LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Analyse spectrale et analyse statistique de la voix humaine parlée

Les cahiers de l'analyse des données, tome 13, n° 1 (1988), p. 99-130

<http://www.numdam.org/item?id=CAD_1988__13_1_99_0>

Les Cahiers de l'Analyse des Données
Vol XIII - 1988 - n° 1 - pp. 99-130

ANALYSE SPECTRALE ET ANALYSE STATISTIQUE DE LA VOIX HUMAINE PARLÉE

[SPECTR. STAT. VOIX]

J.-P. BENZÉCRI

0 Introduction: analyse mathématique et histoire des concepts

Il est généralement admis aujourd'hui que l'étude de la voix humaine parlée, en vue de la synthèse et de la reconnaissance de la parole, doit être fondée sur l'analyse mathématique du signal vocal, assimilé à une suite de nombres, qui est, en bref, la suite des valeurs, mesurées plusieurs dizaines de milliers de fois par seconde, de la pression sonore exercée sur un microphone. Et, de cette analyse mathématique, l'application des méthodes de Fourier constitue la première étape. En 1987, les progrès accélérés de la micro-informatique ont mis une telle étude à la portée d'un chercheur isolé.

Cependant, les conceptions générales que linguistes, physiologistes, physiciens, musiciens et ingénieurs se sont formées de la parole reposent sur les efforts successifs de plusieurs générations de savants qui, depuis le XVIII-ème siècle, ont, par des moyens très divers, tenté de saisir les mots qui volent. Même s'il est hors de notre propos d'écrire de ces recherches une histoire, fût-elle abrégée, nous croyons utile de dire ce que nous savons de l'histoire de concepts, dont la définition est souvent confuse, mêlant à des prémisses implicites les conclusions de théories oubliées ou mal comprises.

Le plan du présent article sera donc le suivant. Aux §§1 et 2, nous dirons de quelle chaîne de traitement physique et numérique nous avons disposé, pour prendre avec la parole, un contact qui, au moins pour le géomètre, peut être appelé un contact direct, explicite. Au §3, nous considérerons les concepts usuels relatifs à la parole, au travers de notre expérience personnelle du traitement numérique, dont le §4 offre quelques exemples, plus ou moins détaillés. Le but étant d'aboutir à l'énoncé de résultats généraux précis. Cependant, l'exposé serait trop aride, et même obscur, si en décrivant la chaîne

de traitement, nous n'anticipions pas sur certains résultats concernant notre objet: la parole. On s'adressera donc au lecteur, en évoquant, d'emblée, maintes questions qu'il ne peut ignorer.

1 La chaîne de traitement utilisée

Notre micro-ordinateur est un Macintosh plus avec 1 mégaoctet de mémoire centrale, un disque dur de 20 mégaoctets, et deux lecteurs de disquette 800k. A l'entrée est un convertisseur d'analogique en digital, distribué dans le commerce, sous le nom de Soundcap, avec un logiciel adapté aux excellentes potentialités conversationnelles du Macintosh: dans la suite, le mot Soundcap désignera tantôt la boîte noire du convertisseur, et tantôt le logiciel. Pour le traitement ultérieur du signal, et principalement en vue de l'analyse des données multidimensionnelles, nous avons écrit un ensemble de programmes en langage Pascal, suivant les normes du compilateur TML. À une brève présentation de ces programmes est consacré le §2.

1.1 Conversion des sons en nombres

Sans entrer dans les détails, disons que le rafraîchissement de l'écran du Macintosh, se fait avec le balayage d'une ligne en 44,93 microsecondes; cette période correspond à une fréquence de 22257Hz, qui est la fréquence d'échantillonnage du Soundcap; la précision de la mesure instantanée est celle d'un entier variant de 0 à 255, soit: un octet.

Il vaut la peine de commenter ces nombres. En bref, la fréquence de 22kHz est satisfaisante pour l'étude minutieuse de la parole; en revanche échantillonner seulement sur un octet (8bits) est une gène, mais non un obstacle insurmontable. Noter par un entier de 0 à 255 un signal dont le signe varie alternativement, équivaut à utiliser l'intervalle (-128, 127) ; ce qui semble garantir une erreur relative inférieure à 1%. En fait, l'amplitude du signal vocal varie grandement au cours du temps. Un réglage du potentiomètre d'entrée du Soundcap, (ou de celui du magnétophone en amont), qui donne au 'a' du mot 'chat' l'amplitude (-100, 100), donnera au bruit consonnantique initial de ce mot une amplitude inférieure à 20; ce qui est juste suffisant, pour décrire avec précision ce bruit. Augmenter la puissance permettra d'être à l'aise dans l'étude du 'ch', mais alors la courbe du 'a' sera écrêtée; ce qui modifie certainement le spectre; et aussi l'impression auditive. Trouver un bon réglage est d'autant plus difficile qu'aux variations d'amplitude, inévitables, d'un phonème à l'autre d'un même mot, s'ajoutent les variations du niveau de la voix au cours d'une phrase et, *a fortiori*, d'un discours.

On sait qu'une fréquence d'échantillonnage de quelque 22kHz, ne permet aucunement d'analyser les sons jusqu'à cette fréquence, mais seulement jusqu'à 11kHz, au plus. De façon précise, considérons un signal sinusoïdal y. de fréquence f, $y= \sin(2\pi ft+\phi)$, échantillonné, (c'est à dire mesuré), aux instants O, T, 2T, 3T, 4T, etc. Si f=1/T, toutes les mesures aurons la même valeur; si

f=1/(2T), on aura une suite alternée de valeurs opposées. D'autre part, augmenter f de 1/T, ne modifie pas les valeurs prises par y aux instants des mesures; et changer le signe de f équivaut à changer la phase (ϕ) en ($\pi-\phi$). De ce fait, un signal de fréquence (1/(2T))+Δf équivaut à un signal de fréquence (1/(2T))-Δf. C'est ce qu'on appelle le phénomène de repli. On ne peut donc déterminer sans confusion le spectre de puissance d'un signal échantillonné avec la période T, que pour celles des fréquences de l'intervalle (0, 1/(2T)) qui ne sont pas, du fait du repli, polluées par des composantes de fréquence supérieure à 1/(2T). Comme il est impossible d'éliminer, par un filtrage physique ou numérique, les composantes dont la fréquence dépasse 1/(2T), sans altérer celles de fréquence peu inférieure à 1/(2T), la bande de fréquence donnée par le Soundcap va de 0 à 9kHz environ.

Cette bande de 9kHz est plus de deux fois plus large que celle du téléphone, qui permet de comprendre une conversation mais non une syllabe isolée, comme on le constate en épelant un nom. En échantillonnant à 45kHz on peut avoir une restitution de haute fidélité. Sans aller jusques là, la bande du Soundcap est pleinement satisfaisante pour l'étude de la parole.

1.2 Examen de la courbe du signal et écoute par segments

Le Macintosh est muni d'un petit haut-parleur, ainsi que d'une sortie permettant de diriger des sons vers un amplificateur externe. Le signal d'un microphone ou d'un magnétophone, une fois digitalisé par le convertisseur et introduit dans le Macintosh, peut soit être écouté immédiatement, soit être saisi dans la mémoire de l'ordinateur. Mais tandis que l'écoute peut se prolonger indéfiniment, la saisie, qui se fait dans la mémoire centrale, ne peut durer plus de 33 secondes sur notre appareil: il faut en effet se souvenir qu'une seconde de signal est convertie en 22k octets. Une fois saisi, le son digitalisé peut être enregistré sur disquette ou disque dur: l'enregistrement direct est impossible, du fait des temps de transfert.

Ecouter au travers de l'ordinateur du son digitalisé n'est pas sans intérêt. D'abord on apprécie la qualité de la restitution; mais surtout, l'écran de l'ordinateur fonctionnant, par Soundcap, comme celui d'un oscilloscope, on peut régler la puissance du signal de telle sorte qu'il n'y ait ni écrétage ni extinction; ce qui, avec une digitalisation sur un seul octet, est, comme on l'a dit, assez délicat. D'autre part, au lieu de se consacrer à l'affichage et à la reproduction du signal reçu, l'ordinateur peut, par une option de Soundcap, afficher des spectres instantanés, calculés par transformée de Fourier. Les transformées sont calculées, au choix, sur des tranches de 124, 256, 512 ou 1024 points. Or 1024 points correspondent à (1/22)s: il faudrait donc calculer 22 spectres par seconde. Malgré une excellente écriture, en langage C, de l'algorithme de transformée rapide, Soundcap ne calcule que 2 spectres par seconde. Avec des tranches de 128 points on voit s'afficher à l'écran près de 20

spectres par seconde, ce qui donne une belle impression de mouvement, mais ne permet pas d'observation précise. L'essentiel de nos observations a donc été fait d'une part, sur des sons déjà saisis, logés dans la mémoire centrale; et, d'autre part, sur les fichiers gardés sur disque et analysés numériquement.

Sur le signal présent en mémoire centrale, (soit qu'il vienne d'être saisi, soit qu'on l'ait rappelé du disque où il était gardé), Soundcap offre d'effectuer de nombreuses opérations. D'abord, la courbe du signal s'affiche; ou plutôt, une partie de cette courbe. En effet, la définition de l'écran du Macintosh permet seulement d'afficher les valeurs par tranches de 500, (nombre des points de l'axe horizontal, transversalement auquel sont portées les valeurs); soit un peu moins de (1/40)s. Il y a, dans la parole commune, de longues plages de quasi-silence, auxquelles correspondent des portions de courbes presque rectilignes, ou faiblement ondulées. Les voyelles offrent des successions de périodes qui, sans être identiques entre elles, se transforment lentement. Ailleurs, on a des courbes en dents de scie, plus ou moins irrégulières. Nous reviendrons sur ces aspects, immédiatement visibles, et qui déjà réservent des surprises.

Cependant, une commande permet de comprimer l'échelle du temps dans un rapport 2, 4, 8,..La courbe ainsi comprimée apparaît vite comme une masse noire, dont on distingue seulement l'enveloppe: mais ce schéma suffit pour choisir, par une commande de la souris, un intervalle de son, (dont la courbe apparaît en nuance inverse: blanc pour noir, noir pour blanc), et qui peut être rejoué seul, une fois ou n fois consécutivement; ou gardé sur disque pour constituer un fichier séparé.

Il est très instructif d'écouter patiemment, (joués en sens direct ou en sens inverse!), des segments de son, dont on fait varier les bornes par degrés imperceptibles. Il s'en faut de beaucoup que ce qu'on entend soit en accord avec une décomposition alphabétique naïve du signal. Isolé du 'a' qui le suit, le début d'une syllabe 'la' s'entend comme une voyelle, dont le timbre est proche du 'u': puis, soudain, à force de prolonger le segment, on entend le roulement du 'la'. D'ailleurs, sans réduction d'échelle, la courbe de ce qu'on sait être le 'l', offre, comme une voyelle, un aspect quasi-périodique.

1.3 Affichage du spectre d'un segment

On a déjà parlé de l'affichage instantané des spectres d'un signal débité, sans interruption, par le convertisseur. Soundcap offre également le calcul de spectres d'une tranche, de 1024 points, du signal présent en mémoire centrale. L'aspect des spectres varie grandement selon que la tranche est prise dans une voyelle, ou dans un bruit consonantique, tel que le 'ch'. Un spectre de type vocalique montre une suite de raies très étroites qui correspondent aux harmoniques de la fréquence avec laquelle se succèdent les quasi-périodes; l'enveloppe de ces raies suggère une ou plusieurs collines ou montagnes, qu'on peut appeler 'formants'; même si ce terme offrira, dans la suite, matière à

d'amples discussions. Dans un spectre de 'ch', etc, au contraire, il n'y a pas de raies, (puisqu'il n'y a pas de quasi-périodes; donc pas de fréquence fondamentale ni d'harmoniques); les maxima d'intensité ont l'aspect d'un amas dense, hérissé de pointes irrégulières.

