				AU REVOIR				ÇA VA ?				TRES BIEN			
0	28	28	31	0	24	61	70	0	26	25	23	0	66	30	77
	0	25	31		0	71	78		0	23	28		0	45	28
		0	24			0	23			0	27			0	83
			0				0				0				0

MERCI				ET VOUS ?				EN VACANCES				A L'ECOLE			
0	38	22	35	0	25	29	31	0	26	25	33	0	23	24	27
	0	24	27		0	24	29		0	30	30		0	26	26
		0	22			0	27			0	37			0	24
			0				0				0				0

J'AI FAIM				J'AI SOMMEIL			
0	27	46	27	0	27	37	44
	0	27	30		0	30	37
		0	22			0	25
			0				0

Fig. D-17c : Scores croisés de reconnaissance entre élocutions du même mot.

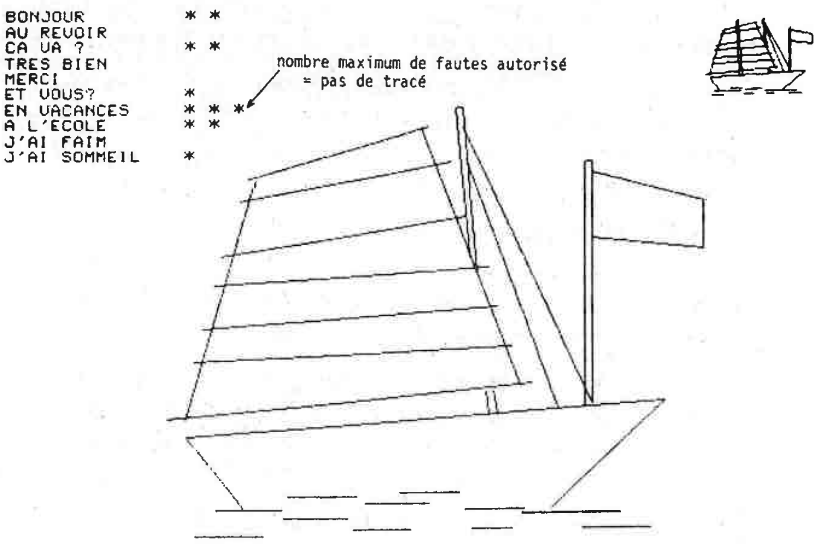


Fig. D-17d : Exercice d'apprentissage d'un vocabulaire de mots par l'élève.

c) la proximité du mot prononcé à une référence ou une autre est déterminante et permet de donner des conseils locaux d'élocution -

On se sert ici du cadrage temporel effectué par la programmation dynamique sur les deux formes concernées pour faire une tentative de différenciation locale entre sons sans faire appel à un algorithme d'étiquetage phonétique.

L'exercice est utilisé pour la différenciation à l'intérieur de séries de sons telles que /i/ - /e/ - /ɛ/ ou /p/ - /b/ - /m/ pour lesquelles l'image labiale est très voisine , ceci grâce au voisement et à la structure spectrale.

Suivant la proximité du mot prononcé au mot attendu ou à un autre, on adresse à l'élève un message d'encouragement ou un conseil local d'élocution en faisant éventuellement référence à des essais antérieurs.

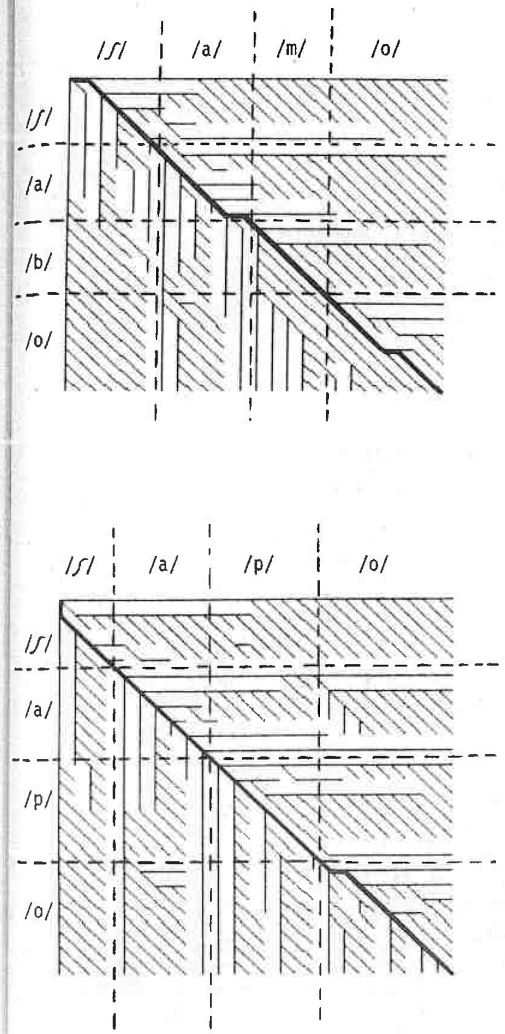
La figure D-18 donne un exemple de performance à partir du vocabulaire de base : {chapeau, chabot, chameau} ,
dans deux situations :

. l'enfant doit prononcer "*chabot*" (figure D-19),

. l'enfant doit prononcer "*chapeau*".

Cet exercice permet, par ailleurs, d'essayer de faire la différenciation voyelle nasale/non nasale, difficile à envisager sans microphone de contact appliqué sur la narine, à condition d'utiliser des mots très courts où l'influence de la voyelle est dominante dans le calcul du score.

Il trouve aussi son application dans les cas d'étirements anormaux de syllabe ou de rupture de souffle due à un effort articulatoire exagéré.



a) mot demandé : "*chabot*"

résultat de la reconnaissance :
1er "*chameau*" (score = 37)
2ème "*chabot*" (45)
3ème "*chapeau*" (64)

message :
"*J'ai compris chameau... Tu dois pouvoir prononcer chabot en te bouchant le nez...*"

b) mot demandé : "*chapeau*"

mot reconnu : "*chapeau*"

message :
"*C'est très bien... Veux-tu recommencer ?...*"

Fig. D-18: Coïncidences entre les prélèvements du mot attendu et du mot reconnu (en gras : le chemin optimal auquel est affecté un score)

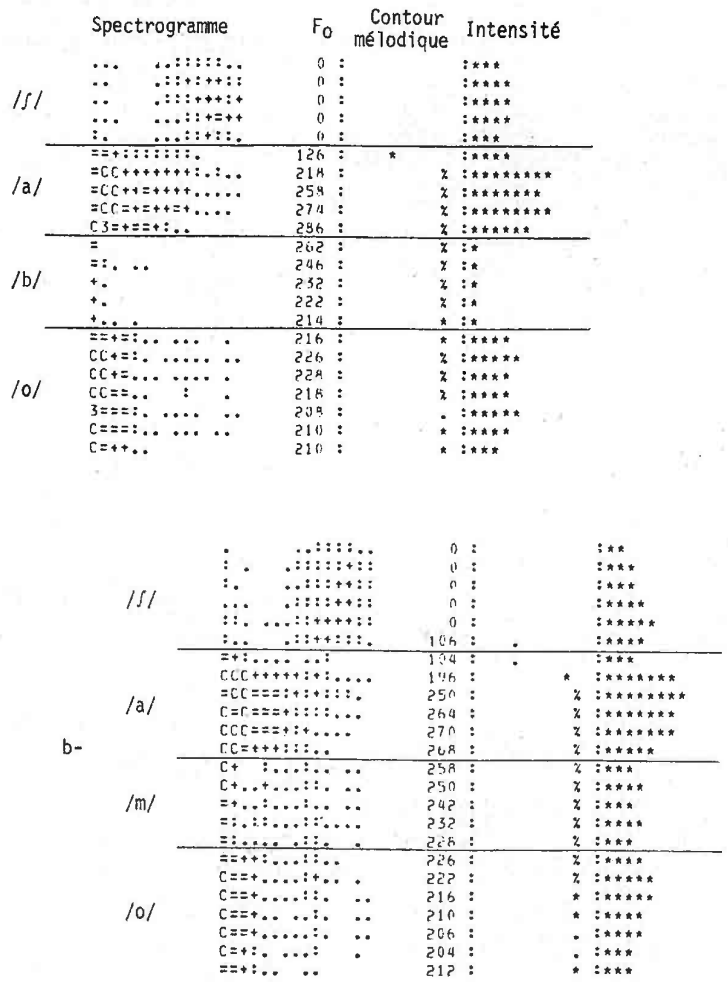


Fig. D-19 : Comparaison dynamique pour conseil d'élocution
a- mot attendu (référence)
b- mot prononcé, identifié comme /ʃ a m o/

2. Décodage analytique

a) généralités -

L'utilisation d'un système de décodage et d'étiquetage phonétiques de la parole est envisagée ici dans un but d'analyse de réponse vocale à rapprocher de l'analyse de réponse en E.I.A.O. et qui se transpose directement au cas de la dictée de mots par exemple.

Un décodage phonétique parfait répondrait aux tenants de l'approche phonématique en rééducation vocale qui se fondent sur l'idée que c'est dans les oppositions entre phonèmes que réside la pertinence linguistique. Bien qu'il ne soit plus évident [STEV - 82] que la segmentation en phonèmes soit la plus favorable à la reconnaissance analytique de la parole, c'est elle qui peut permettre d'aider l'élève au niveau du son élémentaire.

Le programme décrit ci-dessous est réalisé en simulation à partir des faits ou hypothèses suivants :

- le système de décodage phonétique a les performances du système réalisé au laboratoire par M. LAZREK [LAZR - 83], auquel on aurait adjoint une procédure d'adaptation au locuteur étudiée par ailleurs (par apprentissage ou de façon automatique),
- il n'apparaît pas dans l'étiquetage de segments parasites ne correspondant pas à un fait de forme.

Les possibilités que pourrait offrir en rééducation vocale un tel programme sont nécessairement tributaires des travaux fondamentaux en reconnaissance analytique et doivent constituer une motivation supplémentaire pour leur poursuite dans le sens de l'identification des phonèmes.

b) mise en oeuvre -

L'évaluation de la performance de l'élève se fait par comparaison d'un "*treillis d'étiquettes phonétiques*" fourni par le décodeur (et donnant les trois meilleurs choix par chaque segment) à la chaîne correspondant au mot attendu. Dans cette chaîne de référence sont indiquées deux types de particularités dites "*prévues*" : élision ou insertion possibles.

Un algorithme de comparaison dynamique permet la mise en coïncidence optimale en évitant les défauts d'hypothèses de substitution, d'insertion ou de substitution nécessairement séquentielles.

La figure D-20 illustre les résultats fournis par l'algorithme de comparaison. On retient, pour la critique de l'élocution, le chemin le moins pénalisé par les erreurs rencontrées.

