GROUPEMENT des ACOUSTICIENS de LANGUE FRANÇAISE

GROUPE **COMMUNICATION PARLEE**

JOURNEES D'ETUDES

SUR LA PAROLE

.....

AIX-EN-PROVENCE - 1/2 AVRIL 1971

Groupement des Acousticiens de Langue Française
Groupe COMMUNICATION PARLEE

JOURNEES D'ETUDES SUR LA PAROLE

1 - 2 AVRIL 1971
AIX - EN - PROVENCE

Les *JOURNEES D'ETUDES SUR LA PAROLE* qui se sont déroulées les 1 et 2 avril 1971 à l'*Institut de Phonétique de la FACULTE des LETTRES dAIX-en-PROVENCE* ont été organisées grâce à la diligence de Monsieur *M. ROSSI* et avec le soutien du Comité spécialisé : *Intelligence artificielle et Reconnaissance des Formes de l'A.F.C.E.T.*

Vous trouverez dans ce volume les textes complets des exposés et discussions sur les thèmes abordés.

S O M M A I R E

C,
Communications présentées au cours
de la MATINEE du 2 AVRIL 1971

	ALLOCUTION D'INTRODUCTION	M. DREYFUS-GRAF
C/a.	VOCODER NUMERIQUE	M. LAVANANT
C/b.	PRETRAITEMENT ET RECONNAISSANCE DE LA PAROLE - SIMULATION ET REALISATIONS PRATIQUES	MM. HATON & LAMOTTE
C/c.	PARAMETRISATION ET PROCEDURES DE RECONNAISSANCE DE LA PAROLE	MM. GUEGUEN, MAISSIS & PAU
C/d.	SEGMENTATION DE LA PAROLE ET RECONNAISSANCE DES SYLLABES A L'INTERIEUR DES MOTS	M. MERCIER
C/e.	APPLICATION DES TECHNIQUES STATISTIQUES A LA RECONNAISSANCE DE LA PAROLE	M. BERGER-VACHON
C/f.	IDEES GENERALES SUR LA RECONNAISSANCE DES FORMES APPLIQUEE A LA PAROLE	M. ROCHE
C/g.	RECONNAISSANCE DE PHONEMES AU MOYEN D'UNE COHLEE ARTIFICIELLE	M. ALINAT
C/h.	RECONNAISSANCE DE LA PAROLE EN TEMPS REEL	MM. CAELEN, CASTAN & PERENNOU

DISCOURS D'INTRODUCTION

prononcé par Monsieur *PIMONOW*

Président du G.A.L.F.

Quoique la vision permette la réception du débit informationnel, le plus grand atteignant 2 à 3 Mbits/s, l'échange principal d'information entre les humains se fait avec un débit de quelque 50 kbits/s sur le chemin acoustique et notamment au moyen de l'audition et de la parole.

La capacité de parler est aussi, de loin, la plus importante source d'émission des informations que possède un homme.

Evidemment, ce n'est pas chaque parole qui porte des informations.

Les discours de bienvenue ou d'inauguration, comme par exemple celui que je suis en train de faire, ont le plus souvent un débit d'information utile quasiment nul.

De même, en abusant du don de pouvoir converser, on exprime parfois ses soi-disant idées avant de réfléchir, de sorte que les mots émis n'ont aucune valeur et représentent simplement un bruit nuisible.

Mais, à part ces cas exceptionnels et par suite de l'importance primordiale que représente la parole pour les échanges d'information, l'intérêt de son étude est évident.

Ici, je veux souligner que les progrès relatifs à l'information, comme par exemple l'apparition de l'écriture ou la découverte de l'imprimerie par GUTENBERG, étaient capitaux et représentent la base de l'évolution de notre civilisation. Ainsi, toutes les recherches relatives à l'échange, le stockage ou le traitement de l'information, y compris celle relative à la communication parlée, peuvent être considérées comme fondamentales et ayant la plus grande importance. Ce colloque mérite donc la plus vive félicitation et encouragement.

Les études techniques et scientifiques rigoureuses de la parole ont été pratiquement inaugurées par les techniciens des télécommunications qui s'intéressent tout particulièrement aux conditions d'intelligibilité.

En évoluant, l'étude de la parole s'apparente de plus en plus à l'étude de l'information en s'attaquant, d'une part, aux propriétés physiques des signaux et, d'autre part, aux débits, la capacité, la reconnaissance, etc.., où on fait de plus en plus fréquemment appel à l'appareil mathématique et aux expériences de caractère statistique.

Du point de vue physique, la parole est un phénomène vibratoire qu'on définit par les paramètres fréquence et amplitude en fonction du temps, ce qui impose, dans les études, l'analyse spectrale, en particulier l'analyse des phénomènes transitoires.

Un autre aspect d'étude de la parole est basé sur la théorie d'information. Il s'attaque par exemple aux problèmes d'adaptation du débit aux chaînes de transmission ou aux récepteurs, y compris l'oreille, et aussi aux problèmes de la reconnaissance. Si le problème du débit a un caractère quantitatif d'information et sa théorie a assez bien évolué, en particulier grâce aux travaux de SCHANNON, le problème de reconnaissance a un caractère déjà qualitatif et sa théorie est encore assez vierge.

Théoriquement, la reconnaissance des objets, y compris l'objet sonore, tel que la parole, dépend très peu des propriétés physiques des signaux qui représentent l'objet, mais beaucoup plus des relations fonctionnelles entre eux, qui forment les critères de reconnaissance.

Ainsi, l'oreille, par exemple, arrive à comprendre aussi bien la parole naturelle qu'une parole chuchotée où les composantes spectrales discrètes sont absentes. Comme le démontrent les expériences avec le Vocoder, la parole reste intelligible même si l'on la compose de sons musicaux, jamais utilisés par les humains. Toutefois, un traitement des informations aussi complexes qu'exécute, dans un tel cas, notre système nerveux est techniquement sinon impossible, du moins très difficile.

Or, dans le système technique de reconnaissance, il est plus facile d'utiliser directement les signaux comme critère, ou plus exactement de les utiliser après un traitement aussi simple que possible, et non les relations fonctionnelles entre ceux-ci, trop complexes. C'est ce chemin de simplifications qu'essaient de suivre les systèmes techniques et qui sont d'autant plus séduisants que les paramètres physiques des signaux qui composent la parole naturelle, sont enfermés dans des cadres sonores relativement étroits.

Pour terminer, rappelons que, si la parole ne représentait aucun problème scientifique ou technique pour nos ancêtres, avec l'évolution de la science, elle devient un problème de plus en plus complexe, de sorte que l'on ne voit guère le fond de l'étude.

A.

COMMUNICATIONS PRESENTEES AU COURS DE LA JOURNEE DU 1 AVRIL 1971

APERÇU SUR LE TRAITEMENT, PAR L'ANALYSE DE FOURIER, DE LA PAROLE CONSIDEREE COMME UN SIGNAL DETERMINISTE

M. TIMSIT
E.D.F. - Clamart

I - INTRODUCTION

On se propose, dans le présent exposé, de rappeler quelques caractéristiques fondamentales, et, plus encore, quelques limites irréductibles, de la Transformation de FOURIER, utilisée comme outil d'analyse des signaux de parole.

Une hypothèse sera avancée, qui prête largement à contestation : celle du caractère déterministe des signaux de parole. Il convient, toutefois, de tempérer cette contestation par deux éléments : d'une part la difficulté certaine pour le praticien, de tracer avec rigueur la frontière entre l'aléatoire et le déterministe, frontière d'autant plus floue qu'elle dépend apparemment de l'observateur, et, d'autre part, la grande ressemblance, que nous soulignerons dans la conclusion, entre les traitements spectraux effectués matériellement sur ces deux catégories de signaux.

Avant d'entrer plus avant dans le détail, nous voudrions rappeler que, pour traditionnelle et efficiente que soit l'analyse de FOURIER, elle ne saurait épuiser l'information que l'on peut extraire d'un signal donné, et que d'autres modes d'analyse, telle l'étude des statistiques de franchissements de seuils, peuvent avantageusement conforter, ou à la limite remplacer, la décomposition spectrale usuelle. Alors qu'apparaît irréfutable l'intérêt des exponentielles complexes dans l'étude des systèmes linéaires, dont elles sont les fonctions propres, il pourra s'avérer utile de recourir à de tout autres traitements si l'on a en vue, en définitive, de caractériser un signal, par exemple pour le comparer à une référence. A vrai dire, et nous reviendrons sur ce point, il est essentiel de savoir si l'on désire résumer le signal, c'est-à-dire le condenser de façon irréversible, ou si l'on désire au contraire le représenter dans un formalisme différent, qui autorise la reconstitution du signal original.

II - LA TRANSFORMATION DE FOURIER

Résumé des propriétés principales.

a) Rappel sur les séries de FOURIER

Leur utilisation s'appuie sur le théorème suivant :

le signal réel $X(t)$ étant à variation bornée sur $[o, T_o]$, la quantité

$$\hat{X}(t) = \sum_{-\infty}^{\infty} C_n e^{2\pi i \frac{n}{T_o} t}, \text{ où } C_n = \frac{1}{T_o} \int_o^{T_o} X(t) e^{-2\pi i \frac{n}{T_o} t} dt,$$

converge uniformément et en moyenne quadratique vers $X(t)$, en tout point où ce dernier est continu.

Autrement dit, lorsque le développement existe, $X(t)$ est interprétable comme une somme éventuellement infinie de sinusoïdes ayant chacune

1) sa fréquence, toujours multiple de $V_o = \frac{1}{T_o}$,

2) son amplitude, le cas échéant nulle,

3) sa phase.

Bien entendu, le caractère réel de $X(t)$ implique entre les coefficients C_n la relation $C_{-n} = C_n^*$ pour tout n, et, de plus, on peut conserver, à toutes fins utiles, une notation réelle équivalente :

$$X(t) = \sum_{-\infty}^{\infty} a_n \cos 2\pi n V_o t + \sum_{-\infty}^{\infty} b_n \sin 2\pi n V_o t$$

Avant d'aborder l'intégrale de FOURIER, mentionnons le très important théorème de PARSEVAL :

$$\sum_{-\infty}^{\infty} |C_n|^2 = \frac{1}{T_o} \int_o^{T_o} |X(t)|^2 dt,$$

qui a le mérite de permettre le calcul de la puissance véhiculée, aussi bien dans le domaine fréquence que dans le domaine temps, et qui suggère de répartir ladite puissance en "raies" composantes, ce qui s'écrit :

$$S(V) = \sum_{-\infty}^{\infty} |C_n|^2 \delta(V - n V_o),$$

à condition d'admettre sans autre forme de procès l'emploi de la distribution de DIRAC.

b) Intégrale de FOURIER

Ayant en vue de traiter des signaux non périodiques, on peut se demander comment évoluent les composantes précédentes lorsque T_o augmente indéfiniment : l'intuition montre que les raies se rapprochent. Nous introduirons plutôt comme une définition la relation suivante :

$$X(t) = \int_{-\infty}^{\infty} x(V)\, e^{2\pi i V t}\, dV \qquad (1),$$

qui pourra souvent s'inverser en :

$$x(v) = \int_{-\infty}^{\infty} X(t)\, e^{-2\pi i tv}\, dt \qquad (1')$$

Suivant l'éclairage que l'on veut utiliser, on parlera de décomposition de FOURIER, de représentation de FOURIER, ou de transformation de FOURIER pour qualifier l'intégrale (1), et plus généralement de couple (1), (1') ; on écrira par exemple :

$$X(t) \rightleftharpoons x(v)$$

Naturellement l'intégrale (1) ne converge pas toujours. Deux conditions, suffisantes chacune, sont fréquemment citées dans les ouvrages spécialisés :

a) $\int_{-\infty}^{\infty} |X(t)|\, dt < \infty$

b) $\int_{-\infty}^{\infty} |X(t)|^2\, dt < \infty$, que l'on peut symboliser respectivement par :

a) $X \in L$

b) $X \in L^2$

La seconde a une interprétation immédiate : le signal doit être à énergie finie. On aperçoit aussitôt des signaux, très utilisés dans les développements théoriques, qui ne satisfont ni a) ni b) : les sinusoïdes d'une part, les signaux constants de l'autre. Le recours aux distributions de SCHWARTZ permet heureusement de sortir de l'impasse, en attribuant à de tels signaux (le second étant en quelque sorte un cas particulier du premier) une transformée de FOURIER

Notons au passage que les signaux du physicien, de durée et de puissance forcément finies, appartiennent inéluctablement à L^2 ...

