LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

À la recherche de fondements scientifiques pour la reconnaissance de la parole

Les cahiers de l'analyse des données, tome 14, n° 1 (1989), p. 99-116

<http://www.numdam.org/item?id=CAD_1989__14_1_99_0>

Les Cahiers de l'Analyse des Données
Vol XIV - 1989 - n°1 - pp. 99-116.

À LA RECHERCHE DE FONDEMENTS SCIENTIFIQUES POUR LA RECONNAISSANCE DE LA PAROLE

[FOND. REC. PAROLE]

J.-P. BENZÉCRI

1 État de la technique et base scientifique

Il est à peine utile de souligner ici l'extrême difficulté du problème de la reconnaissance de la parole: si la demande pressante d'utilisateurs potentiels ne stimulait les réalisations techniques, sans doute s'accomoderait-on présentement de couvrir les balbutiements de la technique sous l'ample manteau d'un élégant discours théorique, en mettant à contribution les ressources de plusieurs sciences, linguistique, acoustique, psychologie, analyse mathématique...

Mais les spécialistes affirment aujourd'hui que dans la communication entre homme et machine il peut être très utile que celle-ci reçoive de celui-là des commandes vocales; car, d'une part, disposer de son moyen d'expression le plus naturel est pour l'homme, en toute circonstance, un confort appréciable; et d'autre part, dans des cas critiques, de cet avantage peut dépendre le salut...

C'est pourquoi de puissants moyens sont consacrés à la reconnaissance de la parole, notamment en vue d'applications aux systèmes d'armes.

On ne saurait dire qu'on ait ainsi pu prouver le mouvement en marchant: dans la pratique, les machines ne reconnaissent les paroles que comme des signes isolés, provenant d'un locuteur déterminé. Un programme considéré comme faisable, et effectivement réalisé encore qu'avec une fiabilité insuffisante est celui d'une machine adaptative qui, après quelques séances d'apprentissage avec un utilisateur particulier, reconnaît une, voire plusieurs centaines de mots prononcés par celui-ci, isolément, ou, du moins, en ménageant des intervalles de séparation.

Ce n'est pas ici notre propos de spéculer sur l'intelligence artificielle appliquée à la compréhension du discours, même si celui-ci est écrit en clair, et non prononcé de façon plus ou moins confuse. Il est vrai que, dans la vie courante, l'intelligence du contexte est indispensable pour traduire en une suite de formes lexicales une suite imparfaitement articulée de sons; mais, d'une part,

on peut, pour commencer, se placer dans l'hypothèse d'une diction correcte; et, d'autre part, pour défectueuse que soit la prononciation non soignée, on admet généralement qu'elle ne diffère de la diction correcte qu'en ce qu'elle supprime certaines lettres, en abrège d'autres, procède à des substitutions involontaires...

Le problème se trouve ainsi posé de transcrire automatiquement toute chaîne de parole continue en une suite d'éléments, correspondant à ce que l'usage de l'écriture nous a habitués à appeler lettres; et à propos desquels les moins férus de linguistique admettent qu'il pourrait être plus exact d'user du terme de phonème. Au point où nous en sommes, le terme de lettre, manifestement inadéquat, a du moins le mérite de ne pas laisser présumer résolu un problème dont la difficulté semble s'accroître au fur et à mesure des efforts qu'on lui consacre; et il nous est agréable de rappeler que les grammairiens grecs utilisaient le mot στοιχειον, élément, pour désigner les lettres.

Tel est donc, selon nous, le problème scientifique fondamental que la technique de la reconnaissance de la parole nous presse instamment de résoudre: définir les éléments du discours, en faire l'inventaire; et ce de telle sorte qu'il soit possible de découper en segments successifs étiquetés ce phénomène physique continu que produit un sujet qui parle; quitte à aboutir à un inventaire inattendu; qu'il restera à interpréter en termes linguistiques usuels.

2 Une hypothèse de découpage: phones et phonèmes

Placé devant le problème du découpage, on songe d'abord à recourir à l'expérience des linguistes; les plus qualifiés de ceux-ci étant certainement ceux des phonologues qui, ayant la pratique de la linguistique de terrain, ont eu plusieurs fois l'occasion de résoudre, pour des langues ou dialectes dépourvus de tradition orthographique ou orthépique, ce même problème au niveau paradigmatique; c'est-à-dire d'établir la structure du système des phonèmes, même si au niveau syntagmatique, au niveau de l'enchaînement dans le discours, il reste, de l'aveu de tous, beaucoup à faire.

Il importe de nous arrêter ici, d'une part, pour tenter de marquer certaines contraintes que la phonologie ignore, mais qui s'imposent nécessairement quand on cherche à découvrir ce qui est indispensable à la reconnaissance scientifique de la parole; et d'autre part, pour conjecturer ce que, selon la phonologie, on pourrait trouver sous ces contraintes. Ayant toujours manifesté, dans plusieurs exposés de linguistique générale placés en introduction à des travaux de statistique, une certaine réserve vis-à-vis de toute démarche réductrice, nous ne sommes que mieux placé pour avouer que ce que le contact direct avec le signal nous a appris s'est trouvé contredire ce que nous attendions avec confiance de la phonologie.

Une première limite entre la phonologie et la reconnaissance de parole est celle du domaine des faits sur lesquels il est permis de se fonder.

Répondant, dans une communication personnelle, à la conclusion de notre [SPECTR. STAT. VOIX], Claude Hagège s'interroge:

La phonologie fondée ainsi sur les traits acoustiques est-elle *différentielle*? Le linguiste doit traiter l'encodage (parcours onomasiologique) autant que le décodage (parcours sémasiologique). L'auditeur interprète des signifiants que son oreille perçoit et il leur fait correspondre des signifiés; mais le locuteur coule les sens dans des articulations sonores et donc il peut bien définir les formes internes de production aussi.

Il est clair que dans le cadre de la phonologie usuelle, se borner aux traits acoustiques serait une mutilation inopportune; et de même considérer la réception à l'exclusion de l'émission. Mais dans la reconnaissance automatique de la parole, (techniquement non directement dépendante de la synthèse, d'ailleurs plus accessible), il ne peut s'agir que de trouver dans le signal sonore, seul disponible, (à l'exclusion d'autres informations sur l'état de l'appareil phonatoire, éventuellement captées par ailleurs), une structure qui se laisse finalement recoder en unités linguistiques usuelles.

