LES CAHIERS DE L'ANALYSE DES DONNÉES

G. CHARBONNEAU
T. MOUSSA

Traitement numérique du signal acoustique pour une analyse factorielle de la parole

Les cahiers de l'analyse des données, tome 6, n° 2 (1981), p. 187-206

<http://www.numdam.org/item?id=CAD_1981__6_2_187_0>

Les Cahiers de l'Analyse des Données
Vol. VI - 1981 - n° 2 - p. 187-206

TRAITEMENT NUMÉRIQUE DU SIGNAL ACOUSTIQUE POUR UNE ANALYSE FACTORIELLE DE LA PAROLE [PAROLE I]

par G. Charbonneau (1)
T. Moussa (2)

0 La parole est un moyen de communication exceptionnel, tant par sa vitesse, sa souplesse, que sa précision. Elle serait sans doute - et sera peut-être un jour - le moyen privilégié de la relation homme - machine, si sa complexité ne soulevait pas des problèmes considérables quant à sa reconnaissance automatique. La parole a fait l'objet d'études importantes depuis longtemps déjà, mais ce n'est réellement que l'avènement des ordinateurs qui a permis de faire des progrès suffisants pour laisser entrevoir des solutions aux problèmes de reconnaissance.

L'ordinateur a apporté une énorme capacité de stockage, une puissance et une souplesse de traitement incomparables avec les moyens analogiques traditionnels (oscilloscope, analyseur de spectre, etc.).

Dans cet article nous décrivons les méthodes que nous avons utilisées pour obtenir une mise en forme du signal acoustique de parole dans le but d'un traitement statistique par l'analyse factorielle des correspondances. Nous envisageons successivement la numération du signal, puis l'analyse spectrale effectuée sur ce signal en précisant les options choisies.

1 *Numérisation du signal*

1.1 Enregistrement analogique : Pour diverses considérations nous avons choisi d'étudier la langue arabe, mais comme le lecteur pourra lui-même le constater, les méthodes décrites dans cet article sont indépendantes du langage utilisé ou à tout le moins, facilement adaptables à une autre langue. Le texte choisi pour servir de support à notre travail a été tiré d'un journal libanais.

Il a été lu par l'un de nous (T.M.) dans des conditions courantes de prononciation, c'est-à-dire avec un débit normal et une articulation naturelle, sans accentuation ni intonation particulières . Les 373 mots contenus dans l'article de presse ont ainsi été prononcés en 4 minutes, soit un débit moyen de 3 mots /2 secondes. Le texte, lu dans une salle moyennement réverbérante d'environ 50 m^3 a été enregistré à l'aide d'un microphone Beyer type M88N, et d'un magnétophone Revox A77 en 19cm/sm (cf. Annexe 1).

---

(1) *Docteur ès-Sciences, Institut d'Electronique Fondamentale - 91405 Orsay.*

(2) *Docteur 3° cycle, Laboratoire de Statistique. Université Pierre et Marie Curie.*

1.2 Echantillonnage de la parole

1.2.1 Conditions théoriques : L'élaboration analytique des signaux ne se conçoit pas sans les notions de spectre aujourd'hui familières à tous les ingénieurs. Ces notions associées pour toujours au nom de Fourier (qui les a appliquées avec une grande généralité dans sa théorie analytique de la chaleur : 1822) se sont introduites dans le milieu du XVIII-ème siècle, notamment dans les travaux d'Euler.

Le traitement numérique des signaux analogiques est possible grâce à l'équivalence (Wittaker, 1915) entre la connaissance à tout instant S(t) dépendant du temps, et la connaissance de ce même signal à des instants discrets $S(i\,T)$, $i = 1, 2, \ldots, N$, où T est appelé la période d'échantillonnage. Le théorème d'échantillonnage (Shannon, 1962) précise les conditions de validité de l'équivalence : la fonction S(t) doit avoir un spectre en fréquence borné et si F_c est cette borne la fréquence d'échantillonnage F doit être supérieure ou égale à $2F_c$. Si ces conditions ne sont pas respectées, toute fréquence f supérieu- à F/2 est repliée dans la bande (0,F/2) et prend la valeur f_r telle que:

$$f_r = (n+1)F - f \quad \text{si} \quad (n+\tfrac{1}{2})F < f \leq (n+1)F \qquad n = 0, 1, 2, \ldots$$

ou

$$f_r = f - nF \quad \text{si} \quad nF < f \leq (n+\tfrac{1}{2})F \qquad n = 1, 2, 3, \ldots$$

Il importe donc avant tout échantillonnage de faire subir au signal un filtrge passe-bas éliminant aussi soigneusement que possible toute composante supérieure à la demi-fréquence d'échantillonnage choisie.

La bande passante de la parole pour l'essentiel de l'information ne dépasse guère 4000 Hz sauf pour quelques consonnes telles que le "s" ou le "ch". Les filtres passe-bas n'ayant pas une courbe de réponse strictement rectangulaire, il est nécessaire que la fréquence de coupure du filtre soit sensiblement inférieure à la demi-fréquence d'échantillonage. Nous avons choisi de limiter à 7000 Hz la bande passante du signal de parole étudiée et nous avons fixé à 20.480 Hz la fréquence d'échantillonnage, cette valeur étant divisible par une puissance de 2 inférieure ou égale à 4096, ce qui est souhaitable pour l'emploi de la transformée de Fourier rapide (cf. § 2.2). Le filtre Schlumberger FAB24 utilisé (Annexe 1) fournit une atténuation inférieure ou égale à 1 dB dans la bande passante (0 - 7000 Hz) et une atténuation supérieure à 20 dB au dela de 10.240 Hz, ce qui est suffisant pour éviter tout phénomène gênant de repli.

1.2.2 Conversion analogique-numérique (A/N) : Quatre minutes de paroles correspondent à près de cinq millions de nombres à la fréquence d'échantillonnage que nous avons utilisée, ce qui exclut de les stocker directement dans la mémoire centrale de l'ordinateur. Il est donc nécessaire d'écrire ces nombres dans une mémoire de masse, par exemple une bande magnétique qui est un support bien adapté étant donné le caractère séquentiel de l'information et la capacité énorme du stockage (environ 50 millions d'échantillons). L'écriture de la bande magnétique nécessite cependant quelques précautions. D'une part les échantillons sont groupés sur la bande en blocs de longueur finie et le dérouleur s'arrête entre l'écriture de chaque bloc. D'autre part, les dérouleurs de bande numérique ont une vitesse de défilement insuffisamment régulière (les variations peuvent atteindre 2 à 3%), ce qui entraînerait des distorsions dans la cadence d'échantillonnage si les nombres résultant de la conversion analogique numérique étaient envoyés directement sur la bande. De plus, il serait impossible de choisir la fréquence d'échantillonnage puisque celle-ci serait obligatoirement fixée par le débit moyen du dérouleur de bande. Pour toutes ces raisons, il

est nécessaire de découpler le rythme d'écriture des échantillons et celui de leur conversion (qui doit être rigoureusement uniforme). Ce découplage est obtenu à l'aide d'une mémoire tampon (celle de l'ordinateur), divisée en deux tableaux. L'un des tableaux est rempli pendant que l'autre se vide grâce à deux canaux d'accès direct à la mémoire. La vitesse d'entrée des échantillons peut être alors parfaitement uniforme et pilotée par une horloge externe réglable, tandis que la vitesse de sortie vers le dérouleur est irrégulière. Les débits moyens d'entrée et de sortie sont évidemment égaux.

L'organe de conversion A/N est composé d'un échantillonneur-bloqueur Analog Devices SHA 2A et d'un convertisseur Hybrid Systems ADC 591A de 12 bits de précision. Ce temps maximal de conversion d'un échantillon réclame au maximum 4 µs, la fréquence d'échantillonnage maximale sur notre système est donc de 250 kHz si les échantillons sont stockés sur un support très rapide comme un disque. Le dérouleur magnétique que nous possédons limite toutefois à 35 kHz la cadence d'échantillonnage.