La réciprocité de la transformée de FOURIER , quant à elle, n'est pas toujours assurée; elle est toutefois acquise pour les fonctions de L^2 .

QUELQUES PROPRIETES DE LA TRANSFORMEE DE FOURIER

A) Linéarité

$$\sum_{i=1}^{n} a_i X_i(t) \rightleftharpoons \sum_{i=1}^{n} a_i x_i(v) \qquad (a_i \text{ complexe})$$

B) Changement d'échelle

$$X(a\,t) \rightleftharpoons \frac{1}{|a|}\, x\left(\frac{v}{a}\right)$$

C) Translation

en temps $X(t - t_o) \rightleftharpoons x(v) \cdot e^{2\pi i v t_o}$

en fréquence $X(t) \cdot e^{-2\pi i v_o t} \longleftrightarrow x(v - v_o)$

donc $X(t) \cos 2\pi v_o t \longleftrightarrow \frac{1}{2}\left[x(v - v_o) + x(v + v_o)\right]$

D) Dérivation

$$\frac{d^n X(t)}{d t^n} \longleftrightarrow (2\pi i v)^n x(v)$$

E) Symétrie : dans le cas le plus général, $X(t)$ est complexe, et $x(v)$ également. Avec $X = X_1 + i X_2$ et $x = x_1 + i x_2$, on obtient :

$$(1)\quad x_1(v) = \int_{-\infty}^{\infty} \left[X_1(t) \cos 2\pi vt + X_2(t) \sin 2\pi vt\right] dt$$

$$(2)\quad x_2(v) = \int_{-\infty}^{\infty} \left[-X_1(t) \sin 2\pi vt + X_2(t) \cos 2\pi vt\right] dt$$

$$(3)\quad X_1(t) = \int_{-\infty}^{\infty} \left[x_1(v) \cos 2\pi vt - x_2(v) \sin 2\pi vt\right] dv$$

$$(4)\quad X_2(t) = \int_{-\infty}^{\infty} \left[x_1(v) \sin 2\pi vt + x_2(v) \cos 2\pi vt\right] dv$$

Les sypétries proprement dites résultent de ces relations et apparaissent sur la figure jointe en annexe 1.

Dans le cas très courant où $X(t)$ est réel, on note en particulier que l'hermiticité de $x(v)$ rend inutile la connaissance simultanée des composantes aux fréquences positives et négatives, et ceci simplifie les programmes de T.F. sur ordinateur. La quantité complexe $x(v) = x^*(-v)$, peut se représenter de multiples façons, en coordonnées cartésiennes, ou polaires.

On pressent enfin, et ceci est confirmé par l'équation (4), que le caractère réel de $X(t)$ implique une certaine dépendance entre les parties réelle et imaginaire de $x(v)$ [ou encore entre son module et sa phase] ; cette dépendance devient du reste totale (transformation de Hilbert) si l'on exige de surcroît, comme en théorie des filtres, que $X(t)$ soit causal, c'est-à-dire nul pour t négatif.

On trouvera, en annexe 2, quelques exemples de T.F., choisis parmi les plus classiques.

III - ANALYSE SPECTRALE par T.F., dans le cas de signaux déterministes

Nous avons précédemment décomposé $X(t)$, par analyse harmonique, en une superposition, le cas échéant infinie, d'exponentielles complexes $x(v) \cdot e^{2\pi i v t}$. Ayant en vue une évolution de la puissance, on renoncera naturellement à la phase de ces exponentielles. La considération du module $|x(v)|$ conduit au spectre d'amplitude; par élévation au carré, on accède au spectre énergétique $|x(v)|^2$, qui indique comment l'énergie totale du signal, finie puisqu'il appartient à L^2, se distribue entre les différentes bandes élémentaires de fréquence.

A ce stade, le physicien peut ressentir le besoin de parler en termes de puissance et non plus d'énergie, en dépit de ce que, à strictement parler, la puissance moyenne d'un signal fini soit forcément nulle; ceci lui permet notamment d'établir un lien entre sa décomposition spectrale et le niveau global de puissance qu'il mesure au voltmètre efficace (dont la constante d'intégration n'est pas infinie).

L'exemple des signaux périodiques pousse ainsi à généraliser la décomposition précédente jusqu'à y inclure les fonctions à énergie totale infinie, et à puissance moyenne finie, pourvu que la puissance en question soit évaluée sur une durée finie. A la limite, on peut poser la définition précise suivante de la <u>densité spectrale de puissance moyenne</u> :

$$S_{xx}(v) = \lim_{T \to \infty} \frac{|x(v, T)|^2}{2T}$$

Dans cette formule, $x(v, T)$ représente la T.F. du signal $X(t)$, systématiquement annulé en dehors de $[-T, +T]$. Quant à l'appellation de densité, elle permet d'exprimer que la puissance véhiculée dans la bande $[v_1, v_2]$ peut se calculer par intégration de $S_{xx}(v)$ entre ces bornes, et elle évoque en outre le caractère positif de $S_{xx}(v)$.

Comme, dans la pratique, on se fixe une finesse spectrale Δv, on sait qu'elle est susceptible d'être atteinte dès que le signal dure plusieurs $\tau = 1/\Delta v$, par exemple 4τ.

Supposons que l'on désire, physiquement ou par simulation sur calculateur, faire traverser au signal un filtre de réponse impulsionnelle R (t), la sortie de ce filtre, obtenue en convoluant l'entrée avec R (t), sera généralement élevée au carré et intégrée sur une durée T. On voit ainsi que seule une intégration infinie permettrait d'atteindre le spectre du signal.

Là réside, véritablement, le choix crucial à effectuer. L'obtention exacte de S_{xx} (v) nécessite une intégration infinie; or, bien souvent, l'utilisateur souhaite disposer d'une décomposition spectrale évolutive, qui permette de suivre, en quelque sorte en temps réel, les "péripéties" que connaît le signal. Force lui est donc d'adopter une intégration finie, par exemple de 100 ms. Il se contentera, en somme, du passé spectral récent exclusivement, en échange de quoi il pourra observer la variabilité du "spectre" ainsi défini. Certains auteurs distinguent entre le "spectre instantané", intégré sur le passé récent seulement, et le "spectre évolutif", qui prend en compte à chaque estimation tout le passé depuis l'origine. On conçoit que l'expérimentateur, face à ce choix, soit contraint à un compromis entre son désir d'approcher la formule théorique du spectre, et celui d'en observer la déformation progressive. L'incompatibilité que l'on devine entre ces impératifs, auxquels il faut adjoindre celui de laisser se dérouler suffisamment longtemps la réponse R (t) du filtre, peuvent faire l'objet de nombreux développements mathématiques, qui seront examinés dans une autre conférence de cette série, sous le nom de représentation temps-fréquence.

Même en prenant la précaution d'évaluer les spectres instantanés successifs sur des durées compatibles avec la finesse spectrale désirée (1), on observera, en définitive, autant de résultats différents que l'on affichera de constantes d'intégration différentes ... A la limite, on pourra présenter comme extrêmement stable un signal de parole, ou comme typiquement évolutif un signal vibratoire réputé stationnaire, par un choix judicieux de cette constante d'intégration (par exemple 1 mn dans le premier cas, 10 ms dans le second).

Dans le même ordre d'idées, le niveau global acoustique, sur les fluctuations duquel s'appuient des normes en cours d'étude actuellement, présentera des variations d'autant moins accusées que la constante d'intégration associée sera plus élevée : d'où la nécessité absolue de préciser cette constante d'intégration, et même, en toute rigueur, la réponse impulsionnelle complète du circuit intégrateur.

Le champ d'utilisation de la T.F. ne se limite pas à l'analyse spectrale monovoie considérée plus haut : elle permet par exemple de procéder à des traitements croisés très fructueux, destinés à mettre en relief les composantes spectrales communes à deux signaux donnés, le déphasage existant en ces signaux pouvant être fourni pour chaque composante. Il semble, toutefois, que ces traitements croisés n'aient pas encore rencontré de succès très spectaculaire en analyse de la parole, probablement parce que les signaux de parole sont généralement captés comme des scalaires X (t).

IV - CONCLUSIONS

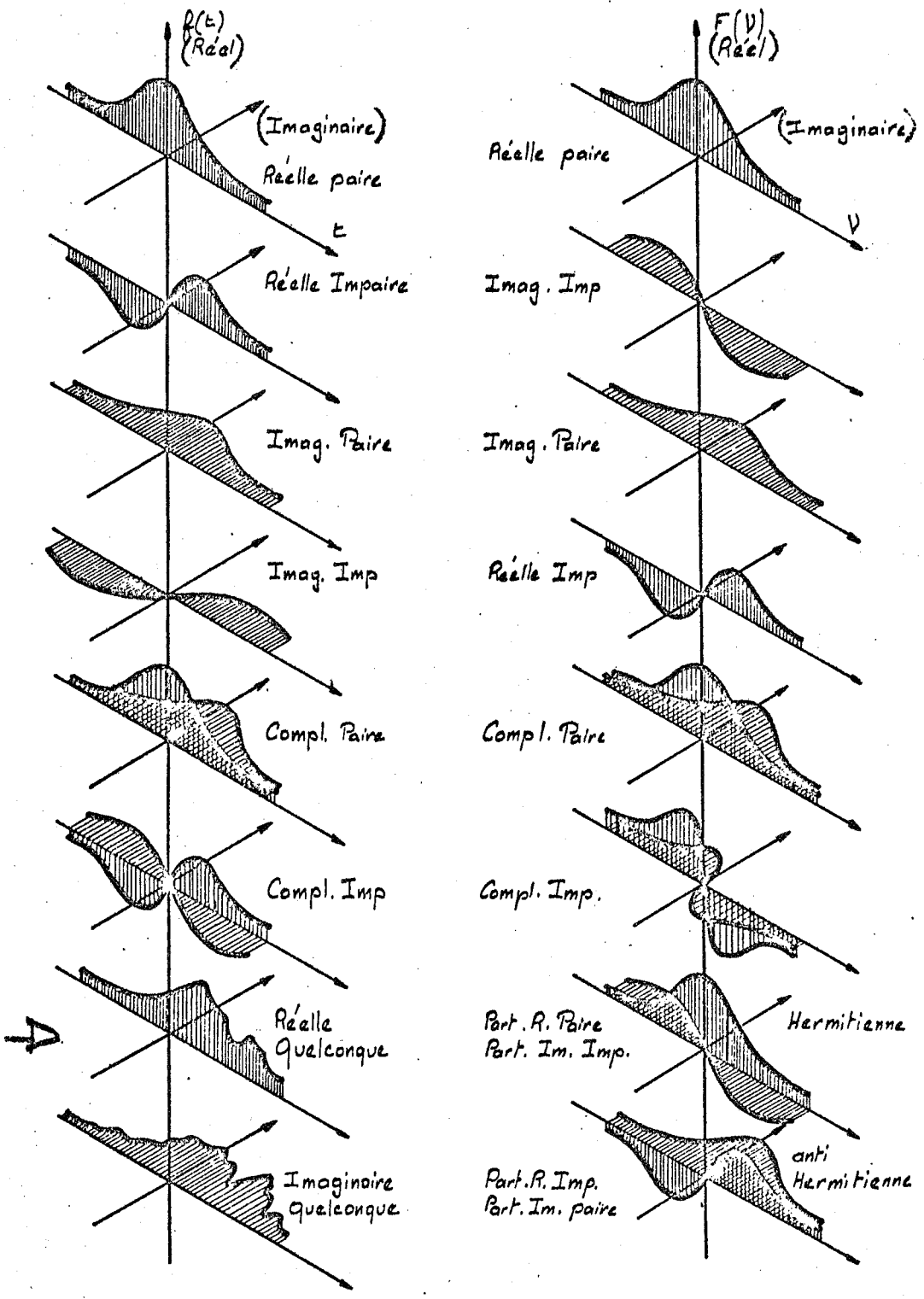
Bien qu'il soit indéniablement utile de connaître les limitations de la méthode d'analyse que l'on emploie, on peut fort bien adopter un point de vue profondément pragmatique. Etant donné un signal X (t), dans la mesure où le problème posé est finalement de le caractériser par quelques paramètres permettant sa reconnaissance - même s'ils appauvrissent l'information de départ -, il importe que le traitement envisagé fasse émerger de ce signal les caractères distinctifs adequats : qué le produit de cette analyse s'identifie ou non à la quantité S_{xx} (v) définie précedemment apparaît, dans ce contexte, totalement accessoire. Hormis des cas précis comme la traversée des systèmes linéaires, où la décomposition spectrale aboutit à une compréhension ou à une prévision meilleures des phénomènes, il semble en effet que la "boite noire" représentant le traitement retenu, doive être appréciée avant tout pour son aptitude à clarifier un problème de reconnaissance, ou encore à définir des normes de nuisance, et non pas pour son aptitude à illustrer la théorie de FOURIER ...