Le fait que les messages parlés sont destinés à des auditeurs qui parlent eux-mêmes implique seulement que les similitudes et les différences acoustiques pertinentes sont bien autres que celles qu'un examen sans *a priori* du signal conduirait à présumer. Nous pensons par exemple que le fait que {b, p, f}aient dans leur mode de production une similitude, une proximité, beaucoup plus grande que celle des sons produits oblige à considérer ceux-ci d'un point de vue particulier.

Un autre problème de limite que rencontre la reconnaissance de la parole et qu'ignore la linguistique classique est celui de la frontière entre parole et non-parole; problème analogue à celui que reconnaît le musicologue Herbert Schindler, quand, après avoir évoqué les notions classiques de fréquence et d'amplitude du son musical, il écrit:

Allerdings ist eine scharfe Grenze zwischen Ton und Geräusch in der Praxis gar nicht zu ziehen.

Volens nolens, l'homme avec la machine se trouvent plongés dans un univers sonore où semblable question cesse d'être purement spéculative.

Ces réserves faites, voici ce que nous proposions dans [VOIX]:

... Il s'agit d'étiqueter des segments successifs de la chaîne parlée. T. Moussa a obtenu sur la segmentation des résultats assez bons. Mais le fond du problème est que ni le découpage ni l'étiquetage ne peuvent correspondre à la transcription classique du discours en une suite de phonèmes... Que le f soit une entité du système de la langue française est une chose; qu'il soit prononcé v, ou p, ou p-h, ou disparaisse en est une autre: seul nous intéresse pour l'instant ce qui est prononcé...

... Selon la terminologie des linguistes, de même qu'on appelle *phonèmes* les unités du projet sonore entre lesquelles existent des différences pertinentes pour le sens, il convient d'appeler *phones* les unités de la réalisation sonore. Cette partition de l'ensemble infini des timbres produits, en un ensemble fini de phones donne à réfléchir! Pourquoi faire une partition plutôt que d'accepter un continuum?

... Au fond, ce problème de la partition a une importance essentielle... ; mais techniquement on est assuré *a priori* qu'il pourra toujours être résolu par une classification automatique qui donne un système fini de classes aux centres desquelles on puisse rattacher tout timbre nouveau, que ces classes soient nettement séparées ou contiguës entre elles.

Nous suggérions qu'on pourrait ainsi discrétiser utilement tout discours à condition de vérifier:

que les classes obtenues s'interprètent bien phonétiquement; et, surtout, qu'en assimilant chaque timbre au timbre moyen de sa classe, on a un codage discret du timbre qui permet une synthèse satisfaisante; la reconstitution d'un discours qui, pour l'oreille, équivaut au discours initial.

Dans [BERGSON], commentant cette proposition du philosophe:

... Mais la vérité est que chaque surcroît d'excitation s'organise dans les excitations précédentes, et que l'ensemble nous fait l'effet d'une phrase musicale qui serait toujours sur le point de finir et sans cesse se modifierait dans sa tonalité par l'addition de quelque note nouvelle.

nous en trouvions l'illustration dans le traitement numérique du signal acoustique:

Le signal sonore est une fonction scalaire du temps: une grandeur physique unidimensionnelle, la pression, porte sur le tympan la parole et la musique. Toutefois, ce n'est pas en tant que phénomène unidimensionnel que le nuage sonore est perçu, car la parole... est codée dans l'oreille interne avant de subir un traitement élaboré dans le cerveau. Même si le codage physiologique n'est connu qu'en partie, ce qu'on en sait suffit à inspirer au mathématicien de calculer par l'intégrale de Fourier un spectre instantané du signal.

Ici, nous rencontrions le dilemne, bergsonien selon nous, d'une largeur de l'instant, car:

En réalité un telle transformation ne peut être à proprement parler instantanée: le calcul résulte d'un compromis entre la précision avec laquelle on veut connaître l'enveloppe du spectre en fonction de la fréquence et la précision avec laquelle on veut le situer dans le temps.

La pratique de l'observation de la parole enregistrée nous a montré qu'ici nous ne faisions pas la part assez belle à Bergson! Il nous est apparu que les

éléments perçus de la parole étaient sentis non consécutivement, mais plutôt simultanément, parallèlement; des traits de plusieurs spectres consécutifs étant combinés dans la perception d'un phonème; et un même spectre pouvant contribuer à plusieurs phonèmes consécutifs. L'écoute de tranches de son dont on déplace, à volonté, les bornes, confirmant, au-delà de toute attente que le son "sans cesse se modifie(rait) dans sa totalité par l'addition de quelque note nouvelle".

3 Phonation selon Raoul Husson et voisement

Dans la communication personnelle déjà citée, Cl. Hagège, ayant trouvé évoqués par nous, les travaux de R. Husson, les apprécie en ces termes:

C'est le type même des recherches que des neurohistologistes devraient poursuivre. Cependant dans la suite de l'article tu ne te sers pas plus avant de ses résultats, probablement parce qu'ils n'ont pas été vérifiés plus tard par des travaux effectués selon la même ligne.

Tel semble être le cas en effet. Les recherches neurohistologiques sont encore plus complexes que les recherches acoustiques, praticables aujourd'hui sur un ordinateur personnel: il n'est donc pas opportun de faire dépendre celles-ci de celles-là. Nous n'aurions pas cité Raoul Husson si, abstraction faite de ses conclusions personnelles, la question qu'il posait en termes suggestifs "la voix sort-elle d'un instrument à anche ou d'une sirène", ne se trouvait rendue à l'actualité quand on cherche le corrélat acoustique de ce que la phonologie appelle voisement.

Répétons brièvement ce qui est dit dans [SPECTR. STAT. VOIX], §3.2: un signal de parole comprend des suites de quasi-périodes et des bruits nullement périodiques (ou plutôt, sans composante périodique comparable au fondamental de la voix). Il est d'ailleurs essentiel, pour la distinction cherchée ci-dessus entre parole et non-parole, de noter au passage que si la *périodicité* du signal est un trait pertinent dont la valeur linguistique est indiscutable, son caractère *approximatif* est requis pour que soit perçue une voix et non un timbre de sonnerie. Ceci dit, il apparaît que la quasi-périodicité s'étend, des voyelles, non seulement aux consonnes liquides, (r, l, m, n), mais au-delà.

Un cas particulièrement frappant, en ce qu'il met en cause la définition même d'une opposition dichotomique, est celui de la 'ceta' espagnole. Par l'histoire et l'orthographe, ce son, généralement reconnu pour non voisé, se rattache au 'zed' qui est voisé. La diachronie, ici comme ailleurs, oblige à postuler une zone de transition. La synchronie dialectale également: un linguiste (dont nous regrettons de ne pas connaître le nom) travaillant dans l'aire du franco-provençal, nous a dit que les variations de prononciation de village à village étaient telles que des locuteurs pouvaient se vanter de faire entre la prononciation des deux sons une distinction absente du parler d'une localité voisine; cependant que le linguiste ne l'entendait ni ici ni là.