Les signaux échantillonnés souffrent d'un bruit inhérent à la conversion A/N qui est le bruit de quantification. La conversion A/N ne fournit que des nombres entiers compris entre 0 et 4095 pour 12 bits de sorte que chaque échantillon est *a priori* entâché d'une incertitude d'arrondi inférieure ou égale à 0,5. Cette incertitude est équivalente à un bruit qui s'évalue généralement (Mathews, 1969) par comparaison avec le plus grand nombre convertible. Pour un convertisseur à 12 bits, le rapport signal sur bruit de quantification (S/B) est ainsi égal à

$$\frac{S}{B} = \frac{4095}{0,5} \text{ soit } 78 \text{ dB}$$

Ce bruit est comparable au niveau de bruit introduit par les meilleurs magnétophones. Il faut tenir compte cependant du fait que cette valeur correspond à un optimum obtenu avec des signaux d'amplitude maximale. Pour des signaux plus faibles, le rapport signal sur bruit diminue. Néanmoins, le signal de parole possède en langage courant une dynamique modérée (de l'ordre de 10 dB), de sorte que le bruit de quantification reste à un niveau négligeable.

2 *Analyse spectrale du signal*

Le signal vocal est une fonction scalaire du temps. Toutefois ce n'est pas en tant que phénomène unidimensionnel que le message sonore est perçu, car la parole est codée dans l'oreille interne avant de subir un traitement élaboré dans le cerveau. Analyser un tel signal est une tâche ardue tant est grand le nombre de paramètres physiques mesurables. Néanmoins, trois paramètres fondamentaux s'imposent (Moles, 1952) : l'intensité, la fréquence et le temps. Quoique non exhaustive, la représentation de l'objet sonore dans un repère formé de ces trois dimensions est féconde, car chacun des plans de référence correspond à un fait perceptif important : le plan dynamique (intensité - temps), le plan harmonique ou spectral (intensité - fréquence) et le plan mélodique (fréquence - temps). Nous avons écrit (Chabrel, Charbonneau, 1976) un programme appelé ASPECT, permettant les multiples traitements nécessaires à l'obtention d'éléments pouvant servir à une analyse statistique féconde de la parole.

2.1 *Le logiciel*

2.1.1 Structure générale du programme ASPECT : La conception du programme ASPECT a été faite avec l'objectif de satisfaire deux conditions. D'une part le programme a été écrit pour pouvoir fonctionner en mode conversationnel, c'est-à-dire qu'il soit capable d'exécuter

immédiatement toute commande de l'utilisateur. D'autre part, pour que cette première condition puisse être satisfaite, il est nécessaire que les diverses directions correspondent à des transformations fournissant des éléments de même forme, de sorte qu'elles soient utilisables dans un ordre quelconque, dans la mesure où celui-ci a un sens physique. ASPECT est écrit en langage FØRTRAN (Dreyfus, 1969) et comporte plus de 5000 instructions découpées en 7 segments dont le programme assure lui-même la gestion.

2.1.2 *Les commandes* : Une commande se présente sous la forme d'un code mnémonique de deux lettres (sauf pour la transformée de Fourier inverse notée F- et la multiplication par la complexe conjuguée notée M*) suivi d'une liste de paramètres séparés par des blancs ou des virgules.

L'utilisateur dispose de cinq types de commandes :

- les commandes de configuration qui permettent de fixer les valeurs de certains paramètres, comme la fréquence d'échantillonnage, la taille des blocs etc.

- les commandes de traitement mathématique qui permettent d'effectuer des opérations élémentaires sur le signal.

- des commandes donnant la possibilité de créer des macro-commandes qui condensent en une seule directive un ensemble quelconque d'instructions et autorisent l'utilisation des boucles.

- des commandes de visualisation sur écran cathodique et de tracé sur papier.

- Enfin, il y a 5 macro-commandes programmées d'avance, qui assurent le calcul et la visualisation des décompositions suivantes du signal :

. Transformée de Fourier d'une partie du signal.

. Sonagramme d'une partie du signal.

. Enveloppe dynamique du son.

. Enveloppe dynamique de chaque harmonique pour les signaux périodiques ou pseudo-périodiques.

. Recherche de la mélodie (pitch) d'une portion de parole.

2.2 *Le calcul des spectres de Fourier*

2.2.1 *Intérêt de l'analyse spectrale* : Depuis longtemps déjà, l'analyse spectrale est apparue comme un puissant moyen d'investigation des signaux sonores (Helmhotz, 1896). Cela est dû essentiellement au fait que les éléments résultant de cette décomposition ne sont pas des termes mathématiques arbitraires. Ils sont en relation étroite avec la perception des sons par l'oreille (Littler, 1965, p. 308). Différentes études ont été réalisées (Greenwood, 1961, Plomp, 1964) sur le pouvoir qu'a l'oreille de séparer une vibration complexe en ses composantes spectrales. D'une part, l'oreille se comporte comme un analyseur de spectre à large bande, peu sélectif, et d'autre part le cerveau prolonge cette première analyse par un traitement aboutissant à une sélectivité remarquable notamment dans la détection de la hauteur (bien supérieure à la discrimination possible avec un analyseur de spectre analogique actuel).

2.2.2 *Les formules et leur calcul* : Le théorème de la transformée de Fourier établit (Schwartz, 1963) la correspondance suivante :

$$F(\omega) = \frac{1}{2}\int_{-\infty}^{+\infty} f(t)\ e^{-j\omega t}\ dt$$

$$f(t) = \int_{-\infty}^{+\infty} F(\omega)\ e^{j\omega t}\ d\omega$$

Cette équivalence définit la représentation spectrale complexe $F(\omega)$ d'une fonction $f(t)$ sous des hypothèses très générales. La transformation ainsi définie n'est que la généralisation de la série de Fourier associée à une fonction de période T lorsque T tend vers l'infini. En pratique, lorsqu'elle représente un phénomène physique réel, $f(t)$ n'est définie que sur un intervalle de temps fini $(0 \leq t \leq T_1)$; de plus $f(t)$ est échantillonnée, l'intervalle T_1 étant divisé en K incréments égaux à $t = \frac{T_1}{K}$, on a alors une transformation de Fourier discrète telle que :

$$f(t_k) = \Delta\omega \sum_{n=0}^{n=N-1} G(\omega_n)\ e^{j\omega_n t_k}$$

avec $$G(\omega_n) = \frac{\Delta t}{2} \sum_{k=0}^{k=K-1} F(t_k) e^{-j\omega_n t_k}$$

et $$\Delta\omega = 2\pi\Delta f = 2\pi\frac{K}{NT}$$

En fait du point de vue mathématique, la transformée de Fourier discrète est un simple exercice d'algèbre linéaire, un changement de base orthonormée dans un espace de dimension finie ; toutefois l'intérêt de ce changement de base provient de la façon dont il tend à la limite vers le développement en intégrale de Fourier lorsqu' on densifie l'échantillonnage. (Il est d'ailleurs de règle en mathématiques appliquées que des calculs qui ne peuvent être faits qu'en dimension finie doivent leur intérêt aux théorèmes limites relatifs aux espaces fonctionnels).

En 1965, Cooley et Tukey (Cooley, Tukey, 1965) ont proposé un algorithme rapide pour le calcul de ces formules, le calcul étant le plus simple si N et K sont pris égaux à une puissance de 2 *. A la fin du calcul, la transformation de Fourier (commande F0) est obtenue sous la forme

$$G(\omega_n) = a_n + jb_n$$

Par multiplication par le complexe conjugué (commande M*) on obtient le spectre en puissance

$$|G(\omega_n)|^2 = a_n^2 + b_n^2$$

2.2.3 Choix des variables : Dans une analyse spectrale, les variables dont dispose l'utilisateur sont :

- la fréquence d'échantillonnage F
- la fréquence maximale qui puisse être détectée F_{max}
- la résolution de fréquence Δf
- l'intervalle de temps Δt séparant deux échantillons
- la durée T_1 de la fraction de signal analysée

---

* *En fait le principe de la méthode se trouve déjà dans Runge (1903); pour un historique complet de ce remarquable algorithme, cf. Cooley et coll. (1967).*

- le nombre de points N sur lesquels porte l'analyse (longueur d'un bloc).