Conservant ce même point de vue empirique, on conviendra enfin que le caractère aléatoire ou déterministe d'un signal, aussi tranché et important soit-il dans les ouvrages théoriques, n'intervient pas ici de façon fondamentale : rien n'empêche l'utilisateur qui se trouve en présence d'un signal, de lui imposer dans tous les cas le traitement qu'il a conçu a priori, à charge pour lui d'exploiter et d'interprêter à bon escient, dans un deuxième temps, le fruit de ce traitement.

(1) ou encore d'adapter la finesse spectrale à la cadence d'analyse requise.

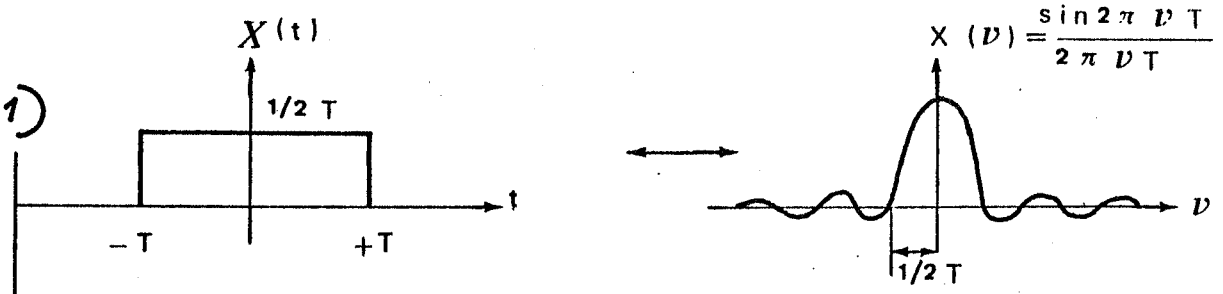
N.B. Compte tenu de l'abondance de la littérature sur le sujet, nous ne citerons que deux ouvrages. Le premier, datant de 1962, rassemble avec une très grande clarté les principaux résultats théoriques relatifs au cas déterministe :
PAPOULIS, The Fourier Integral - Mc Graw Hill, New-York.
Plus au goût du jour, sous l'angle informatique, le second développe de façon très détaillée les tenants et aboutissants de l'algorithme F F T :
ENOCHSON, Programming and analysis for digital time series data - Shock and vibration information center, U.S. Dept of Defense, 1968.

ANNEXE 1. D'après J. MAX.

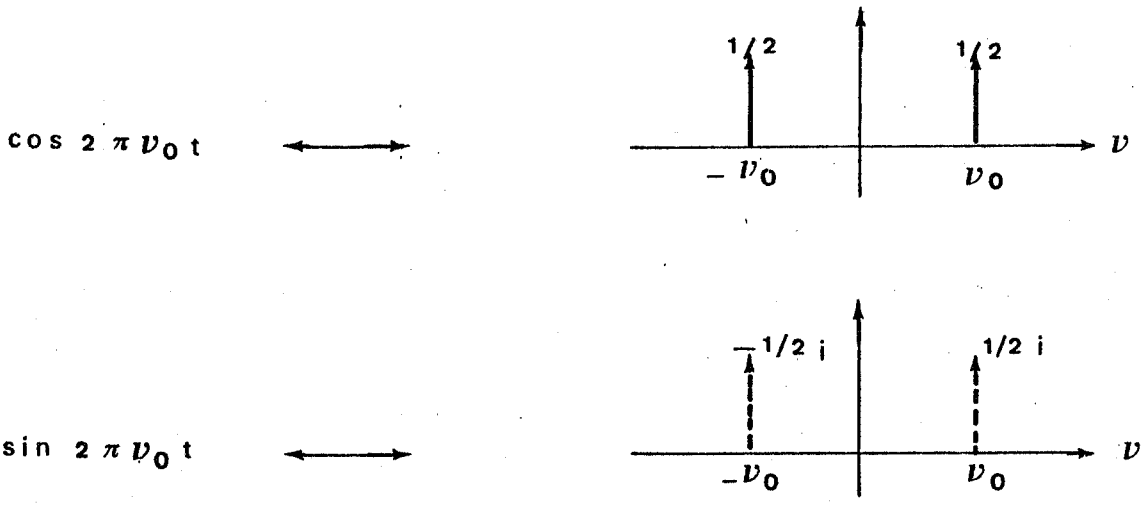


PROPRIETES DE SYMETRIE DES FONCTIONS ET DE LEURS TRANSFORMEES DE FOURIER

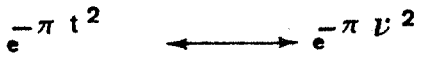
1)


X (t)
1/2 T
t
– T
+T
X (ν) = sin 2π ν T / 2π ν T
ν
1/2 T

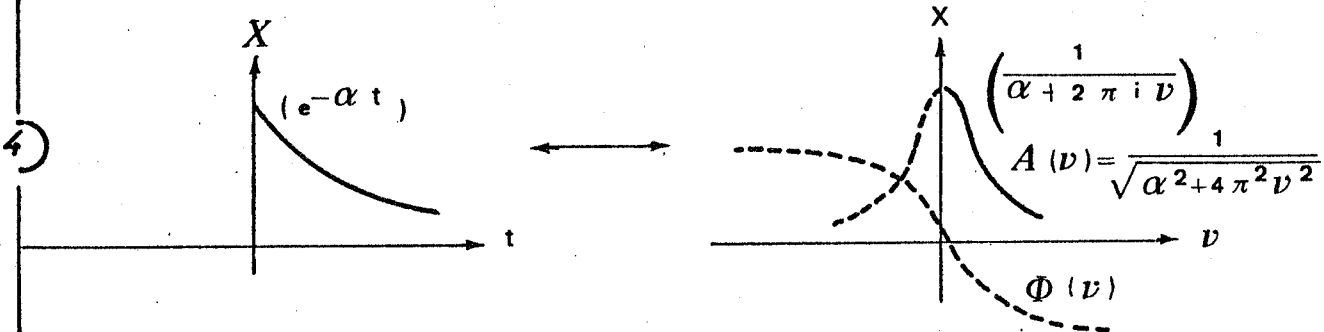
2)

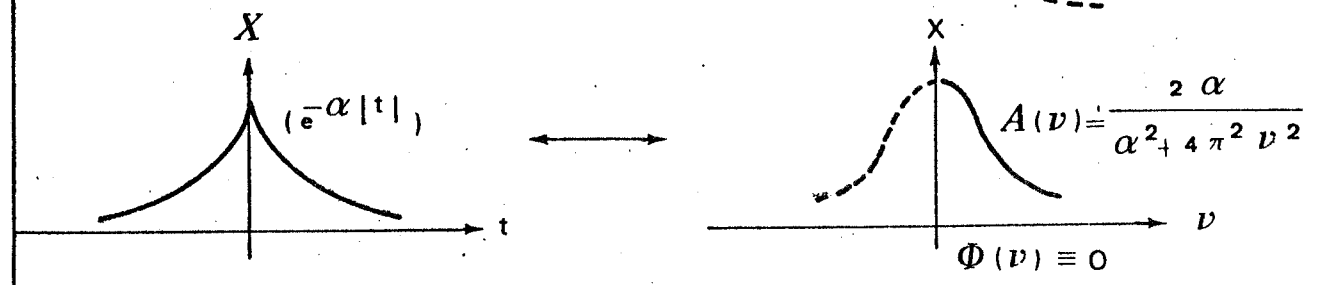

cos 2 π ν0 t
1/2
1/2
ν
– ν0
ν0
sin 2 π ν0 t
– 1/2 i
1/2 i
ν
–ν0
ν0

3)

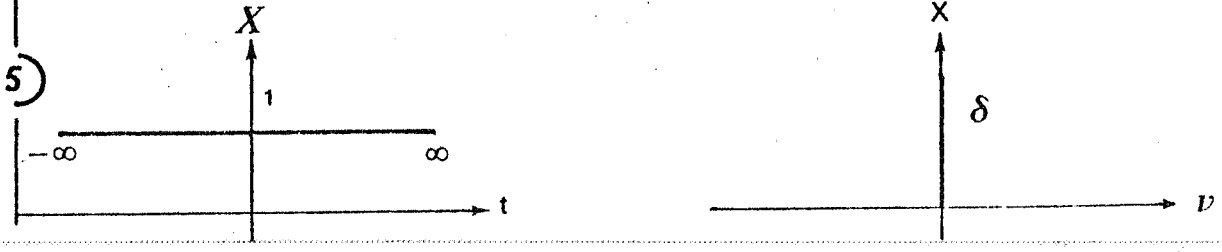

e^{–π t²}
e^{–π ν²}

4)


X
(e^{–α t})
t
X
(1 / (α + 2 π i ν))
A (ν) = 1 / √(α² + 4 π² ν²)
ν
Φ (ν)


X
(e^{–α |t|})
t
X
A (ν) = 2 α / (α² + 4 π² ν²)
ν
Φ (ν) ≡ 0

5)


X
1
–∞
∞
t
X
δ
ν

J. MAX - La notion de transformée de *FOURIER* est homogène à la notion de filtre car les filtres ont nous parlons sont des filtres *linéaires* (au sens de *BLANC-LAPIERRE*) dont les fonctions propres sont des sinusoïdes ou exponentielles complexes.

D. ABENSOUR demande des précisions sur la définition de la "Discret *FOURIER* Transform" et évoque les problèmes soulevés par l'échantillonnage d'un signal temporel.

B. ESCUDIÉ - L'irréversibilité dont on vient de parler vient du fait suivant : il y a une infinité de signaux de même densité spectrale. S'il y a une infinité de signaux de même "spectre instantané" (définition quelconque), cette infinité là est bien plus *restreinte* que la première, ce qui permet la "reconnaissance".

J.P. PECKELS évoque les problèmes posés par la localisation d'une source de bruit parmi d'autres dans une salle des machines et la localisation d'un locuteur dans une assemblée (voir *H. LANGE* dans KORRELATIONS ELEKTRONIK publié en Allemagne de l'Est).

P. DEMAN - Dans le spectre que vous avez défini, il y a un grand nombre de spectres différents car on a réalisé une convolution suivie par une élévation au carré et une intégration. Or, le spectre observé dépend à la fois du signal et des caractéristiques du filtre d'analyse.

R. CARRÉ - Le problème de l'analyse doit être posé sous deux formes :
- dans le cas de l'étude de l'émetteur, tout type de traitement peut être effectué en cherchant à conserver toutes les informations ;
- dans le cas de l'étude du signal par rapport au récepteur, l'analyse revient à l'étude d'une simulation de ce récepteur.

B. ESCUDIÉ - Les travaux de *SUGA* ont permis de montrer que l'oreille des mammifères (chauve-souris) réalise des analyses "temps-fréquence" différentes de la transformation de *FOURIER*. Ceci permet de réaliser des reconnaissances de signaux.

J. MAX - L'analyse de *FOURIER* et l'analyse par filtre sont la même chose. L'analyse de *FOURIER* est la convolution avec une sinusoïde de fréquence f, de $-\infty$ à $+\infty$. Comme on ne doit pas intégrer aussi longtemps, on a un filtre dont la réponse impulsionnelle est une sinusoïde tronquée, donc une réponse en fréquence qui a une valeur non nulle.

P. DEMAN - Dans le cas d'une transformation de *FOURIER*, on obtient pour chaque fréquence une seule valeur. Dans une "convolution" à partir d'une batterie de filtres, on obtient pour chaque fréquence une série de valeurs en fonction du temps plus riche en informations, permettant ainsi de prendre une décision après une durée finie.
L'application de l'analyse de *FOURIER* entraînerait le report de la décision dans l'infini du temps et condamnerait à ne pas transmettre d'informations.