On pourrait penser qu'il s'agit d'une simple variation dans la puissance relative de la composante de basse fréquence du spectre. Mais l'examen des courbes suggère tout autre chose: le 'c' du mot 'gracias' montre une courbe quasi périodique aussi ample que le 'z' d'une autre langue; la différence est dans le caratère très anguleux de la sinusoïde approximative de 'gracias'. Il n'est d'ailleurs pas très vraisemblable qu'une distinction pertinente se fonde sur une différence de puissance, dans la mesure, où, ainsi que nous l'avons noté, la perception de la parole s'accommode bien d'un contrôle de tonalité qui bouleverse les proportions relatives des composantes du spectre.

Il y aurait donc des signaux quasi périodiques voisés; et d'autres voilés... les premiers rentrant dans la conception classique d'un appareil phonatoire, instrument à anche; et les seconds requerrant un mécanisme sous-jacent plus persistant, qui pourrait être la *sirène* de R. Husson. Il apparaît, en tout cas, qu'une notion de base de la phonologie se révèle, pour l'acoustique fort complexe.

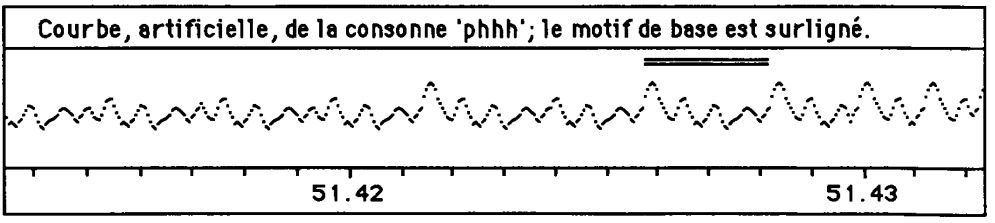
4 Perception recombinée de traits vocaliques et consonantiques

4.1 De la voyelle 'u' à la diphtongue 'ui'

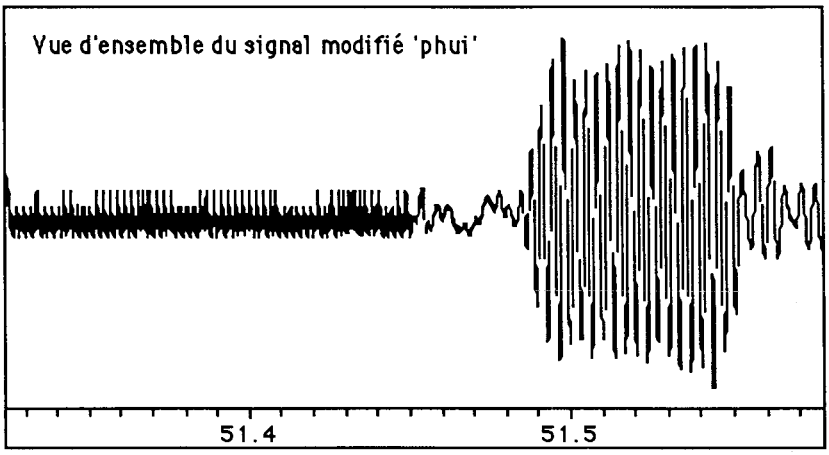
L'expérience qui nous a fourni le présent exemple part de la syllabe 'cri' du mot "écrite", saisi dans l'enregistrement d'une conférence.

Dans une certaine mesure, la consonne initiale 'cr' rappelle le 'tr' d'Arletty, dans "un très joli costume", tel que nous l'avons décrit dans [SPECTR. STAT. VOIX], §4.1. Un bruit consonantique prolongé, irrégulier, est perçu comme une occlusive si on en écoute seulement un tranche de 2 centièmes de seconde avant la voyelle; avec 1 ou 2 dixièmes de seconde, on perçoit une occlusive suivie d'une fricative.

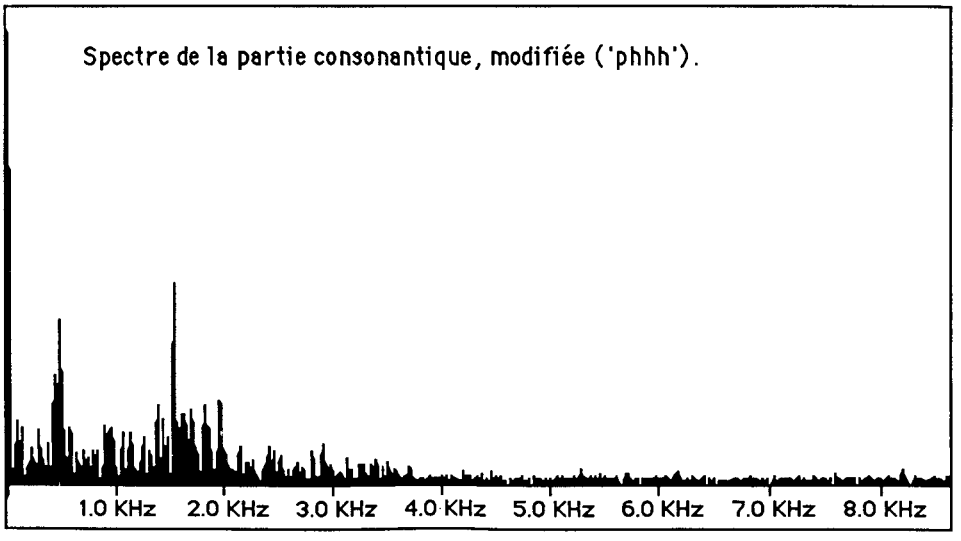
Ce qui nous arrêté d'abord dans cette syllabe 'cri' est la complexité du timbre consonantique: selon que l'on débute plus ou moins tôt l'écoute du bruit consonantique avant la voyelle, on peut entendre 'pi' ou 'ti'. Après avoir supprimé un petit segment de bruit, qui nous paraissait responsable de la perception d'un 'p', nous avons en effet entendu 'ti' ou 'cri'; et non 'pi'.