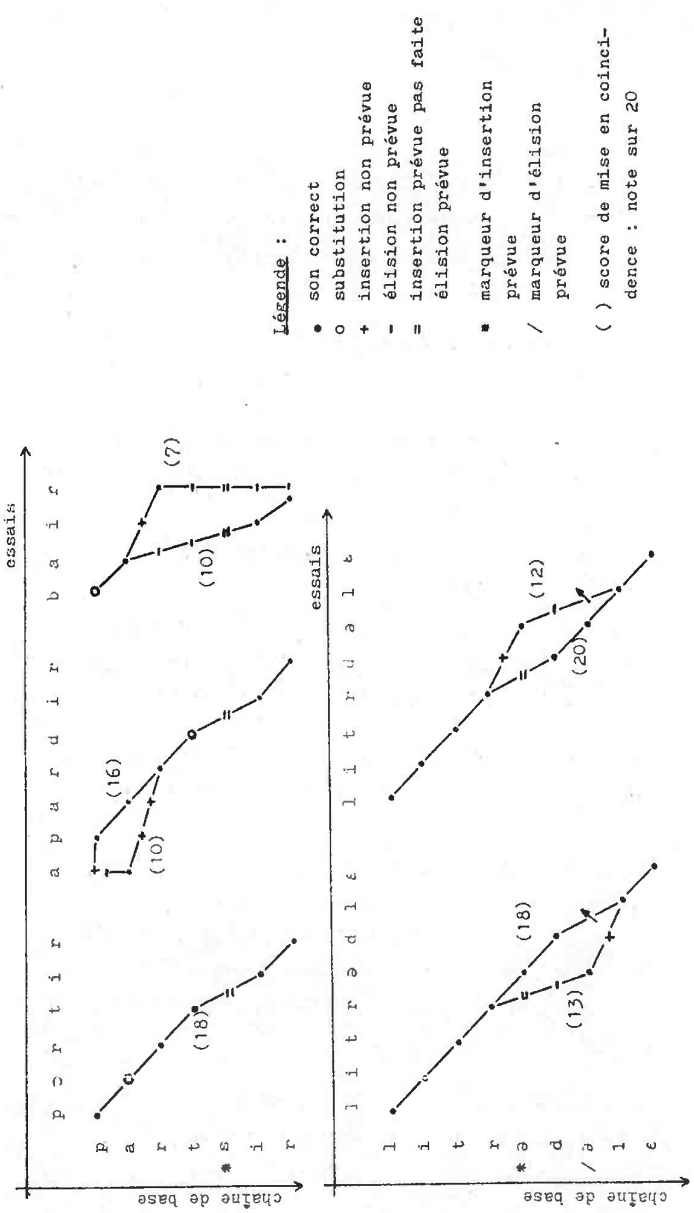


Fig. D-20 : Reconnaissance analytique : analyse de réponse.

L'évaluation du coût des chemins se fait grâce à une "*matrice de confusion*" tenant compte de deux facteurs :

- les performances du décodeur,
- l'importance donnée aux fautes d'élocution.

Pour ce faire, l'ensemble des phonèmes du français a été divisé en classes entre lesquelles on n'autorise pas de confusion : voyelles orales, sons nasaux ou nasalisés, consonnes voisées, consonnes non voisées et l'ensemble des liquides et des semi-voyelles. A l'intérieur de ces classes sont définies des sous-classes dépendant fortement du décodeur et telles que :

- entre sous-classes différentes d'une même classe, la confusion est assez sévèrement pénalisée,
- à l'intérieur d'une même sous-classe, les pénalités de confusion sont modulables, une valeur nulle correspondant à un échec du système dans la différenciation des deux sons.

La détermination du chemin optimal de coïncidence entre la chaîne de référence et le treillis issu de l'élocution de l'élève permet :

- le calcul d'une note (entre 0 et 20), à partir des coûts attribués aux insertions prévues ou non, aux élisions prévues ou non (coût fixe) et aux substitutions (coûts variables fournis par la matrice triangulaire de confusion),
- le repérage de la succession des erreurs pour l'édition de messages.

La note brute obtenue est ensuite modulée suivant un taux de sévérité prenant en compte l'âge et l'acquis de l'élève.

Les messages éventuels sont de trois types :

- indication des insertions et élisions.
  Dans le cas où les erreurs sont "*prévues*" dans la chaîne de référence, on ne donne l'indication que pour une note élevée. Par ailleurs, en cas de note faible, on insiste sur les sons bien identifiés,
- conseils articulatoires en cas de substitutions différenciables par le décodeur, en particulier :
  . opposition son voisé / non voisé,
  . opposition voyelle nasalisée / non nasalisée,
  . différenciation s /ʃ ,
  . opposition son plosif / tenu
  et, en plus, indication sur les durées et le niveau d'intensité en cas d'anomalie,
- encouragements avec référence possible aux essais antérieurs ("*attention, l'essai précédent était meilleur*", "*c'est beaucoup mieux*", ...).

En cas de difficulté répétée (mauvaise appréciation de l'acquis, difficulté pour l'élève à transposer les aptitudes acquises localement à l'énoncé plus général de mots ou de phrases), on accède à une procédure spéciale permettant d'entraîner l'enfant à des exercices élémentaires au niveau syllabique par exemple.

La structure du programme, après les opérations de calcul de la note et l'élaboration des messages, peut être schématisée ainsi :

```
Si note très bonne  alors mention de félicitation
                          nouveau mot
                    sinon si note correcte  alors mention d'encouragement
                                                  messages d'erreur
                                                  nouvel essai (note, mes-
                                                                sages)
                                            sinon messages sons corrects
                                                  procédure spéciale
                          finsi
finsi
```

La procédure spéciale a elle-même la structure suivante :

```
Tant que nbresuccès < seuil  faire
   nouvel essai (note, messages)
   si note correcte ou amélioration  alors mention d'encouragement
                                          indication des sons corrects
                                          messages d'erreur principaux
                                          nbre succès := nbre succès + 1
                                     sinon mention "essai ne compte pas"
                                          messages d'erreur principaux
                                          nbre succès := minimum
   finsi
fin tant que
```


CHAPITRE 2

EXPERIMENTATION

I - CONDITIONS D'EXPERIMENTATION

Une version préliminaire de SIRENE a été testée dans le cadre d'un mémoire de certificat de capacité orthophoniste [DUTE - 79], au cours d'un travail d'une année au laboratoire.

Nous avons pu, en dépit de certaines difficultés (d'ordre administratif ou matériel), travailler avec trois groupes d'enfants : enfants malentendants scolarisés en école primaire spécialisée, jeunes gens malentendants intégrés en CES dans une classe de 6ème "normale", enfants entendants d'origine étrangère ayant quelques difficultés spécifiques dans l'apprentissage du français. Dans ce dernier cas, il était question pour la classe d'où venaient les enfants d'acquérir le français comme langue seconde.

Pour des raisons pratiques, notre échantillon ne comportait ni déficient auditif profond, ni enfant de moins de 8 ans.

Pour les deux groupes d'enfants malentendants, nous avons procédé à une séance d'enregistrement préalable (en expression spontanée et en lecture) de façon à faire un bilan orthophonique sommaire avant d'entreprendre les séances de rééducation vocale.

Nous ne détaillerons pas le contenu des séances et les conclusions de nature purement orthophonique que l'on peut trouver dans le mémoire cité plus haut. L'ensemble des exercices présentés en D.1 a pu être expérimenté (excepté l'utilisation envisagée plus tard du décodage phonétique) dans une version qui, par la suite, a pu évoluer.

Sans avoir d'emblée construit de programmes de progression spécifique, ce qui n'était pas l'objet de notre travail, nous avons cherché à respecter quelques idées majeures :

- provoquer ou conserver chez l'enfant le désir de communiquer,
- accorder le maximum d'importance à la perception auditive,
- s'intéresser ensuite à l'intelligibilité et à la qualité du message.

Sur ce dernier point, on peut dans un but de simplification distinguer deux stratégies d'approche pour l'apprentissage de la parole :

- une approche analytique du mot, essentiellement phonématique, où l'accent est mis sur l'articulation successive des phonèmes composants et qui néglige, dans un premier temps, et la prosodie et les phonèmes de coarticulation,
- une approche synthétique qui s'oppose à la précédente.

Aucune de ces méthodes ne nous a paru raisonnable. Il est nécessaire d'appliquer une solution intermédiaire moins catégorique mais qui suppose de la part de l'éducateur une conduite de la progression par approximations successives, en quelque sorte, pour atteindre une meilleure intelligibilité du message. Cela suppose aussi, au niveau du bilan orthophonique, un repérage complet des erreurs articulatoires et suprasegmentales (ainsi que des défauts dus à une mauvaise intégration des règles phonologiques et linguistiques) et la définition d'une hiérarchie des facteurs qui agissent sur l'intelligibilité et

le naturel de la voix, ce qui est tâche difficile. Citons, à ce propos, la mise au point *"d'exercices structuraux et de microdialogues"* établis pour l'enseignement de l'intonation française [DICR - 1971].

II - REFLEXIONS DIVERSES

1. Possibilités et limites de l'aide visuelle

De plus en plus sollicités par les jeux visuels, le mouvement et la couleur, tous les enfants ont montré, en plus de l'aisance devant microphone et écrans, un désir de participation technique, un attrait pour les formes visuelles proposées et une stimulation certaine par les *"notes"* et les messages d'évaluation.

Il est pourtant important de prendre quelques précautions dans l'utilisation prolongée de l'aide visuelle. Avec le retour visuel systématique, l'enfant se trouve placé dans une situation d'apprentissage qui n'est pas "normale". Il est prudent d'éviter que l'enfant ne se focalise sur la contreréaction extrinsèque au détriment du développement des processus d'anticipation (*"feedforward"*) [LING - 1977]. Nous rejoignons l'avis de cet auteur quand il propose que la contreréaction visuelle soit retardée petit à petit, jusqu'à un degré où l'évaluation automatique ne se fait plus qu'en fin d'essai.

Enfin, si nous avons pu enregistrer des progrès ponctuels chez les enfants après quelques séances de travail, il est sûr que nous n'avons pu apprécier leur incidence à plus long terme. La durabilité du résultat ne peut être garantie que par l'acquis dont nous avons parlé plus haut. D'après les expériences relatées dans la littérature, un effet favorable nécessite pour devenir durable une longue période

d'entraînement. Les progrès sont de toute façon supérieurs à ceux que l'on obtient avec des méthodes traditionnelles, moyennant un travail équivalent de la part du spécialiste.

Ce travail du spécialiste doit permettre que ce que dit l'enfant ne soit pas négligé au profit de la façon dont il le dit. La technologie la plus sophistiquée ne peut remplacer le savoir-faire et la psychologie du logopède pour qui le sujet forme un tout dont il faut tenir compte.

Dans ce même ordre d'idées, la logique de progression suivie dans l'apprentissage est de première importance. Dans une première phase, il est utile d'identifier les sons que l'enfant est capable de produire en s'imitant lui-même ou en imitant le maître et de déterminer l'ordre dans lequel les exercices doivent être proposés. Pour exiger une performance de l'enfant, il faut être certain que les aptitudes requises ont été correctement établies. Cependant, le fait de pouvoir associer plusieurs paramètres dans un même exercice peut lui donner une certaine souplesse. Par ailleurs, le fait de pouvoir multiplier les formes-témoins correspondant à des situations vécues ou facilement assimilables par l'enfant permet d'éviter la monotonie et de contraster les formes kinesthésiques associées. Enfin, la possibilité d'auto-apprentissage, motivante et favorisant la multiplicité des essais, n'est valable qu'après mise en place de l'exercice par le maître qui doit s'assurer que l'analyse automatique ne laisse pas passer des erreurs que lui saurait repérer, et aussi que l'élève ne détourne pas l'exercice de son objet. Ceci est souligné également dans un travail d'enseignement de l'intonation par aide visuelle que nous avons eu l'occasion d'encadrer [SALL - 80].