Ces variables vérifient les relations suivantes :

$$F = 1/\Delta t$$

$$F_{max} = F/2$$

$$T_1 = N\Delta t = N/F$$

$$\Delta f' = F/N = 1/T_1$$

$$\Delta f.N/\Delta t = 1$$

Cette dernière relation est l'expression du principe d'incertitude en acoustique : il est impossible de faire une analyse fine simultanément en temps et en fréquence.

En réalité, il n'y a pas deux variables indépendantes : F et N. La fréquence d'échantillonnage étant fixée à 20.480 Hz pour satisfaire la bande passante du signal vocal, il ne reste de réellement variable que le nombre de points N définissant la taille des blocs servant au calcul du spectre en puissance de Fourier.

Nous nous sommes efforcés de concilier deux exigences contradictoires : d'une part, avoir une résolution en fréquence aussi bonne que possible (impliquant N grand) et d'autre part ne pas avoir des blocs d'une durée trop longue pour ne pas risquer de noyer les sons brefs (p, t, k,q, par exemple). Ces exigences conduisent à ne considérer que deux tables de blocs, N = 512 qui correspond à une durée T_1 = 25 ms assurant une résolution Δf = 40 Hz, et N = 1024 correspondant à T_1 = 50 ms et Δf = Hz.

2.2.4 Spectres successifs et spectres décalés : Pour obtenir une évolution du spectre de fréquence au cours du temps, nous avons calculé les spectres instantanés portant sur des blocs successifs de N points (512 ou 1024). Nous avons essayé également de ne pas prendre les blocs successifs mais des blocs de 1024 points concécutifs décalés de 512 en 512, et aussi de 256 en 256. L'étude de ces spectres en puissance nous a montré qu'on obtenait la représentation la plus fidèle avec des spectres décalés de 256. En effet, l'analyse d'une fraction du signal revient à considérer que le signal complet a été multiplié par une fenêtre temporelle rectangulaire valant 1 pendant la durée examinée et 0 ailleurs. Dans ces conditions la transformée de Fourier est la convolution de la transformée de Fourier du signal par celle de la fenêtre. Les bords de la fenêtre provoquent ainsi des blocs gênants dans les spectres en fréquence (Rabiner et Gold, 1975). Pour réduire l'influence du phénomène nous avons utilisé la fenêtre de Hanning (Kaiser, 1966) qui s'écrit

$$h(n) = \frac{1}{2}\left(1 - \cos\frac{2\pi n}{N}\right) \qquad n = 0, 1, 2, \ldots, N$$

Si on utilise cette fenêtre qui annule le signal aux bords de la fenêtre, certains phénomènes brefs risquent d' être atténués fortement. Le décalage de 256 en 256 élimine cet inconvénient car chaque phonème se trouve toujours convenablement placé dans au moins un bloc.

3 *Codage des spectres pour l'analyse factorielle*

3.0 Principe d'application de l'analyse des correspondances : Pour chacun des N points, le programme de transformée de Fourier rapide donne N composantes $\{G(\omega_n), n = 1, \ldots, N\}$ avec $\omega_n = F/N*(n-1)$ Hz.

Comme le signal est réel, les amplitudes complexes $G(\omega_n)$ sont liées entre-elles par la relation

$$G(\omega_n) = G(\omega_N - \omega_n)$$

La phase ne contribuant pas au message, seuls comptent les carrés des modules $|G(\omega_n)|^2$, ce qui réduit l'information à N/2 nombres:

$$|G(\omega_n)|^2 , \quad n = 1, 2, \ldots, N/2 .$$

Pour des blocs de N = 1024 points, cela fait 512 nombres à traiter d'un point de vue statistique. Cependant, les études antérieures de la parole et l'usage des vocoders à canaux indiquent qu'il est possible de réduire fortement cette quantité de variables. Le principe de notre traitement est de découper les spectres de puissances en tranches consécutives qui sont l'analogue des canaux des codeurs usuels ; à cette différence éventuelle près que nous pouvons modifier les limites de ces canaux comme des paramètres du programme de traitement, alors que dans le traitement analogique une telle modification ne peut se faire sans changer les batteries du filtre.

Comme de règle en analyse des correspondances, nous constituerons un tableau I × J ; il est commode d'appeler les " i " *individus* et les " j " *variables* ; en fait ici " i " un spectre et " j " un canal ; k(i,j) étant la puissance du spectre " i " dans le canal " j " ; ainsi qu'on l'explique en détail ci-dessous.

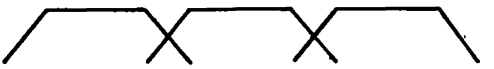
3.1 *Découpage en canaux des spectres* : Chaque spectre instantané présente une suite de maxima correspondant aux harmoniques successifs d'un fondamental dont la fréquence est variable pour un même sujet au cours du discours et variable d'un sujet à un autre dans une plage comprise approximativement entre 100 et 400 Hz (100 à 200 Hz pour un homme, 200 à 400 Hz pour une femme).

D'autre part, ces maxima sont pondérés par une enveloppe spectrale dont les bosses, appelées formants sont représentatives dans une certaine mesure des différents phonèmes.

Pour suivre le profil de cette enveloppe spectrale nous avons découpé le spectre en canaux de nombre et largeur variables. Le découpage a été effectué d'une part en conciliant empiriquement les spectres des divers phonèmes (cf. figure 1.a) et d'autre part en faisant le calcul d'un "spectre moyen" par la moyenne arithmétique terme à terme de 2000 spectres consécutifs (cf. figure 1.b). Les deux méthodes ont abouti à un découpage similaire compte-tenu de la limitation à 20 ou 30 canaux qui semblait raisonnable pour assurer une représentation détaillée sans toutefois faire interférer trop les variations de la mélodie (fréquence fondamentale). Ce découpage est porté sur les figures.

Cependant la bande passante d'un canal n'est pas un créneau nettement découpé mais une fonction présentant un large maximum plat entre une ascension et une descente rapides : ainsi pour éviter le caractère parfois arbitraire des bornes et suivre avec plus de finesse modèle du spectre, nous avons chargé les canaux en les faisant empiéter entre eux :

exemple 1 : trois canaux consécutifs à front raide



exemple 2 : trois canaux consécutifs à front incliné

Numéro de canal	*rang dans le spectre*	*fréquence borne sup (Hz)*	*largeur de la bande*
1	1- 5	80	80
2	6- 10	180	100
3	11- 17	320	140
4	18- 24	460	140
5	25- 31	600	140
6	32- 37	720	120
7	38- 44	860	140
8	45- 53	1040	180
9	54- 60	1180	140
10	61- 73	1440	260
11	74- 80	1580	140
12	81- 91	1800	220
13	92-102	2020	220
14	103-115	2280	260
15	116-126	2500	220
16	127-180	3580	1080
17	181-202	4020	440
18	203-212	4220	200
19	213-228	4540	320
20	229-266	5300	760
21	267-307	6120	820
22	308-220	6380	260
23	321-329	6560	180
24	330-354	7080	520
25	355-512	10240	3160

Nous définissons pour chaque canal c une fonction de bande $\varphi_c(n)$ en sorte que

$$\forall\, n \in e[1, \ldots, \tfrac{N}{2}] \quad \Sigma\{\varphi_c(n) \mid c \in C\} = 1$$

pour un canal à bande abrupte la fonction φ_c serait une fonction en créneau; pour nos canaux la fonction est continue avec des fronts inclinés.

La puissance passant dans un canal c étant

$$\Sigma\{|G(\omega_n)|^2 . \varphi_c(n) \mid n = n_1, \ldots, n_2\}$$

où n_1 , n_2 sont les bornes du canal.

Différents empiétements de canaux ont été testés (avec des pourcentages d'élargissement de 0 à 40% de la bande du canal). Ils n'ont pas apporté de différences très significatives (e.g. les analyses factorielles faites sur les tableaux I × J sont semblables) ce qui confirme la validité du découpage et la stabilité des résultats de l'analyse des correspondances.