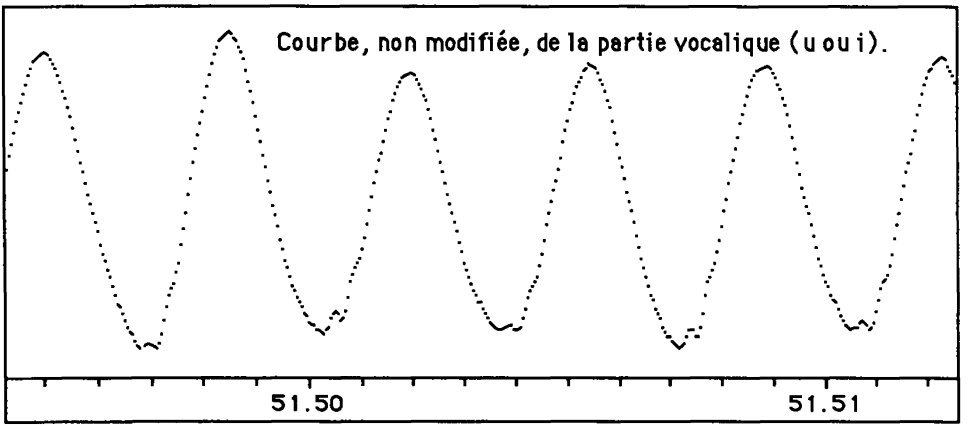
Nous avons alors entrepris d'étudier pour lui-même ce petit segment, ou motif, du 'p' dans ses rapports avec la voyelle qui le suit. À cette fin, nous



avons créé une copie indépendante de la syllabe 'pi'; puis recopié le motif consonantique 'p' jusqu'à obtenir plus de 2 dixièmes de seconde de bruit. Ce bruit, étant rigoureusement périodique, produisait un effet de timbre, ou de vibreur qui gênait notre expérimentation. Nous avons donc patiemment détruit la périodicité en effaçant des dents de la courbe du bruit, en en racourcissant d'autres: finalement, il en est résulté un bruit dont le spectre ne présente aucune structure de raie, et qui s'entend comme un souffle fort: 'phfhhhfhhh'.



Notre attention s'est alors portée sur la syllabe que ce bruit constitue avec les quasi-périodes vocaliques qui le suivent. Dans son ensemble, cette syllabe est perçue comme 'phui'. Ceci nous conduit à considérer pour elle-même la partie vocalique.

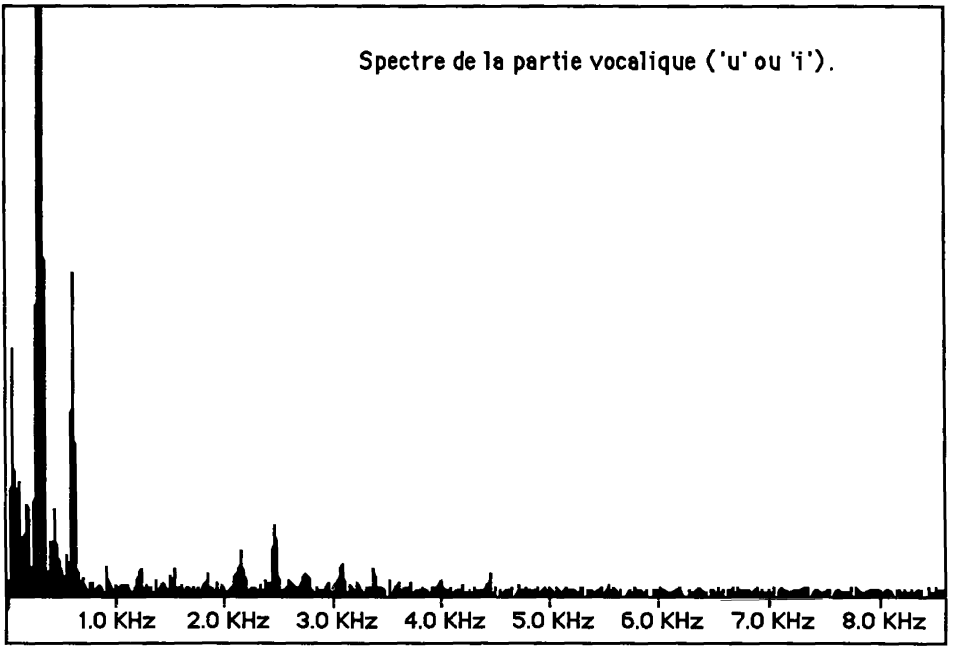


Toute seule, celle-ci ne s'entend pas comme une diphtongue 'ui', mais bien comme une voyelle simple, dont le timbre n'est pas nettement celui du 'i' mais se rapprocherait du 'u', comme dans certaines prononciations de la voyelle 'ы' du russe. D'ailleurs, dans le contexte du mot "écrire", quel que soit le segment de signal rejoué, nous n'avons jamais entendu une diphtongue, mais seulement un 'i'.

La courbe du signal vocalique offre l'aspect d'une sinusoïde ébréchée, dentelée, au niveau de ces minima. Ces irrégularités se voient bien sur le spectre, comme une raie voisine de 2500 Hz. Autant qu'on peut en juger par l'observation directe de la courbe, il n'y a pas d'évolution importante dans le spectre de la voyelle; et c'est pourquoi nous expérimentons sur le contexte, pour préciser la variabilité du timbre vocalique que l'on perçoit.

Avec le contexte initial copié sur la syllabe 'cri', on perçoit 'pi'. Il en est ainsi tant que le segment de bruit artificiel écouté a une durée de l'ordre du centième de seconde. Puis, à un seuil de durée qui, comme il est de règle pour de tels phénomènes, varie non seulement avec l'observateur mais avec l'observation en cours, on perçoit 'pui'; et, avec un dixième de seconde de bruit consonantique, la consonne occlusive 'p' s'étend, se prolonge par une aspiration qui nous paraît être plus proche d'un 'h' que d'un 'f'.