2. Réflexions critiques d'ordre psychopédagogique [DUTE - 79] dans l'utilisation de l'aide visuelle

a) matériel didactique

Le matériel doit répondre à différents critères d'évaluation :

i - un critère de dimension sémantique concernant le rapport existant entre le fait à visualiser et sa symbolisation. A ce propos, nous avons distingué :

- les symboles concernant directement la parole : contour, figure, etc. Il est impératif de leur donner une cohérence sémantique de façon à favoriser le lien entre le code visuel et les schémas moteurs. La relation image-élément doit absolument être biunivoque,
- les symboles relatifs à des notions supplémentaires, placées comme contexte du tracé (imagerie, décor) ou jouant un rôle dans l'exercice (tunnel, labyrinthe ...). Les images doivent être suffisamment explicites pour l'élève et leur affichage doit être absolument indépendant de toute production vocale,
- les symboles traduisant la réussite ou l'échec, les messages de conseils et d'évaluation. Pour éviter toute confusion, on peut leur réserver une zone de l'écran et utiliser largement la couleur et le graphisme (la référence "trop faible / trop fort" étant évoquée par exemple systématiquement par un médaillon dans un angle de l'écran),

ii - un critère de dimension syntaxique relatif au rapport des images visuelles entre elles. Le symbole obtenu doit résulter de la mise en place d'une relation entre différents symbolismes cohérents. Par exemple, le tracé évoluant dans le temps traduisant la corrélation entre intensité et hauteur doit résulter de l'imbrication des symboles des deux paramètres pris isolément (forme prégnante au sens de la *Gestalttheorie*),

iii - un critère de dimension pragmatique : c'est la conception même du système (informatique, interactif, modulaire...) qui peut garantir la condition de souplesse et de fiabilité d'utilisation définie en collaboration avec les utilisateurs futurs.

b) modifications dans la situation d'apprentissage

Un certain nombre de facteurs apparaissent comme favorables à la progression des enfants :

- l'attrait de la nouveauté qui risque de s'amoindrir avec la multiplication des jeux électroniques et le développement de l'enseignement assisté par ordinateur et qu'il faudra tenter de maintenir,

- le contexte de jeu dans lequel l'enfant est placé. Il ne faut cependant pas perdre de vue l'objectif fixé. L'élève ne doit pas pouvoir contourner la difficulté de l'exercice par un artifice quelconque ce qui est du ressort du traitement du signal, ni être tenté d'en détourner le but ce qui relève de la pédagogie,

- la répétition des exercices individuellement mais aussi l'émulation collective qu'il ne faudrait pas faire disparaître dans le groupe,

- l'encouragement suscité par des prises de conscience simples telles que celle de l'action des productions vocales sur l'environnement. La dépendance totale du visuel cependant doit être absolument évitée,

- la participation active de l'enfant au processus pédagogique (choix du contexte de l'exercice...) et la présentation de messages individualisés. La machine ne doit pas être "maternelle" mais l'enfant doit sentir qu'il est directement concerné et actif. La référence aux essais antérieurs est en ce sens intéressante ; elle nécessite une infrastructure supplémentaire (mémoire auxiliaire individuelle, gestion de base de données sonores) qui ne pose aucun problème.

c) incidence sur l'organisation de la matière à enseigner

Nous ne développerons pas cet aspect qui relève du travail de l'orthophoniste pour lequel chaque cas est un cas particulier. Nous signalerons seulement quelques points généraux.

L'accent mis sur la rééducation des paramètres prosodiques [CARTO - 76], [DUTE - 79], [SALL - 80] correspond à l'idée actuelle que la prosodie est le cadre essentiel dans lequel les traits segmentaux sont organisés dans le temps. Le matériel proposé permet un enseignement en parallèle des différents composants de la parole dans un programme personnalisé. La constitution d'une *"base de données-apprenant"* doit aider le maitre dans sa tâche d'orientation à donner à la rééducation. Cette base doit comprendre :

- le graphe des exercices antérieurs avec indication des performances,
- des données vocales obtenues en cours d'exercice correspondant à la meilleure performance et aux plus récentes,
- une zone de données choisies pour une étude en différé dans un double but : recherche des caractéristiques à rééduquer, estimation subjective de l'évolution de l'intelligibilité.

L'extension des exercices ponctuels au cas de la parole continue se fait par l'intermédiaire des exercices sur les mots et sur les contours intonatifs pertinents. A ce propos, il nous paraît que les possibilités de mémorisation d'images numérisées et les techniques de gestion d'écran (dans le cadre d'un éditeur d'images) ont un rôle important à jouer dans la définition d'un matériel plus complet, en regard de l'important matériel didactique imagé, utilisé et souvent fabriqué par l'orthophoniste. Un travail très sérieux est à faire dans ce domaine, dans une collaboration entre informaticiens et enseignants spécialisés.

Les essais effectués dans notre phase d'expérimentation et le rôle potentiel de l'image dans un système plus vaste nous ont conduit à envisager le cas du très jeune enfant [DUTE - 81] qui se pose de plus en plus souvent grâce au diagnostic précoce de la surdité. On sait que l'intonation, liée au côté affectif de la parole, est la première acquisition véritablement linguistique sur laquelle se greffent les autres systèmes. L'enfant entendant normalement saisit les schémas intonatifs très tôt, entre 6 et 10 mois. La perception s'affine ensuite pour arriver au niveau segmental. Chez le petit sourd dont on attend qu'il devienne émetteur, le passage du babil, support de l'expression émotive, à la voix, support d'un message, est extrêmement difficile. Nous avons alors recherché le moyen de valoriser ses premières productions vocales et de fonder sur elles une progression dans l'éducation de la parole. Nous proposons pour cela des séries d'exercices faisant largement appel à une imagerie symbolique s'intéressant dans l'ordre aux points suivants :

- la prise de conscience de l'émission vocale : opposition son/silence,
- la tenue de l'émission sonore et, par opposition, les sons brefs comme les bruits ou les syllabes,
- les structures rythmiques suivant un modèle proposé : suite de sons du type "*pa*" ou petite phrase associée à une image,
- la hauteur de la voix, d'abord sans contrainte puis, si nécessaire, en respectant les limites imposées par la condition d'intelligibilité et enfin la sensibilisation au rôle informatif de l'intonation,
- les différences d'intensité et les utilisations pertinentes de ces différences,
- l'articulation des voyelles en dernier lieu.

CHAPITRE 3

DÉVELOPPEMENT

I - DEFINITION D'UNE VERSION DE SIRENE SUR MICROPROCESSEUR

1. Introduction

La maquette actuelle de SIRENE, écrite en Fortran IV sur mini-ordinateur Mitra 125 en mode 15, a permis de montrer l'utilité du système pour les orthophonistes et les professeurs de jeunes sourds et d'effectuer déjà de nombreux tests avec des enfants.

Néanmoins, cette maquette présente plusieurs inconvénients dus essentiellement à l'obsolescence de la technologie disponible au moment du lancement du projet :

- faible espace mémoire centrale disponible, limitant les performances et les développements ultérieurs,
- taille de l'ensemble de l'installation ne permettant pas le transport de systèmes SIRENE sur les lieux d'utilisation (écoles, institutions, voire domiciles),
- complexité du frontal acoustique permettant l'acquisition et l'analyse de la parole (essentiellement "Vocoder" analogique et détecteur de fondamental),
- coût prohibitif du système.

L'évolution technologique, et en particulier l'apparition de microprocesseurs et de circuits VLSI spécialisés de bonnes performances, nous ont amené à repenser la conception de SIRENE de façon à parvenir à une version :

. portable,
. fiable,
. de faible coût.

Micro-SIRENE apparaît alors comme un système susceptible d'être acquis par les écoles, voire par des parents de jeunes sourds, de façon à augmenter le temps consacré par chaque enfant à l'entraînement vocal.

La conception de Micro-SIRENE comporte trois phases :

- définition de l'architecture générale et choix du microprocesseur,
- réalisation d'un système d'analyse du signal vocal à l'aide de circuits VLSI programmables,
- transport et adaptation du logiciel.

Nous décrivons, ci-dessous, les caractéristiques principales de ces différents points.

2. Schéma général du système

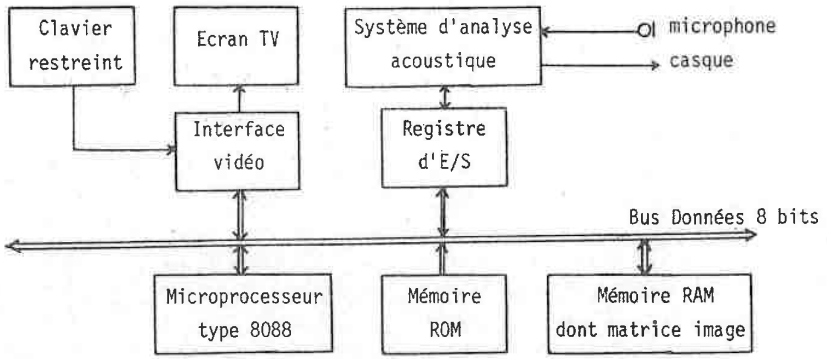
Le fonctionnement du système est contrôlé par un microprocesseur qui assure par ailleurs certaines fonctions d'analyse. Un microprocesseur 8 bits INTEL 8085 avait été initialement prévu, en particulier pour l'ensemble des fonctions concernant les paramètres prosodiques. Une puissance de calcul un peu plus grande apparaît cependant nécessaire, c'est pourquoi un "faux" microprocesseur 16 bits (puissance de calcul sur 16 bits. bus de données 8 bits) tel que INTEL 8088 constitue le compromis idéal.

Le système comprend donc :

- un processeur central INTEL 8088 (un coprocesseur arithmétique 8087 peut être adjoint pour des analyses nécessitant beaucoup de calculs),

- une mémoire centrale :
  . ROM (programmes de jeux) 40 Ko,
  . RAM (données acoustiques) 20 Ko,
- une interface vidéo couleur pour attaquer un moniteur de télévision ordinaire (prise Péritel),
- une interface série pour un clavier de commande,
- une interface rapide pour la connexion avec le système d'analyse acoustique qui réalise les fonctions :
  . d'acquisition du signal vocal,
  . de filtrage,
  . d'analyse spectrale,
  . de détection du fondamental de la voix.

Le schéma ci-dessous représente l'ensemble de la configuration.



Configuration de micro-SIRENE
Figure D-21

Une telle configuration peut être obtenue sur un micro-ordinateur du commerce type IBM PC, SIRIUS (Victor), etc. avec, en plus, une unité de disquettes pour le développement de nouveaux logiciels. Le système réalisé par IBM [DEST - 84] pour la rééducation des paramètres prosodiques et expérimenté à l'INJS de Paris et à l'étranger donne un exemple de matériel réalisé sur IBM PC.