3.2 *Les individus (spectres)* : Chacun des spectres décalés de 256 en 256 points constitue un individu d'une analyse factorielle. Il convient donc qu'il soit muni d'une étiquette désignant aussi précisément que possible le(s) phonème(s) contenu(s) dans les N points considérés. Si l'origine de certains phonèmes comme les plosives est nette , dans de nombreux phonèmes, elle est plutôt floue.

Nous avons utilisé des étiquettes comportant 4 symboles, chacun de ces symboles représente un quart du bloc (i.e. pour des blocs de 1024 points, le premier symbole désigne les 256 premiers points , le deuxième les 256 suivants, etc.). Ainsi l'étiquette AAAB correspond à l'apparition d'un "B" à la suite d'un "A" dans le dernier quart du bloc.

Pour procéder à l'étiquetage, nous avons observé le signal échantillonné lui-même, mais nous (T.M.) avons surtout identifié les individus par audition du son. Pour cela, nous disposons d'une commande (PL) dans le programme ASPECT permettant la restitution du son par conversion numérique-analogique d'un nombre donné de blocs consécutifs commençant ou se terminant au choix par un bloc déterminé. Le son est envoyé directement à un casque qui isole l'auditeur des bruits environnants. Quoique longue et fastidieuse la méthode permet l'étiquetage des blocs avec une précision satisfaisante.

Nous avons choisi de supprimer les silences qui ne pouvaient qu'alourdir le temps de l'analyse sans réellement apporter de contribution appréciable. Pour décider si un spectre correspond à un silence ou non, nous avons fixé à 7, la limite du poids total du spectre, c'est-à-dire la somme de toutes les composantes contenues dans le spectre. Le poids des spectres varie de 0 à 10.000 environ de sorte que le poids délimitant le silence correspond à 1% approximativement du poids du son le plus intense. En dépit de ce seuil très bas, le nombre de blocs ainsi éliminés est substanciel puisqu'il est de l'ordre de 15%.

3.3 *Temps de traitement* : L'échantillonnage du signal étant effectué, les différentes étapes du traitement se décomposent de la façon suivante :

- chargement du bloc
- application de la fenêtre de Hanning
- calcul de la Transformée de Fourier
- calcul du spectre de puissance
- calcul du poids des C canaux

L'ensemble de ces opérations réclame (toutes opérations d'entrée sortie inclues) 7,96 secondes pour un bloc de 1024 points, soit avec les conditions optimales que nous avons décrites précédemment, à savoir des blocs de 1024 points décalés de 256 en 256, 100 blocs correspondent à 50 secondes. Ce traitement nécessite donc un temps 160 fois plus long que le temps réel. Néanmoins, les calculs pourraient être faits avec des circuits électroniques appropriés en un temps avoisinant le temps réel avec la technologie actuelle.

4 *Conclusion*

A partir des spectres de puissance codés comme il a été indiqué auparavant, l'analyse factorielle du tableau $I \times J$ fournit une description du signal de parole d'une remarquable précision (bien supérieure à la description par formants). Cette analyse peut être utilisée de façons diverses dans les différentes étapes pouvant mener à la reconnaissance automatique du langage, notamment dans la segmentation en phonèmes.

Les premiers résultats que nous publierons dans un prochain article laissent espérer d'importants progrès dans le traitement automatique de la parole continue. Sans anticiper sur ces résultats nous pouvons signaler qu'il est possible d'atteindre déjà un taux d'erreur dans la segmentation en phonèmes d'environ 5% en se fondant uniquement sur des indices acoustiques.

ANNEXE I

CONFIGURATION DES SYSTEMES INFORMATIQUES ET ANALOGIQUES

Unité Centrale

Hewlett Packard 2100 A, 32 Kmots de 16 bits, 1 registre d'entrée, 2 registres accumulateurs, addition, multiplication en virgule flottante cablées, 2 canaux d'accès direct à la mémoire autorisant une vitesse de transfert maximale de 1020400 mots par seconde. 52 périphériques peuvent être reliés simultanément.

Mémoires de masse

1) Une unité de disques HP 7900 A à têtes mobiles, qui comprend 2 disuqes, l'un fixe, l'autre amovible comportant chacun 200 pistes de 48 secteurs. La capacité totale de stockage est de 5 millions d'octets et la vitesse de transfert moyenne est de 312000 octets/seconde.

2) Un dérouleur de bande magnétique HP 7970 E à pistes compatible IBM. La densité d'enregistrement est de 1600 bits par inch et la vitsse de défilement de la bande 45 inchs par seconde (variations à court terme de cette vitesse inférieure ou égale à ± 3%) Temps d'arrêt et temps de départ : 8,333 ms. La vitesse de transfert maximale (pour des blocs de longueur infinie) est de 72000 octets par seconde.

Convertisseurs

1) Numérique analogique : 2 canaux indépendant comprenant chacun 1 convertisseur Analog Devices à 12 bits DAC 12 QM. Le temps de conversion est inférieur ou égal à 1 µs. Les deux canaux peuvent être couplés pour donner une précision de 13 bits sur un seul canal.

2) Analogique numérique : 1 canal comprend 1 échantillonneur Analog Devices SHA-2 A (temps d'échantillonnage 500 nanosecondes) et un convertisseur Hybrid Systems ADC 591 A-12 A de 12 bits de précision. Le temps de conversion minimal est de 3,5 µs.

Autres périphériques

- Une imprimante General Electric (Termi Net 340)
- Une console SINTRA servant de pupitre de commande
- Un traceur BENSON 1102
- Un lecteur de ruban perforé HP 2748 A lisant jusqu'à 500 octets par seconde
- Un perforateur de ruban (HP 8100 A) perforant jusqu'à 500 octets par seconde
- Une unité de visualisation (HP 1300 A) ayant une définition d'image de 1/256 en abscisse et en ordonnée.

L'ensemble du système informatique fonctionne sous DOS (Disc Operating System).

Matériel analogique spécifique au traitement du signal sonore

- Horloge externe : C'est un générateur sinusoïdal variable de 1 Hz à 1 MHz (Philipps 5160). La fréquence du signal est pralablement divisée par 10 (gamme utile : C, 1 Hz à 100 kHz) et le signal transformé en impulsions rectangulaires de 200 ns de largeur et de même période que la sinusoïde. Cette suite d'impulsions détermine la cadence d'échantillonnage.

- Filtres : 2 canaux indépendants comprenant chacun un filtre passe bas fixe construits suivant le schéma fourni par Dunbar (Mathews 1969, p. 30), ayant une bande passante à 1 dB de Ø à 7000 Hz et une atténuation supérieure à 50 dB à 10000 Hz

1 filtre réglable de 0,1 Hz à 100 kHz à 2 voies indépendantes ayant chacune une atténuation de 24 dB par octave (Schumberger FAB 24).

Magnétophones

Revox A 77 - Rapport signal-bruit = 60 dB environ

Amplificateur

Revox A 78 : 2 40 watts (4 - 3 ohms)

Microphone

Boyer électrodynamique M88N

Baffles

2 enceintes Acoustic Research AR3

Les figures présentent les spectres individuels de 6 phonèmes, ainsi qu'un spectre moyen.

Les figures qui suivent présentent les spectres individuels de 6 phonèmes, ainsi qu'un spectre moyen.

Sur les figures 1a, 1b, 1c on a les spectres de trois phonèmes consonantiques voisés ; dans ces trois spectres, le fondamental et quelques uns de ses harmoniques sont nettement visibles. On notera que l'importance relative de ces harmoniques n'est pas du tout la même : d'où dans le profil sonore de grandes différences sur lesquelles se fonde l'analyse factorielle pour discriminer les sons. Les sons B et D objets des figures 1b et 1c ont en arabe une prononciation voisine de celle des sons français correspondants. Le son ayn (fig. 1a) parfois représenté dans les transcriptions par deux a consécutifs, (ex : aabd - allaḥ) et fortement grasseyé, montre une suite remarquable de 9 harmoniques après le fondamental.