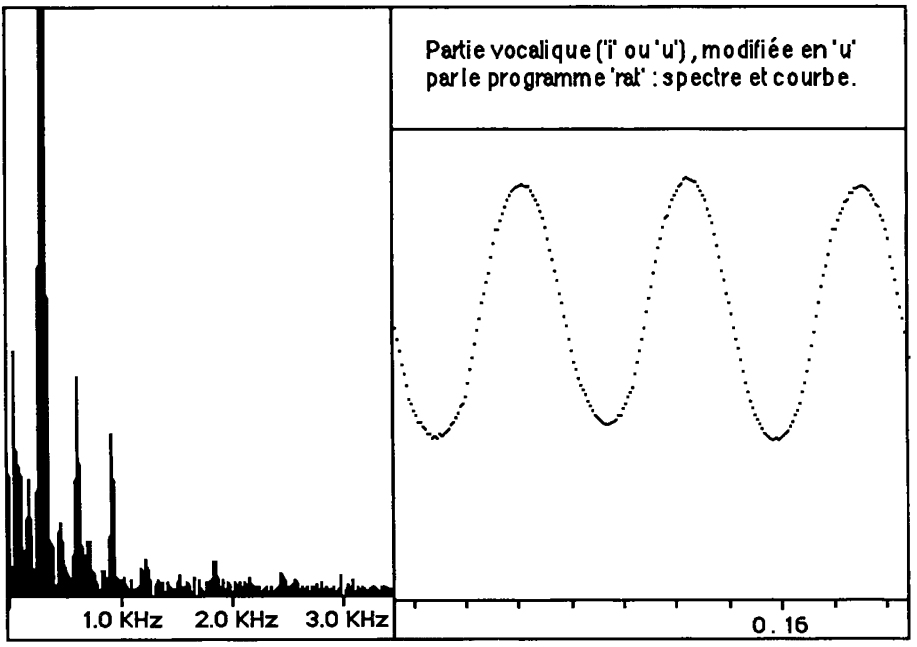
Nous pensons pouvoir conclure que la perception d'une diphtongue, même si elle n'est pas due exclusivement à un effet de contexte, requiert ce contexte



comme une condition indispensable. D'ailleurs, il ne s'agit pas d'un simple effet de durée; car tout contexte ne favorise pas la perception d'une diphtongue: c'est ce qu'on a constaté dans le contexte de la syllabe non modifiée 'cri', où il n'apparaît pas de diphtongue; et où, d'autre part, le timbre de la voyelle est perçu plus nettement comme étant un 'i', (sans hésitation entre 'i' et 'u'), peut-être du fait du spectre propre au bruit consonantique de 'cr'.

Il nous semble raisonnable d'attribuer à la suggestion qu'offre le spectre du bruit consonantique 'phhh' que nous avons créé, le dédoublement du son vocalique en une partie initiale, perçue 'u', et une partie finale, perçue 'i'. C'est pourquoi nous avons intitulé le présent §: "Un exemple de perception recombinée de traits vocaliques et consonantiques"; et croyons offrir ici une illustration des vues de Bergson, dans l'interprétation qui en est donnée à la fin du §2.

D'ailleurs, du point de vue de la phonologie classique, une affinité entre 'u' et 'p', du fait de la place de l'articulation, n'est aucunement inacceptable. En revanche, la nécessité de combiner des éléments de plusieurs spectres successifs pour rendre raison de la perception, qu'on pourrait croire propre à chacun de



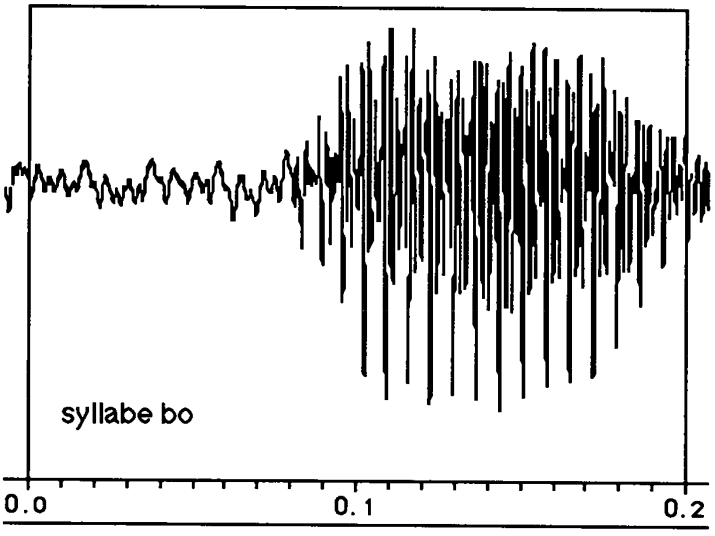
ceux-ci, donne au problème de la reconnaissance de la parole une complexité que l'on n'attendait pas *a priori*..

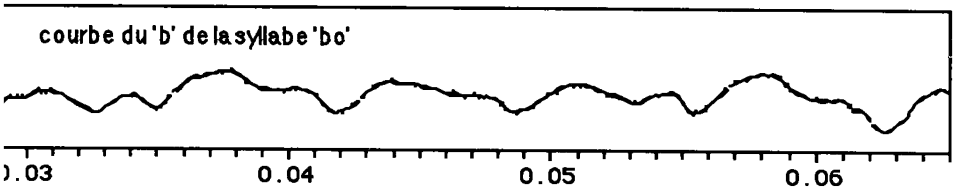
Afin de mieux cerner le timbre de la partie vocalique, on a modifié la courbe de celle-ci par le programme 'rat', déjà cité dans [SPECTR. STAT. VOIX], §§4.1 et 4.5. En éliminant les indentations de la courbe, on parvient à supprimer le maximum du spectre compris entre 2 et 3 kHz; la courbe modifiée n'est toutefois pas une sinusoïde parfaite car il subsiste, avant 1 kHz, 2 harmoniques du fondamental. Dans tous les contextes, le son perçu se rapproche nettement de 'u'; et l'on n'entend plus de diphtongue.

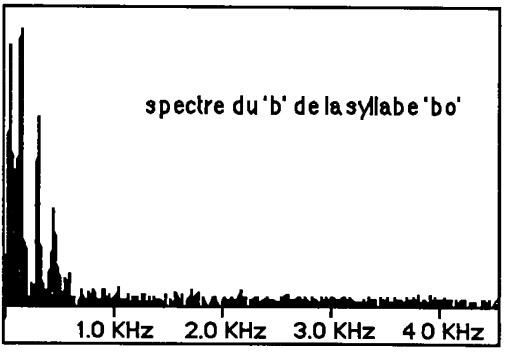
4.2 De la voyelle 'o' à la syllabe 'to'

L'expérience rapportée ici part d'une syllabe 'bo', extraite d'un enregistrement en langue italienne. Comme l'atteste le graphique, (obtenu par copie d'écran Soundcap avec compression de l'échelle du temps), la syllabe se compose de deux parties bien distinctes dont chacune dure environ (1/10) s.