R. CHOCHOLLE - On est parti de l'idée que l'oreille faisait une analyse de *FOURIER* parce qu'on connaissait les filtres qui fonctionnent sur un mode sinusoïdal. Mais on peut se demander si ceci est absolument exact ; on ne pourrait pas, par exemple, percevoir des battements ; d'autre part, on ne sait pas comment on pourrait distinguer des transitoires, des clics suivant leur forme.
Quels autres procédés d'analyse peut-on prévoir ? Où trouver des informations sur les autres calculs d'analyse ?

R.A. GUEDJ - Pourquoi la transformée de *FOURIER* à propos de traitement du signal de la parole ? Il est peut-être utile de rappeler l'analogie existant avec l'application de la logique symbolique dans les sciences humaines, illustrée par l'histoire racontée par le Professeur *Y. BAR-HILLEL* :

" Un homme ivre cherche sa clé sous un lampadaire. Pourquoi cher-
" chez-vous votre clé ici ; est-ce là que vous l'avez perdue ? lui
" demande-t-on. Non, répond-il, mais c'est là qu'est la lumière. "

Les propriétés de linéarité de l'opérateur de *FOURIER* sont la lumière. Le signal de parole n'est pas nécessairement linéaire.

UNE NOUVELLE METHODE D ' ANALYSE SPECTRALE

J . M A X
Centre d'Etudes Nucléaires - Grenoble

1. DENSITE SPECTRALE ENERGETIQUE (DSE) - DENSITE SPECTRALE DE PUISSANCE

1.1. APPROCHE INTUITIVE DE LA DSE : filtrage

Considérons une grandeur $x(t)$ évoluant en fonction d'une variable indépendante t, le temps par exemple.

Considérons un filtre de fréquences passe-bande F (supposé idéal) de largeur de bande $\Delta\nu$, centré sur la fréquence ν. Un tel filtre a un gain unité pour toutes les fréquences comprises dans la bande : $\nu - \frac{\Delta\nu}{2}$, $\nu + \frac{\Delta\nu}{2}$, et un gain nul pour toutes les fréquences extérieures à cet intervalle (figure 1).

Filtrons $x(t)$ dans le filtre F, soit $x_F(t)$ la sortie de ce filtre (figure 2).

Mesurons alors la puissance moyenne de $x_F(t)$ qui sera :

$$\Pi_{xx}(\nu,\Delta\nu) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \left(x_F(t)^2 \right) dt \ .$$

On remarque déjà que la seule différence entre puissance et énergie provient du facteur $1/T$.

Si $\Delta\nu$ tend vers zéro, $\Pi_{xx}(\nu,\Delta\nu)$ peut être considéré comme la différentielle d'une grandeur $P_{xx}(\nu)$.

$$\lim_{\Delta\nu \to 0} \Pi_{xx}(\nu,\Delta\nu) = dP_{xx}(\nu) \ .$$

On peut écrire : $$d\nu \cdot \frac{dP_{xx}(\nu,\Delta\nu)}{d\nu} = S_{xx}(\nu)\, d\nu \ .$$

$S_{xx}(\nu)$ est appelé densité de puissance ou densité spectrale énergétique de $x(t)$.

Du fait que, pour définir $S_{xx}(\nu)$ nous avons fait tendre la largeur de bande $\Delta\nu$ du filtre F vers zéro, on conçoit que cette densité spectrale énergétique ne soit pas accessible car un filtre de bande infiniment étroit ne laisserait passer qu'une puissance infiniment petite, donc non mesurable (sauf bien sûr dans le cas où le signal $x(t)$ comporterait une composante périodique à la fréquence ν).

La seule grandeur que l'on puisse mesurer sera la puissance du signal transmis par un filtre de largeur spectrale $\Delta\nu$. Soit :

$S_{xx}(\nu_0,\Delta\nu) = \int_{\nu_0-(\Delta\nu/2)}^{\nu_0+(\Delta\nu/2)} S_{xx}(\nu)\, d\nu$ (figure 3). Cela revient à faire sur $S_{xx}(\nu)$ un échantillonnage au moyen d'un échantillonneur moyenneur.

En répétant cette opération de filtrage + mesure de la puissance moyenne pour différentes valeurs de la fréquence centrale du filtre ν_1, ν_2, ..., ν_k, ..., ν_n, on obtiendra n points de la densité spectrale énergétique (figure 4).

$\int_0^\infty S_{xx}(\nu)$ est la puissance totale du signal $x(t)$.

$\int_{\nu_1}^{\nu_2} S_{xx}(\nu)\, d\nu$ est la puissance du signal $x(t)$ contenue dans la bande de fréquences ν_1, ν_2.

FIGURE 1

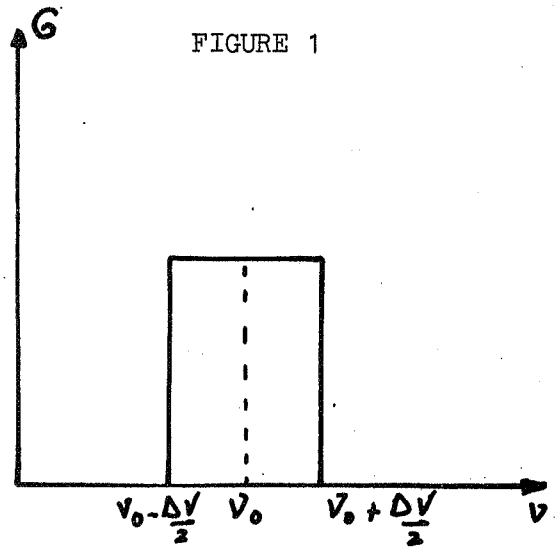


FIGURE 2

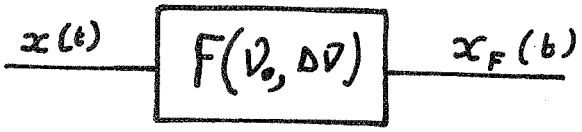


FIGURE 3

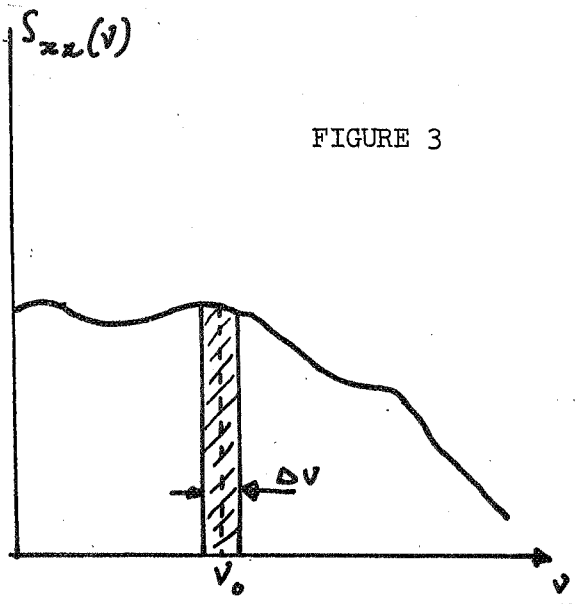
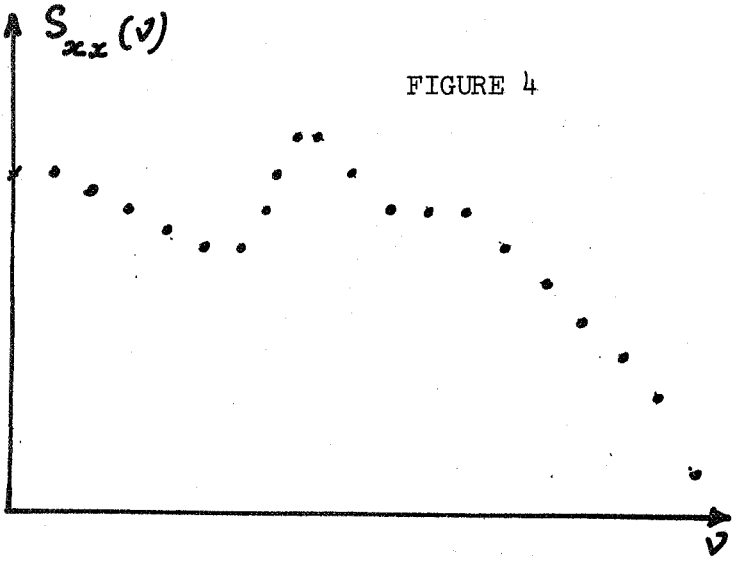


FIGURE 4



De même que dans le cas d'un signal x(t), on a défini une densité spectrale $S_{xx}(\nu)$, dans le cas de deux signaux x(t) et y(t), on peut définir non seulement leurs densités spectrales propres (ou autospectre) $S_{xx}(\nu)$ et $S_{yy}(\nu)$, mais deux densités spectrales d'interaction (ou interspectres) $S_{xy}(\nu)$ et $S_{yx}(\nu)$.

Considérons l'une de ces densités spectrales d'interaction $S_{xy}(\nu)$: cette densité spectrale d'interaction comprend une composante "en phase" ou "réelle" et une composante "en quadrature" ou "imaginaire".

La composante "en phase" s'obtiendra en filtrant x(t) et y(t) dans deux filtres identiques $F(\nu,\Delta\nu)$ et en mesurant la puissance moyenne :

$\frac{1}{T}\int_0^T x_F(t)\, y_F(t)\, dt = R\left(S_{xy}(\nu)\right)$. La composante "en quadrature" s'obtiendra en effectuant la même opération après avoir déphasé $y_F(t)$ de $\frac{\pi}{2}$:

$\frac{1}{T}\int_0^T x_F(t)\, y_F(t) \quad \underline{/ - \pi/2} \quad = I\left(S_{xy}(\nu)\right)$; et l'on a la relation :

$$S_{xy}(\nu) = S^*_{yx}(\nu) .$$

* signifie : complexe conjugué.

1.2. SECONDE APPROCHE DE LA DENSITE SPECTRALE : à partir de la transformée de Fourier des signaux

Considérons le signal (x)t qui se déroule dans le temps ; considérons une "tranche" de durée θ de ce signal, qui s'écoule donc de kθ à (k + 1)θ ; nous appellerons $x_k(t,\theta)$ cette tranche :

$x_k(t,\theta) = x(t)$ pour $k\theta < t < (k+1)\theta$,

$x_k(t,\theta) = 0$ pour t extérieur à cet intervalle.

On peut définir la puissance moyenne de $x_k(t,\theta)$ qui sera :

$$\frac{1}{\theta} \int_{k\theta}^{(k+1)\theta} \left(x(t) \right)^2 dt = \frac{1}{\theta} \int_{-\infty}^{\infty} x_k(t,\theta)\, dt .$$

$x_k(t,\theta)$ a une transformée de Fourier $X_k(\nu,\theta)$.

$$x_k(\nu,\theta) = \int_{-\infty}^{\infty} x_k(t,\theta)\, e^{-2\pi j\nu t}\, dt = \int_{k\theta}^{(k+1)\theta} x(t) e^{-2\pi j\nu t} dt .$$

D'après le théorème de PARSEVAL :

$$\frac{1}{\theta} \int_{-\infty}^{\infty} \left[x_k(t,\theta)\right]^2 dt = \frac{1}{\theta} \int_{0}^{\infty} \left[X_k(\nu,\theta)\right]^2 d\nu .$$

Par définition de la densité spectrale

$$\frac{1}{\theta} \int_{-\infty}^{\infty} \left[x_k(t,\theta)\right]^2 dt = \frac{1}{\theta} \int_{0}^{\infty} S_{x_k x_k}(\nu,\theta)\, d\nu ,$$

d'où :

$$S_{x_k x_k}(\nu,\theta) = \frac{\left[X_k(\nu,\theta)\right]^2}{\theta} .$$

On sait par ailleurs que, puisque la "tranche" de signal considérée a une durée θ, $X_k(\nu,\theta)$, donc la densité spectrale sera obtenue avec une résolution au plus égale à $\Delta\nu = 1/\theta$.

Autrement dit la durée θ de la "tranche" de signal considérée fixe la largeur de bande Δν du filtre équivalent.