Sur les figures 1d, 1e, 1f on a des sifflantes non voisées , caractérisées par un bruit intense dans la bande des hautes fréquences . Pour le CH (fig. 1f) la bande de bruit est nettement plus basse que pour le S (fig. 1e) ; le sad (s emphatique propre à l'arabe) se situe un peu plus bas que le S; de plus en basse fréquence (0 - 1000 Hz) l'harmonique 4 est nettement visible, alors qu'il disparaît dans le S.

La figure 2 , moyenne de l'ensemble des spectres, montre la répartition globale de l'énergie sonore ; c'est d'après cette figure qu'on peut en première approximation choisir les bornes des canaux suivant lesquels on code les spectres dans l'analyse factorielle.

Figure 1.a

Le spectre représenté est une fraction d'un ayn ع dont la durée totale est d'environ 165 ms. Cette fraction se situe 100 ms après le début du phonème ع dans le mot اسعف

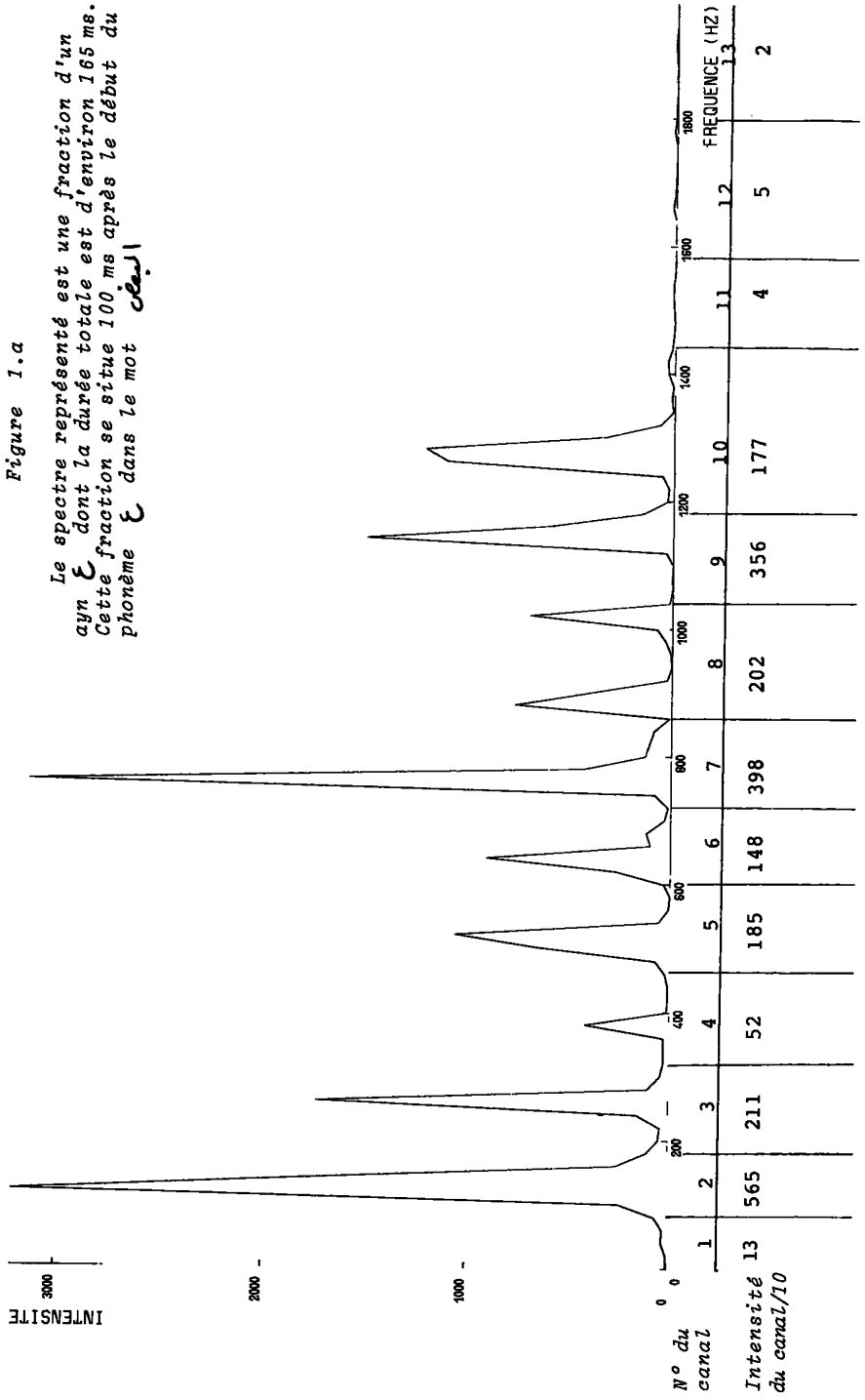
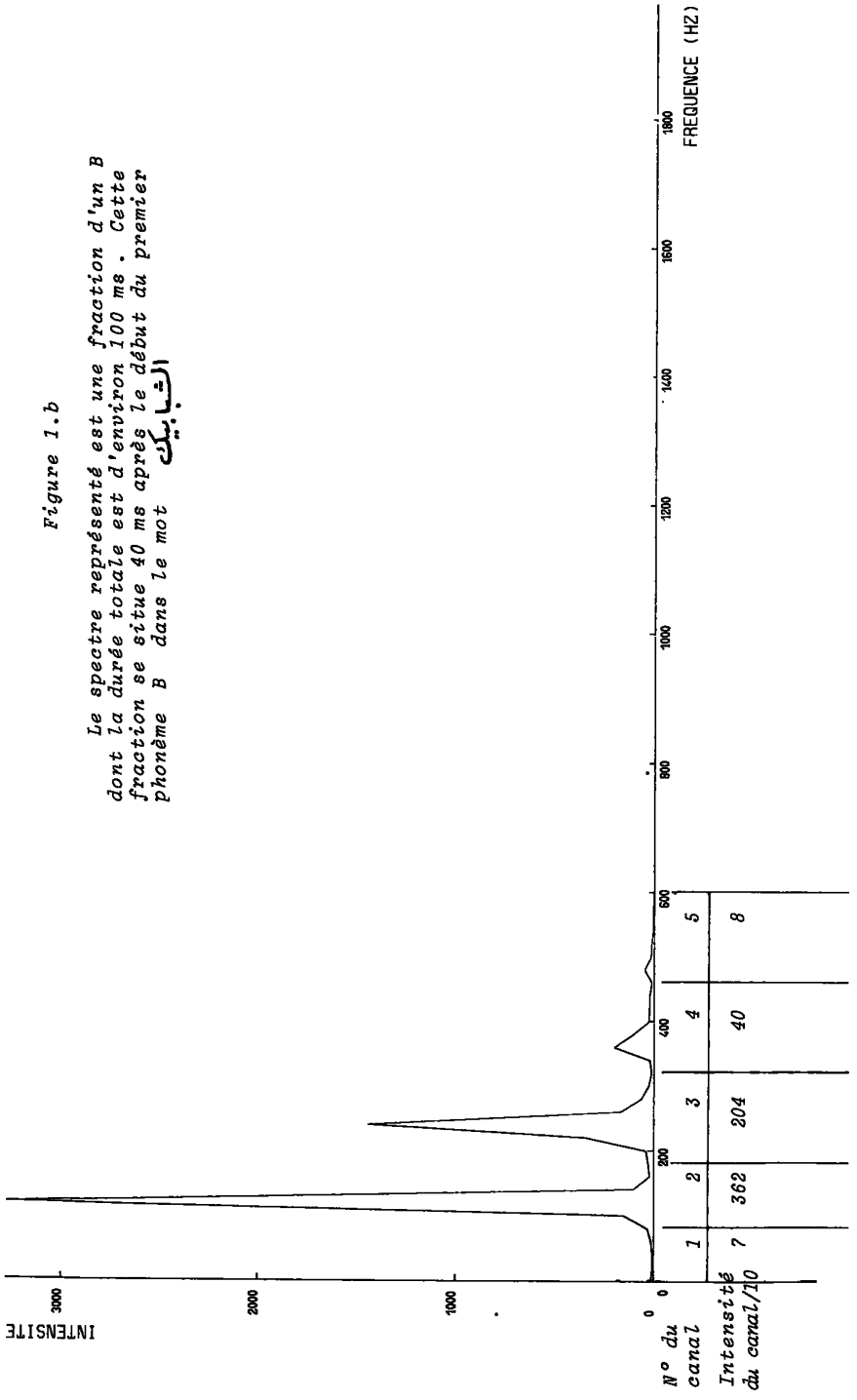