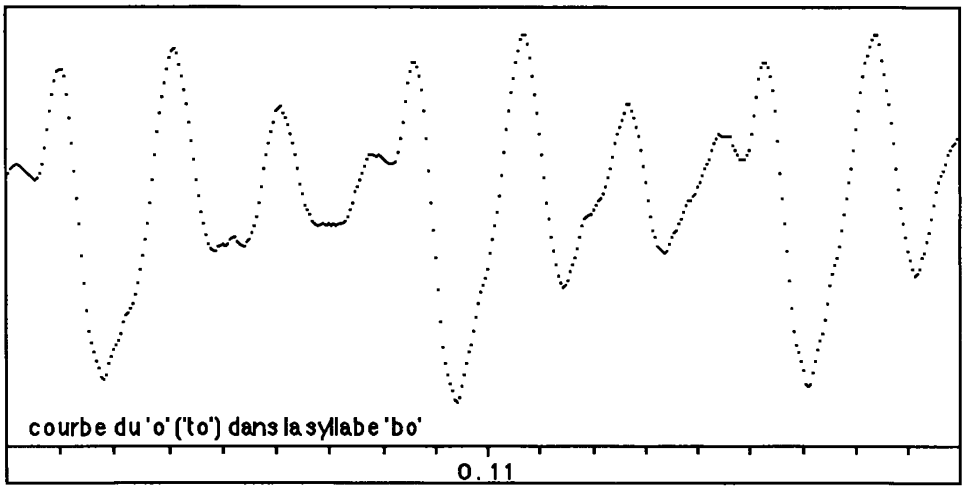
La courbe du 'b' montre une suite d'ondulations quasi périodiques, sans aucune indentation. Corrélativement, on a un spectre de raies tout entier compris en dessous de 500 Hz, sans aucune composante de haute fréquence.

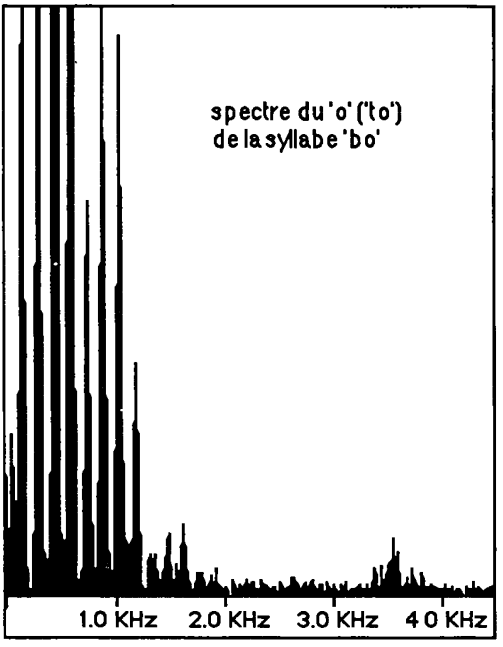






Relativement à celle du 'b' qui le précède, l'amplitude du 'o' apparaît grande sur le graphique: cela semble naturel, dans la mesure où 'b', malgré sa structure quasi périodique, n'est pas une voyelle mais a le rôle de consonne. Quant à l'intonation, on s'assure, en mesurant la longueur des quasi-périodes, qu'elle est montante sur le 'b' puis descendante sur le 'o'.





Mais s'agit-il bien d'un 'o'? En écoutant seule la partie présumée vocalique de la syllabe 'bo', on perçoit une syllabe 'to': fait d'autant plus surprenant que le 'b' n'offre aucune des caractéristiques du bruit d'un 't'; et qu'aucun segment de

bruit ne s'interpose entre les deux parties de la syllabe.

Considérons cependant le spectre du 'o'. Après un spectre de raies, dont presque toute la puissance est avant 1000 Hz, mais qui s'étend un peu au delà de 1500 Hz, on a, entre 3 kHz et 4 kHz, un pic de bruit qui pourrait être celui d'un 't' (ou d'un 's'). En effet, la courbe du 'o' a, dans ses quasi-périodes, quelques dentelures; et, surtout, les pointes de certains maxima et minima sont très aiguës. Ce sont ces caractères que la transformation de Fourier exprime par un pic de bruit.

Il paraît naturel d'attribuer à ce pic la perception d'une consonne 't'; qui, ne pouvant être placée en même temps que la voyelle 'o', est située avant celle-ci pour former une syllabe 'to'; du moins en l'absence du 'b' initial qui masque le pic.

À titre de vérification, écoutons, rejouée dans l'ordre inverse des temps, la syllabe 'bo' toute entière ou sa moitié vocalique, (entendue 'to' dans le sens direct). Il n'y a plus trace du 't'! Notre hypothèse est ici qu'à l'écoute directe, la suggestion du 't' par le pic de bruit est renforcée par le mouvement descendant de l'intonation sur le 'o'; mouvement qui suggère une détente après une explosion.

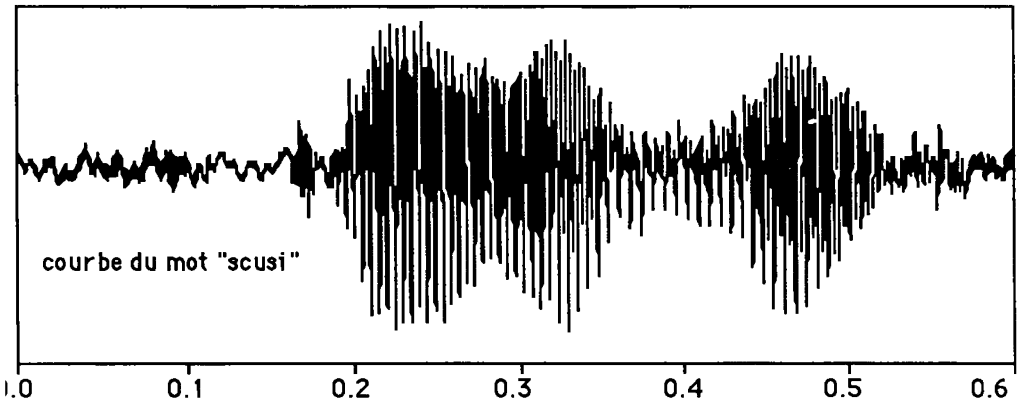
Afin de mettre à l'épreuve cette nouvelle hypothèse, il faudra modifier la courbe du 'o' pour en inverser l'intonation; ce qui se peut faire soit en permutant des quasi-périodes; soit en allongeant les premières de celles-ci et surtout (ce qui est plus facile) en racourcissant les dernières; sans que ces modifications altèrent la forme générale des quasi-périodes, donc l'enveloppe spectrale.