3. Carte d'analyse acoustique du signal vocal

Le but de cette réalisation était d'obtenir un périphérique d'entrée de parole, facilement adaptable à différents ordinateurs et prenant en charge un premier traitement du signal, en vue d'applications de reconnaissance et d'analyse de parole.

Il a été ainsi réalisé dans notre équipe de recherche [SANC - 83] une carte d'acquisition et de traitement, équipée de plusieurs processeurs NECμPC 7720. Cette carte a été réalisée au standard Intel Multibus, suffisamment général pour que son adaptation à d'autres standards soit possible. D'autres réalisations sont d'ailleurs à l'étude dans notre équipe (standard Versabus pour Motorola Exormacs, standard IBM PC, ...).

Les logiciels de traitement du signal, de segmentation et de préétiquetage phonétique font actuellement l'objet d'une étude, dans le cadre d'une thèse de docteur-ingénieur.

II - TRANSPORT DU LOGICIEL SIRENE ET COMPLEMENTS

Compte-tenu des disponibilités sur le marché, il semble judicieux de transposer le logiciel de SIRENE, actuellement écrit en Fortran IV, dans le langage BASIC. L'apprentissage répandu de ce langage devrait rendre possible le développement par certains utilisateurs de nouveaux jeux adaptés à leurs besoins.

Il est envisageable d'adjoindre au système de base les compléments suivants permettant :

- d'automatiser les tests de perception grâce à la génération de sons de nature diverse : purs, à bande étroite ou large bande, etc. et la restitution de parole numérisée,

- d'implanter des tests de performance systématiques,

- d'adjoindre des modules d'apprentissage de la parole signée à l'usage des maîtres et des familles,

- d'implanter des outils pédagogiques tels que le tableau "*Bliss*", le graphique "*Logo*" ou d'autres systèmes picto-graphiques.

Dans tous les cas, le transport de la version actuelle en Fortran sur un micro-ordinateur est une tâche qui ne pose pas de problème particulier. Nous engageons actuellement une phase de prédéveloppement industriel dans le cadre d'une aide ANVAR (réalisation de cinq prototypes de SIRENE sur micro-ordinateur grand public à des fins d'évaluation sur des sites de rééducation vocale d'enfants sourds ou d'enseignement des langues).

III - EXTENSIONS DU SYSTEME

1. Domaines autres que celui de la surdité

Les différentes caractéristiques du système SIRENE présentées plus haut font qu'il est permis d'envisager son application à d'autres domaines que l'apprentissage de la parole par les déficients auditifs tels que :

a) l'apprentissage du français comme langue étrangère : nous avons eu l'occasion d'observer la prise de conscience par des enfants asiatiques, grâce au retour visuel, de certaines oppositions phonologiques que leur oreille n'était pas à même d'apprécier et le succès après auto-apprentissage dans l'élocution différenciée de /s/ et /ʃ/ par exemple. Un travail réalisé au CNET [LEBR - 81] propose l'évaluation automatique des performances dans la correction du /h aspiré/ et des diphtongues chez des sujets anglophones. Citons également le système *MicroLEA* appliquant la reconnaissance de la parole à la correction phonétique [JANO - 82]. De tels systèmes peuvent, avec avantage, être associés à un ensemble où l'image, donc l'aspect visuel, aurait sa place,

b) l'apprentissage des langues étrangères sans modification du système pour ce qui concerne la prosodie ou les mots, et moyennant une adaptation aux particularités articulatoires pour la mise en évidence des défauts de prononciation. Les contacts établis avec des centres d'enseignement des langues étrangères confirment l'intérêt des pédagogues pour un outil de ce type,

c) la rééducation de certains troubles de l'articulation, surtout au niveau des diverses formes de sigmatisme, ce qui rejoint l'exemple cité en a), ou encore la correction des dysphonies, des tendances exagérées à l'assourdissement ou à la nasalisation,

d) l'introduction des entrées-sorties vocales dans un projet d'Enseignement Assisté par Ordinateur : c'est ce que nous précisons au paragraphe suivant.

2. Enseignement Assisté par Ordinateur

Notre travail trouve un prolongement dans un projet établi au CRIN [BENN - 84] dans lequel notre équipe s'intéresse à l'introduction du son et de l'image en Enseignement Assisté par Ordinateur grâce à des éditeurs spécialisés. Nous pouvons rappeler les points suivants :

a) réalisation d'un ensemble de logiciels de traitement et de gestion de données sonores, participation à la réalisation d'une carte d'entrées-sorties vocales,

b) mise au point du système SIRENE qui est en quelque sorte un système d'E.A.O. spécialisé grâce aux caractéristiques suivantes :

- . dans la relation didacticiel-enseignant : manipulation d'une configuration standard par un utilisateur non averti, modification et extension possibles du didacticiel de base, incidence sur l'organisation de la matière à enseigner des possibilités de suivi de bilan, etc.,
- . dans la relation didacticiel-apprenant : possibilité d'auto-apprentissage, évaluation automatique, constitution et mémorisation d'un bilan individualisé accompagné d'une base de formes vocales servant de référence d'une séance à l'autre...

L'introduction du son en E.A.O. peut couvrir un vaste champ d'application. Pour ce qui nous concerne directement, nous mentionnerons :

- au niveau des entrées-sorties vocales : l'envoi de messages et d'encouragements (parole, musique numérique...) ou le retour auditif des performances vocales de l'apprenant,

- au niveau des exercices : la reproduction d'une structure rythmique, la tenue de note, les exercices mélodiques, la lecture de mots. Ajoutons que l'algorithme de comparaison de chaînes discuté en D.1.IV.2., dans le but de fournir des conseils articulatoires, peut se transposer directement à la dictée écrite de mots. Il est, dans ce dernier cas, astucieux de conserver, dans le lexique des mots ou phrases de référence, leur transcription phonétique mise en parallèle de façon à pouvoir émettre une double appréciation : jugement de la performance, d'une part, explication de particularités orthographiques, d'autre part.

CONCLUSION

L'exposé de notre travail sur l'analyse des voix et l'éducation de la parole se termine ici. Ce travail s'est développé principalement grâce aux interactions constantes avec les travaux de recherche de notre équipe au CRIN et aux contacts bibliographiques ou directs avec nombre de personnes concernées par ces domaines, ce qui nous a permis d'en acquérir une bonne vue d'ensemble et de proposer des solutions sur deux plans.

Le premier concerne l'analyse des voix dans le triple objectif de l'aide au diagnostic médical, de l'aide à l'orientation à donner à la rééducation vocale, le cas échéant, et de l'appréciation des performances et des progrès. Nous pensons par exemple que la recherche en phoniatrie peut trouver un point d'ancrage dans le plan d'analyse que nous proposons, la nécessité d'outils d'appréciation objective et de caractérisation de la voix se faisant particulièrement sentir dans ce domaine.

Le prolongement de ce premier travail fut la réalisation du système SIRENE d'aide à l'éducation vocale, sur mini-ordinateur. Nous voulons insister à nouveau sur le fait qu'il a été conçu comme une assistance au rééducateur, orthophoniste ou enseignant spécialisé, et qu'il ne prétend en rien se substituer à lui. Nous avons d'ailleurs voulu que la phase d'expérimentation faite au laboratoire, dans le cadre d'un mémoire en orthophonie, garantisse la primauté de la pédagogie sur la technique.

Il fait partie de nos projets dans un avenir relativement proche, conjointement avec des entreprises, de proposer des prototypes de SIRENE sur micro-ordinateur et d'en faire une évaluation sur des sites spécialisés dans la rééducation orthophonique ou l'apprentissage des langues étrangères.

Il nous semble important de faire remarquer à nouveau que l'aide extrinsèque apportée par un système tel que SIRENE place le sujet, principalement dans le cas des enfants, dans une situation d'apprentissage et non de communication réelle et que l'intelligence de son utilisation et le choix des exercices doivent tendre vers la transposition des efforts et des progrès à la relation directe avec autrui.

Nous terminerons en disant que la mise au point de matériels de ce type donne une valeur et une motivation supplémentaires à la recherche fondamentale dans les domaines intéressant la parole sous tous ses aspects. Nous pensons en particulier aux travaux sur le décodage phonétique, dont nous avons étudié une application de façon simulée seulement, et à l'aide apportée sur ce plan par la construction de systèmes à base de connaissances, comme SYSTEXP développé au laboratoire.

L'achèvement d'un travail de plusieurs années peut être à la fois un aboutissement et un point de départ ; notre souhait est que le nôtre puisse jouer ce double rôle.

BIBLIOGRAPHIE

[AIMA - 74] AIMARD P., "L'enfant et son langage", SIMEP Editions, Villeurbanne, 1974.

[ASKE - 78] ASKENFELT A. et al., "Electroglottograph and contact microphone for measuring vocal pitch", STL Quaterly Progress and Status Report, Stockholm, Sweden, January 1978.

[ASKE - 80] ASKENFELT A. and SJÖLIN A., "Voice analysis in depressed patients : Rate of change of fundamental frequency related to mental state", STL Quaterly Progress and Status Report, Stockholm, Sweden, October 1980.

[ATAL - 71] ATAL B.S. and HANAUER S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", JASA, Vol. 50, N° 2, P. 2, pp. 637- 655, August 1971.

[BART - 75] BARTH S., "Application des procédés de reconnaissance automatique de la parole à l'aide aux déficients auditifs profonds", Thèse E.N.S.P., 1975.

[BART - 78] BARTH S. et ADDA G., "Les aides visuelles à l'apprentissage de la parole", Rapport d'activité I.N.J.S., Chambéry, 1976-1978.

[BART - 81] BARTH S. et al., "Vers un enseignement de la parole assisté par ordinateur", Medical Informatics Europe, Toulouse, 1981.

[BAST - 77] BASTET L. et al., "Détection de mélodie par un modèle fonctionnel du système auditif périphérique", 7èmes J.E.P., GALF, Aix-en-Provence, 1977.

[BATE - 82] BATE E.M. et al., "A speech training aid for the deaf with display of voicing, frication and silence", Proc. IEEE ICASSP, Paris, Mai 1982.

[BAUD - 80] BAUDRY M. et al., "Un système microprocesseur d'aide à l'éducation de la parole", Congrès AFCET Informatique, Nancy, Novembre 1980.

[BEKE - 59] BEKESY G.v., "Similarities between hearing and skin sensations", Psy. Rev., 55, pp. 1-22, 1959.

[BEKE - 60] BEKESY G.v., "Experiments in Hearing", New-York, Mc Graw Hill, 1960.

[BENN - 84] BENNANI M., HATON M.C. et PIERREL J.M., "Parole, graphique et E.A.O. Bilan d'une expérience et perspectives", Actes du 1er colloque Sc. Francophone sur l'E.A.O., Lyon, Septembre 1984.

[BOE - 71] BOE L.J. et RAKOTOFIRINGA H., "Exigences, réalisation et limite d'un appareillage destiné à l'étude de l'intensité et de la hauteur de la voix", Revue d'Acoustique, N° 14, pp. 104-113, 1971.