Figure 1.b

Le spectre représenté est une fraction d'un B dont la durée totale est d'environ 100 ms . Cette fraction se situe 40 ms après le début du premier phonème B dans le mot الثعابين



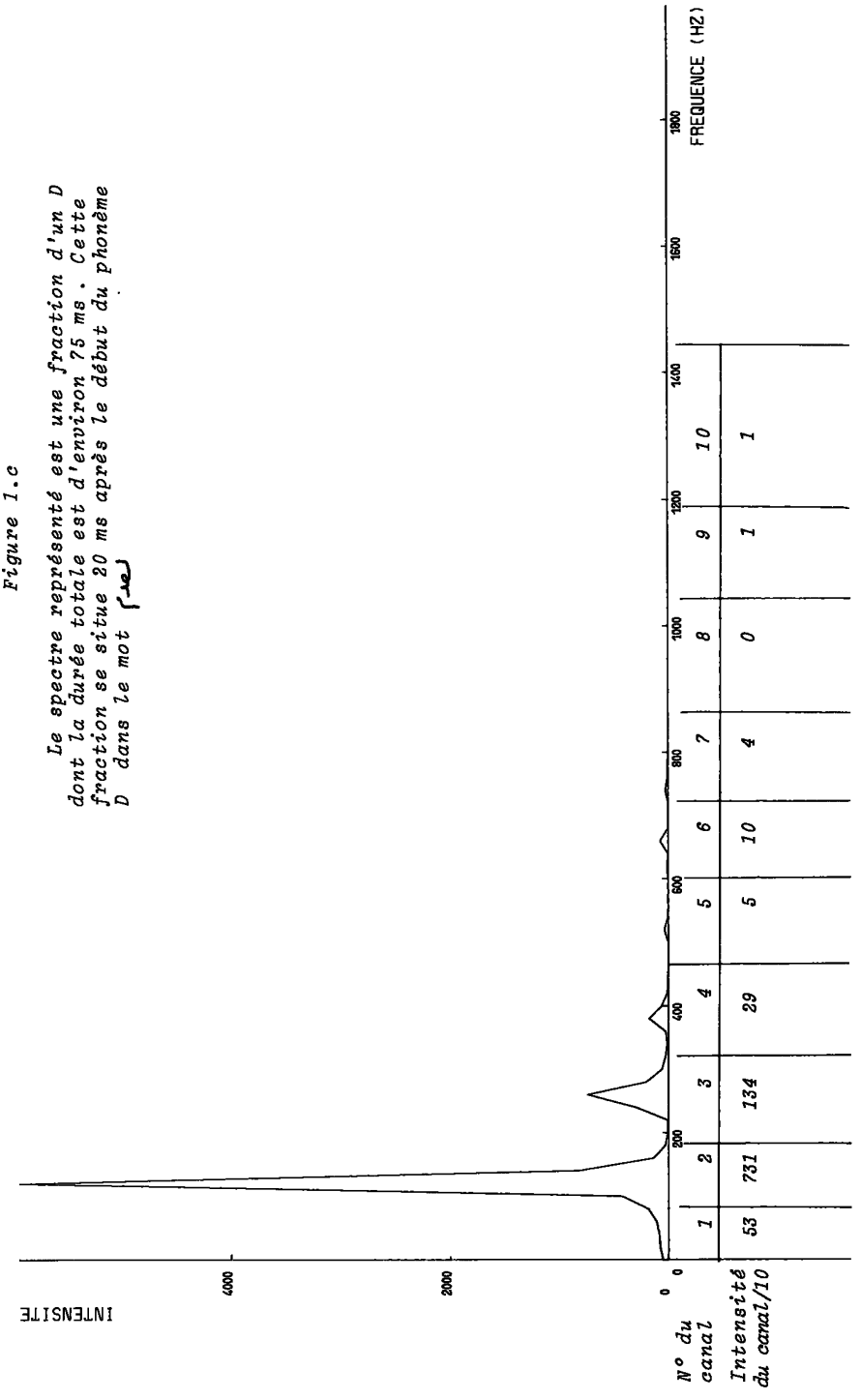
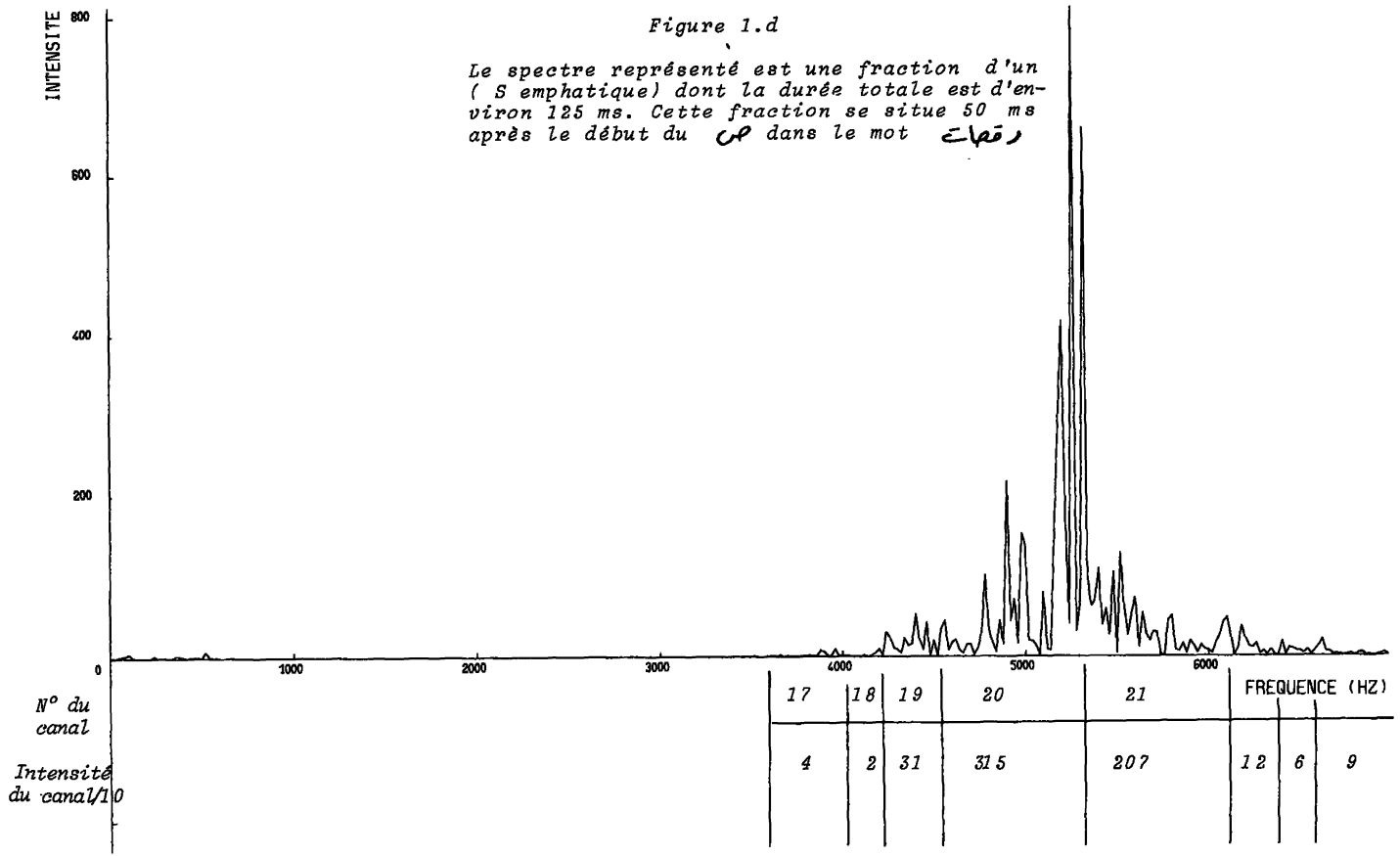