Si l'on veut mesurer la densité spectrale avec la même résolution, non plus sur une seule tranche θ, mais sur une durée du signal $T = n\theta$ on calculera $X_{x_k x_k}(\nu)$ pour chaque tranche et on en fera la moyenne :

$$S_{xx}(\nu,\theta) = \frac{1}{n} \sum_{k=1}^{n} S_{x_k x_k}(\nu,\theta) = \frac{1}{n} \sum_{k=1}^{n} \frac{\left[X_k(\nu,\theta)\right]^2}{\theta} .$$

$X_k(\nu,\theta)$ est en général une grandeur complexe : $X_k(\nu,\theta) = R_{x_k x_k}(\nu,\theta) - j\, I_{x_k x_k}(\nu,\theta)$ d'où

$$S_{xx}(\nu,\theta) = \frac{1}{n} \sum_{k=1}^{n} \left[\left(R_{x_k x_k}(\nu,\theta)\right)^2 + \left(I_{x_k x_k}(\nu,\theta)\right)^2 \right].$$

Cette approche de la densité spectrale énergétique permet une extension immédiate au cas des interspectres :

à $x_k(t,\theta)$ correspond $X_k(\nu,\theta) = R_{x_k x_k}(\nu,\theta) - j\, I_{x_k x_k}(\nu,\theta)$,

à $y_k(t,\theta)$ correspond $Y_k(\nu,\theta) = R_{y_k y_k}(\nu,\theta) - j\, I_{y_k y_k}(\nu,\theta)$.

Les spectres croisés seront :

$$S_{x_k y_k}(\nu,\theta) = X_k(\nu,\theta) \cdot Y_k^*(\nu,\theta),$$

$$S_{y_k x_k}(\nu,\theta) = X_k^*(\nu,\theta) \cdot Y_k(\nu,\theta),$$

d'où

$$R_{x_k y_k}(\nu,\theta) = R_{x_k x_k}(\nu,\theta) \cdot R_{y_k y_k}(\nu,\theta) + I_{x_k x_k}(\nu,\theta) \cdot I_{y_k y_k}(\nu,\theta),$$

$$J_{x_k y_k}(\nu,\theta) = R_{x_k x_k}(\nu,\theta) \cdot I_{y_k y_k}(\nu,\theta) - R_{y_k y_k}(\nu,\theta) \cdot I_{x_k x_k}(\nu,\theta),$$

et :

$$R_{x_y}(\nu,\theta) = \frac{1}{n} \sum_{k=1}^{n} R_{x_k y_k}(\nu,\theta),$$

$$J_{xy}(\nu,\theta) = \frac{1}{n} \sum_{k=1}^{n} I_{x_k y_k}(\nu,\theta).$$

1.3. TROISIEME APPROCHE DE LA DENSITE SPECTRALE A PARTIR DES FONCTIONS DE CORRELATION

Considérons encore le signal $x(t)$; on définit sa fonction d'autocorrélation :

$$C_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)\, x(t-\tau)\, dt.$$

En fait, comme on opère toujours sur une durée limitée, ce que l'on obtient n'est pas $C_{xx}(\tau)$, mais :

$$C_{xx}(\tau,T) = \frac{1}{T} \int_0^T x(t)\, x(t-\tau)\, dt,$$

T étant la durée totale de connaissance du signal $x(t)$.

On aura néanmoins, pour plus de simplicité :

$$C_{xx}(\tau) = \frac{1}{T} \int_0^T x(t)\, x(t-\tau)\, dt.$$

D'après le théorème de WIENER-KINCHINE la transformée de Fourier de $C_{xx}(\tau)$ est la densité spectrale énergétique de $x(t)$:

$$F\left[C_{xx}(\tau)\right] = \int_{-\infty}^{\infty} C_{xx}(\tau)\, e^{-2\pi j \nu \tau} d\tau = S_{xx}(\nu).$$

De même, à partir de deux signaux x(t) et y(t) on définit deux fonctions d'intercorrélation :

$$C_{xy}(\tau) = \frac{1}{T} \int_0^T x(t)\ y(t - \tau)\ dt ,$$

$$C_{yx}(\tau) = \frac{1}{T} \int_0^T y(t)\ x(t - \tau)\ dt ,$$

avec $C_{xy}(\tau) = C_{yx}(-\tau)$.

A partir de ces fonctions d'intercorrélation, toujours d'après le théorème de WIENER-KINCHINE, on obtient par transformée de Fourier les interspectres :

$$S_{xy}(\nu) = F\left[C_{xy}(\tau)\right] = \int_{-\infty}^{\infty} C_{xy}(\tau)\ e^{-2\pi j\nu\tau} d\tau ,$$

$$S_{yx}(\nu) = F\left[C_{yx}(\tau)\right] = \int_{-\infty}^{\infty} C_{yx}(\tau)\ e^{-2\pi j\nu\tau} d\tau ,$$

$$S_{xy}(\nu) = S^{*}_{yx}(\nu) .$$

Largeur du filtre équivalent :

La fonction de corrélation étant calculée pour des retards allant non de $-\infty$ à $+\infty$, mais de $-\tau_n$ à $+\tau_n$ (avec $\tau_n \leq T/50$ en général), la largeur du filtre équivalent sera :

$$\Delta\nu = \frac{1}{2\tau_n} .$$

On a donc ainsi une troisième voie, indirecte, pour atteindre la densité spectrale ; mesure de la fonction de corrélation puis transformation de Fourier de la fonction de corrélation.

2. PRECISION SUR LA MESURE DE LA DENSITE SPECTRALE

Dans toutes ces approches de la densité spectrale, il a été question de valeurs moyennes, qui, théoriquement, sont définies par l'opérateur :

$$\frac{1}{T} \int_0^T (\quad)\ dt \qquad\qquad \lim T \to \tau .$$

Les mesures que l'on fait répondent à l'opération :

$$\frac{1}{T} \int_0^T (\quad)\ dt, \text{ T étant fini.}$$

En considérant donc $\frac{1}{T} \int_0^T (\quad)\ dt$ au lieu de la limite de cette quantité pour $T \to \infty$, on commet une erreur, dite erreur d'estimation.

La variance de cette erreur, peut, en première approximation s'exprimer sous la forme :

$$\varepsilon = 1/\sqrt{\Delta\nu\ T} ,$$

$\Delta\nu$ étant la largeur de bande du filtre équivalent, et T la durée totale utile du signal.

3. MESURE DE LA DENSITE SPECTRALE

Mesurer une densité spectrale revient donc à obtenir, par un moyen ou un autre, un nombre suffisant de valeurs discrètes de $S(\nu)$ pour différentes valeurs de ν: $\nu_1, \nu_2 \ldots \nu_k \ldots \nu_n$.

On conçoit que, selon le degré de complexité technique du dispositif de mesure mis en oeuvre il puisse y avoir plusieurs façons d'obtenir les résultats.

3.1. MESURE POINT PAR POINT

Le signal $x(t)$ de durée T étant préalablement enregistré, en mesure $S_{xx}(\nu_1)$, puis on reproduit $x(t)$ et on mesure $S_{xx}(\nu_2)$, on rejoue à nouveau $x(t)$ de manière à mesurer $S_{xx}(\nu_3)$ et ainsi de suite jusqu'à $S_{xx}(\nu_n)$.

On voit que ce procédé, s'il est théoriquement excellent, est très long, puisque pour avoir N points de spectre, il faut répéter N fois le signal enregistré de durée T, et que par le fait même, il ne permet pas d'avoir le résultat de l'analyse d'une manière immédiate ; il faut attendre au moins NT après la fin du signal $x(t)$.

3.2. MESURE AUTOMATIQUE

Le signal durant toujours T, on calcule $S_{xx}(\nu_1)$ sur une durée θ, puis $S_{xx}(\nu_2)$ sur une durée θ et ainsi de suite jusqu'à la $n^{ème}$ (filtre suiveur).

Ici on a l'"impression" d'avoir le résultat immédiat mais on utilise pour chaque point une durée θ du signal, θ étant au plus égal à T/N ; alors que l'on dispose d'un signal de durée T, on en utilise en fait que la durée $\theta = T/N$.

De ce fait, la précision sur $S_{xx}(\nu)$ au lieu de conduire à une erreur :

$$\varepsilon = \frac{1}{\sqrt{T\,\Delta\nu}} \qquad \text{conduit à} \quad \varepsilon' = \frac{1}{\sqrt{\theta\,\Delta\nu}} = \frac{N}{\sqrt{T\,\Delta\nu}} = \varepsilon\sqrt{N}\ ;$$

l'erreur est donc : $\sqrt{T/\theta} \simeq \sqrt{N}$ fois plus grande.

On peut également opérer par "tranches" (Batch mode). On met en mémoire une durée θ du signal ; sur cette durée θ on calcule n points du spectre, ce calcul dure un temps θ' ; lorsque le calcul est fini on vide la mémoire et on y remet une durée θ du signal ; on calcule à nouveau n points du spectre ; on fait ceci K fois, et on fait la moyenne des résultats ; cela revient donc à avoir opéré sur un signal de durée $K\theta$, mais cela a pris une durée totale $K\theta + K\theta'$; or le signal a une durée utile de T :

$$T = K\theta + K\theta'.$$

On voit donc que dans ce cas, le taux d'utilisation du signal sera $\frac{\theta}{\theta + \theta'}$ avec : $n(\theta + \theta') = T$.

Si θ' est important devant θ, cela conduira à une précision faible.

Au lieu de $\varepsilon = \frac{1}{\sqrt{\Delta\nu\ T}}$ on aura $\varepsilon' = \frac{1}{\sqrt{\Delta\nu . n\theta}}$.

Si $\theta' = \alpha\theta$, $n\theta(1 + \alpha) = T$, d'où $\varepsilon' = \varepsilon . \sqrt{1 + \alpha}$.

Tout cela revêt une très grande importance car dans de nombreux cas, on ne dispose pas d'une durée T du signal aussi grande que l'on veut ; on dispose d'une durée T de signal bien déterminée, et il faut l'utiliser au mieux. Dans le cas par exemple d'un essai en vol d'avion, lors d'un passage à un régime critique, on peut faire durer l'essai 30 secondes ; il n'est évidemment pas alors possible de n'utiliser que 1/100 ou 1/10 de cette durée, car cela conduirait à une précision inacceptable, il faut alors opérer en temps réel.

3.3. MESURE AUTOMATIQUE EN TEMPS REEL

Cela signifie que l'on n'opère pas en temps différé d'une part, et que d'autre part on utilise la totalité du signal utile, c'est-à-dire que la durée T qui intervient dans le calcul de la précision est la durée totale du signal et non une fraction de celle-ci.

Un analyseur en temps réel serait schématisé d'une manière simple par n analyseurs monopoints associés en parallèle, n ensembles (filtres + mesure de puissance) associés en parallèle (voir ci-après en 5.2.).

4. IMPORTANCE DE LA MESURE DE LA DENSITE SPECTRALE

La mesure des densités spectrales (ou spectres de puissance) est une opération que l'on rencontre de plus en plus fréquemment dans beaucoup de domaines, qu'il s'agisse de la densité spectrale propre (autospectre, ou simplement spectre) ou de la densité spectrale d'interaction (interspectre).

Rappelons quelques domaines d'applications :

- Automatisme
  - étude des bruits
  - identification de processus (mesure de fonctions de transfert)
- Etude des phénomènes turbulents et vibratoires
  - bruit de turbulences
  - analyse de vibrations

  } aéronautique, automobile, industrie chimique....
- Médecine, biologie - par exemple :
  - analyse fréquentielle d'électroencéphalogramme
  - analyse fréquentielle de rhéogrammes (mesure de la circulation sanguine par rhéographie)
- Physique nucléaire
  - étude de plasma
  - contrôle de réacteurs nucléaires
  - étude de milieux
- astrophysique, radioastronomie
- sonars radar

Dans la plupart de ces domaines, il est presque toujours indispensable d'obtenir la densité spectrale énergétique d'une manière automatique et en temps réel.

5. PROCEDE DE MESURE DE LA DENSITE SPECTRALE PAR FILTRAGE

Ce procédé relève de la première approche décrite en 1.1. ci-dessus.