On se souvient que, dans l'affichage continu, le choix était offert, pour les tranches analysées, entre diverses longueurs variant de 128 à 1024, en progression géométrique de raison 2. C'est l'occasion de discuter des avantages des divers choix. Sans faire d'exposé mathématique, nous rappellerons que, dans la transformation de Fourier du signal écchantillonné, la largeur du spectre est, (à la perte près due au filtrage nécessaire pour obvier au repli, cf §1.1), la moitié de la fréqence d'échantilonnage; tandis que le pas en fréquence, (i.e. l'intervalle entre deux valeurs successives pour lesquelles est donnée la valeur de la transformée) est l'inverse de la durée de la tranche transformée. Il est impossible d'avoir une bonne définition en fréquence sans allonger la tranche; et celà se conçoit, sans faire de calculs, car, en particulier, la plus faible fréquence accessible doit correspondre à une période dont la longueur ne sorte pas des limites de la tranche analysée. Dans notre cas, la structure de raies des spectres vocaliques devient confuse avec une longueur de tranche inférieure à 512 points: en effet, une tranche de 256 points dure (1/86)s; et la durée d'une quasi-période n'est pas beaucoup plus courte, au moins pour une voix d'homme. Mais, d'autre part, l'examen direct de la courbe montre que sur une tranche de 1024 points, le signal a rarement un aspect stable; en sorte qu'aucun choix n'est pleinement satisfaisant.

2 Programmes en langage Pascal pour l'analyse et la synthèse des signaux sonores sous forme numérique

L'examen direct des spectres affichés par Soundcap est très suggestif, mais ne permet pas de comparaisons précises; il permet encore moins d'effectuer des classifications automatiques, en vue de la reconnaissance automatique de la parole. D'autre part, même si Soundcap peut, notamment par inversion du sens, coupage et collage, créer des séquences sonores, sur lesquelles on vérifiera certaines hypothèses, il reste beaucoup à faire pour engendrer, modifier ou grouper des sons. C'est pourquoi on a dû écrire un ensemble de programmes, dont les principaux sont décrits ci-dessous, dans leurs grandes lignes.

Dans la suite, nous exposerons quelques légitimes objections qu'on peut faire à l'usage constant de la transformation de Fourier; mais, puisque toutes les analyses mathématiques que nous avons faites jusqu'ici reposent sur cette transformation, c'est par elle que doit commencer la description des programmes.

2.1 Transformation de Fourier

Nous avons dit que Soundcap peut garder sur disque des fichiers de son. Du point de vue du contenu, ces fichiers sont des fichiers d'octets (nombres de 0 à 255), qu'on interprète en introduisant quelque peu arbitrairement un zéro de référence, (fixé à 128). Dans le langage TML Pascal, on n'a pas de fichiers d'octets; mais il est facile de relire les fichiers de son comme des fichiers d'entiers, en décodant chaque entier comme une suite de deux nombres positifs, (et non comme un nombre unique avec signe). On peut alors effectuer la transformation de Fourier.

Pour celà, on utilise, comme de règle, l'algorithme dit 'rapide': reste à préciser à quelles tranches de son on l'applique, et à quelles transformations ultérieures on soumet les résultats. On sait que l' algorithme ne s'emploie commodément que pour une tranche dont le nombre de points est une puissance de 2. Soundcap nous a permis d'apprécier les résultats obtenus avec des tranches de 128 à 1024 points. Des travaux de T. Moussa, on peut conclure qu'il n'est pas suffisant de décrire la parole par les spectres d'un ensemble de tranches *consécutives* de 1024 points: car certaines consonnes, telles le qaf de l'arabe, se manifestent par des traits sonores brefs, susceptibles de disparaître à la jonction de deux tranches. C'est pourquoi T. Moussa utilise des fenêtres empiétantes de 1024 points, chacune étant décalée de 256 points sur la précédente. On a ainsi toute l'information pertinente; non seulement les enveloppes spectrales qui caractérisent les phonèmes, (même si la reconnaissance de ceux-ci ne se fait pas actuellement sans erreurs), mais encore l'intonation, (qui, comme on l'a dit au §1.3, se montre, sur les spectres vocaliques, par l'espacement des raies). Cependant, le coût est élévé: chaque point est pris dans 4 tranches; et l'on doit, pour chaque tranche construire l'enveloppe spectrale par un algorithme complémentaire.

De notre propre expérience, nous concluons que la description des enveloppes spectrales est satisfaisante si l'on prend, pour les voix d'homme, des tranches consécutives de 256 points; et, pour les voix de femme, des tranches de 128 ou 256 points ($\approx$(1/80)s). En effet, cette longueur est adaptée à celle des pseudopériodes, (plus brèves chez les femmes dont le timbre est aigu). Pour justifier notre assertion, il a fallu, d'une part, introduire une variante dans le programme de transformation de Fourier; et, d'autre part, recourir à l'analyse des données. Expliquons d'abord en quoi consiste cette variante.

Si l'on découpe, sans précaution particulière, une tranche d'une courbe de son, la première valeur de la tranche est, généralement, nettement différente de la dernière; la fonction périodique engendrée par répétition indéfinie de la tranche est donc une fonction discontinue. Or c'est à cette fonction que s'applique véritablement la transformation de Fourier, et la discontinuité se manifeste dans le spectre par des composantes parasites de haute fréquence, (analogues à celles de la transformée d'un signal carré). Pour éliminer ces composantes, l'usage est de multiplier la tranche par une fonction en 1+sin, s'annulant aux deux

extrémités, et appelée 'fenêtre de Haning'. (On pourrait également retrancher de la tranche une fonction linéaire choisie pour que le signal ait même valeur aux deux extrémités).

Cependant, dans l'analyse d'un son vocalique, découpage arbitraire et fenêtre ne sont que des pis aller: l'objet visé est la fonction périodique engendrée par répétition indéfinie d'une quasi-période. Il faudrait prendre pour tranche une quasi-période, et analyser celle-ci sans introduire de fenêtre, (multiplication par 1+sin), puisqu'il n'y a pas de discontinuité à compenser. La difficulté est qu'une quasi-période ne compte qu'exceptionnellement 128 ou 256 points; de plus, nous n'avons pas écrit jusqu'à présent d'algorithme de découpage automatique des quasi-périodes. Voici donc ce que nous faisons. Une quasi-période est découpée à vue, grâce à Soundcap, et gardée comme un fichier séparé. Lors de l'analyse, la fonction est intrapolée sur 128 ou 256 points; et le spectre, (calculé sans introduire de fenêtre), est remis à l'échelle, (compte tenu de ce que, en bref, intrapoler équivaut à simuler un taux d'échantillonnage modifié). On peut alors, en passant par l'analyse des données, comparer, au véritable spectre d'une quasi-période, les spectres de tranches de 128 ou 256 points, découpées sans art... De ces comparaisons, nous avons conclu qu'il était satisfaisant de prendre des tranches de 256 points.

2.2 Analyse des correspondances et classification ascendante hiérarchique (CAH); codage des spectres

Nous ne reprendrons pas l'exposé de méthodes bien connues et déjà appliquées par T. Moussa à l'analyse des profils spectraux. En l'état actuel des programmes, on peut, sur Macintosh, analyser un tableau de correspondance 1000×100, en 2 heures environ; pour l'analyse des spectres, nous n'avons pas dépassé 40 colonnes, avec un temps de (1/2)h. La classification ascendante hiérarchique requiert environ 3h pour 1000 individus, décrits par 20 facteurs au plus; le temps de calcul dépend peu du nombre de facteurs utilisés, parce que l'algorithme de calcul des distances, (comme somme des carrés des différences des facteurs), arrête la sommation dès qu'est dépassé le seuil de comparaison. Reste à dire comment on utilise les algorithmes de base dans l'étude des spectres.

Au départ, une tranche de 128 points fournit un spectre à 64 composantes; (le nombre est divisé par 2 parce qu'en bref, dans un spectre de puissance, les phases ne sont pas prises en compte). Du fait du filtrage physique du Soundcap, les composantes de haute fréquence sont dépourvues de sens; et jusqu'à présent, toutes nos analyses se sont limitées aux composantes 1 à 40, soit une bande de ≈7000Hz. Avec des tranches de 256 points on a, de même, des spectres à 128 composantes; mais, compte tenu de notre objectif, les spectres ont été ramenés à 64 canaux, dont on a retenu les 40 premiers, comme pour les tranches de 128 points. (Une description plus fine des spectres n'aurait d'intérêt que pour des tranches de 512 ou 1024 points, en vue d'étudier l'intonation).

Sur les spectres tels quels, l'analyse et la classification donnent des résultats peu satisfaisants, en ce sens que des spectres très dissemblables peuvent être rangés dans une même classe.Il y a, à cela, plusieurs causes, dont nous corrigeons les effets progressivement. D'abord, la puissance des premiers canaux, (<500Hz), est prédominante, particulièrement si, afin d'économiser l'espace on code les composantes des spectres sur un seul octet, ce qui fait totalement disparaître les composantes faibles. On corrige ce défaut en multipliant les premières composantes par un coefficient d'extinction qui rejoint 1 vers 1000Hz, (canal 6 ou 7). Les coefficients peuvent être choisis pour égaliser les poids des premiers canaux, calculés par un programme spécial, ou pris sur le listage usuel d'a des cor. Ces corrections *ad hoc* , ont une justification psychophysique: la sensibilité de l'oreille aux basses fréquences est faible, en sorte que la puissance perçue n'est pas, sur toute l'étendue du spectre, proportionnelle à la puissance physique que nous calculons par transformation de Fourier.

Ensuite, il n'est pas utile de retenir 40 canaux de largeur ≈175Hz: cette précision est même nuisible, en haute fréquence, car elle met entre deux canaux contigus une différence qui ne correspond à rien dans la perception. C'est pourquoi, d'une part, on a mis, dans les spectres, non les valeurs calculées mais des moyennes pondérées entre canaux contigus; d'autre part, on peut, lors d'une analyse, grouper les canaux en blocs, dont les limites sont choisies d'après les résultats d'une analyse antérieure, ou d'une CAH sur les 40 canaux.

Mais il y a plus. Une distance, telle que celle du χ^2, calculée en fonction des différences des composantes, n'est pas fidèle aux véritables similitudes entre spectres. En effet, pour une telle distance, il y a, entre 0 et 4, la même différence qu'entre 4 et 8; alors que la perception se fonde plutôt sur la règle du tout ou rien: la seule présence de puissance sonore vers 2000Hz suffisant, par exemple, à ditinguer un 'u' (français) d'un 'ou'. Il s'impose donc de coder les spectres en découpant en classes la puissance contenue dans une dizaine de canaux; ce qui requiert le choix de seuils. Ce choix peut être aidé, comme dans tout autre domaine, par l'examen d'un histogramme. Mais, en définitive, le choix doit être conforme à la valeur perçue des sons; le seuil doit se placer à la limite, pourtant inconstante, entre ce qui est perçu comme 'tout' et ce qui est perçu comme 'rien'; ou, plutôt, il faut introduire une modalité moyenne où varie cette limite.

La nécessité d'un codage par classes et la difficulté qu'on éprouve à fixer les limites de celles-ci, apparaissent également si l'on critique la notion même de profil spectral. Il est vrai qu'on peut faire varier la puissance sonore de 1 à 100, (différence de 20dB) sans modifier les qualités perçues du message sonore; et ceci justifie qu'on considère non les valeurs brutes des composantes d'un spectre, mais les composantes du profil; (destinées à être ensuite découpées en classes). Mais d'autre part, on peut agir en toute liberté sur les réglages de

tonalité des graves et des aigus, sans modifier le message perçu: ce qui prouve que la valeur précise du rapport entre deux composantes importe peu; et nous rappelle la nécessité de choisir des seuils.

Enfin, notamment pour les bruits consonantiques, un canal peut être masqué par un autre: ainsi, la superposition d'un 'ch' et d'un 's' s'entend 'ch'; comme si le bruit du 's' était masqué par celui, de fréquence plus basse, du 'ch'.

2.3 Classes et étalons; étiquetage du message

Le programme de CAH permet d'effectuer un partition de l'ensemble des individus, définie soit par les nœuds les plus hauts, soit, plus généralement, par des nœuds spécifiés. On serait pleinement satisfait si chaque classe correspondait à un phonème de la langue étudiée. Plus précisément, puisque la phonologie a enseigné aux linguistes à ranger, sous un même phonème des sons différents, pourvu que cette différence ne serve pas à distinguer entre des mots de sens distincts, on voudrait reconnaître les "phones", qui sont, en bref, des subdivisions des phonèmes, homogènes quant au son. Cet objectif n'est pas directement accessible non plus, dans la mesure où la valeur perçue d'une tranche de son dépend du contexte. La classification devrait seulement réaliser, (d'après les spectres), une partition de l'ensemble des tranches de son telle qu'en substituant au message sonore initial la suite des numéros de classe des tranches, on ne perdît rien de l'information phonémique; ou, plus concrètement, une partition telle qu'en substituant aux tranches successives dont se compose un message, des tranches de même classe spectrale, (et réalisées avec la même intonation), on obtînt un message équivalent. Cet objectif taxinomique intermédiaire une fois atteint, il resterait à découvrir les lois de l'interprétation contextuelle, en termes de phones et de phonèmes, des suites d' éléménts ainsi étiquetés. Pour avancer dans cette voie, il convient de modifier et compléter le programme de CAH.