[BOOT - 72] BOOTHROYD A., "Sensory aids research project. Clarke School for the Deaf", in G. Fant (Ed.) "Speech Communication ability and profound deafness", A.G. Bell Assoc. for the Deaf, 1972.

[BOOT - 73] BOOTHROYD A., "Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display", IEEE Trans. Audio Electroacoust., Vol. Au-21, N° 3, pp. 274-278, June 1973.

[BOOT - 76] BOOTHROYD A. and DAMASHEK M., "The development of small speech training aids for the deaf", Rapport S.A.R.P., The Clarke School for the Deaf, Northampton, Mass., USA, February 1976.

[BORE - 60] BOREL-MAISONNY S., "Langage parlé et écrit", Atlas des gestes de la méthode de lecture", Tome 1, Ed. Delachaux Niestlé, 1960.

[BORR - 68] BORRILD K., "Experience with the design and use of technical aids for the training of deaf and hard of hearing children", Amer. Ann. Deaf, 113, pp. 168-177, 1968.

[BOST - 73] BOSTON D.W., "Synthetic facial communication", Brit. J. Audiology, 7, pp. 95-101, 1973.

[BOUR - 83] BOURIGAULT G., "Communication gestuelle et pédagogique spécialisée", Revue Gén. Enseignement des Déficients Auditifs, N° 2, 1983.

[BROO - 83] BROOKS P.L. and FROST B.J., "Evaluation of a tactile vocoder for word recognition", J.A.S.A., Vol. 74, N° 1, July 1983.

[BRUN - 66] BRUNER J.S., "Toward a Theory of instruction", Cambridge, Mass., Belknap Press, 1966.

[CAEL - 77] CAELEN J. et CAZENAVE P., "Mesure du fondamental par filtrage variable", 7èmes J.E.P., GALF, Aix-en-Provence, 1977.

[CARA - 76] CARAYANNIS G. and GUEGUEN C. "The Factorial linear Modeling. A Karhunen-Loeve Approach to Speech Analysis", IEEE ICASSP, Philadelphia, USA, April 1976.

[CARL - 82] CARLSON R. et al., "BLISS Communication with speech on text output", Proc. IEEE ICASSP, pp. 747-750, Paris, 1982.

[CARR - 80] CARRE R., "Modèles auditifs et modèles de production de parole", 10th International Congress on Acoustics, Sydney, 1980.

[CARTA - 61] CARTAN H., "Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes", Hermann ed., Paris, 1961.

[CARTO - 76] CARTON P., "Les structures intonatives du français : leurs réalisations et leur acquisition par des enfants déficients auditifs", Thèse Professorat Jeunes Sourds, Paris, 1976.

[CHAN - 74] CHANDRA S. and LIN W.C., "Experimental Comparison between stationary and non-stationary formulations of linear prediction applied to voiced speech analysis", IEEE Trans. ASSP-22, pp. 403-415, 1974.

[CHEV - 73] CHEVRIE-MULLER C. et DECANTE P., "Etude de la fréquence fondamentale en pathologie", Bulletin d'Audiophonologie, Vol. 3, N° 2, pp. 147-194, 1973.

[CHEV - 77] CHEVRIE-MULLER C. et DECANTE P., "Programme pour le traitement automatique des données obtenues par extraction du fondamental de la parole. (Application à la pathologie de la prosodie en psychiatrie et en neurologie)". 8èmes J.E.P., GALF, Aix-en-Provence, Mai 1977.

[CHOL - 82] CHOLLET G. et GAGNOULET C., "On the evaluation of speech recognizers and data bases using a reference system", Proc. IEEE ICASSP, pp. 2026-2029, Paris, 1982.

[CHOU - 78] CHOUARD C.H., "Entendre sans oreille", Collection "Un homme et son métier", Laffont, 1978.

[CHOU - 81] CHOUARD C.H. et FUGAIN C., "Indication et résultat de l'implant cochléaire à multi-électrodes", Bulletin de liaison de la Recherche en Informatique et Automatique, INRIA, N° 74, 1981.

[CHRI - 77] CHRISTI R., LHOTE E. et al., "Quelques problèmes posés par l'introduction de la linguistique dans l'analyse des troubles d'articulation et de parole", Phonetica, N° 34, pp. 423-445, 1977.

[COHE - 68] COHEN M.L., "The ADL Sustained Phoneme Analyzer", American Annals of the Deaf, Vol. 113, pp. 247-252, 1968

[CONT - 77] CONTINI M. et BOE L.J., "Contribution à l'étude quantitative de l'évolution de la fréquence laryngienne dans la phrase énonciative du français", Bulletin de l'Inst. de Phonétique de Grenoble, Vol. II, 1973.

[COOL - 65] COOLEY J.W. and TUKEY J.W., "An algorithm for the machine calculation of complex Fourier series", Math. Comp., 19, 1965.

[CORN - 67] CORNETT R.O., "Cued Speech", Am. Ann. Deaf, 112, pp. 3-13, 1967.

[CORN - 77] CORNETT R.O. et al., "Automatic Cued Speech", Research Conference on Speech Processing Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[COUR - 81] COURVILLE L. et al., "Utilisation d'un microprocesseur dans un système d'électropalatographie", 12èmes J.E.P., GALF, Montréal, Mai 1981.

[CRIC - 74] CRICHTON R.G. and FALLSIDE F., "The development of a deaf speech training aid using linear prediction analysis", Speech Communication Seminar, Stockholm, August 1974.

[DAME - 83] DAMESTOY J.P., "Les performances des cepstres en reconnaissance de la parole", Rapport de DEA, CRIN, Université de Nancy I, 1983.

[DAVI - 76] DAVIS B., "Computer Evaluation of Laryngeal Pathology Based on Inverse Filtering of Speech", SCRL Monograph, N° 13, 1976.

[DAVID - 57] DAVID E.E. Jr., "Signal Theory in Speech Transmission", IRE Trans. on Circuit Theory, CT-3, pp. 232-244, 1957.

[DAVIS - 82] DAVIS S.B. and MERMELSTEIN P., "Comparison of Parametric Representations for Monosyllabic Word Recognition System", IEEE ICASSP, Paris, 1982.

[DEGU - 76] DEGUCHI T. and KUROKI S., "Frequency discrimination ability of hard of hearing children", JAS Japan, 32, N° 1, pp. 26-27, 1976.

[DELL - 82] DELLER J.R. Jr., "Evaluation of Laryngeal Dysfunction Based on Features on an Accurate Estimate of the Glottal Waveform", IEEE ICASSP, Paris, 1982.

[DEMO - 77] DE MORI R. et al., "A Syntactic Procedure for the Recognition of Glottal Pulses in Continuous Speech", Pattern Recognition, 9, pp. 181-189, 1977.

[DESE - 75] DE SERPA-LEITÃO A. and GALYAS K., "Measuring nasality ? A status report", S.T.L., Stockholm, Quaterly Progress and Status Report, January 1975.

[DEST - 81] DESTOMBES F. et al., "Automatic assistance to speech training for deaf children", Euromicro, September 1981.

[DIBE - 81] DI BENEDETTO M.D., "Conception et développement de systèmes informatiques de traitement de la parole pour l'aide aux malentendants", Thèse de Docteur-Ingénieur, Université de Paris IX, Décembre 1981.

[DIBE - 82] DI BENEDETTO M.D. et al., "Phonetic Recognition to Assist Lip-reading for Deaf Children", Proc. IEEE ICASSP, pp. 739-742, Paris, Mai 1982.

[DICR - 71] DI CRISTO A., "L'enseignement de l'intonation française : exercices structuraux pour la classe et le laboratoire", le français dans le monde, N° 80, 1971.

[DIMA - 83] DI MARTINO J., HATON J.P. et HATON M.C., "Evaluation d'algorithmes en reconnaissance automatique de la parole", Proc. 11th Int. Congress on Acoustics, Paris, Juillet 1983.

[DOLA - 55] DOLANSKY L.O., "An instantaneous pitch period indicator", J.A.S.A., 27, N° 1, pp. 62-72, 1955.

[DOLA - 65] DOLANSKY L.O. et al., "Teaching intonations and inflections to the Deaf", Cooperative Research Project N° S-281, Northeastern University, Boston, 1965.

[DOUR - 74a] DOURS D. et al., "Analyse temporelle du signal de parole comparée à l'analyse fréquentielle du point de vue de la reconnaissance", 5èmes J.E.P., GALF, Orsay, Mai 1974.

[DOUR - 74b] DOURS D. et FACCA R., "Méthode de segmentation et d'analyse par traitement direct du signal vocal. Application à la classification et à la reconnaissance des voyelles et des consonnes", Thèse de Docteur-Ingénieur, Université de Toulouse, Nov. 1974.

[DRUC - 77] DRUCKER H. et al., "Microprocessor-based signal processing for the perceptually deaf", IEEE ICASSP, pp. 255-256, Hartford, USA, May 1977.

[DUTE - 79] DUTEL M.M., "SIRENE : Essai de validation d'un système d'aide visuelle pour les déficients auditifs", Mémoire de capacité d'orthophoniste, Faculté B de Médecine, Nancy, Octobre 1979.

[DUTE - 81] DUTEL M.M., HATON M.C. et FRUMHOLTZ M., "Une extension des aides visuelles au domaine de l'éducation précoce de l'enfant sourd", Colloque "La surdité du premier âge, dix ans après", Besançon, Décembre 1980, Bulletin d'Audiophonologie, Vol. 3, 1981.

[EDMO - 74] EDMONDSON W.H., "Preliminary experiments with a new vibrotactile speech training aid for the deaf", S.C.S. Seminar, Stockholm, August 1974.

[EDMO - 77] EDMONDSON W.H., "Speech and the deaf child : some ideas for discussion", Conference on Speech. Analyzing Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[ELMA - 77] EL MALLAWANI I. et ZURCHER J.F., "Détecteur numérique de mélodie", 8èmes J.E.P., GALF, Aix-en-Provence, 1977.

[ENGE - 75] ENGELMANN S. and ROSOV R., "Tactual hearing experiment with deaf and hearing subjects", J. Exceptional Children, 41, pp. 243-253, 1975.

[ERBE - 72] ERBER N.P., "Speech-Envelope Cues as an Acoustic Aid to Lip-reading for Profoundly Deaf Children", J.A.S.A., Vol. 51, N° 4, pp. 1224-1227, April 1972.

[ERBE - 74] ERBER N.P., "Auditory, visual and auditory visual recognition of consonants by children with normal and impaired hearing", JSHR 17, pp. 99-112, 1974.

[ERBE - 77] ERBER N.P., "Speech perception by profoundly deaf children", Conf. on Speech Analyzing Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[FABR - 57] FABRE P., "Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation : glottographie de haute fréquence ; premiers résultats", Bull Ac. Nat. Méd., pp. 66-69, 1957.

[FANT - 60] FANT G., "Acoustic Theory of Speech Production", Mouton and co, Ed., 'S-Gravenhage, 1960.