5.1. PRINCIPE (figure 5) - ANALYSEUR MONOPOINT

On fait passer le signal x(t) dans un filtre dont la réponse impulsionnelle $h_{\nu_0,B}(t)$ a une transformée de Fourier qui se rapproche le plus possible du filtre idéal, c'est-à-dire qui aurait la valeur 1 dans l'intervalle $\nu_0 - \Delta\nu/2$, $\nu_0 + \Delta\nu/2$ et qui serait nulle à l'extérieur.

$$F\left[h_{\nu_0,\Delta\nu}(t)\right] = H_{\nu_0,\Delta\nu}(\nu).$$

Soit $x_F(t)$ le signal ainsi filtré. On fait passer ce signal $x_F(t)$ dans un quadrateur qui donne $\left[x_F(t)\right]^2$, puis dans un dispositif moyenneur qui délivre en sortie :

$$\frac{1}{T}\int_0^T \left[x_F(t)\right]^2 dt = S_{xx}(\nu_0, \Delta\nu).$$

Il faut évidemment répéter ces opérations de filtrage, quadrature, intégration, pour autant de valeurs de ν que l'on désire de points sur le spectre $S_{xx}(\nu)$ qui est obtenu ainsi sous forme échantillonnée. Le fait de répéter cette opération N fois pour avoir N points du spectre, nécessite un enregistrement préalable des signaux et exclut évidemment d'opérer en temps réel.

FIGURE 5

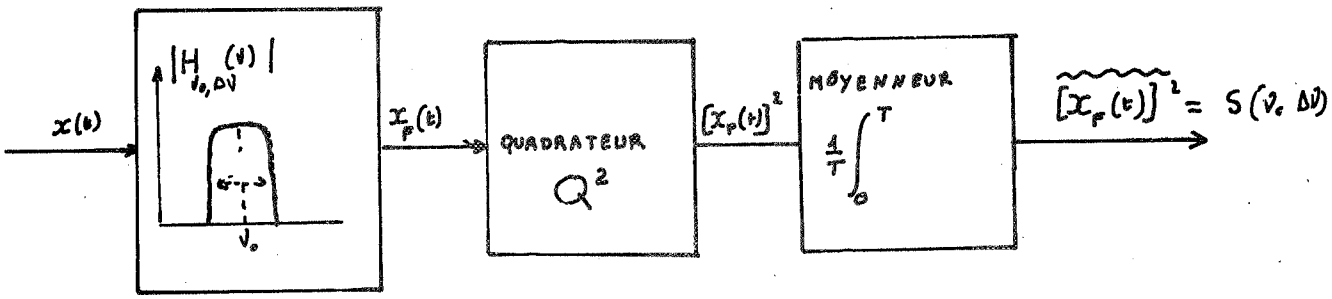
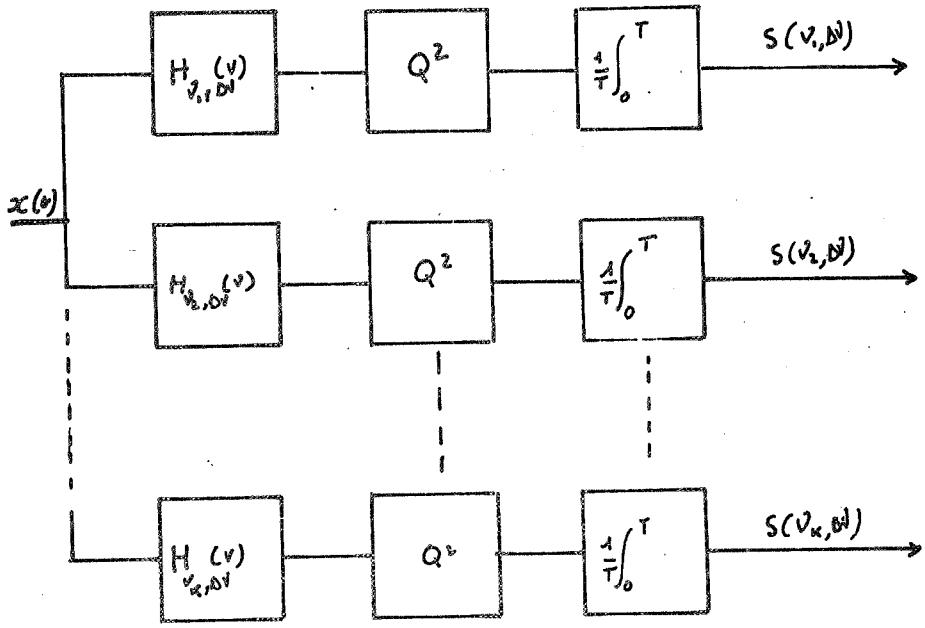


FIGURE 6



5.2. DISPOSITIF A BATTERIE DE FILTRES (figure 6)

Si l'on veut opérer en temps réel, cela peut se faire, soit par batterie de filtres, soit par compression de temps.

Il faut réaliser N filtres de largeur de bandes B rigoureusement égales et dont les fréquences centrales ν_1, ν_2..ν_n soient réglables, les gains de tous ces filtres étant rigoureusement égaux.

C'est là une solution qui présente d'énormes difficultés de réalisation et dont l'application ne peut être envisagée que dans quelques cas très particuliers.

La difficulté essentielle réside dans la quasi-impossibilité de réaliser simplement n filtres de largeur de bande identique, ne diffèrant que par la fréquence centrale ; si l'on admet d'utiliser des filtres compliqués, avec oscillateurs locaux notamment, le prix d'un tel ensemble devient prohibitif.

5.3. ANALYSEUR DE DENSITE SPECTRALE A COMPRESSION DE TEMPS

On numérise le signal à l'entrée, et on met en mémoire une tranche du signal de K échantillons, représentant une durée θ. On relit ensuite cette mémoire à vitesse beaucoup plus grande, on décode en analogique le signal mis en mémoire et on le filtre dans un filtre $F(\nu_0, \Delta\nu)$ suivi d'un quadrateur et d'un moyenneur. A chaque relecture de la mémoire on fait varier ν_0, on a ainsi après n relectures, n points du spectre de la tranche de durée θ. Pour faire varier ν_0 il est commode d'utiliser un même filtre en amont duquel est associé un oscillateur local.

On peut également filtrer la sortie numérique de la mémoire dans un filtre numérique.

Bien entendu, comme l'on opère sur une "tranche" de signal de durée θ, il faudra, pour avoir une précision suffisante recommencer N fois, de telle sorte que $N\theta = T$.

Si les n relectures de la mémoire peuvent être exécutées pendant une période de T_e d'échantillonnage du signal x(t), le système opère en temps réel.

Un inconvénient de ce dispositif à compression de temps est que, à cause même de la compression de temps, il faut que les échantillons de x(t) entrent dans la mémoire à une cadence très inférieure (500 fois au moins) à la cadence à laquelle ils sortent de la mémoire lors de la lecture accélérée. Ceci limite la largeur de bande des signaux que l'on peut analyser (50.000 Hz environ).

5.4. CAS DES INTERSPECTRES

Ce dispositif à filtrage présente de très gros inconvénients dès lors que l'on veut mesurer des interspectres, car il est alors nécessaire de doubler une partie importante de l'appareil. (figure 7, page A/b/10.)

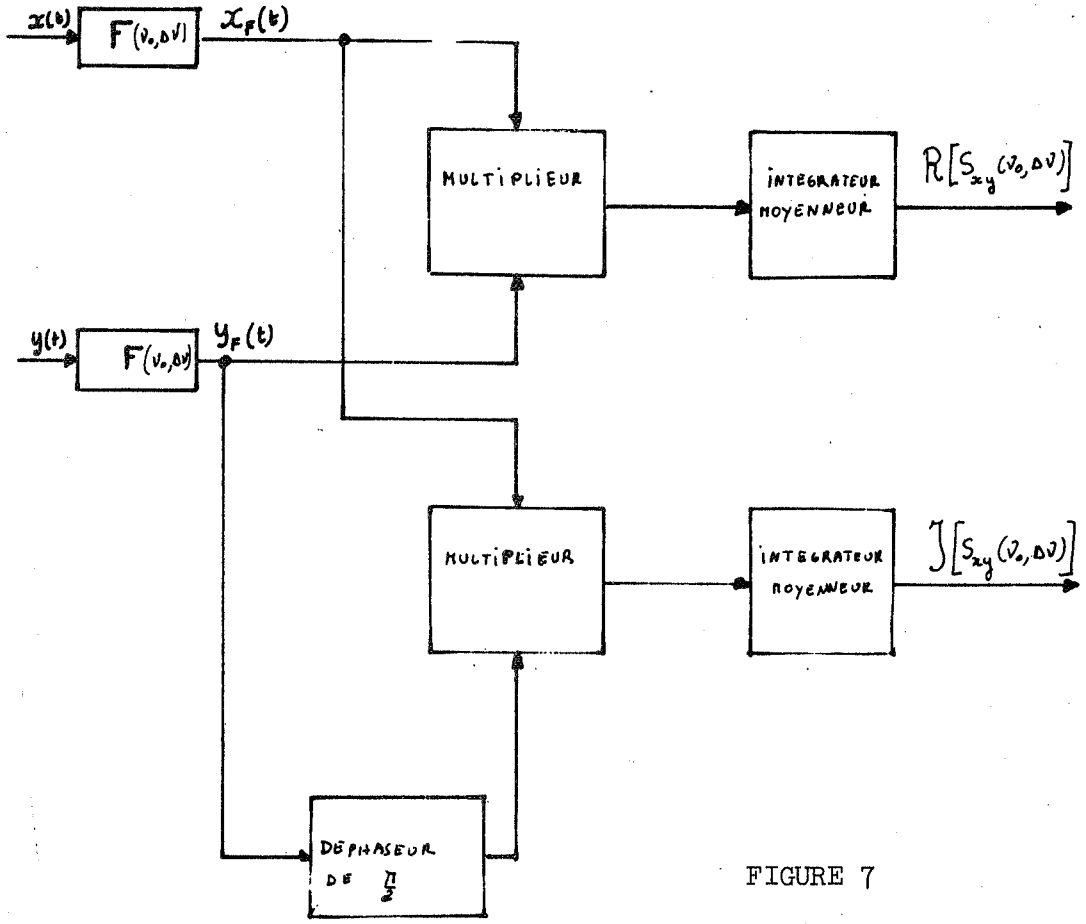


FIGURE 7

6. PROCEDE DE MESURE DE LA DENSITE SPECTRALE ENERGETIQUE PAR TRANSFORMATION DE FOURIER DES SIGNAUX

L'avantage essentiel de cette méthode est qu'elle permet d'opérer entièrement en numérique en utilisant la grande puissance des calculateurs numériques.

On introduit dans la mémoire du calculateur n échantillons du signal x(t), ces n mots correspondent à une "tranche" de durée θ du signal.

On calcule la transformée de Fourier (partie réelle et partie imaginaire) de cette tranche et on en déduit (voir en 1.2.) le spectre.

Bien entendu, pour avoir une précision statistique suffisante il faut répéter cette opération un nombre de fois suffisant K, de telle sorte que $K\theta$ corresponde à une durée T suffisante. Cette méthode a été surtout utilisée jusqu'ici pour le traitement en temps différé de signaux enregistrés sur bande magnétique numérique.

Pour réduire le temps de calcul, COOLEY et TUCKEY ont mis au point un algorithme (dit F.F.T.) perfectionné ces dernières années par M. et Mme CONNES (observatoire de MEUDON).

Le calcul de la partie réelle et de la partie imaginaire de la transformée de Fourier de chaque tranche de durée θ nécessitent un grand nombre d'opérations. Si l'on veut opérer en temps réel en ligne, directement sur le signal, il faut que toutes ces opérations puissent être effectuées pendant une période d'échantillonnage du signal x(t). De cette façon et de cette façon seulement on pourra traiter en ligne toute l'information contenue dans le signal.

Ceci bien entendu va limiter la largeur de bande que l'on peut espérer traiter.

Avec les calculateurs programmés actuels, et en utilisant les algorithmes F.F.T. on ne peut pas dépasser, en temps réel une largeur de bande de quelques centaines de Hertz.

En utilisant les techniques de microprogrammation (sous programmes câblés) on gagne un facteur de 5 à 10 sur la largeur de bande des signaux que l'on peut traiter en temps réel.

La solution la plus rapide, mais encore chère dans l'état actuel de la technologie serait de réaliser un calculateur spécialisé de transformée de Fourier utilisant des mémoires rapides en circuits intégrés.