Figure 1.c

Le spectre représenté est une fraction d'un D dont la durée totale est d'environ 75 ms . Cette fraction se situe 20 ms après le début du phonème D dans le mot صدر

Figure 1.d

Le spectre représenté est une fraction d'un (S emphatique) dont la durée totale est d'environ 125 ms. Cette fraction se situe 50 ms après le début du ص dans le mot رقصات



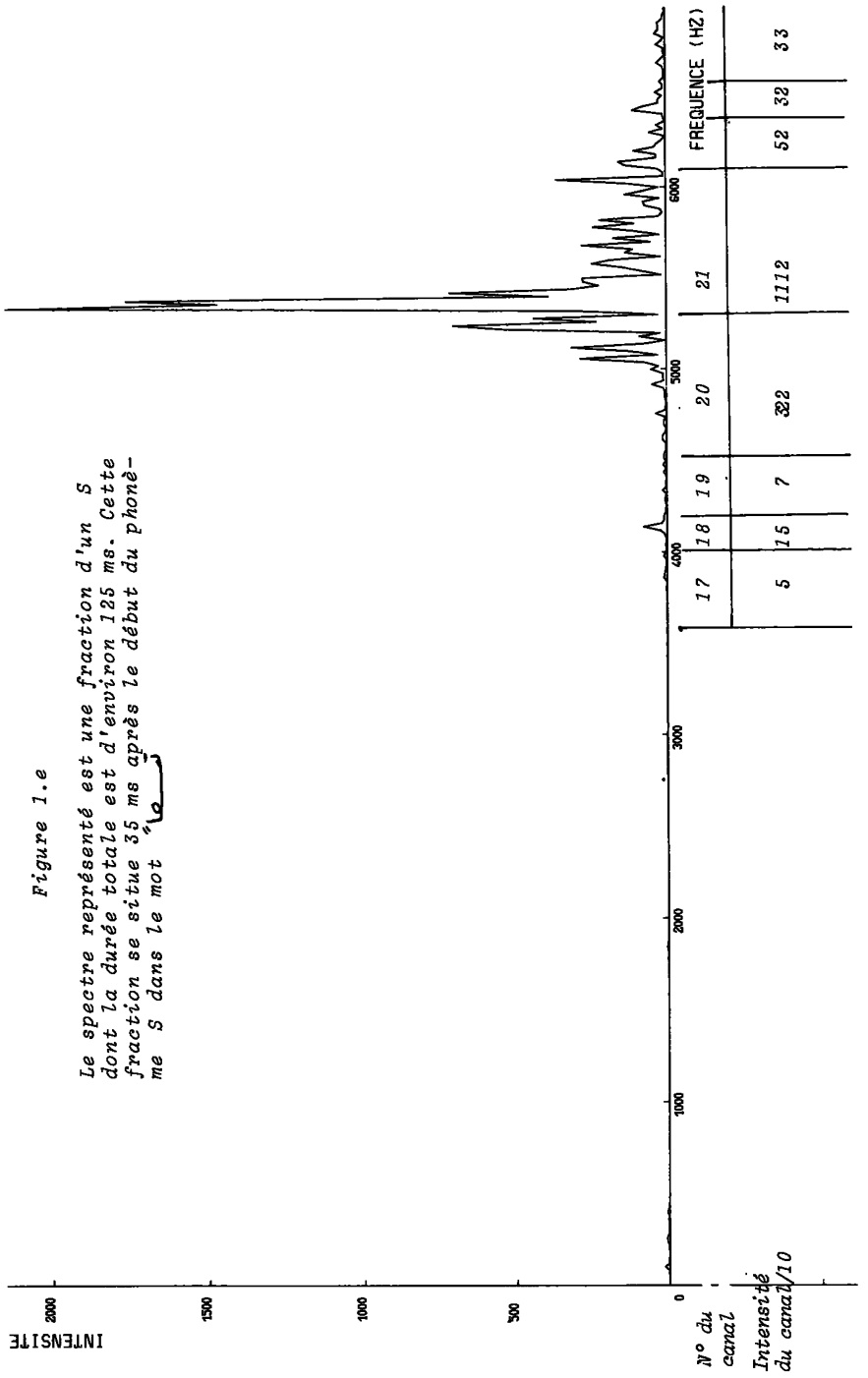


Figure 1.e

Le spectre représenté est une fraction d'un S dont la durée totale est d'environ 125 ms. Cette fraction se situe 35 ms après le début du phonème S dans le mot صـ

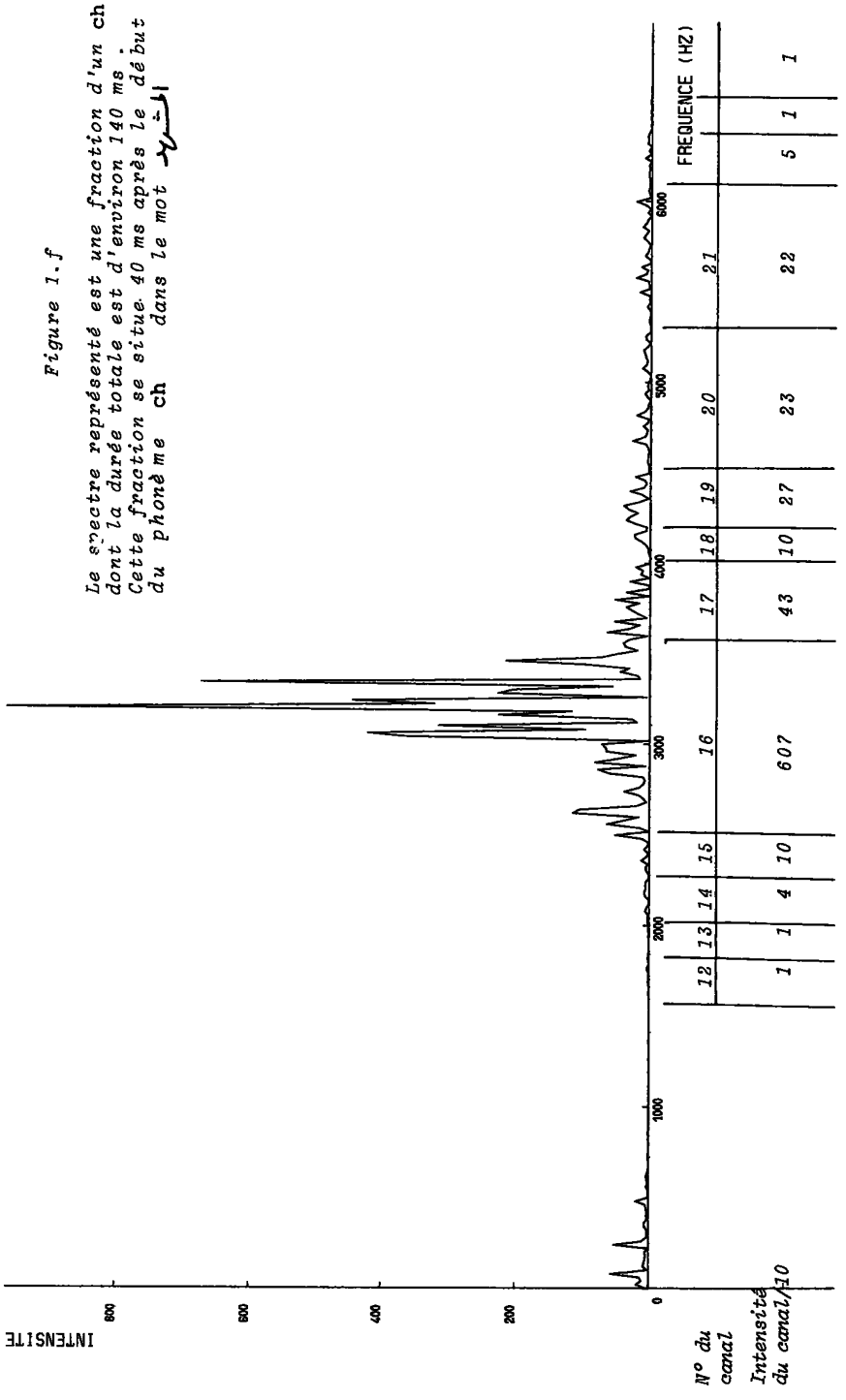


Figure 1.f

Le spectre représenté est une fraction d'un ch *dont la durée totale est d'environ 140 ms. Cette fraction se situe 40 ms après le début du phonème* ch *dans le mot* الشرب