[FARD - 81] FARDEAU M., "Prothèse auditive par implant cochléaire", Bulletin de liaison de la Recherche en Informatique et Automatique, INRIA, N° 74, 1981.

[FLAN - 72] FLANAGAN J.L., "Speech Analysis Synthesis and Perception", 2nd ed., Springer-Verlag, New-York, 1972.

[FLET - 76] FLETCHER S.G., "Speech Proficiency, Nasality Reduction and Cleft Palate", Biocommunication Research Reports, University of Alabama in Birmingham, Vol. 1, N° 1, 1976.

[FLET - 79] FLETCHER S.G. et al., "Use of linguapalatal contact patterns to modify articulation in a deaf adult speaker", Biocommunication Research Reports, Vol. 2, N° 1, Birmingham, University of Alabama, USA, October 1979.

[FOUR - 76] FOURCIN A.J. and ABBERTON E., "The Laryngograph and the Voiscope in Speech Therapy", XVIth Int. Cong. of Logopedics and Phoniatrics, pp. 116-122, 1976.

[FUJI - 81] FUJISAKI H., "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. Acoustical Analysis and Physiological Interpretations", 4th FASE Symposium, Venise, 1981.

[FURT - 71] FURTH H.G. and YOUNISS J., "Formal operations and language : a comparison of deaf and hearing adolescents", Int. J. of Psychology, 6, 1971.

[FURT - 73] FURTH H.G., "Deafness and learning : a psychosocial approach", Belmont, Wadsworth, 1973.

[GABU - 82] GABUS J.C., "Application of technical aids to the C.P. child", Conférence invitée, IFIP-IMIA Congress, Haïfa, Israël, Nov. 1981, in "Uses of Computers in aiding the disabled", J. Raviv, ed., 1982.

[GARD - 70] GARDE E., "La Voix", P.U.F., 1970.

[GENT - 80] GENTIL M., "Labialité en français. Etude phonétique et aspects physiologiques", Thèse de doctorat de 3ème cycle, Grenoble, 1980.

[GESC - 70] GESCHEIDER G.A., "Some comparisons between touch and hearing", IEEE Trans. M.M.S., 11, pp. 28-35, 1970.

[GOFF - 67] GOFF G.D., "Differential discrimination of frequency of cutaneous mechanical vibration", J. Exp. Psych., 72, pp. 294-299, 1967.

[GOLD - 69] GOLD B. and RABINER L.R., "Parallel processing techniques for estimating pitch periods of speech in the time domain", J.A.S.A., Vol. 46, pp. 442-448, 1969.

[GOLDB - 72] GOLDBERG A.J.,"A Visual Feature Indicator for the Severely Hard of Hearing", IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, N° 1, pp. 16-23, March 1972.

[GOLDS - 76] GOLDSTEIN M.H.Jr and STARK R.E., "Modification of vocalizations of preschool deaf children by vibrotactile and visual displays", J.A.S.A., 59, pp. 1477-1481, 1976.

[GRAI - 77] GRAILLOT P. et BOE L.J., "Relation entre l'évolution de la fréquence laryngienne et de l'intensité pour la phrase énonciative du français", 8èmes J.E.P., GALF, pp. 193-199, Aix-en-Provence, 1977.

[GRAI - 81] GRAILLOT P. et EMERARD F., "Prothèse vocale à l'usage des handicapés moteurs déficients de la parole", Bulletin de liaison de la Recherche en Informatique et Automatique, INRIA, N° 74, 1981.

[GUBR - 77] GUBRYNOWICZ et al., "Evaluation de l'état pathologique des cordes vocales d'après l'analyse des variations du fondamental", 8èmes J.E.P., GALF, Aix-en-Provence, Mai 1977.

[GUEG - 72] GUEGUEN C.J. et MAISSIS A.H., "Un système d'aide aux sourds profonds", Xèmes Assises Nationales de la Prothèse Auditive, Paris, Octobre 1972.

[GUER - 77] GUERIN B. et BOE L.J., "La régulation de la vibration des cordes vocales : simulation à l'aide d'un modèle à deux masses", Actes des 8èmes J.E.P., GALF, pp. 37-41, Aix-en-Provence, 1977.

[GUER - 78] GUERIN B., "Contribution aux recherches sur la production de la parole. Etude du fonctionnement de la source vocale. Simulation d'un modèle", Thèse d'Etat, Université de Grenoble, 1978.

[GUTT - 70] GUTTMAN N. et al., "Articulatory training of the deaf using low-frequency surrogate fricatives", JSHR, 13, pp. 19-29, 1970.

[HAME - 83] HAMER R.D. et al., "Vibrotactile masking of Pacinian and non-Pacinian channels", JASA, 73, 4, April 1983.

[HANS - 68] HANSEN V.M., "Speech education with deaf children", UNESCO Seminar, Denmark, August-September 1968.

[HATO - 74a] HATON J.P., "Contribution à l'analyse, la paramétrisation et la reconnaissance de la parole", Thèse d'Etat, Univ. de Nancy I, 1974.

[HATO - 74b] HATON J.P. and HATON M.C., "Some statistical feature preselection methods and their properties", ACUSTICA, 31, N° 5, pp. 281-284, November 1974.

[HATO - 75] HATON M.C. et HATON J.P., "Essai de caractérisation des voix d'enfants sourds par analyse polynomiale de la mélodie", Actes des 6èmes J.E.P., GALF, Toulouse, Mai 1975.

[HATO - 76] HATON M.C. et HATON J.P., "Une méthode de représentation du signal vocal en base adaptative", Actes des 7èmes J.E.P., GALF, Nancy, Mai 1976.

[HATO - 77] HATON M.C. et HATON J.P., "Analyse et rééducation des paramètres prosodiques chez l'enfant sourd", Actes des 8èmes J.E.P., GALF, Aix-en-Provence, Mai 1977.

[HATO - 79] HATON M.C., "An interactive system for the acquisition, reduction and visualization of spectral speech data", Communication aux Journées sur la Classification, Paris, Mai 1979, Abstract in Biometrics.

[HATO - 81] HATON M.C. et HATON J.P., "Computer-Aided Speech Analysis and Training for the Hearing Impaired", IFIP-IMIA Congress, Conférence invitée, Haïfa, Israël, J. Raviv, ed., November 1981.

[HESS - 81] HESS W.J., "Algorithmes et méthodes pour la détermination du fondamental", 12èmes J.E.P., GALF, Montreal, Mai 1981.

[HIKI - 74] HIKI S. et OIZUMI J., "Speech Synthesis by Rule from Neurophysiological Parameter", Speech Communication Seminar, Stockholm, August 1974.

[HIKI - 76] HIKI S. and KAGAMI R., "Some properties of formant frequencies of vowels uttered by hearing-impaired children", 91st Meeting of the Ac. Soc. of Am., Washington, DC, April 1976

[HOLB - 70] HOLBROOK A. and CRAWFORD G.H., "Modification of speech behavior of the deaf (hypernasality)", Conference of Executives of American Schools for the Deaf, St Augustine, Florida, April 1970.

[HOLB - 71] HOLBROOK A., "Modification of speech behavior with preschool deaf children by means of spectrum control", AOEHI Bulletin, 1971.

[HOLM - 72] HOLM C., "La parole de l'enfant sourd du point de vue phoniatrique", Surdité du Premier Age, Colloque Int. d'Audiophonologie, Besançon, Novembre 1972.

[HOUD - 73] HOUDE R.A., "Instantaneous visual feedback in speech training for the deaf", Annual Convention of the American Speech and Hearing Association, Detroit, Michigan, USA, October 1973.

[JAKO - 63] JAKOBSON R. et al., "Preliminaries to Speech Analysis", Cambridge, Mass., MIT Press, 1963.

[JAKO - 68] JAKOBSON R., "Child Language Aphasia and Phonological Universals", The Netherlands, 1968.

[JAMA - 81] JAMART P., "Des outils pédagogiques pour une pédagogie appliquée à l'adolescent déficient auditif", Thèse pour l'obtention du CAEJDA 2ème degré, I.N.J.S. St Jacques, Paris, 1981.

[JANO - 82] JANOT-GIORGETTI M.T., "Expériences en reconnaissance de la parole. Application à l'apprentissage des langues : le système MicroLEA", Thèse d'Etat, UNiversité de Nancy 1, 1982.

[JOHA - 66] JOHANSSON B., "The use of the transposer for the management of the deaf child", J. Int. Audiology, 5, pp. 362-371, 1966.

[JOHN - 83] JOHNSON H.W. and BURRUS C.S., "The design of optimal DFT algorithms using dynamic programming", IEEE Trans. ASSP, Vol. 31, N° 2, April 1983.

[JONE - 67] JONES C., "Deaf Voice. A Description Derived from a Survey of the Literature", The Volta Review, pp. 507-540, October 1967.

[JOSP - 79] JOSPA P., "Propagation sonore dans le conduit vocal non stationnaire", Rapport d'activité de l'Inst. de Phonétique de Bruxelles, Décembre 1978 - Novembre 1979.

[KAKI - 76] KAKITA Y. et HIKI S., "A study of laryngeal control for pitch change by use of anatomical structure model", Proc. IEEE ICASSP, pp. 43-46, Philadelphia, Pa., USA, 1976.

[KALI - 72] KALIKOW D.N. and SWETS J.A., "Experiments with computer-controlled displays in second language learning", IEEE Trans. Audio Electroacoust., Vol. 20, N° 1, pp. 23-28, 1972.

[KENT - 76] KENT R.D., "Anatomical and neuromuscular maturation of the speech mechanism : evidence from acoustic studies", J.S.H.R., 19, pp. 421-447, 1976.

[KIM - 74] KIM B. and FUJISAKI H., "Measurement of mandibular control in vowels and its relevance to the articulatory description of the vowel systems of Korean and Japanese", Speech Communication Seminar, Stockholm, August 1974.

[KING - 82] KING A. et al., "A Speech Display Computer for Use in Schools for the Deaf", Proc. ICASSP 82, pp. 755-758, May 1982.

[KIRM - 73] KIRMAN J.H., "Tactile Communication of speech : a review and an analysis", Psych. Bull., Vol. 80, N° 1, pp. 54-74, 1973.

[KISU - 76] KISU S. et al., "Correction of vowel articulation by use of articulatory trainer", Tech. Group. Speech, Acous. Soc. Japan, Paper s 75-53, 1976.

[KÖST - 77] KÖSTER J.P., Conférence à l'Institut de Phonétique de Nancy, 12 janvier 1977.

[KRIN - 63] KRINGLEBOTN M., "Experiments with some visual and vibrotactile aids for the deaf", Am. Ann. Deaf, 113, pp. 311-7, 1963.

[LAMO - 75] LAMOTTE M. et VIGNERON C., "Utilisation d'un diviseur de fréquences audibles en éducation orthophonique", 6èmes J.E.P., GALF, Toulouse, Mai 1975.

[LAND - 77] LANDERCY A., "Variations à court terme de la fonction de source vocale", Actes des 8èmes J.E.P., pp. 29-33, Aix-en-Provence, 1977.

[LAZR - 83] LAZREK M., "Décodage acoustico-phonétique en compréhension automatique de la parole continue", Thèse de 3ème cycle, Université de Nancy 1, Juin 1983.