A une classe, on associe communément son centre de gravité. Si ce centre peut être assimilé à un individu, il fournit un étalon acceptable. Mais quand intervient un codage par classes, il est difficile d'assimiler à un individu, une moyenne dont les composantes non nulles s'étalent sur plusieurs modalités de chacune des variables qu'on a découpées. Pour obtenir un étalon ayant les caractères d'un individu réel, le plus sûr est de prendre celui des individus de la classe qui est le plus proche de son centre, dans l'espace engendré par les axes factoriels utilisés, c'est à dire au sens de la distance entre individus (spectres) après codage.

Les étalons sont utiles à plus d'un titre. D'abord tout spectre, même s'il n'appartient pas à l'ensemble des individus sur lesquels on a fait la classification, peut être assimilé à l'étalon dont il est le plus proche. Ainsi, à un message sonore quelconque, on substituera, (cf *supra*), la suite des spectres étalon auxquels est ratachè chacune de ses tranches. Il est commode de présenter

cet étiquetage comme un sonogramme numérique, où, à chaque tranche de son, corespondent deux lignes successives, donnant l'une le spectre initial et l'autre l'étalon qu'on lui a substitué. La lecture de ces sonogrammes est plus facile si on remplace par des blancs les composantes faibles des spectres; ou si, poussant plus loin la schématisation, on résume par une valeur centrale chaque bloc de composantes non nulles d'un spectre, ce qui revient à interpréter les spectres en terme de formants.

D'autre part, nous avons dit qu'une CAH sur 1000 individus nous prend 3 heures; le temps étant proportionnel au carré du nombre des individus, il est difficile d'en prendre plusieurs milliers. Or, pour explorer la diversité des sons de parole, (ce qu'il faut faire, selon nous en traitant des discours de plusieurs langues), on doit classer plus de 1000 spectres. Une solution assez satisfaisante nous paraît être de procéder hiérarchiquement. On prépare, par exemple, 4 fichiers de 500 spectres; chacun de ceux-ci est, en une heure de calcul, réduit à 125 étalons; soit, au total, 500 étalons dont l'ensemble est lui même soumis à la CAH. Il est préférable qu'à l'étape intermédiaire les étalons retenus soient de véritables individus, plutôt que des moyennes, car de moyenne en moyenne, on aboutirait à des spectres étalés sans caractère linguistique.

Il importe d'apprécier dans quelle mesure les spectres, de toute origine, rattachés à chaque étalon se ressemblent entre eux. A cette fin, on peut créer, pour chaque étalon un fichier de ses spectres; et les imprimer, (comme on imprime un sonogramme), ou même faire entendre les tranches de son correspondantes, les unes après les autres, chacune avec un bref contexte.

3 La parole: notions physiologiques, phonétiques, linguistiques et analyse numérique du signal

Entre 1920 et 1930, le Professeur H. Bouasse, longtemps justement apprécié, pour son sens physique autant que pour son style incisif, a publié un traité dont le tome II, intitulé "Instruments à vent" contient 3 chapitres consacrés respectivement à la "Voix humaine", la "Théorie des voyelles", la "Phonétique expérimentale". Nous suivrons l'ordre de ces chapitres, pour placer dans une perspective historique les concepts de la théorie de la parole, aujourd'hui confrontés à l'analyse numérique.

3.1 Production de la voix humaine

Il ne nous appartient pas de décrire, dans toute sa complexité, le larynx humain; mais la structure quasi périodique des sons vocaliques est inséparable de leur origine, qu'il nous faut donc considérer ici.

Au plus simple, les voies aériennes peuvent être schématisées comme suit. A une extrémité, le réservoir des poumons, dont l'accès est constitué par le conduit de la trachée; à l'autre, les cavités de la bouche et du nez qui confluent et s'enfoncent en formant le pharynx; entre les deux, le larynx, organe complexe,

formé de plusieurs pièces cartilagineuses articulées, sur lesquelles s'insèrent des muscles revêtus de membranes. Au milieu du parcours laryngé, l'air passe par une fente dont les deux lèvres musculeuses, appelées cordes vocales, peuvent être plus ou moins tendues et rapprochées selon la disposition des diverses pièces du larynx.

Aucune théorie ne met en doute le fait que la voix chantée, comme la parole sonore, (à la différence de la parole chuchotée, dont on traitera ensuite), trouve son origine dans le mouvement des lèvres vocales qui s'ouvrent et se ferment avec la fréquence fondamentale du son émis. La question est de savoir comment ce mouvement est provoqué et réglé. Le larynx étant un organe vivant, dont la ressemblance avec un tuyau à anche est manifeste, les explications proposées diffèrent principalement quant au rôle attribué à la spécificité du vivant. Il y a trois théories principales, qu'on peut, en bref, associer aux noms de Ferrein, Ewald et Husson; et ces théories peuvent être conjuguées.

Au milieu du XVIII-ème siècle, alors qu'on trouve, dans les orgues, des jeux d'anches battantes imitant le chant de la "voix humaine", et que se perfectionnent les automates, Ferrein, (1741) expérimente sur le larynx mort. Nous citons, d'après Bouasse, une description, qui offre au lecteur un aperçu de la complexité des articulations du larynx:

"Pour faire sonner le larynx, il faut serrer entre le pouce et l'index les cartilages aryténoïdes l'un contre l'autre (rétrécir la glotte) et souffler de bas en haut dans la trachée artère à la faveur d'un tuyau de 4 à 5 lignes [9 à 11 mm.] de diamètre... La poitrine a peine à fournir au larynx du bœuf, du cochon,... je me sers alors d'un soufflet... Lorsque je veux donner une plus grande tension aux cordes vocales et faire monter le son, je presse le cartilage *scutiforme* [thyroïde] sur la partie antérieure du cartilage *annulaire*" [cricoïde]...

Ferrein admet que les lèvres vocales sont mises en vibration par le vent, qu'elles forment une anche bilabiale entretenue par le jet d'air. Ferrein, comme après lui Muller et Helmoltz, assimile cette anche à une anche membraneuse réalisée à l'aide de lames en caoutchouc.

Cependant, dans le larynx mort, seul vibre le bord des lèvres; les "cordes", explique Bouasse, seraient le ligament et la muqueuse. D'ailleurs, Ferrein doit créer des tensions bien supérieures à celles auxquelles peut être soumis l'organe vivant. Selon Ewald, la masse musculeuse, (proprement appelée *corde*), participe à la vibration, avec des caractéristiques élastiques qui dépendent de l'état de tension du muscle. Pour expliquer en quel sens le muscle contracté est tendu, Bouasse propose de cette propriété biomécanique un modèle purement physique, qui nous intéresse en ce qu'il suggère l'étude de milieux continus comportant une source d'énergie distribuée, comme l'est le métabolisme pour les muscles.

"Fixons une hélice de laiton par les deux bouts: elle vibre transversalement avec une certaine fréquence. Dans cette hélice, faisons passer un courant électrique: les spires s'attirent; *la fréquence de la vibration transversale croît*; tout se passe comme si la tension, (force de rappel à la position d'équilibre), était augmentée."

En 1913, (in *Pflügers Archiv für die gesamte Physiologie*, n°152), Ewald publie la description de *Polsterpfeifen* (tuyaux à bourrelets), qu'il a construits comme un modèle du larynx. En bref, le passage de l'air dans le tuyau est interrompu par deux bourrelets de caoutchouc jointifs, gonflés avec une pression variable. Sous la force du courant, les bourrelets s'écartent en vibrant transversalement avec une fréquence d'autant plus élévée que la pression de gonflage est plus forte.

Cependant, pour Raoul Husson, (que nous citons d'après une coupure d'un journal médical du début des années 1960),

"la soi-disant 'vibration' des cordes vocales <n'est> qu'une activité neurogène rapide et rythmée commandée par les influx moteurs tombant pendant la phonation sur les cordes vocales...La voix est ainsi produite par un mécanisme banal de sirène... ce n'est aucunement l'air qui fait mouvoir les cordes vocales, mais uniquement les influx moteurs... Et la fréquence de la voix est... imposée à chaque instant par celle des salves d'influx".

On trouve déjà dans Bouasse des arguments contre l'existence de contractions musculaires neurogènes d'une fréquence aussi élévée que celle de la voix; l'originalité de Husson est d'avoir opposé à ces arguments un faisceau d'observations neurohistologiques nouvelles. Aujourd'hui, 20 ans après la mort de R. Husson, sa théorie n'est guère admise: mais dans l'article *Phonation* de l'*Encyclopædia Universalis*, B. Vallancien prône une théorie myoélastique qui complète celle d'Ewald, en tenant compte des influx nerveux (arc réflexe sensitivo-moteur du nerf laryngé supérieur) et aussi du glissement de la muqueuse, (à laquelle était dévolu le rôle de "corde" dans l'expérience de Ferrein; ainsi que dans la théorie muco-ondulatoire de Perello, Barcelone, 1962).

3.2 Hauteur du fondamental et formants, dans la théorie des voyelles

Si on observe, à l'écran du Macintosh, un signal de parole, en le parcourant dans toute sa longueur, comme Soundcap permet de le faire, on remarque, (outre des lignes de quasi-silence aux ondulations peu marquées), d'une part, des portions de courbe de grande amplitude, formées de quasi-périodes régulières qui se succèdent en se déformant lentement; et d'autre part, des portions de faible amplitude, irrégulières, certaines en dents de scie, d'autres marquées d'oscillations serrées. Nous sommes convenus d'appeler les premières "sons vocaliques", et les deuxièmes "bruits consonantiques", même

si, comme on le verra dans la suite, les consonnes liquides (r, l, n, m) sont plutôt des sons vocaliques, et que parfois un bruit se superpose au son.

A l'œil nu, il semble assez facile de découper avec précision les quasi-périodes d'un son vocalique, en repérant un maximum ou un minimum particulièrement aigu (cf §2.1). En même temps que leur forme, la durée des quasi-périodes varie graduellement. Si on se fie à la durée exacte des segments qu'on a découpés, (et cela est légitime, dans la mesure où cette durée varie très régulièrement), on peut calculer une fréquence instantanée du fondamental; et la variation de cette fréquence, au cours du temps, donne une courbe d'intonation, ou mélodie.

Toute autre est la méthode numérique fondée sur l'analyse de Fourier. On part d'une tranche de 1024 points (ou 2048...) qui généralement ne peut être découpée en un nombre exact de sous-périodes, mais doit en contenir plusieurs. Et, ainsi qu'on l'a dit, on a pour la tranche un spectre formé de raies, dont l'espacement régulier donne la fréquence du fondamental. Cette méthode a le mérite de pouvoir être mise en algorithme, comme l'a fait notamment T. Moussa dans sa thèse; mais, pour la détection de la mélodie, la méthode visuelle est bien plus sensible et précise. Le mieux serait, sans doute de déterminer selon Fourier un période approchée du fondamental; puis d'utiliser ce résultat pour découper automatiquement le signal en quasi-périodes, par exemple en partant d'un extremum local aigu isolé, et recherchant le suivant, là où on doit le touver d'après la période approchée déjà connue.

Du point de vue mathématique, la méthode de Fourier est absolument universelle; un siècle après Helmoltz et Lord Raleigh, tous les physiciens sont convaincus de cette vérité, (même si Bouasse n'en était pas pénétré..). Mais la question reste posée de savoir si, pour l'explication causale d'un phénomène donné, l'application automatique de cette méthode sépare toujours au mieux les constituants du système étudié. Par exemple, un signal formé d'une dizaine de quasi-périodes successives peut être engendré en superposant à un fondamental des partiels qui ne sont pas exactement ses harmoniques (fréquences f, $2f+\varepsilon$, $3f+\varepsilon'$..). Il vaut la peine, à ce propos, de chercher, chez Bouasse, (copieusement cité ou paraphrasé ci-dessous), l'histoire de la notion de formant, si généralement admise aujourd'hui.