4.3 Consonne sonore, consonne liquide ou voyelle ?

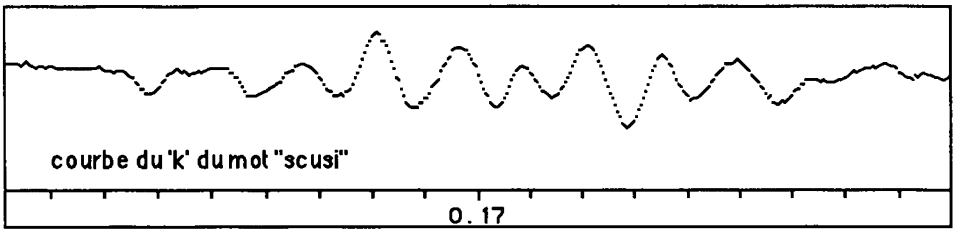
Du même enregistrement que la syllabe 'bo', objet du §4.2, nous avons extrait le mot italien "scusi". On sait que la deuxième lettre 's' de ce mot se prononce 'z': c'est cette consonne sonore qui a pu être entendue 'l' (consonne liquide) ou 'é' (voyelle), sans subir pour cela d'autre modification que d'être enfermée au sein de sous-segments diversement découpés dans le mot "scusi".

Il vaut la peine de parcourir la courbe du mot "scusi" en en prenant quelques spectres, (obtenus à l'écran par commande de Soundcap qui calcule sur des tranches de 1024 points; soit environ (1/20) s).

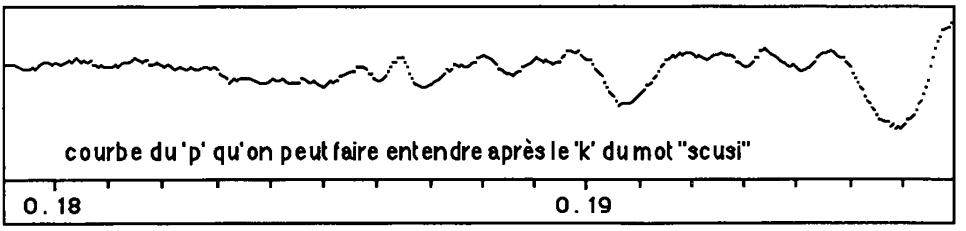
On a d'abord, sur plus d'un dixième de seconde, le bruit du 's' initial. Si l'on coupe le début du mot en gardant moins d'un centième de seconde de ce bruit, le début du mot est perçu 'kou'; entre 0,01 s et 0,02 s, on a 'tkou'; vers 0,03 s ou 0,04 s le début perçu est 'tskou'; avec plus de 0,05 s de bruit, on a bien pour début 'skou'. De telles gradations entre occlusive et fricative se rencontrent fréquemment.



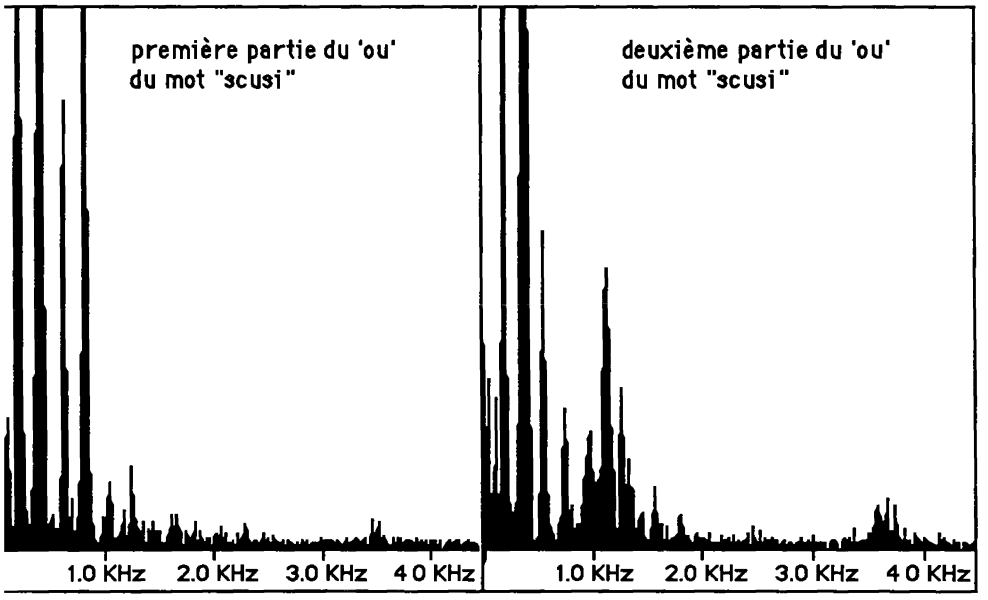
La consonne 'k' apparaît produite par un train d'onde dont la durée est inférieure à (1/100) s; avec des oscillations de ≈ 0,0015 s; soit une fréquence de ≈ 700 Hz.



Si l'on commence l'écoute du mot après ce train d'onde, on entend 'pouzi'. Il apparaît que le son 'p' provient d'un train d'onde, moins régulier que celui du 'k' et de fréquence environ 2 fois plus élevée. Nous signalons ici ce détail, qui ne fait pas l'objet principal du présent §, parce que nous l'avons déjà rencontré ailleurs.



Sur la courbe du mot "scusi", à échelle comprimée, la voyelle 'ou' s'étend approximativement de 0,2 s à 0,35 s et comprend 2 parties juxtaposées.



Il vaut la peine de considérer juxtaposés les spectres de ces deux parties qui, dans leur contexte naturel, ne produisent aucunement l'effet d'une diphtongue.

Le spectre de la 1-ère partie, que nous appellerons 'ou1', consiste principalement en un massif de quelques raies, avant 1 kHz. On peut le comparer au 'o' du §4.2: 'ou1' s'étend en fréquence moins loin que 'o' vers l'aigu (ce qui est assez connu); mais l'intonation du 'o' est plus basse que celle de 'ou1'; (les raies spectrales sont plus serrées pour celui-là que pour celui-ci).

Le spectre de la 2-ème partie, 'ou2', est fort différent. On peut le diviser en un massif, avant 500 Hz, et deux pics; dont le 1-er est centré peu après 1 kHz; et le 2-ème entre 3 et 4 kHz. Plus précisément, le 2-ème pic est dû à un poudroiement de bruit qui affecte les dernières quasi-périodes de 'ou2'.