[LEBR - 81] LE BRAS J., "Utilisation de la reconnaissance automatique de la parole pour l'apprentissage des langues", Thèse de 3ème cycle, Université de Rennes, 1981.

[LENY - 79] LE NY J.F., "La sémantique psychologique", P.U.F., 1979.

[LERN - 52] LERNER R.M. and FANO R.M., "Vocoder", Quaterly Progress Report, Research Laboratory of Electronics, M.I.T., 1952.

[LETE - 83] LETELLIER Ph., "Transmission d'images à bas débit pour un système de communication téléphonique adapté aux sourds", Thèse de Docteur-Ingénieur, Université de Paris-Sud, Centre d'Orsay, Septembre 1983.

[LEVI - 72] LEVITT H., "Acoustic Analysis of Deaf Speech Using Digital Processing Techniques", IEEE Transactions on Audio and Electroacoustics, Vol. AU-20, N° 1, pp. 35-41, March 1972.

[LEVI - 73] LEVITT H., "Speech Processing Aids for the Deaf : An Overview", IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, N° 3, pp. 269-272, June 1973.

[LHOT - 72] L'HOTE E., "Le glottographe et ses applications phonétiques", Bulletin d'Audiophonologie, Vol. 2, N° 6, pp. 17-37, 1972.

[LIBE - 67] LIBERMAN A.M. et al., "Perception of the speech code", Psy. Rev., 74, 1967.

[LIND - 71] LINDBLOM B., "Phonetics and the description of language", Proceedings 7th International Congress Phon. Sci., pp. 22-28, 1971.

[LING - 64] LING D., "Implications of hearing aid amplification below 300 CPS", Volta Review, 66, pp. 723-724, 1964.

[LING - 75] LING D. and CLARKE B.R., "Cued Speech : an evaluation study", Am. Ann. Deaf, 120, pp. 480-488, 1975.

[LING - 77] LING D., "Models for speech training with non auditory feedback", Conference on Speech-Analysis Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[LORA - 75] LORAND P. et al., "Un système destiné à la rééducation des déficients auditifs profonds : le système P.A.R.M.E., 6èmes J.E.P., GALF, Toulouse, Mai 1975.

[LUND - 78] LUNDMAN M., "Technical Aids for the Speech Impaired. An international Survey on Research and Development Projects", compiled for the Swedish Institute for the Handicapped, ICTA Information Centre Stockholm, March 1978.

[MAED - 77] MAEDA S., "Sur les corrélatifs physiologiques de la fréquence du fondamental de la parole", Actes des 8èmes J.E.P., pp. 43-50, GALF, Aix-en-Provence, Mai 1977.

[MAIS - 73] MAISSIS A., "Le traitement de l'information acoustique, étape fondamentale de la reconnaissance de la parole", Thèse d'Etat, Paris VI, 1973.

[MAKH - 75] MAKHOUL J.I., "Linear prediction : A tutorial review", Proc. IEEE, 63, pp. 561-580, 1975.

[MALM - 68] MALMBERG B., "Manual of Phonetics", North-Holland, 1968.

[MARK - 72] MARKEL J.D., "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 367-377, December 1972.

[MARK - 73] MARKEL J.D. and GRAY A.H., "On autocorrelation Equations as Applied to Speech Analysis", IEEE Trans. Audio Electroacoust., AU-21, N° 2, April 1973.

[MARK - 76] MARKEL J.D. and GRAY A.H. Jr., "Linear Prediction of Speech", Springer Verlag, 1976.

[MART - 77] MARTIN P., "Un analyseur-visualiseur de mélodie à microprocesseur", 8èmes J.E.P., GALF, Aix-en-Provence, Mai 1977.

[MARTI - 66] MARTINET A., "Eléments de linguistique générale", Colin, 1966.

[MARTO - 68] MARTONY J., "On the correction of the voice pitch level for severely hard of hearing subjects", Amer. Ann. Deaf, 113, pp. 195-202, 1968.

[MARTO - 70] MARTONY J., "Visual aids for speech correction : summary of three years experiences", In G. Fant (Ed.), Speech Communication Ability and Profound Deafness, Washington, pp. 345-349, 1970.

[MARTO - 74] MARTONY J., "Some psychoacoustic tests with hearing impaired children", STL-Quaterly Progress Status Report, 2-3, pp. 72-89, 1974.

[MARTO - 77] MARTONY J., "Vowels, Pitch, Intonation, and Fricatives of Deaf Persons", Paper presented at the Research Conference on Speech Processing Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[MASS - 72] MASSARO D.W., "Preperceptual images, processing time, and perceptual units in auditory perception", Psy. Rev., 79, pp. 124-145, 1972.

[MASS - 80] MASSARO D.W. and ODEN G.C., "Evaluation and integration of acoustic features in speech perception", J.A.S.A., Vol. 67, N° 3, March 1980.

[MERM - 67] MERMELSTEIN P., "Determination of the Vocal-Tract Shape from Measured Formants Frequencies", J.A.S.A. 41, pp. 1283-1294, 1967.

[METT - 71] METTAS O., "Les techniques de la phonétique instrumentale et de l'intonation", P.U.B., Maloine ed., Paris, 1971.

[MILL - 56] MILLER G.A., "Langage et Communication", P.U.F., Paris, 1956.

[MILLE - 76] MILLER J.D. et al., "Preliminary research with a three-channel vibrotactile speech-reception aid for the deaf", in G. Fant (Ed.) Speech Communication, Vol. 4, New York : Wiley, pp. 97-103, 1976.

[MUSS-1844] MUSSET (de) A., "Pierre et Camille", 1844.

[NEIS - 67] NEISSER U., "Cognitive Psychology", Appleton-Century-Crofts, 1967.

[NICK - 73] NICKERSON R.S. and STEVENS K.N., "Teaching Speech to the Deaf : Can a Computer Help ?", IEEE Transactions on Audio and Electroacoustics, AU-21, pp. 445-455, 1973.

[NICK - 74] NICKERSON R.S. et al., "A computer-based system of speech training aids for the deaf : a progress report", BBN Report, N° 2901, September 1974.

[NICK - 75] NICKERSON R.S., "Speech training and speech reception aids for the deaf", Bolt, Beranek and Newman Inc., Report N° 2980, 1975.

[NICK - 76] NICKERSON R.S., KALIKOW D.N. and STEVENS K.N., "Computer-Aided Speech Training for the Deaf", Journal of Speech and Hearing Disorders, Vol. 41, N° 1, pp. 120-132, February 1976.

[NOLL - 64] NOLL A.M., "Short-Time" Cepstrum" Pitch Detection", J.A.S.A., 36, p. 1030, 1964.

[OPPE - 69] OPPENHEIM A.V., "A speech analysis-synthesis system based on homomorphic filtering", J.A.S.A., Vol. 45, February 1969.

[OSBE - 79] OSBERGER M.J. and LEVITT H., "The effect of timing errors on the intelligibility of deaf children's speech", J.A.S.A., 66, pp. 1316-1624, 1979.

[OSBE - 81] OSBERGER M.J. et al., "Computer-assisted Speech Training for the Hearing Impaired", J.A.R.A., XIV, pp. 145-158, Fall, 1981.

[PHIL - 68] PHILLIPS N.D. et al., "Teaching of intonation to the deaf by visual pattern matching", Amer. Ann. Deaf, 113, pp. 239-246, 1968.

[PICK - 63] PICKETT J.M., "Tactual Communication of Speech Sounds to the Deaf : Comparisons with Lip-reading", J.S.H.D., 28, pp. 315-330, 1963.

[PICK - 68] PICKETT J.M. and CONSTAM A., "A Visual Speech Trainer with Simplified Indication of Vowel Spectrum", American Annals of the Deaf, Vol. 113, pp. 253-258, 1968.

[PICK - 69] PICKETT J.M., "Some applications of speech analysis to communication aids for the deaf", IEEE Trans. Audio. Electroacoust., AU-17, pp. 283-289, 1969.

[PINS - 62] PINSON E.N., "Pitch-Synchronous Time-Domain in Estimation of Formart Frequency and Bandwidths", J.A.S.A., Vol. 35, N° 2, pp. 1264-1273, August 1962.

[PIST - 84] PISTER C., "Adaptation au locuteur par apprentissage automatique. Application à un système de reconnaissance automatique de la parole", Thèse de 3ème cycle, Université de Nancy 1, 1984.

[PLAN - 60] PLANT G.R.G., "The Plant-Mandy Voice Trainer. Some notes by the designer", Teacher of the Deaf, 58, pp. 12-15, 1960.

[PLAN - 83] PLANT G. and RISBERG A., "The transmission of fundamental frequency variations via a single channel vibrotactile aid", STL Quaterly Progress and Status Report, Stockholm, October 1983.

[POTT - 48] POTTER R.. and PETERSON G.E., "The perception of vowels and their movements", J.A.S.A., 20, pp. 528-535, 1948.

[POVE - 74] POVEL D.J., "Development of a Vowel Corrector for the Deaf", Psychol. Res. 37, pp. 51-70, 1974.

[PRON - 47] PRONOVOST W., "Visual aid to speech improvement", J.S.H.D., 12, pp. 387-391, 1947.

[PRON - 68] PRONOVOST W. et al., "The Voice Visualizer", American Annals of the Deaf, Vol. 113, pp. 230-238, 1968.

[QUIG - 64] QUIGLEY L.F. et al., "Measuring palatopharyngeal competence with the nasal anemometer", Cleft Palate J., pp. 304-313, 1964.

[RISB - 68] RISBERG A., "Visual Aids for Speech Correction", Amer. Ann. Deaf, 113, pp. 178-194, 1968.

[RISB - 75] RISBERG A., "Some comments on the development of new technical aids for the deaf", 6èmes J.E.P., GALF, Toulouse, Mai 1975.

[ROJO - 78] ROJO-TORRES M., "Le rythme pré-musical dans l'éducation de l'enfant sourd", Rev. Gén. Ens. Déficients Auditifs, N° 2, pp. 53-65, 1978.

[ROSS - 74] ROSS M.J. et al., "Average Magnitude Difference Function Pitch Extractor", IEEE Trans. ASSP-22, pp. 353-362, 1974.

[ROTH - 73] ROTHENBERG M., "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing", JASA, Vol. 53, pp. 1632-1645, 1973.

[ROTH - 77] ROTHENBERG M. et al., "Vibrotactile frequency for encoding a speech parameter", JASA, Vol. 62, pp. 1003-1012, 1977.

[ROUF - 82] ROUFS J.A.J. et al., "Some experiments on sharpness in relation to contrast bearing on electronic optical imaging", IPO Annual Progress Report, Eindhoven, Netherlands, 1982.

[RUCH - 82] RUCH E., "Carte d'E/S vocale pour la reconnaissance et la synthèse", Rapport de DESS, Université de Nancy 1, 1982.

[SAKO - 78] SAKOE H. and CHIBA S., "Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. ASSP-26, 1, pp. 43-49, 1978.