Bouasse fait état de trois théories des voyelles. Selon la première, dite "Théorie du rapport des intensités des harmoniques", une voyelle est caractérisée par des rapports déterminés entre les intensités des harmoniques de la note sur laquelle elle est émise. (Par exemple, pour telle voyelle, l'harmonique 2 aurait une intensité considérable..). Selon cette théorie, une voyelle peut être transposée, comme un accord joué à un niveau ou à un autre du clavier d'un piano. Bouasse sait que cette théorie est fausse, car il a expérimenté en jouant à des vitesses différentes un même enregistrement phonographique: 'u', rejoué à

une vitesse plus grande, peut devenir 'i'. Restent deux autres théories, dont nous avons hérité.

Selon la "Théorie du renforcement", issue de Helmoltz, la voyelle est caractérisée par le renforcement d'un ou de plusieurs harmoniques de la note sur laquelle elle est émise. Cet harmonique ou ces harmoniques sont voisins de sons dont les fréquences, fixes et déterminables une fois pour toutes (vocables), caractérisent la voyelle.Les harmoniques, présents dans le son laryngien, sont renforcés par la résonance des cavités supralaryngiennes, la bouche en particulier, dont la disposition varie selon la voyelle prononcée. Bouasse juge cette théorie acceptable, mais il y fait cette objection . L'une des vocables de l'i a une fréquence supérieure à 2000Hz; si on chante 'i' sur ut_2=128, un harmonique de rang au moins égal à 16 devra être renforcé; et on peut douter que cet harmonique se trouve dans l'ut_2 avec une intensité qui permette de reconnaître un 'i'. À la vérité, quelle que soit la vraie nature du son laryngien, (son d'anche, né de la vibration des cordes, ou son de sirène, créé par les interruptions périodiques du courant d'air), ce son son primitif ne peut être que très riche en harmoniques, car, au laryngoscope, on voit que les lèvres vocales ne se séparent que pendant une faible fraction du temps et qu'elles se referment brusquement à chaque période.

Selon la théorie des formants, (due à Willis), la voyelle est caractérisée par un ou plusieurs sons de hauteur invariable (formants) qui se superposent au fondamental et aux harmoniques de la note sur laquelle elle est émise. Bouasse voit la proximité des deux théories; mais, dit-il, tandis que, selon Helmoltz, les harmoniques renforcés sont seulement voisins des vocables, chez Willis les formants sont fixes, donc généralement inharmoniques de la note sur laquelle la voyelle est émise.Quand il en vient à l'explication physique, Bouasse dit que le formant est un son de cavité buccale relancé *in tempo*, à chaque période du son du larynx. Et en faveur de l'origine buccale des sons vocaliques, il donne cet argument que dans la voix chuchotée, il n'y a pas de son laryngien, donc pas d'harmonique dont on puisse invoquer le renforcement...

En réalité, d'une part l'ensemble des voies aériennes, des poumons jusqu'à la bouche et au nez, forme un seul système physique dont on ne peut arbitrairement découpler les parties pour en décrire les oscillations. Et d'autre part, compte tenu de la validité universelle de l'analyse de Fourier, une succession de trains d'onde de fréquence f (formant) relancés périodiquement avec une fréquence f' (fondamental), constitue finalement une fonction de fréquence f', décomposable en harmoniques du fondamental. Et la voix chuchotée est produite par le filtrage d'un bruit blanc, comme la voix sonore l'est par le filtrage du son laryngien.

La forme de la cavité buccale (communiquant éventuellement avec la cavité nasale) détermine des résonances plus ou moins larges sous l'enveloppe

desquelles se logent les harmoniques d'un fondamental laryngien, ou se colore le bruit blanc produit par le frottement de l'air au niveau du pharynx. Mais on continue à parler de fréquence des formants de chaque voyelle, alors qu'il s'agit plutôt de profils spectraux, (notion qui intéresse l'analyse des correspondances, malgré toutes les précautions qui s'imposent dans le codage: cf §2.2).

Cependant, l'observateur attentif, qui suit à l'écran la déformation des quasi-périodes d'un son vocalique, en consultant de temps en temps la transformée de Fourier que lui offre Soundcap, apprend à reconnaître dans tel groupe d'ondelettes, renouvelé à chaque période, l'origine d'un formant vers 1200Hz; et dans le poudroiement qui déchiquette un sommet, un formant de plus haute fréquence. Ce qui ramène au point de vue de Bouasse sur la théorie de Willis, et incite à chercher, au delà de représentations mathématiques complémentaires, certaine partie du système physique où se produirait tel formant.

Enfin, pour l'oreille, est essentielle la non-périodicité des voyelles (qu'on peut décrire aussi comme une non-harmonicité des partiels): si l'on répète 100 fois, par une commande de Soundcap, une quasi-période découpée dans un son vocalique, on n'entend pas une voyelle, mais un timbre de sonnerie. Au contraire, il est assez facile de synthétiser une fonction numérique, qui, en s'écartant du strict schéma périodique, prenne l'apparence d'une voyelle acceptable.

3.3 Analyse mathématique du signal et notions phonétiques

Empruntons à Bouasse une définition de la phonétique qui nous paraît décrire, avec exactitude et concision, l'activité de cette science, au cours du premier quart du XX-ème siècle:

"La *phonétique* étudie les sons en tant que phénomènes linguistiques (*phonèmes*); dès la plus haute antiquité (Indiens, Grecs) on s'est efforcé de les classer. La phonétique moderne ne diffère de celle des temps passés que par l'emploi de procédés expérimentaux qu'ils ne connaissaient pas. La *phonétique expérimentale* étudie les mouvements connexes de toutes les parties mobiles laryngiennes et supralaryngiennes (langue, lèvres, voile du palais, mâchoire) dans l'articulation des groupes des sons qui constituent le langage".

Le terme de phonème reçoit aujourd'hui un sens différent. Dans l'esprit du Prince de Trubetskoj, la phonologie vise à décrire les phonèmes sans considérer le détail des sons, mais par un nombre minimum de *traits pertinents*; qui sont les caractères nécessaires et suffisants pour les distinctions que fait une langue donnée. Ainsi, en français, 's' peut être prononcé avec un défaut de langue sans qu'aucune confusion en puisse résulter; mais, en anglais, il y a deux phonèmes, puisque les deux mots 'sing' et 'thing' ont des sens différents. Quoique la détermination du système phonémique d'une langue suive, en principe, une procédure systématique, le résultat de l'opération n'est pas nécessairement

unique: c'est que l'on réunit, sous un même phonème, non seulement toutes les réalisations individuelles d'un même projet sonore, ou *phone*, mais des sons relevant de projets distincts, et que les principes de la phonologie amènent à regrouper: les *allophones* du Phonème.

A ceux qui, comme nous, déplorent de voir régner, chez les linguistes, une conception aussi schématique des sons du langage, nous proposerons d'abord, comme une justification du radicalisme de la phonologie, cette citation extraite d'un *Traité de prédication* paru en 1931 aux *Éditions du Cerf*:

"Certains auteurs se refusent à distinguer un E moyen ou un O moyen <entre ouvert et fermé>... Pour nous, non seulement il y a un *e* moyen, un *o* moyen; mais dans les phrases suivantes,...: j'avais fait ce legs à mon curé; — Octave a volé ce broc à un pauvre, les sons *e* et *o* présentent quatre timbres nettement distincts".

L'éminent auteur du Traité, est, au demeurant, bien averti des périls auxquels exposent les subtilités de la phonétique, lui qui note ailleurs qu "une oreille fine perçoit bien des nuances intermédiaires, mais qu'on ne peut préciser et qui prêteraient à des divergences d'opinions inextricables".

C'est cependant dans le fourré *inextricable* des sons naturels du langage que se trouve nécessairement engagée l'étude physico-mathématique de la parole, avec cette difficulté supplémentaire que la typologie des sons doit être faite d'après le seul signal acoustique, sans recourir aux mouvements des organes émetteurs, dont l'étude est, selon Bouasse, comprise dans la *phonétique expérimentale.*

Considérons d'abord sur quelles bases la linguistique contemporaine décrit le système des phonèmes d'une langue.

Selon Cl. Hagège et A. Haudricourt, (in *La phonologie panchronique*, p32)," On peut... dire, en simplifiant beaucoup, que, du point de vue acoustique, la succession est constituée de bruits initiaux, qui correspondent aux consonnes de début de syllabe, puis de la partie la plus audible, qui est la voyelle ou le sommet syllabique, enfin , le cas échéant, de bruits de relâchement final, qui définissent la ou les consonnes fermant la syllabe".

Si l'on suppose qu'on a pu définir conjointement, dans cette voie, la notion de syllabe et la distinction entre consonnes et voyelles, la typologie se poursuit, principalement, en termes d'organes émetteurs, et non en termes acoustiques. Les voyelles sont classées d'après la forme du résonateur buccal, (déterminée par la langue, les lèvres, la communication établie ou non avec les fosses nasales..). Pour les consonnes, on se place dans le schéma physique suivant: l'air expiré parcourt le tractus vocal en rencontrant des obstacles. Le niveau de l'obstacle principal est le point d'articulation; la consonne est occlusive ou

fricative selon qu'il y a, ou non, interruption du flux d'air par l'obstacle; elle est voisée ou sourde selon que les cordes vocales vibrent ou non.

Or, pour qui examine le signal, les distinctions majeures ne coïncident pas exactement avec celles que la phonologie met en relief. La première opposition n'est pas entre consonne et voyelle, mais entre son sourd et son voisé. Sera dit voisé un son qui, marqué par la périodicité du mouvement des cordes vocales, se présente comme une suite de ce que nous avons appelé des *quasi-périodes*, (qp). Dans la voix chuchotée, il n'y a pas de sons voisés; il y a, cependant, des voyelles et des consonnes, même si, comme le note Bouasse, on ne distingue plus entre deux consonnes qui, telles 'p' et 'b', sont normalement distinguées par le trait pertinent du voisement. D'autre part, des consonnes, 'l', 'm',... , peuvent être constituées d'une suite de qp; bien plus, si l'on écoute isolée une de ces consonnes, ou qu'on interrompt après elle la reproduction du message, on perçoit, ordinairement, une véritable voyelle. Ainsi, l'opposition physique entre voisé et non voisé reste à interpréter linguistiquement par rapport au contexte, avant qu'on puisse parler de voyelle et de consonne. Hagège et Haudricourt ne nous contrediront pas, eux qui écrivent, (*op. laud.* pp. 22-23) :

"La nécessité des voyelles,..., est liée à un fait physique: pour qu'un son soit différencié, il faut qu'il soit perçu..... la prononciation sonore et orale des voyelles correspond au maximum d'audibilité; aussitôt après les voyelles viennent, selon ce même critère, les consonnes dites sonantes, c'est à dire les semi-voyelles, les nasales et les liquides, normalement sonores."

Dans la voix usuelle, (non chuchotée), le bruit consonantique, (ie ce qui n'est pas constitué de quasi-périodes) devrait correspondre aux consonnes sourdes, occlusives ou fricatives. Nous croyons qu'il en est bien ainsi, l'opposition entre occlusive et fricative étant exprimée acoustiquement par une opposition de durée: un même bruit, de spectre donné, peut être perçu comme 'ts', 's' ou 't', selon le contexte, la variation de la puissance et la durée.

Le cas des consonnes voisées est complexe: elles consistent en une suite de quasi-périodes; et c'est pourquoi, en basse fréquence, avant 500Hz, leur spectre comporte une ou plusieurs raies. Cependant, en haute fréqence, par exemple à 3kHz, il peut y avoir non des raies (harmoniques supérieurs du fondamental), mais des pics, plus ou moins larges, de bruit. Pour les consonnes voisées, comme pour les sourdes, l'opposition entre occlusive et fricative est liée à la durée.

Enfin, il n'est pas douteux que le timbre des voyelles, comme le point d'articulation des consonnes, se manifestent par des caractères de l'enveloppe spectrale, en d'autres termes par des formants. Mais, dans les spectres de raies comme dans les spectres de bruit, il faut prendre garde à diverses formes de masquage. Un bruit de haute fréquence peut être masqué par un bruit de

fréquence plus basse; un même son vocalique hétérogène peut, selon le contexte, être perçu comme une diphtongue ou comme une voyelle unique.

Dans le cas des voyelles, il nous est apparu que le son perçu pouvait être déterminé non par la valeur centrale ou le maximum de la bande correspondant à un formant, mais par la place exacte du front de ce formant. Par exemple, 'é' et 'i' ont une bande s'étendant au delà de 2500Hz. Le front ascendant de la bande du 'é' est vers 2000Hz; celui du 'i' vers 2500Hz. Là semble être la différence pertinente; le développement de la bande vers l'aigu étant irrelevant. Du point de vue physiologique, il n'est pas invraisemblable qu'existent sur la membrane basilaire, comme sur la rétine, des détecteurs de contraste propres à distinguer ces fronts.