7. ANALYSE SPECTRALE PAR CORRELATION

7.1. PRINCIPE

Cette méthode est basée sur le théorème de WIENER-KINCHINE qui montre que la densité spectrale est la transformée de Fourier de la fonction de corrélation :

$$S_{xx}(\nu) = F\left[C_{xx}(\tau)\right],$$

à cause de la propriété de parité des fonctions d'autocorrélation l'autospectre n'a pas de composante imaginaire :

$$S_{xx}(\nu) = \int_{-\infty}^{\infty} C_{xx}(\tau) \cos 2\pi\nu\tau \, d\tau ;$$

dans le cas de l'interspectre de deux signaux :

$$S_{xy}(\nu) = \int_{-\infty}^{\infty} C_{xy}(\tau) \cos 2\pi\nu\tau \, d\tau - j \int_{-\infty}^{\infty} C_{xy}(\tau) \sin 2\pi\nu\tau \, d\tau .$$

7.2. ECHANTILLONNAGE

Les corrélateurs automatiques en temps réel actuels donnent la fonction de corrélation $C_{xy}(\tau)$ sous forme échantillonnée pour des valeurs discrètes du retard τ, multiples de la période d'échantillonnage des signaux T_e.

$$C_{xy}(\tau) = \sum_{m=-n}^{n} C_{xy}(m,T_e) \cdot \delta(\tau - mT_e) .$$

On démontre [ ] que les parties réelles et imaginaires de la transformée de Fourier de $C_{xy}(\tau)$ s'obtiennent rigoureusement en calculant :

$$T_e \sum_{m=-n}^{n} C_{xy}(mT_e) \cos 2\pi\nu \, mT_e = R\left[S_{xy}(\nu)\right] .$$

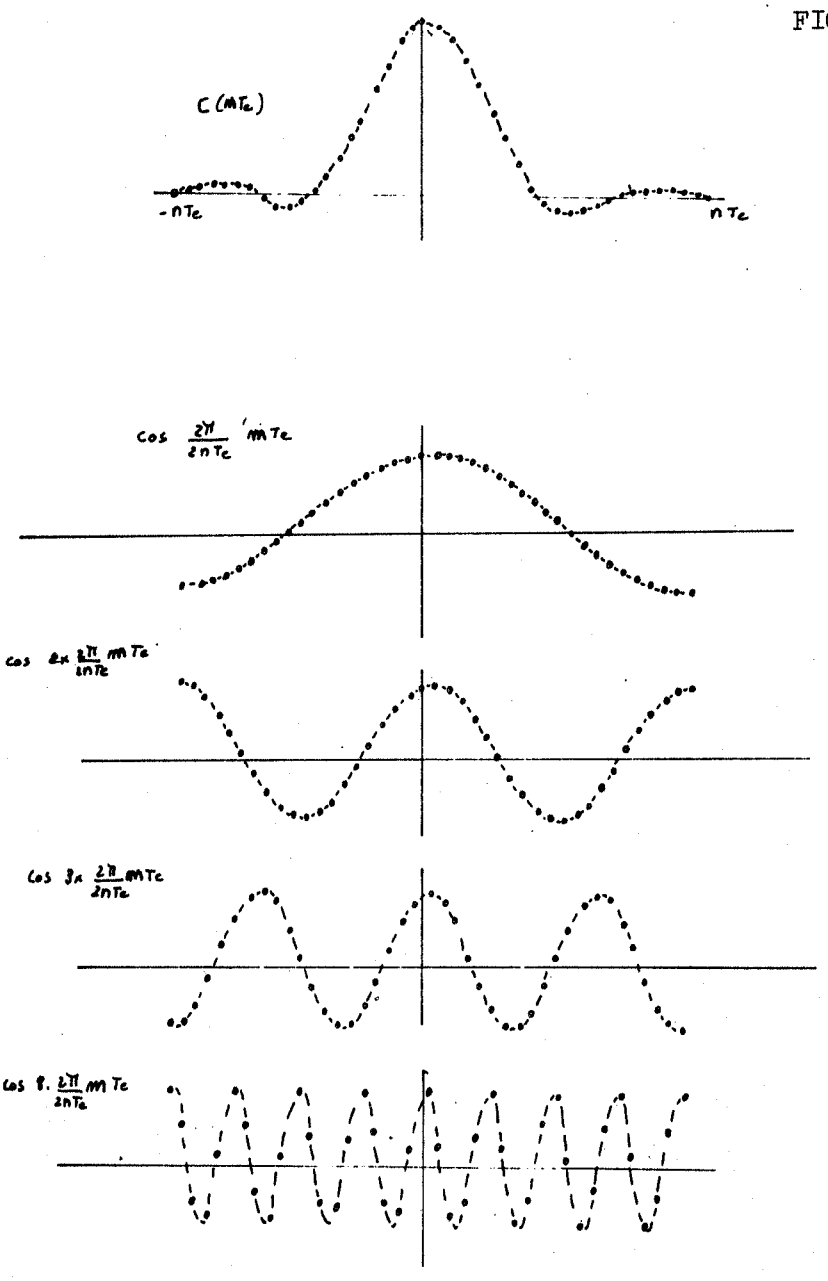
$$T_e \sum_{m=-n}^{n} C_{xy}(mT_e) \sin 2\pi\nu mT_e = I\left[S_{xy}(\nu)\right] .$$

D'où le résultat capital :

Pour obtenir la valeur exacte de la densité spectrale correspondant à la fréquence ν il suffit de faire la somme des produits, échantillon à échantillon, de la fonction de corrélation échantillonnée et de la fonction sin $2\pi\nu\tau$ (ou cos $2\pi\nu\tau$) échantillonnée à la même cadence de $C_{xy}(\tau)$.

IL EST DONC INUTILE D'INTERPOLER LA FONCTION DE CORRELATION. Il suffit de considérer les points successifs de $C_{xy}(mT_e)$ et de les multiplier chacun par la valeur correspondante de la fonction cosinus 2 mT_e (ou de la fonction sinus (figure 8)).

FIGURE 8

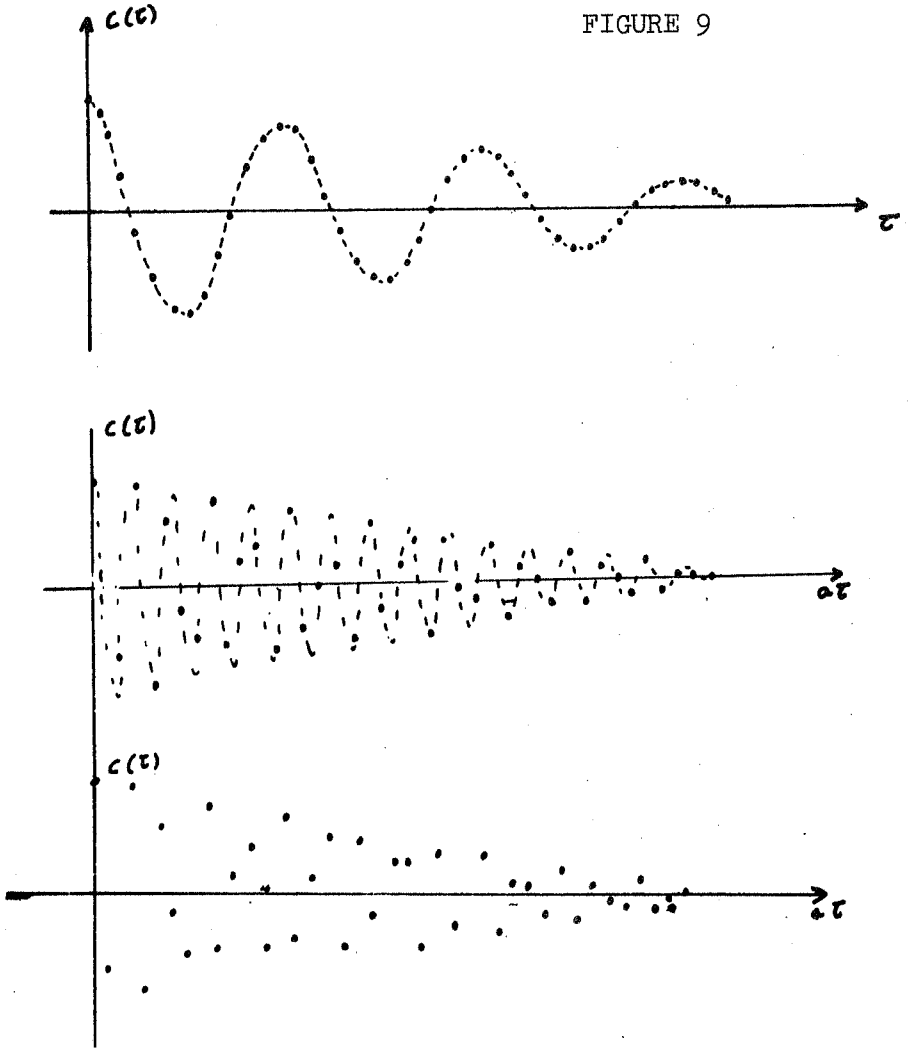


7.3. FINESSE DU FILTRE EQUIVALENT

La fonction de corrélation n'est pas connue pour des valeurs du retard τ allant de $-\infty$ à $+\infty$. Les corrélateurs donnent un nombre de points limite correspondant à autant de retards différents. Avec un corrélateur à n points, l'intervalle entre chaque point étant T_e, la fonction de corrélation est connue de $-\tau_M$ à $+\tau_M$ avec $\tau_M = (n-1)T_e$. Si la fonction de corrélation est connue dans l'intervalle $-\tau_{M'}+\tau_{M'}$ de valeur $2(n-1)\,T_e = 2nT_e$, on démontre que la finesse d'analyse ne peut être meilleure que $1/2n'T_e$. On a donc intérêt à avoir $2n'T_e$ aussi

grand que possible ; n' étant fixé on ne peut agir que sur T_e ; il faudra prendre T_e aussi grand que possible, donc F_e aussi petit que possible, F_e restant compatible avec le théorème de SCHANNON. Aussi curieux que cela paraisse, pour obtenir le spectre avec la meilleure définition possible, on a intérêt à avoir sur la fonction de corrélation des points aussi espacés que possible (figure 9).

FIGURE 9



7.4. FREQUENCES ANALYSEES

Puisque la finesse d'analyse la meilleure est d'environ $\frac{1}{2NT_e} = \frac{1}{2(n-1)T_e}$ il est inutile de calculer le spectre pour les fréquences inférieures à $\frac{1}{2(n-1)T_e}$ ou $\frac{1}{2nT_e}$ qui est peu différent.

La plus basse fréquence analysée sera : $\nu_1 = \frac{1}{2nT_e}$.

Donc le premier point du spectre sera : $S_{xy}(\nu_1) = S_{xy}\left(\frac{1}{2nT_e}\right)$.

Pour la même raison, il est inutile de chercher des points du spectre espacés de moins $1/2nT_e$.

Le second point sera donc : $S_{xy}(\nu_2) = S_{xy}\left(\frac{2}{2nT_e}\right)$.

Puisque, par définition, $S_{xy}(\nu) \equiv 0$ pour $|\nu| \geq \frac{Fe}{2}$ on s'arrêtera dans le calcul des différents points, à la fréquence ν_r telle que :

$$\nu_r = \frac{r}{2nT_e} = \frac{1}{2T_e} \text{ ,}$$

d'où r = n : avec un corrélateur à n points, on peut donc calculer n points du spectre.

N.B. Nous avons pris $\frac{1}{2nT_e}$ au lieu de $\frac{1}{2(n-1)T_e}$ parce que le nombre n, nombre de points du corrélateur est le plus souvent un nombre "rond" (50, 100, 150, 200).

7.5. ECHANTILLONNAGE DES FONCTIONS COSINUS ET SINUS

La plus basse fréquence à considérer étant : $\nu_1 = \frac{1}{2nT_e}$, on aura, pour la partie réelle du spectre, par exemple :

$$R_{xy}(\nu_1) = R_{xy}\left(\frac{1}{2nT_e}\right) = T_e \sum_m C_{xy}(mT_e) \cos 2\pi \frac{1}{2nT_e} mT_e \text{ ,}$$

le nombre de valeurs distinctes de m étant 2n - 1 :

m = - (n-1), -(n-2)... -2, -1, 0, 1, 2, ... (n-2), (n-1).

Ces valeurs vont couvrir une période entière du cosinus (ou du sinus) échantillonné à la fréquence $F_e = 1/2nT_e$.