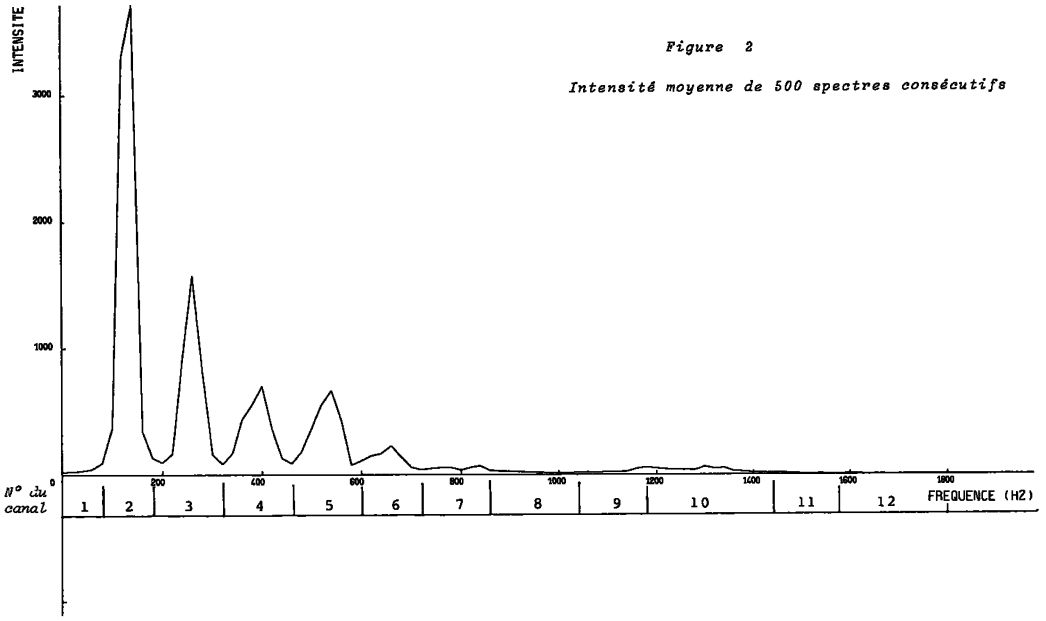
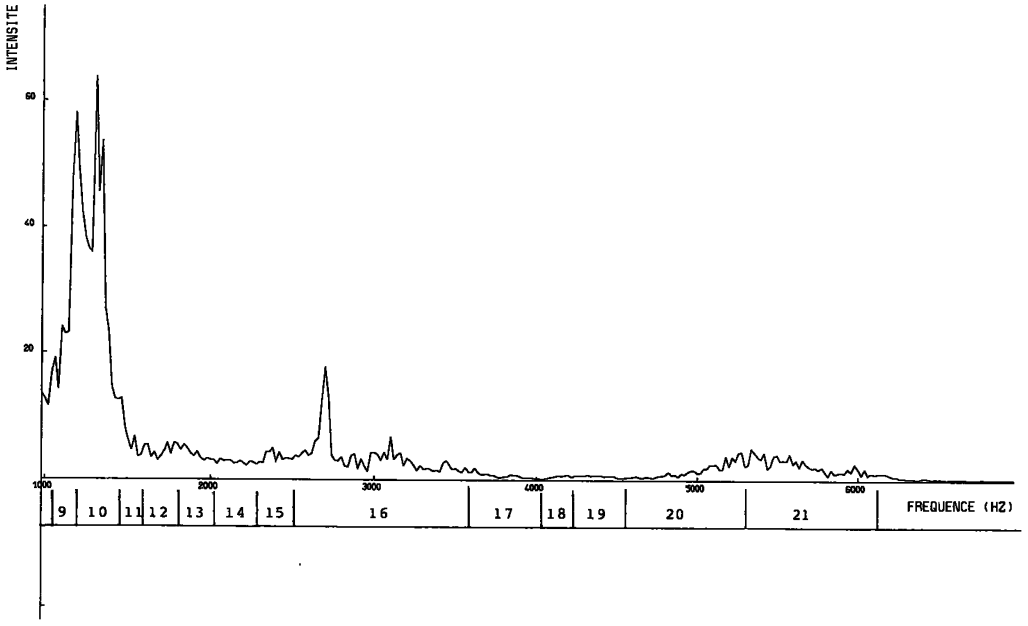


Figure 2

Intensité moyenne de 500 spectres consécutifs



BIBLIOGRAPHIE

S. CHABREL, G. CHARBONNEAU : Un système d'analyse de sons par ordinateur fonctionnant en mode conversationnel *Onde Electrique*, Vol. 8-9, p. 358 (1976)

G. CHARBONNEAU : *Réalisation d'un système intégré d'analyse et de synthèse de sons par ordinateur. Application de ce système à l'étude de l'attribut de hauteur sonore et de sa perception*. Thèse de doctorat. Orsay (1976)

J. W. COOLEY, P.A.W. LEWIS & P.D. WELCH : Historical notes on the Fast Fourier Transform *Proc. IEEE* Vol 55 n° 10 (oct. 1967)

J. W. COOLEY, J.W. TUKEY : An algorithm for the machine calculation of complex Fourier series. *Math. Computations*, Vol 19, pp 297-301 (1965)

M. DREYFUS : *FØRTRAN IV*. Dunod Ed., Paris (1969)

J.B. FOURIER : *Théorie analytique de la chaleur*. Paris (1822)

D.D. GREENWOOD : Critical Badwidth and the frequency coordinates of the basilar membrane. *J. Acoust. Soc. Am.*, Vol 13, pp 1344-1356 (1961)

H. HELMHOLTZ : *Die Lehre von der Tonempfin dungen als physiologische Grundlage für die Theorie der Musik* Ed. Vieweg und Sohn, Braunschweig (5° édition) (1896)

ou

H. HELMHOLTZ : *On the sensations of tone* (traduction anglaise par A.J. Ellis) Dover Public., New-York (1954)

J.F. KAISER : *System analysis by digital computer* Kuo and Kaiser Eds., New-York, Wiley (1966)

T.S. LITTLER : *The physics of the ear* Pergamon Press Ltd, London (1965)

M.V. MATHEWS : *The technology of computer music* M.I.T. Press, Cambridge, Mass. (1969)

A. MOLES : *La structure physique du signal sonore*. Thèse Bibliothèque de la Sorbonne, Paris (1952)

R. PLOMP : The ear as a frequency analyzer *J. Acoust. Soc. Am.*, Vol 36, pp 1628-1636 (1964)

L.R. RABINER, B. GOLD : *Theory and Applications of digital signal processing* Prentice Hall, Inc., Englewood Cliffs, New Jersey p. 88 (1975)

C. RUNGE : *Zeit für Math and Physik*, Vol 48, p. 443 ; (1903) et : *ibid*, Vol 53, p. 117 ; (1905)

C.E. SHANNON, W. WEAVER : *The mathematical theory of communication* University of Illinois Press (1962)

L. SCHWARTZ : *L'intégrale de Fourier Méthodes Mathématiques de la Physique* Vol 5, pp 1-34 Centre de Documentation Universitaire , Paris (1963)

J.M. WHITTAKER : *Interpolatory function theory* Cambridge University Press, London (1915)