A défaut d'un système acoustique cohérent de la phonétique générale et de son application à la phonologie des diverses langues, nous voulons du moins présenter, au §4, quelques exemples qui nous ont suggéré les hypothèses proposées ci-dessus.

4 Analyses commentées de quelques sons

Un son de parole peut être considéré à trois niveaux: comme une production, réalisée dans le temps par les organes vocaux convenablement disposés; comme un message perçu par l'auditeur; comme un phénomène acoustique.

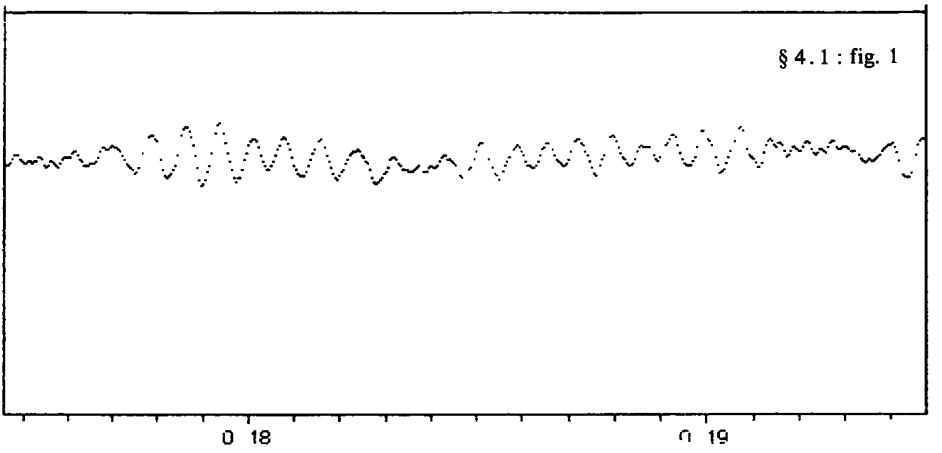
Le niveau acoustique est central: il faudrait savoir par quels traits acoustiques se manifestent les divers traits pertinents que phonologistes et phonéticiens n'ont jusqu'ici clairement décrits qu'au niveau de l'émission; il faudrait savoir suivant quelles lois combinatoires les traits acoustiques successifs sont perçus. On pourrait alors effectuer la synthèse de la parole à partir d'éléments minimaux, et non de diphones, comme on le fait principalement aujourd'hui; et la reconnaissance de la parole serait assise sur des bases analytiques certaines.

Afin d'atteindre ces objectifs, nous croyons utile d'étudier la parole avec minutie, en conjugant l'observation et l'expérience, comme le botaniste le fait pour les plantes. Les exemples qui suivent montrent comment nous procédons; ils viennent d'une Cassette Radio France intitulée: *Arletty, de Courbevoie aux enfants du paradis*, et d'une Cassette Hachette : *Paul Valéry, Pages choisies*, où *La fileuse* est lue par Marguerite Perrin.

Dans le présent §, nous n'avons pas cru possible de recourir à une notation phonétique ou phonologique arrêtée; mais, comme nous avons commencé de le faire dans les précédents §, nous prenons simplement les lettres de l'alphabet, avec leur valeur usuelle en français, pour désigner et décrire les faits observés. Nous utilisons le signe " pour citer les mots ou les syllabes; et le signe ' pour les sons ou les lettres : "très", "un", 's', 'ts', etc...

4.1 "Un très joli costume", dit par Arletty

En examinant la courbe et rejouant des segments, on peut, avec quelque certitude, interpréter en terme de lettres ou de syllabes les parties successives du signal. L'essentiel étant formé de quasi-périodes, on peut commencer par situer dans le discours les intervalles de bruit.



Un premier intervalle de bruit se place après "un"; il est formé de trains d'une dizaine d'ondes, (fig1), dont la fréquence est d'environ 1500Hz, séparés par de brefs passages de faible amplitude à une fréquence beaucoup plus élévée, (cf spectre, fig2). L'hypothèse se présente que ce bruit peut être le 'r', assez rapeux de "très": on entendrait une attaque en 't', semblable à celle d'un 's', (bruit de haute fréquence), qui, débuté sans transition, s'entend 'ts'.

Après plusieurs essais, pour distinguer dans ce bruit des éléments pertinents en le modifiant, on constate que sa mise en silence pur et simple laisse subsister le 'tr', avec toutefois beacoup moins de puissance et de netteté. Si, alors, on écoute en débutant dans le silence créé, (et non dans le "un"), le 'tr' disparaît totalement. On conclut de celà que l'interruption du son entre "un" et 'è' incite à percevoir un 'r', suggéré, sans doute, par le timbre bruité du 'è'.

Revenons au signal original, avec le bruit du 'tr' Soundcap permet de permuter ou de répéter les tranches plus ou moins fortes de ce bruit. En plaçant devant 'è' une brève tranche de bruit fort, on peut entendre "tè", sans le frottement du 'r'. Au contraire si on parvient à faire croître le bruit lentement et régulièrement, l'attaque en 't' disparaît.

Joué à l'envers, le signal s'entend "èrun", sans 't'. Pour entendre "ètrun", il suffit d'interposer après le "un" une assez longue plage de silence, (cf "téré" et "tété", au §4.4).

On a un deuxième intervalle de bruit à l'intérieur du mot "costume"; ce bruit qui, sur la courbe, apparaît comme une bande poudreuse dessinant une sinusoïde, est le 's'. Mettons-le en silence: on entend "cotume". Où est le 't'? Si on débute l'écoute après "co" on entend seulement "ume". Il apparaît donc que l'interruption entre 'o' et 'u' suffit ici à faire percevoir un 't'.

Entre "très" et "oli" le son 'j' occupe quelques quasi-périodes dont la courbe est denticulée. Un de nos programmes, (appelé "rat", parce qu'il façonne une courbe par la souris), permet de modifier la courbe, (en cliquant sur des triplets de points pour introduire des arcs de parabole bien lisses), et d'entendre le son correspondant. Si l'on polit les premières qp seulement, on entend la consonne 'd'; si on les polit toutes, on a "boli".

Le début du mot "costume", est marqué par un train d'onde unique, (assez semblable à ceux qui précèdent "très"), placé entre deux plages de silence, la première longue, la seconde brève. Nous avons généralement retrouvé un tel train d'onde pour les 'k' ou les 'c', non seulement du français mais d'autres langues.

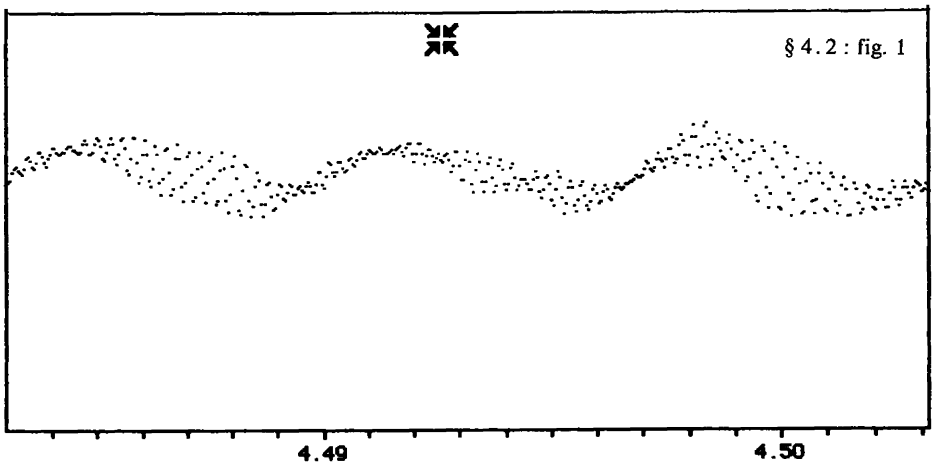
4.2 "zir" dans "plaisir"

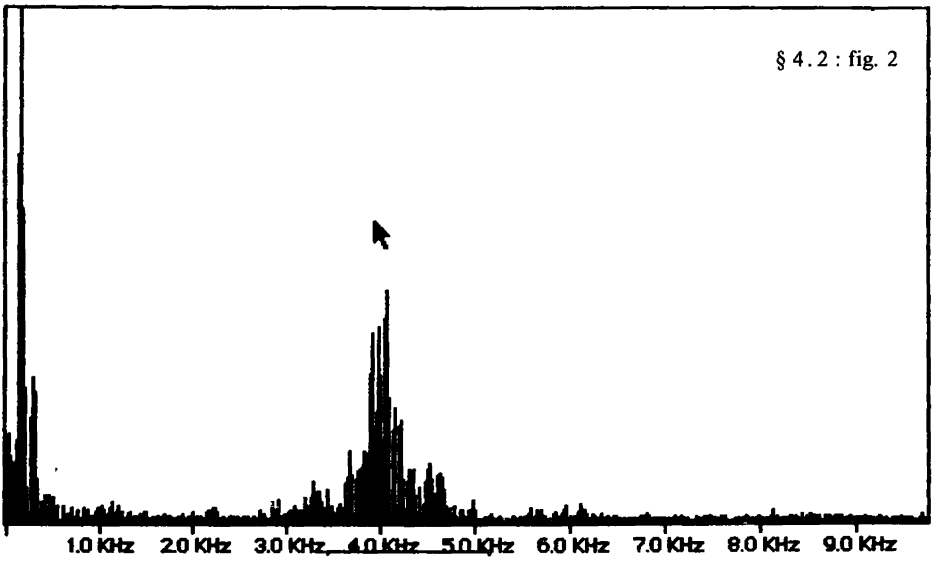
La syllabe "zir", sur laquelle nous concentrerons notre attention, peut être décomposée en quatre parties , que nous désignerons par les lettres 'z', 'i', 'è', 'r'; le choix de ces lettres permettant de décrire ce que l'on perçoit en écoutant tout ou partie de la syllabe joué en sens directe ou inverse.

La courbe du 'z' est une sinusoïde bruitée, (fig1). L'analyse spectrale, (fig2), sépare nettement un fondamental pur, presque sans harmonique, vers 150Hz, d'un monticule de bruit, qui culmine à 4kHz, mais s'étend de 3 à 5kHz. Cependant, sur la courbe, le bruit n'apparaît pas simplement superposé au

fondamental vocalique, et comme indépendant de celui-ci: il y a, sur chaque période de la sinusoïde, comme un ventre de bruit après un silence. Ceci nous rappelle une proposition de Bouasse: le bruit serait relancé *in tempo* par le fondamental.