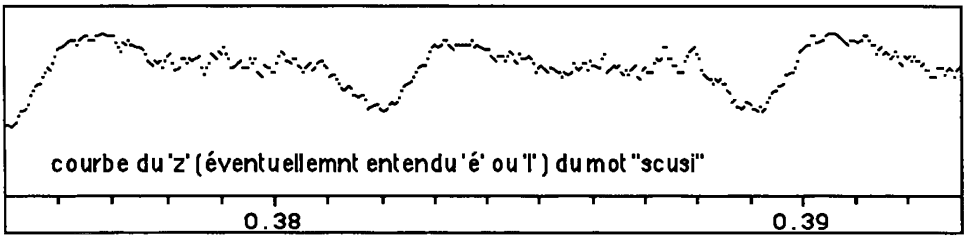
Alors que 'ou1' isolé s'entend franchement 'ou', 'ou2' isolé paraît être une voyelle de timbre incertain, entre 'ou', 'u' et 'i'. À l'écoute inverse, du fait de la place du pic de bruit, l'ensemble du 'ou' est perçu 'tou': c'est le cas notamment si l'on inverse le segment s'étendant de 0,20 s à 0,35 s.

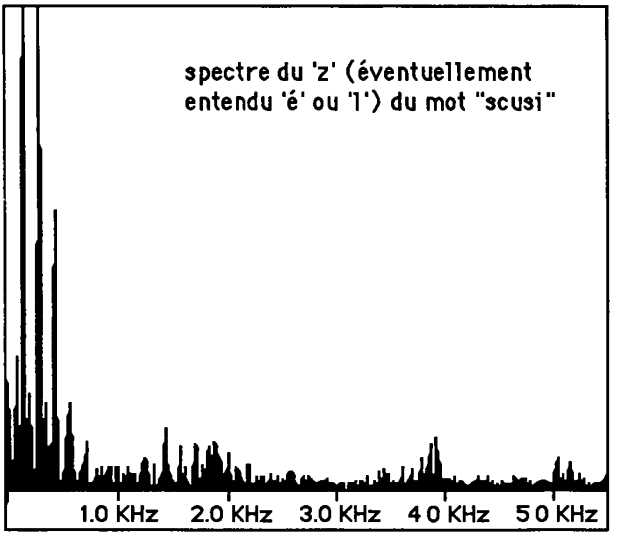
Après le 'ou', (cf. courbe d'ensemble du mot "scusi" à échelle comprimée), l'amplitude du signal s'abaisse nettement, et ne s'élève qu'après 0,43 s pour atteindre un dernier maximum vers 0,7 s. Il est facile de vérifier qu'il s'agit ici de la succession du 'z' et du 'i'.

Mais il s'en faut de beaucoup que la perception du 'z' soit stable et constante. Voici, schématisé, ce que nous avons perçu, en fonction de l'intervalle écouté:

```
{0,44 - 0,52} : 'i'      ;   {0,30 - 0,38} : 'oué' , ou 'oui'  ;
{0,40 - 0,52} : 'li'     ;   {0,30 - 0,40} : 'ouli', ou 'ouri' ;
{0,37 - 0,52} : 'zi'     ;   {0,30 - 0,48} : 'ouzi'  ;
{0,33 - 0,52} : 'ouzi'   .
```

On ne peut rendre compte de l'ensemble de ces observations sans attribuer, suivant les cas, à tout ou partie des quasi-périodes du 'z' la valeur d'une consonne sonore ('z'), d'une consonne liquide ('l' ou 'r') ou d'une voyelle ('é' ou 'i').





Examinons la courbe et le spectre du 'z'. La courbe montre une suite de quasi-périodes analogues à celles du 'b' du §4.2, mais fortement bruitées.

Le spectre comprend plusieurs maxima successifs, d'inégale importance: le 1-er, avant 500 Hz; le 2-ème, près de 2 kHz; le 3-ème, vers 4 kHz.

Selon ce que nous savons des consonnes sonores, lesquelles doivent comprendre un bruit consonantique, (plus ou moins aigu selon le point d'articulation), associé à une basse vocalique, le 1-er et le 3-ème pic pourraient ensemble produire un 'z'; (le fréquence élevée du 3-ème pic, et plus encore celle du 4-ème qui le suit, suggérant un 's', qui est la consonne non voisée à laquelle correspond le 'z').

Les deux premiers pics, d'autre part, produiraient soit une voyelle telle que 'é'; soit une liquide: 'l', voire 'r'.

5 Analyse et reconstruction perceptive de la chaine parlée

Afin de déterminer, d'après un signal sonore et son spectre, ce qui peut être perçu, on songe d'abord à établir un schéma en termes de formants, de blocs spectraux plus ou moins étroits marqués ou non de raies.

Ce schéma ne peut s'interpréter sans fixer des seuils, au-dessous desquels un formant ne sera pas perçu. Ces seuils sont relatifs non seulement à la puissance globale du spectre considéré, mais aussi à la disposition des formants; car le masquage est essentiel (la très haute fréquence étant, notamment, masquée par la haute).

Mais il y a plus: la perception d'un segment dépend grandement du contexte phonétique (pour ne rien dire du contexte sémantique): traits et formants sont refusés ou acceptés pour être combinés et interprétés selon une dynamique de la syllabe qui, en partie selon les suggestions de l'intonation et de l'enveloppe de puissance, postule des segments successifs ayant rang de consonnes de liquides ou de voyelles.

Nous disposons de programmes qui reconnaissent les maxima successifs des spectres; et avons effectué des analyses factorielles sur des profils spectraux recodés en tenant compte de seuils dont il semble aisé de varier le choix (cf. [NOT. SON]). La détection la plus fine de la mélodie est, croyons-nous, celle fondée sur le découpage visuel du signal en quasi-périodes; mais ce découpage n'a pas été rendu automatique; la disposition des raies, déjà utilisée par T. Moussa, fournit une première approximation.

Cependant la difficulté majeure est de se former une conception globale de ce que, faute d'un terme plus précis, nous avons appelé *dynamique de la syllabe*.

Références bibliographiques

[SPECTR. STAT. VOIX]: *CAD*, Vol XIII n°1, pp. 99-130; 1988.

[BERGSON]: *CAD*, Vol VII n°4, pp. 395-412; 1982.

[VOIX]: *CAD*, Vol VIII n°2, pp. 181-186; 1983.

[NOT. SON]: *Programmes de traitement du son: Notice d'utilisation.*

Remerciements

Les recherches de l'auteur, exposées dans le présent travail ainsi que dans [SPECTR. STAT. VOIX], ont bénéficié d'un soutien de la DRET dans le cadre du contrat 86.1158.