[SALL - 80] SALLES J.L., "Une expérience de rééducation de l'intonation à l'aide d'un indicateur de mélodie", Thèse du second degré de l'Enseignement de Jeunes Sourds, C.T.O.P., Fougères, 1980.

[SANC - 77] SANCHEZ C. et MESSENET G., " L'Observateur", Rapport d'activité de l'équipe RFIA du CRIN, Nancy, 1977.

[SANC - 78] SANCHEZ C. et al., "Etude et utilisation des indices acoustiques et des traits pour la segmentation et la reconnaissance phonétique de la parole", 9èmes J.E.P., GALF, Lannion, 1978.

[SANC - 83] SANCHEZ C., HANSER P. et HATON J.P., "Environnement logiciel et matériel pour l'analyse acoustico-phonétique du signal vocal", Séminaire GALF-GRECO "Analyse du signal de parole", Paris, Décembre 1983.

[SARG - 82] SARGENT D.C., "Rhythmic cues aid lip readers", IEEE Spectrum, pp. 46-49, April 1982.

[SAUN - 76] SAUNDERS F.A. et al., "Hearing substitution : a wearable electrotactile vocoder for the deaf", Meeting of the Amer. Ass. for the Advancement of Science, Boston, February 1976.

[SAUS - 76] SAUSSURE (de) F., "Cours de linguistique générale", Payot, 1976.

[SAVI - 70] SAVIN H.B. and BEVER T.C., "The nonperceptual reality of the phonemes", J. of Verbal Learning and Verbal Behavior, 9, pp. 295-302, 1970.

[SAWA - 78] SAWASHIMA M., "Dynamic Palatography", Project N° 43, Technical aids for the speech impaired, Compiled by M. Lundman, The Swedish Institute for the Handicapped, March 1978.

[SCHA - 72] SCHAFER R.W., "A survey of digital speech processing techniques", IEEE Trans. Audio Electroacoust., Vol. AU-20, N° 1, pp. 28-35, March 1972.

[SCHI - 80] SCHIRA M., "Le langage complété Cornett pour l'éducation de l'enfant sourd, Guide pour les parents", Mémoire d'orthophonie, Nancy, Octobre 1980.

[SCHO - 81] SCHOENTGEN J., "L'analyse acoustique de la voix et ses applications cliniques", Rapport d'activités de l'Institut de Phonétique de Bruxelles, Belgique, Août 1981.

[SCHU - 70] SCHULTE K., "Experimental comparison between three oscilloscope phoneme visualizing systems", In G. Fant (Ed.) Speech Communication Ability and Profound Deafness, Washington, D.C., pp. 355-358, 1970.

[SCHU - 72] SCHULTE K., "Phonemetransmitting Manual System (PMS)", In G. Fant (Ed.) Speech Communication Ability and Profound Deafness, pp. 255-260, Washington, D.C., 1972.

[SCHU - 81] SCHULTE K., "The Fonator-System. A vibroacoustic aid for speech and communication", Technical Paper, Päd Hochschule, Heidelberg, R.F.A., 1981.

[SEGU - 82] SEGUI J., "Percevoir les sons et accéder au lexique : une illustration de la recherche en psycholinguistique", Actes du Colloque "Domaines et Objectifs de la Recherche Cognitive", Pont-à-Mousson, Avril 1982.

[SERI - 74] SERIGNAT J.F., "Contribution aux recherches sur la communication parlée. Travaux sur le vocoder à autocorrélation. Etude et simulation d'un vocoder à prédiction linéaire", Thèse de Docteur-Ingénieur, Grenoble, 1974.

[SHER - 77] SHERRICK C.E. and CHOLEWIAK R.W., "Matching Speech to Vision and Touch", Research Conference on Speech Analyzing Aids for the Deaf, Gallaudet College, Washington, D.C., May 1977.

[SONE - 60] SONESSON B., "On the anatomy and vibratory pattern of the human vocal folds with special reference to a photo-electrical method for studying vibratory movements", Acta Otolaryngologica Supplements, 156, pp. 1-80, 1960.

[SPAR - 79] SPARKS D.W. et al., "Investigating the MESA (Multipoint Electrotactile Speech Aid) : the transmission of connected discourse", J.A.S.A. 65(3), pp. 810-815, March 1979.

[SPEN - 81] SPENS K.E., "Tactile speech communication aids for the deaf : a comparison", STL Quaterly Progress and Status Report, Stockholm, January 1981.

[STAR - 71] STARK R.E., "The use of real-time visual displays of speech in the training of a profoundly deaf, non-speaking child : a case report", J.S.H.D., 36, pp. 397-409, 1971.

[STEA - 77] STEARNS W.P. et al., "Quantitative measurement techniques for fitting hearing instruments using selective spectrum filtering", IEEE Int. Conf. ASSP, Hartford, pp. 248-256, May 1977.

[STEI - 46] STEINBERG J.C. and FRENCH N.R., "The portrayal of visible speech", J.A.S.A., 18, pp. 4-18, 1946.

[STEL - 76] STELT (van der) J.M., "Une comparaison du rythme de la parole entre des enfants sourds à la naissance et des enfants qui sont devenus sourds", Colloque Prélangage 3, Besançon, Novembre 1976.

[STEV - 75] STEVENS K.N. et al., "A miniature accelerometer for detecting glottal waveforms and nasalization", J.S.H.R., 18, pp. 594-599, 1975.

[STEV - 80] STEVENS K.N. et al., "Studies of speech production by children with disorders of speech production", RLE Progress Report N° 122, Page 155, January 1980.

[STEV - 82] STEVENS K.N., "Toward a feature-based model of speech perception", IPO Annual Progress Report, Eindhoven, Netherlands, 17, 1982.

[STEW - 76] STEWART L.C., LARKIN W.D., HOUDE R.A., "A Real-Time Spectrograph with Implications for Speech Training for the Deaf", Int. Conf. on Acoustics, Speech and Signal Processing, Canterbury Press, Rome, NY, 1976.

[TEST - 77] TESTON B. et ROSSI M., "Un système de détection automatique du fondamental et de l'intensité", 8èmes J.E.P., GALF, Aix-en-Provence, 1977.

[TEST - 78] TESTON B., "La détection de l'intensité dans la parole : problèmes et méthodes", Travaux de l'Institut de Phonétique d'Aix-en-Provence, Vol. 5, 1978.

[THOM - 68] THOMAS I.B., "Real-time visual display of speech parameters", Proc. N.E.C., 24, pp. 382-387, 1968.

[THOM - 70] THOMAS I.B. et al., "Articulation training through visual speech patterns", The Volta Review, 72, pp. 310-318, 1970.

[TITZ - 79] TITZE I.R. and TALKIN D.T., "A Theorical Study of the Effects of Various Laryngeal Configurations on the Acoustics of Phonation", JASA, 66, N° 1, pp. 60-74, 1979.

[TONG - 83] TONG Y.C. et al., "Psychophysical studies evaluating the feasibility of a speech processing strategy for a multiple-channel cochlear implant", J.A.S.A., Vol. 74, N° 1, July 1983.

[TRAU - 74] TRAUNMULLER H., "A visual lipreading aid", Speech Communication Seminar, Stockholm, Vol. 4, August 1974.

[TRAU - 80] TRAUNMULLER H., "A tactual speech communication aid", J. Commun. Dis., 13, pp. 183-193, 1980.

[UPTO - 68] UPTON H.W., "Wearable Eyeglass Speechreading Aid", American Annals of the Deaf, Vol. 113, pp. 222-229, 1968.

[VALL - 73] VALLENCIEN B., "Essai d'interprétation des courbes de micro-mélodie ou mélodie articulatoire", Bulletin d'Audiophonologie, Vol. 3, N° 2, pp. 41-57, 1973.

[VOIR - 74] VOIRON G., "Contribution à l'étude objective de la voix des déficients auditifs moyens et profonds avec l'utilisation d'un indicateur de mélodie", Thèse E.N.S.P., Paris, 1974.

[WAKI - 73] WAKITA H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", IEEE Trans., AU-21, pp. 417-427, October 1973.

[WATA - 75] WATANABE A. and KISU S., "Articulatory trainer for vowels by inverse filter", Tech. Group Speech, Acous. Soc. Japan, Paper s 74-51, 1975.

[WATA - 76] WATANABE A. and OKAMURA H., "Speech trainer for correction of intonation and its effect to hard of hearing children", Tech. Group Speech Acous. Soc. Japan, Paper s 75-54, 1976.

[WATA - 78] WATANABE A. et al., "Color Display System of Connected Speech for Deaf Subjects", Joint Meeting of Acoustical Societies of America and Japan, Hawaï, November 1978.

[WILL - 72] WILLEMAIN T.R. and LEE F.F., "Tactile pitch displays of the deaf", IEEE Trans. Audio Electroacoust., AU-20, 1, pp. 9-16, 1972.

[WITC - 56] WITCHER C.M., "Vocotac (Sensory replacement project)", Quaterly Progress Reports, Research Laboratory of Electronics, M.I.T., January-October 1956.

[WOOD - 71] WOOD M.L., "Computer generated spectrograms and cepstrograms", M. Sc. Thesis, M.I.T., June 1971.

[WOUT - 76] WOUTS F., "L'Alphabet des Kinèmes Assistés", Rev. Gén. Enseignement Déf. Auditifs, N° 3, Paris, 1976.

[YAKI - 76] YAKITA Y. and HIKI S., "Investigation of laryngeal control in speech by use of thyrometer", J.A.S.A., Vol. 59, N° 3, March 1976.

[YOUN - 62] YOUNG T.Y. and HUGGINS W.H., "Representation and analysis of signals, pt. III, Discrete orthonormal exponentials", Proc. Nat. Electronics Conf., pp. 10-18, October 1962.

[ZURC - 77] ZURCHER J.F., "La mesure du fondamental par la détection de crêtes : techniques employées, résultats", 8èmes J.E.P., GALF, Aix-en-Provence, Mai 1977.

[ZWIC - 81] ZWICKER E. et FELDTKELLER R., "Psychoacoustique. L'oreille, récepteur d'information", Masson, Paris 1981.

NOM DE L'ETUDIANT : Madame HATON Marie-Christine

NATURE DE LA THESE : Doctorat d'Etat ès sciences mathématiques

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le - 7 FEV. 1985 n° 261

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

UNIVERSITE NANCY I
LE PRESIDENT
R. MAINARD

RESUME

Dans le cadre de recherches plus générales sur le traitement automatique de la parole, nous étudions son application au domaine de l'éducation vocale assistée par ordinateur.

Après avoir placé notre travail dans le contexte de la communication parlée et de l'acquisition du langage, nous présentons les techniques suivant lesquelles nous accédons à différents paramètres pertinents et à différents modes de représentation de la parole.

Nous traitons ensuite de l'analyse des voix en temps différé dans le double point de vue de leur caractérisation et de l'évaluation des progrès.

Nous terminons par la description du système SIRENE d'aide visuelle à la production vocale, la relation d'une phase d'expérimentation auprès d'enfants malentendants et les perspectives de développement.

MOTS-CLES

Traitement automatique de la parole - Education vocale - Système SIRENE - Analyse de la voix - E.A.O. spécialisé - Apprentissage des langues.