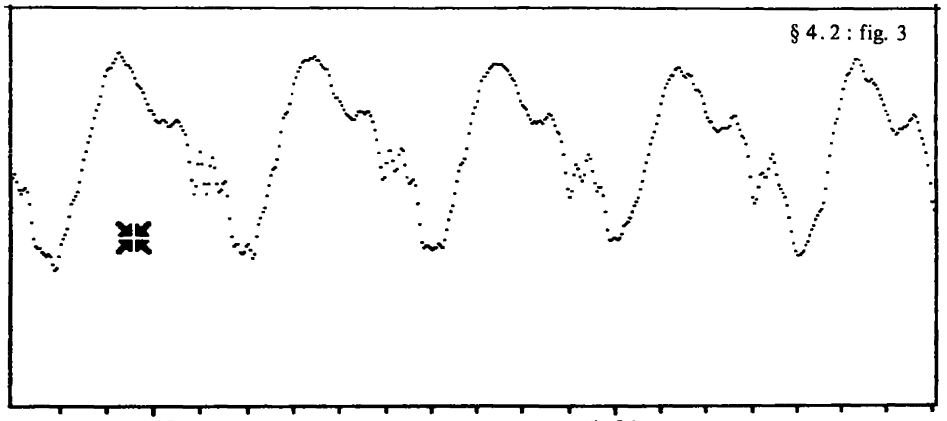
L'écoute partielle permet ici de confirmer la parenté, affirmée par les linguistes, entre 'd', dentale occlusive sonore; et 'z', dentale fricative sonore. Si

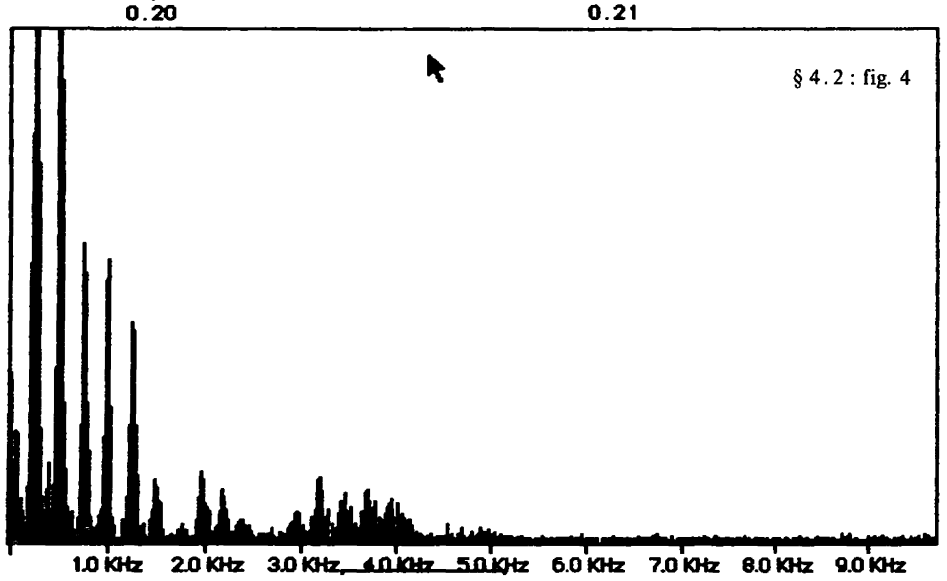




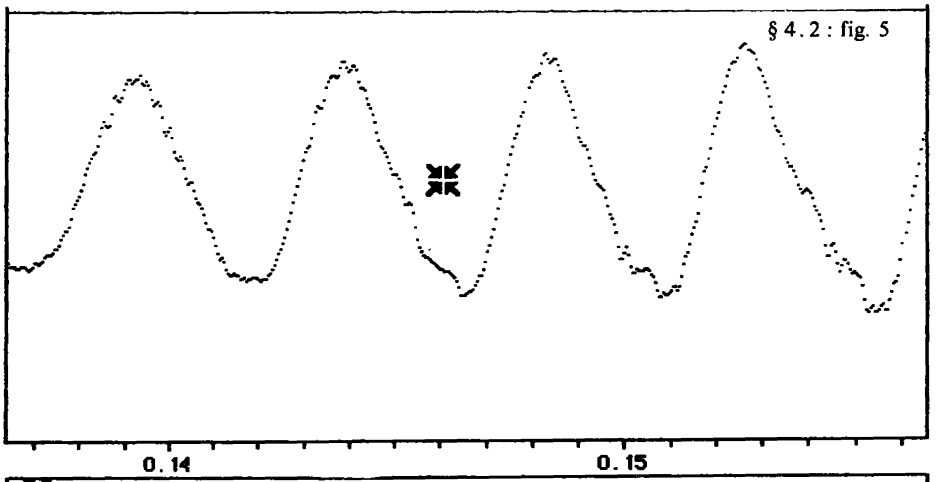
l'on joue une seule quasi-période du 'z', on entend "dir": ainsi, au niveau acoustique, l'opposition entre occlusive et fricative équivaudrait à une opposition de durée. L'effet est peu modifié si on prolonge la courbe du 'z' en répétant 6 fois l'une de ses dernières qp, particulièrement bruitée: selon le nombre de qp joué on a 'zir' ou 'dir'.

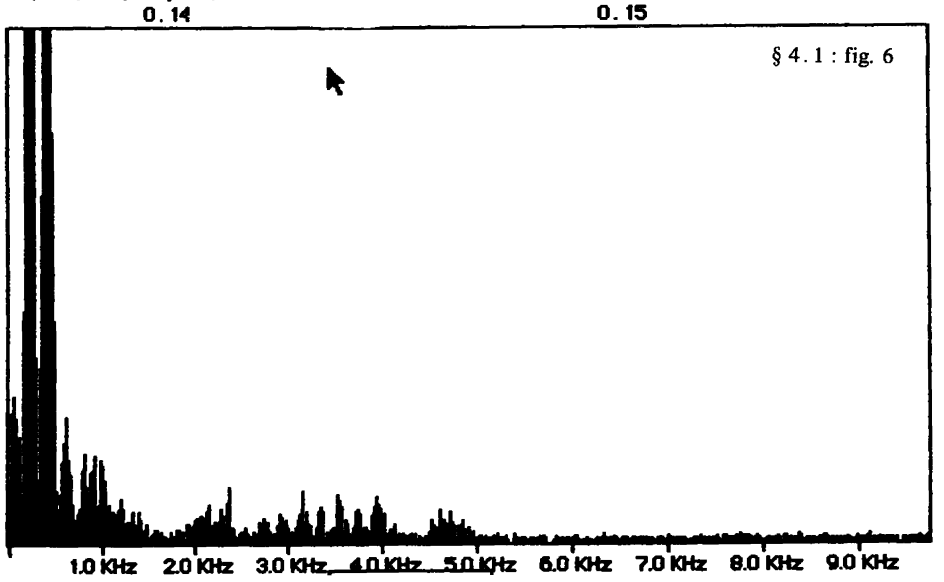
Après les deux voyelles 'i' et 'è' que nous décrirons ensuite, le 'r' se présente comme un bruit, succession assez hétérogène de trains de dents de scie. Cette fricative sourde ressemble au 'tr' de "un très joli"; et en effet, à l'écoute inverse, on entend "tréz".





Si, à l'écoute directe, on supprime le "r" final, on a la surprise de percevoir un "ziè", au lieu du "zi" attendu. La plage vocalique notée 'i' 'è' a une courbe mélodique très marquée, avec une descente de plus de 2 tons suivie d'une remontée égale: la durée des qp part de ≈85, atteint ≈115 avant la fin du 'i' puis revient à ≈88 à la fin du 'è'. La lettre 'è', (fig3 et 4), décrit ce qu'on perçoit quand ce son est entendu seul. Au contraire, la valeur du son noté 'i', (fig5 et 6), n'est pas nette; mais en composition il vaut 'i' ou jod.





A l'écoute inverse, on a, symboliquement, la succession 'r' 'è' 'i' 'z'. L'ensemble est perçu comme "tréz"; en débutant à la fin du 'r', on a "téz". En supprimant 'z', on a "tré", (ou "té"). En supprimant 'i' et 'z', on a "trè". Ainsi, la succession 'è' 'i' est perçue 'é'.

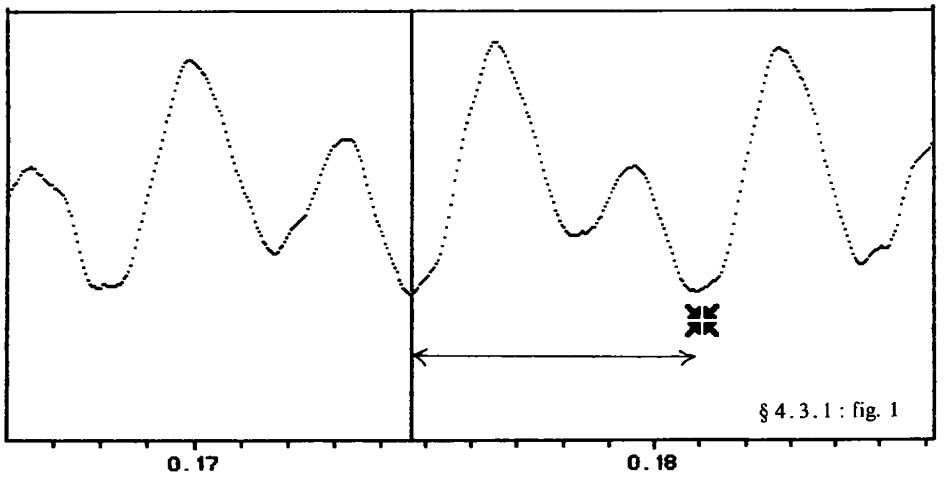
écoute directe "z"i"è"r"		"r"è"i"z" écoute inverse		
ziè	zir	trè	tré	tréz
diè	dir	tè	té	téz
iè	ir	è	é	éz

Le tableau ci-joint schématise tout ce qu'on a perçu, dans l'écoute directe ou inverse, en tranchant, éventuellement, des segments au début ou à la fin.

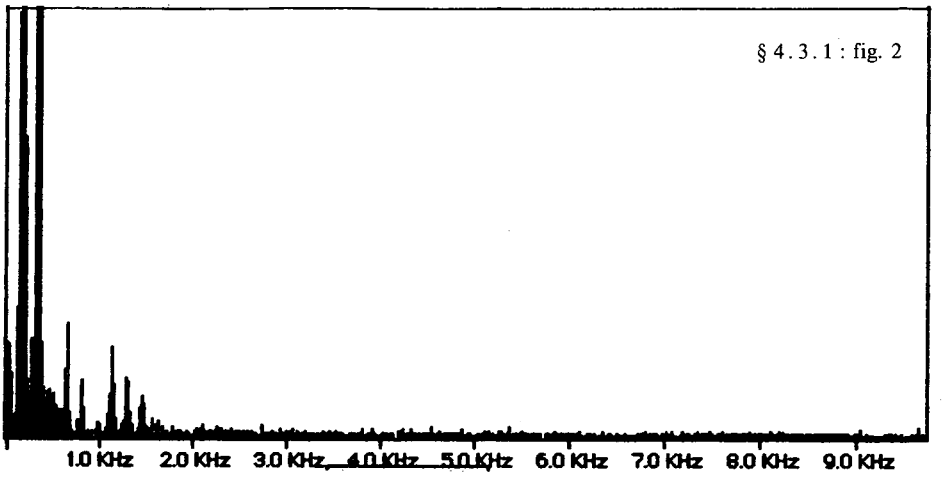
4.3 'm' naturel et 'm' artificiel

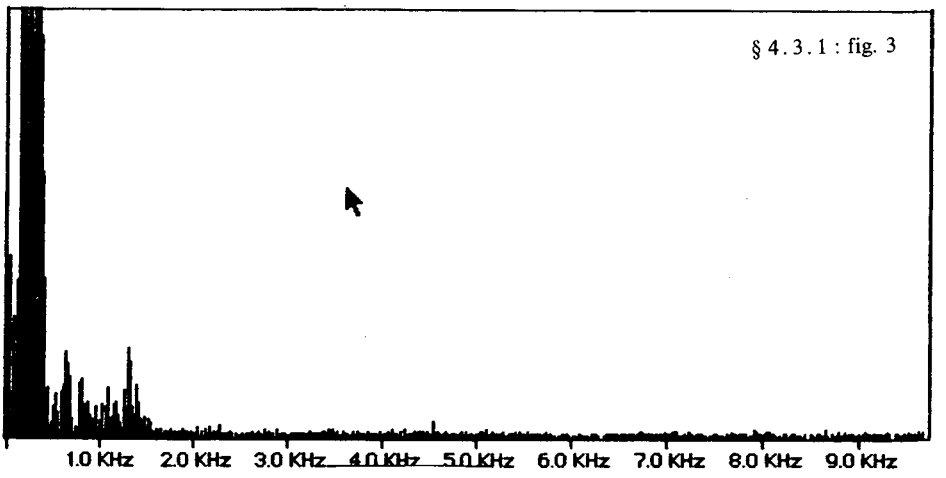
4.3.1 'm' dans "en tandem avec moi"

Ce 'm' apparaît comme une suite de quasi-périodes vocaliques, (fig1), dont certaines sont plus ou moins bruitées; le spectre, (fig3), n'est pas net: il comporte une bande puissante avant 400Hz et quelques pics de bruit jusqu'à 1500Hz. Afin d'éprouver si cette irrégularité importait à la perception du 'm', nous avons substitué à celui-ci la répétition d'une des qp les moins bruitées, (marquée sur la fig1): aucun changement perceptible n'en est résulté. Il faut toutefois noter qu'il ne s'agit pas d'une répétition rigoureuse, mais de qp raccordées sur l'écran par collage, grâce au Soundcap; et c'est pourquoi les raies du spectre de notre 'm' artificiel, (fig2), ont une certaine largeur. Entendu seul, le 'm' artificiel, suite de qp, est perçu comme une voyelle: le timbre pourrait être 'œ' ou 'u'.



§ 4.3.1 : fig. 1



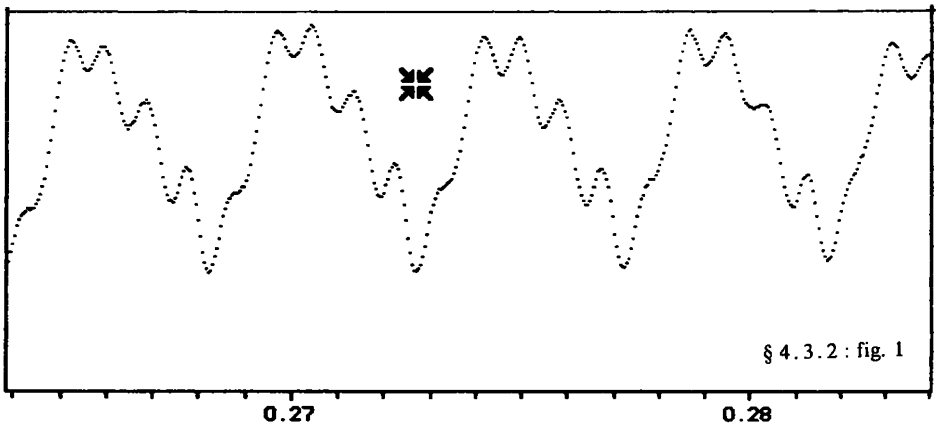


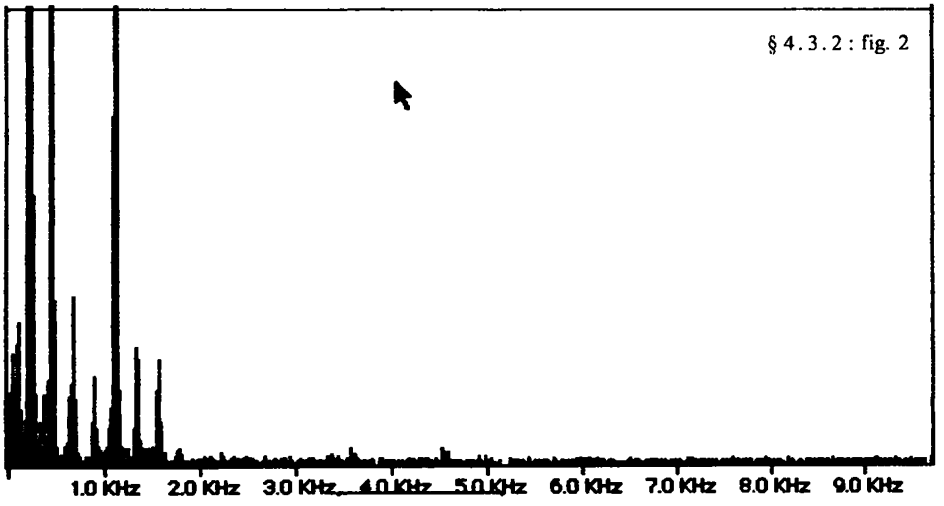
4.3.2 'm' dans "je lui ai dit: et moi"

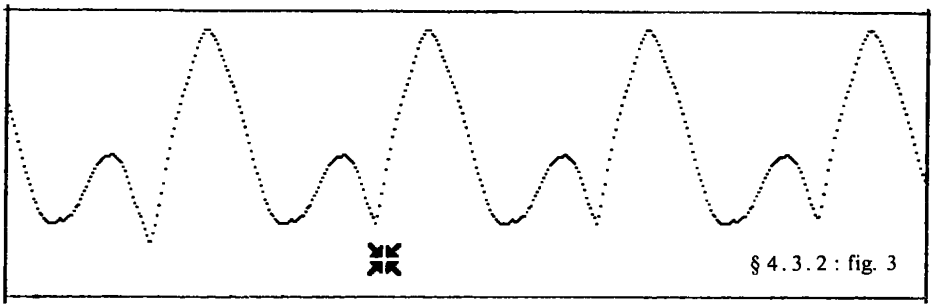
Voici un 'm' bien différent du précédent quant au dessin, (fig1, page suivante), non bruité et régulier, même si les qp successives se déforment: le spectre, (fig2), offre une succession de raies très nettes. On a tenté de substituer à cet 'm', par collage, le 'm' artificiel du §4.3.1. Tel quel, le résultat est mauvais: les qp du 'm' artificiel sont trop longues pour le contexte où on les insère; il se produit une brusque descente de l'intonation. Afin de corriger ce

défaut, on a, sans précaution particulière, retranché, (par la fonction "effacer" de Soundcap), une partie de chaque qp. La courbe du son ainsi modifié, (fig3; où on reconnaît une partie de la période marquée sur la fig1 du §4.3.1), présente des angles aigus; mais la qualité du produit n'en est aucunement affectée: on entend un "èmoi", que nous ne savons pas distinguer du son naturel. La fig4 donne le spectre de cet 'm' substitué, qu'on comparera à ceux d'autres 'm'.

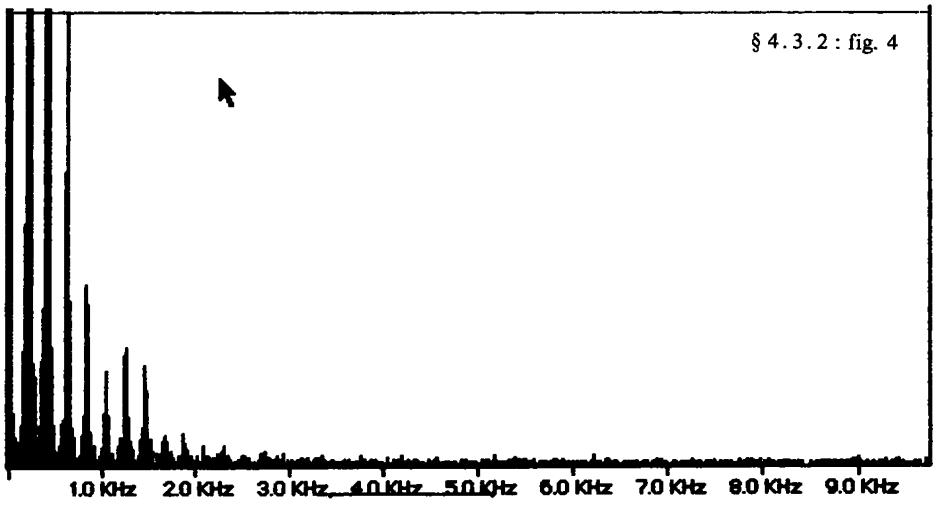
Une copie d'écran, (fig5), montre à une échelle comprimée les deux "èmoi", destinés à être entendus l'un après l'autre, voire répétés alternativement plusieurs fois, afin de les comparer. A l'œil, le signal artificiel se distingue par le bloc rectangulaire des qp répétées presque sans variation.



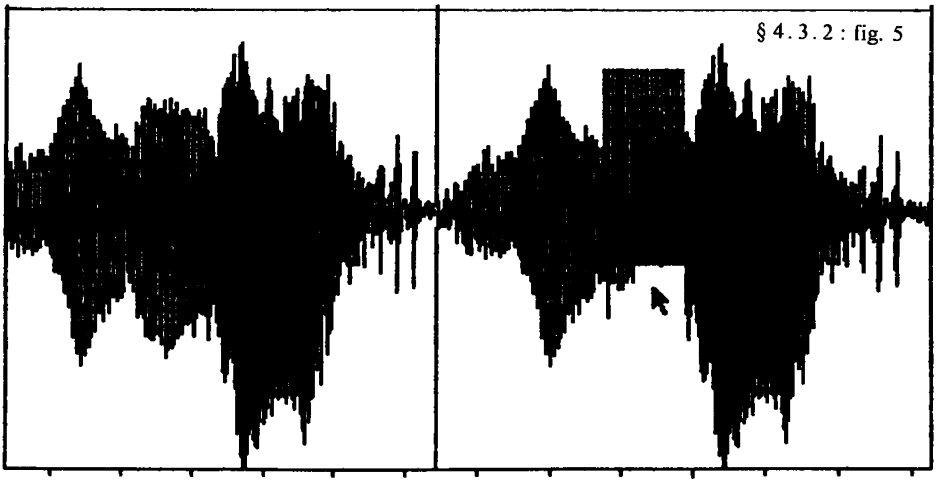




§ 4.3.2 : fig. 3



§ 4.3.2 : fig. 4



§ 4.3.2 : fig. 5

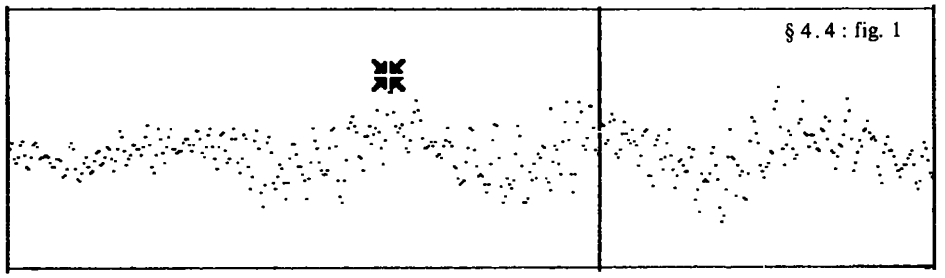
4.4 "Mystérieuse" par Marguerite Perrin

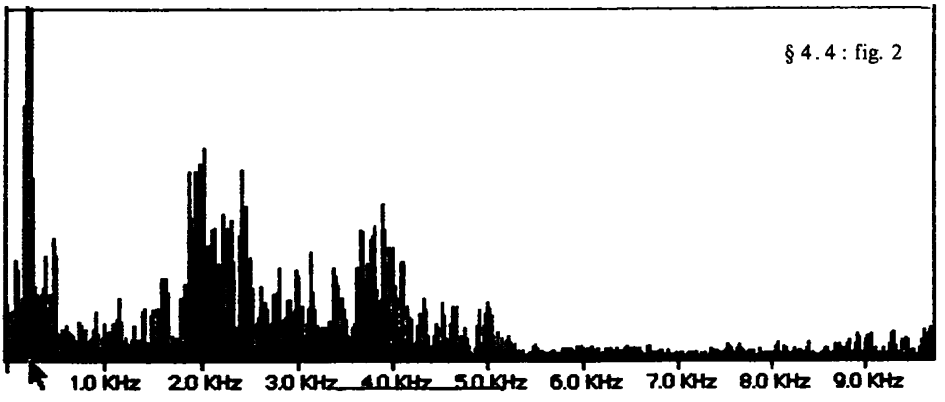
Nous ferons seulement quelques observations sur le segment "térieuse".

Nous avons entendu "téré" pour la syllabe "té" répétée deux fois de suite, sans interposer de silence; et "tété", en interposant un silence. A l'écoute inverse, on perçoit "zœiré": mis à part le 't' initial, qu'on devine peut-être comme un 's' très bref, la réversibilité est parfaite, (une fois n'est pas coutume).

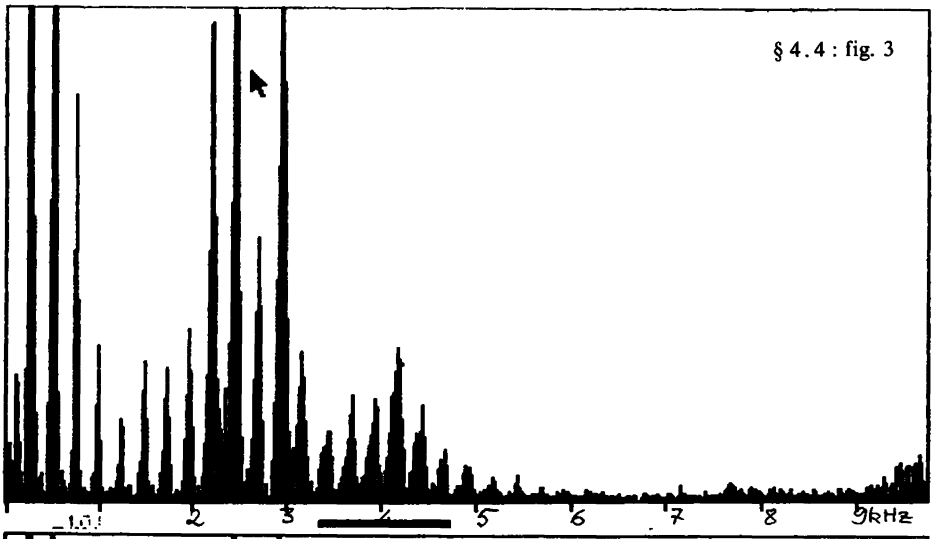
On peut observer, sur le son 'z', ce qui a été vu au §4.2: courbe semblable, à ceci près que la fréquence du fondamental est plus aiguë, à 220Hz : 10 périodes successives d'une sinusoïde très bruitée durent un peu moins de (1/20)s. Si, en écoute inverse, on débute après 6 périodes, on entend "dœ".

Le 'r' diffère de ceux rencontrés chez Arletty, ("très", "plaisir"): c'est, comme le 'z', une sinusoïde bruitée, (fig1); autrement dit, une fricative voisée; mais avec une importance relative du fondamental qui semble moindre, et un autre timbre de bruit, (fig2). Afin d'expérimenter sur cette fricative, partons du son inversé. On peut, par les commandes du Soundcap, multiplier le 'r', (le prolonger), lui donner une courbe de puissance, (en prenant soin d'éviter les modulations insolites qui perturbent la perception: cf §4.5). Le son vocalique 'é' précédé de 'r' prolongé à puissance constante, s'entend "kré"; 'i' suivi d'un silence puis d'une attaque brusque du 'r', avec ensuite une puissance décroissante s'entend "ik"; avec le 'é' on a "iké". Il apparaît que, par le spectre du bruit, ce 'r' est une gutturale.

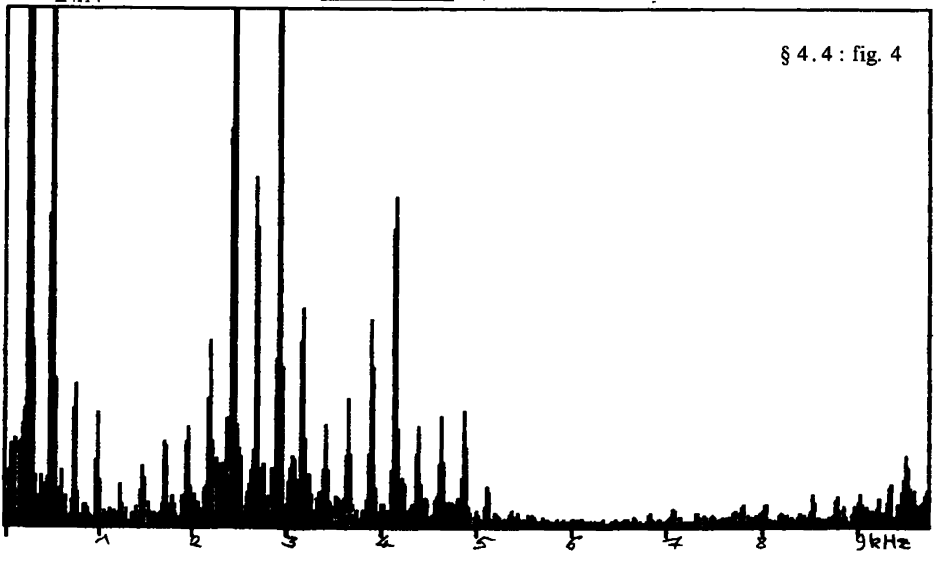




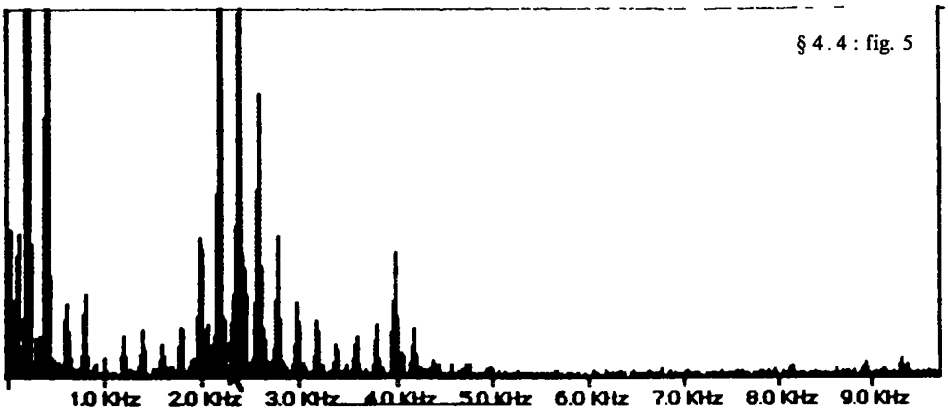
Il est bon de comparer les spectres des deux voyelles 'é' et 'i', (de "térieuse"). Tous deux, (fig 3 et 4), présentent trois formants: le premier avant 1kHz, le second entre 2kHz et 3kHz; le troisième vers 4kHz. Compte tenu d'autres enregistrements, la différence pertinente nous paraît être dans la largeur, plus grande pour 'i', de l'intervalle qui sépare les formants 1 et 2; et, plus particulièrement, dans le front initial du formant 2, (vers 2200Hz pour 'é' et vers 2500Hz pour 'i').



§ 4.4 : fig. 3



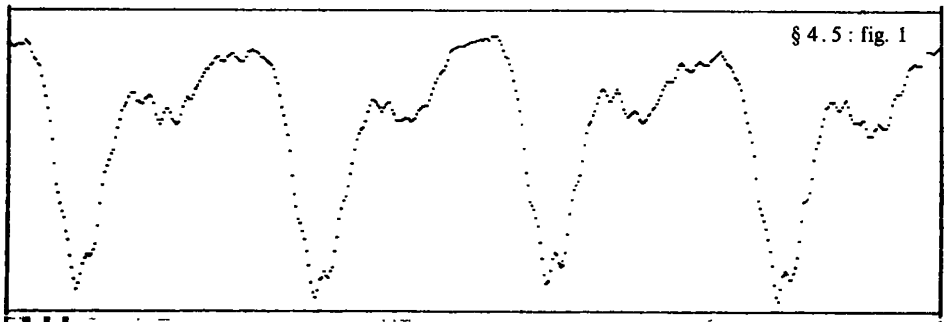
§ 4.4 : fig. 4

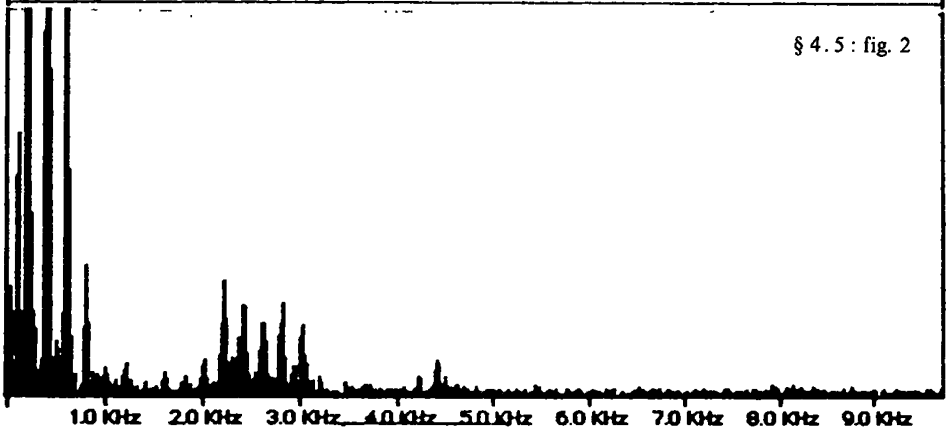


D'autre part, la voyelle 'ü' peut, elle aussi, présenter un spectre à trois formants, (fig5): le formant 1 se terminant à 500Hz; éventuellement le formant 3 manque.

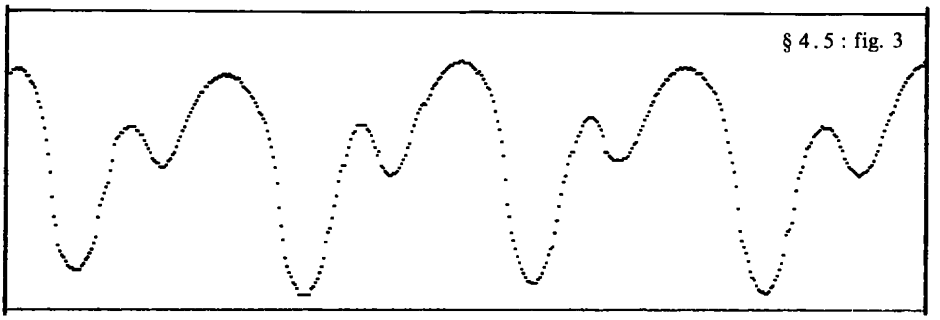
4.5 Courbe du son naturel et courbe arrondie

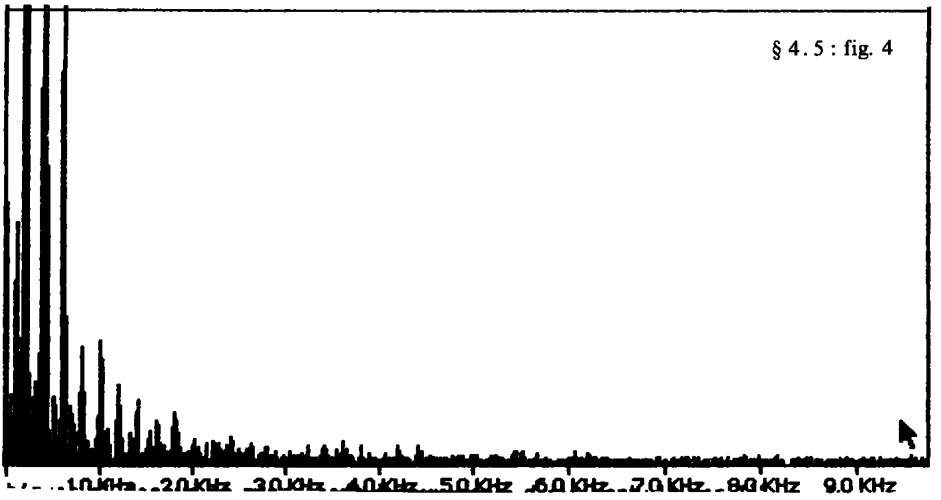
Les graphiques considérés se rapportent à deux sons: d'une part, (fig1, 2), le son 'é', extrait d'une lecture par Marguerite Perrin de "la fileuse", de Paul Valéry; d'autre part, (fig 3, 4), le son 'é*', obtenu en arrondissant la courbe de 'é' par le programme "rat", (déjà cité, §4.1); ce dernier son s'entend plutôt 'eu'.





Si l'on tente de comparer les deux sons en les faisant jouer immédiatement l'un après l'autre par concaténation, on entend un effet de roulement qui pourrait être comparé à un chant d'oiseau. Cet effet n'est pas dû à la différence entre 'é' et 'é*', comme on peut s'en assurer en juxtaposant deux exemplaires de 'é*', ce qui produit le même effet. Mais nous l'attribuons à la mélodie: sur un exemplaire de 'é', ou de 'é*', la durée des quasi-périodes décroît quelque peu: la première durant 115 points, les suivantes 112, les dernières 109, (le tout couvrant 1390 points, ie 1390/ 22000 s).





Afin de vérifier cette hypothèse, on a concaténé 4 exemplaires de 'é*', en veillant seulement à ce que les raccords aient la forme de périodes normales: d'où le fichier 4é*p, (p = pur). Ce fichier produit un net effet de roulement. On a ensuite retouché un exemplaire de 'é*' par coupure, voire par collage, en retranchant ou insérant des points isolés, pour donner à toutes les périodes une durée de 110 points(?). Puis on a créé 4é* par concaténation de 4 exemplaires de 'é*' ainsi retouché: un certain roulement subsiste, mais bien moindre que dans 4é*p. On peut également répéter 50 fois une quasi-période de é*: ainsi, il n'y a plus aucun roulement.

Si on répète plusieurs fois 4é*p ou 4é*, on entend un martellemnt correspondant au raccord de 2 exemplaires successifs; pour supprimer ce martellement, il suffit de veiller à ce qu'au niveau du raccord se forme une quasi-période acceptable; ce que nous avons fait. Il ne subsiste alors qu'un roulement, sans martellement.

5 Remarque finale

Nous n'avons pu, en quelques dizaines de pages , présenter la totalité de nos observations, qui elles mêmes sont loin de constituer un parcours d'ensemble des sons de la voix humaine parlée. Nous espérons du moins avoir convaincu quelques lecteurs de ce que l'on peut aujourd'hui construire une phonologie fondée, non sur les traits pertinents du processus d'émission, mais sur ceux du signal acoustique.