Pour la fréquence suivante $\frac{2}{2nT_e}$ on aura :

$$R_{xy}(\nu_2) = R_{xy}\frac{2nT_e}{2} = T_e \sum_m C_{xy}(mT_e) \cos \frac{2\pi}{2n} \cdot 2m \text{ ;}$$

on voit que les (2n-1) valeurs de $\cos \frac{2\pi}{2n} \cdot 2m$ font partie de l'ensemble des (2n-1) valeurs de :

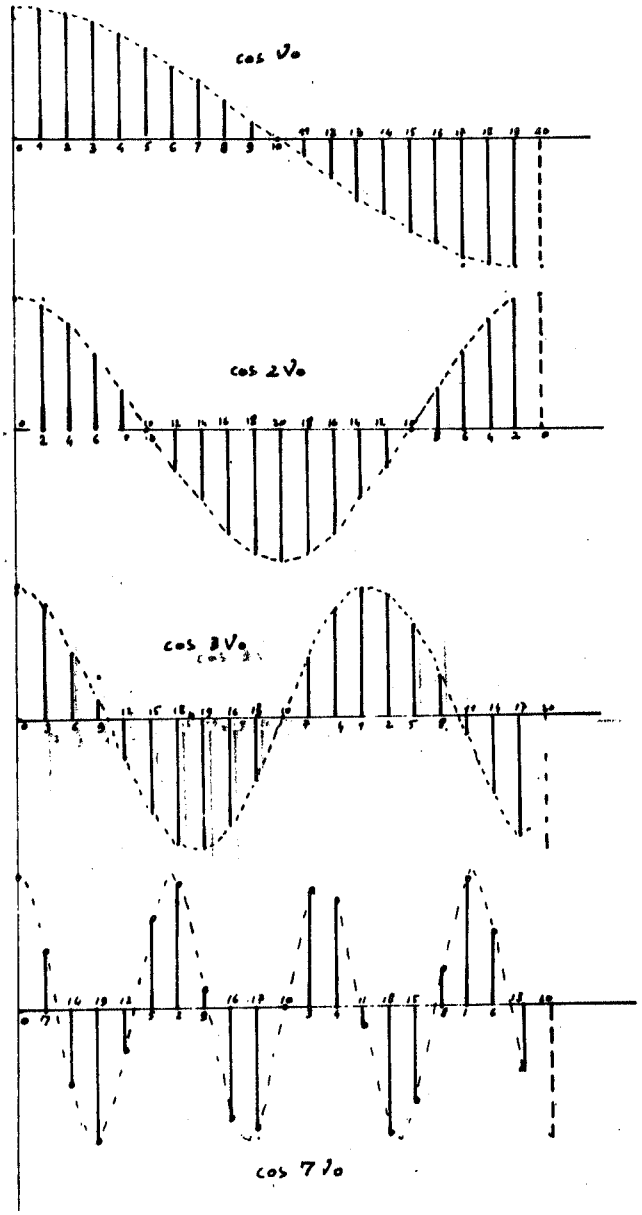
$$\cos \frac{2\pi}{2n} \cdot m \text{ .}$$

Il en serait de même pour toute autre valeur de la fréquence ν_k à laquelle correspond :

$$\cos \frac{2\pi}{2n} \cdot Km \text{ .}$$

Si donc, on dispose en mémoire des 2n-1 valeurs de $\cos \frac{2\pi}{2n} \cdot m$, on dispose par le fait même de toutes les valeurs $\cos \frac{2\pi}{2n} \cdot Km$ (ou sinus $\frac{2\pi}{2n} \cdot Km$). On conçoit l'importance de cette remarque qui, au moyen d'une table contenant 2n valeurs de la fonction cosinus $\frac{2\pi}{2n} \cdot m$ pour m entier tel que $|m| \leq n$, permet de générer toutes les fonctions cosinus ou sinus nécessaires à l'obtention du spectre. Les propriétés de symétrie des fonctions sinus et cosinus permettent d'ailleurs de ne mettre en mémoire que n ou même n/2 valeurs distinctes. C'est la raison pour laquelle on a défini la 1/2 période du cosinus sur un nombre pair (n) et non sur n-1.

FIGURE 10



La figure 10 montre comment, à partir de n échantillons du cosinus, on obtient la fréquence ν_1 , en lisant les n points par un, dans l'ordre 0, 1, 2, 3... On obtient la fréquence ν_2 en les lisant 2 par 2 soit 0,2,4,6... On obtient la fréquence ν_r en les lisant de r en r soit 0,r,2r...

7.6. REALISATION

7.6.1. MEMOIRE D'ENTREE

La fonction de corrélation (auto ou intercorrélation) fournie par le corrélateur est numérisée et mise en mémoire (si cette mémoire n'existe pas dans le corrélateur, ce qui est le cas pour certains appareils)

C'est à partir de cette fonction de corrélation mémorisée que le transformateur de Fourier va fonctionner.

7.6.2. GENERATEUR DE COSINUS OU DE SINUS

La valeur des n points de la fonction cosinus sont mis en mémoire une fois pour toutes et on attaque cette mémoire par l'adresse de l'échantillon cherché.

Nous avons choisi de mettre le cosinus en mémoire par mots de 8 bits. (Cette mémoire est une matrice à diodes ; très prochainement nous utiliserons des mémoires mortes en technique - L.S.I.). A chaque adresse correspond donc un mot de 8 bits qui se traduit par l'état des 8 sorties de la mémoire.

7.6.3. MULTIPLIEUR

Nous utilisons un multiplieur numérique réalisé à l'aide d'additionneurs binaires en circuit intégré (multiplieur à décalage-addition).

7.6.4. INTEGRATEUR

L'intégration est réalisée à l'aide de mémoires à bistables en circuits intégrés.

7.6.5. FONCTIONS DE PONDERATION

Dans le cas défavorable où l'on veut analyser une densité spectrale d'une part très étendue en fréquences et d'autre part présentant des "pics" étroits, le pouvoir séparateur peut être insuffisant et le graphe $S(\nu)$ que l'on obtient présente des oscillations parasites qui peuvent prêter à confusion.

L'application d'une fonction de pondération sur la fonction de corrélation, si elle n'accroît pas le pouvoir séparateur, a du moins l'intérêt de lisser ces oscillations parasites ; on saura alors que tous les pics que l'on obtient correspondent à la réalité.

La fonction de pondération vaut 1 pour le premier point C(o) et 0 pour le dernier $C(\tau_M)$; entre ces points, nous avons adopté la loi :

$$\frac{\sin \pi \frac{mT_e}{\tau_M}}{\pi \frac{mT_e}{\tau_M}} ,$$

(pondération de Fauque-Berthier-Max).

On peut aussi utiliser les pondérations bien connues HANNING et HAMMING.

On réalise cette pondération en utilisant une mémoire morte pour la fonction de pondération et un multiplieur analogue à celui utilisé en 8.5.3. qui va multiplier chaque point de la fonction de corrélation par la valeur correspondante de la fonction de pondération.

7.6.6. CALCULS ANNEXES

Pour chaque fréquence, on calcule donc partie réelle et partie imaginaire, et de là on calcule sans difficultés le module du spectre et la phase, ce qui permet de tracer en sortie, soit le diagramme de NYQUIST, soit le diagramme de BODE.

7.7. ANALYSEUR DE SPECTRE

L'analyseur de spectre complet se compose donc d'un corrélateur qui donne les fonctions d'autocorrélation ou d'intercorrélation et d'un transformateur de Fourier.

Le même type de transformateur de Fourier peut être associé aux différents types de corrélateurs existant :

- corrélateurs numériques basses fréquences fabriqués par INTERTECHNIQUE ou S.A.I.P. (sous licence C.E.A.)
- corrélateurs moyenne et haute fréquence fabriqués par S.A.I.P. (sous licence C.E.A.) qui peuvent traiter des signaux jusqu'à 5 Mégahertz.
- corrélateur très haute fréquence (jusqu'à 50 MHz) réalisé au Laboratoire d'Electronique et de Technologie de l'Informatique du Centre d'Etudes Nucléaires de GRENOBLE.

Le même ensemble corrélateur + transformateur de Fourier permet de mesurer des autospectres ou des interspectres sans aucune complication.

Un tel ensemble fournit également les fonctions de corrélation (auto et intercorrélation) ce qui présente de très gros avantages dans beaucoup de cas.

Par exemple, en identifications de processus, la fonction d'intercorrélation entrée-sortie (lorsque l'entrée est excitée par un signal convenable) fournit la réponse impulsionnelle du système, alors que l'interspectre entrée-sortie donne la fonction de transfert.

Dans les études de turbulence la fonction d'auto-corrélation joue aussi un grand rôle et permet d'avoir une mesure des vitesses d'écoulement.

Le prix de revient d'un tel appareil est, à perfomances égales, très inférieur aux dispositifs utilisant les deux autres approches (par filtrage et par transformée de Fourier directe) et il est le seul par son principe à pouvoir traiter en temps réel des signaux de 0 à 50 Mégahertz.

La figure 11 représente un transformateur de Fourier destiné à être couplé sur un corrélateur quelconque.

Les figures 12 et 12 bis représentent un analyseur de spectre autonome (corrélateur et transformateur de Fourier).

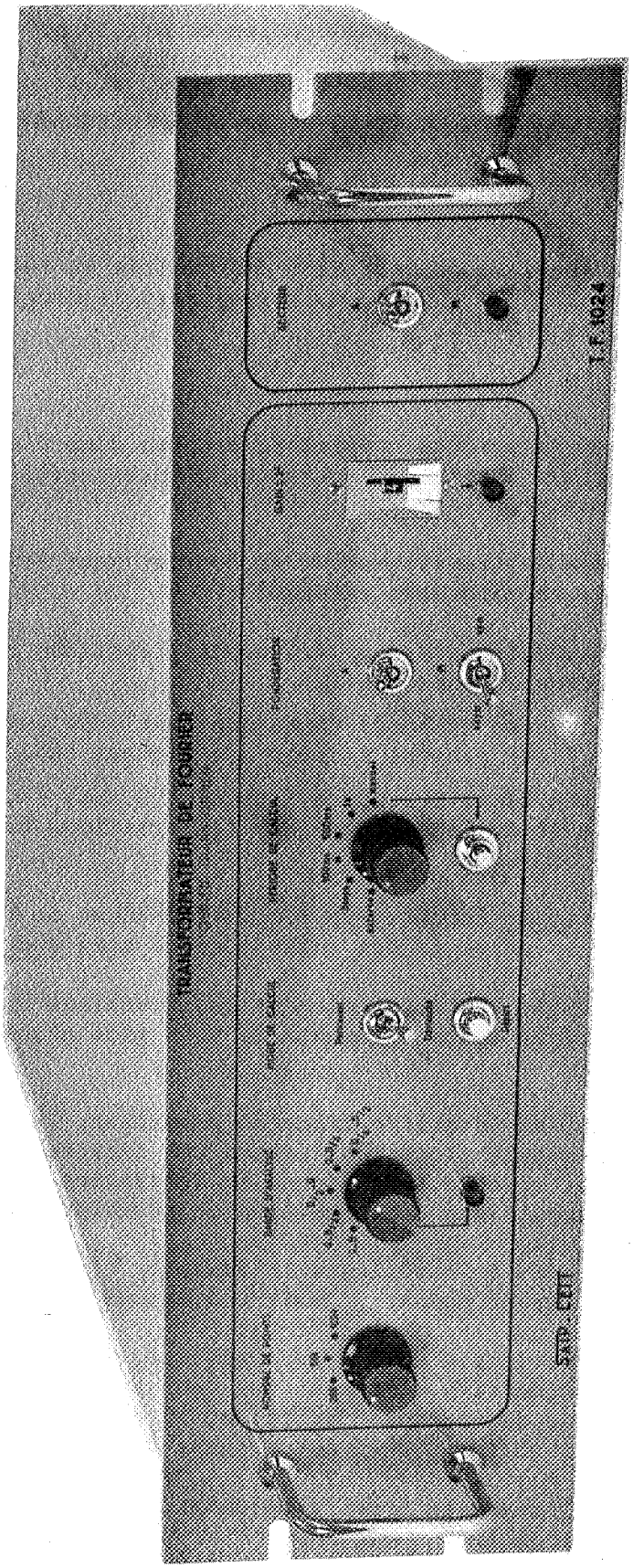


FIGURE 11

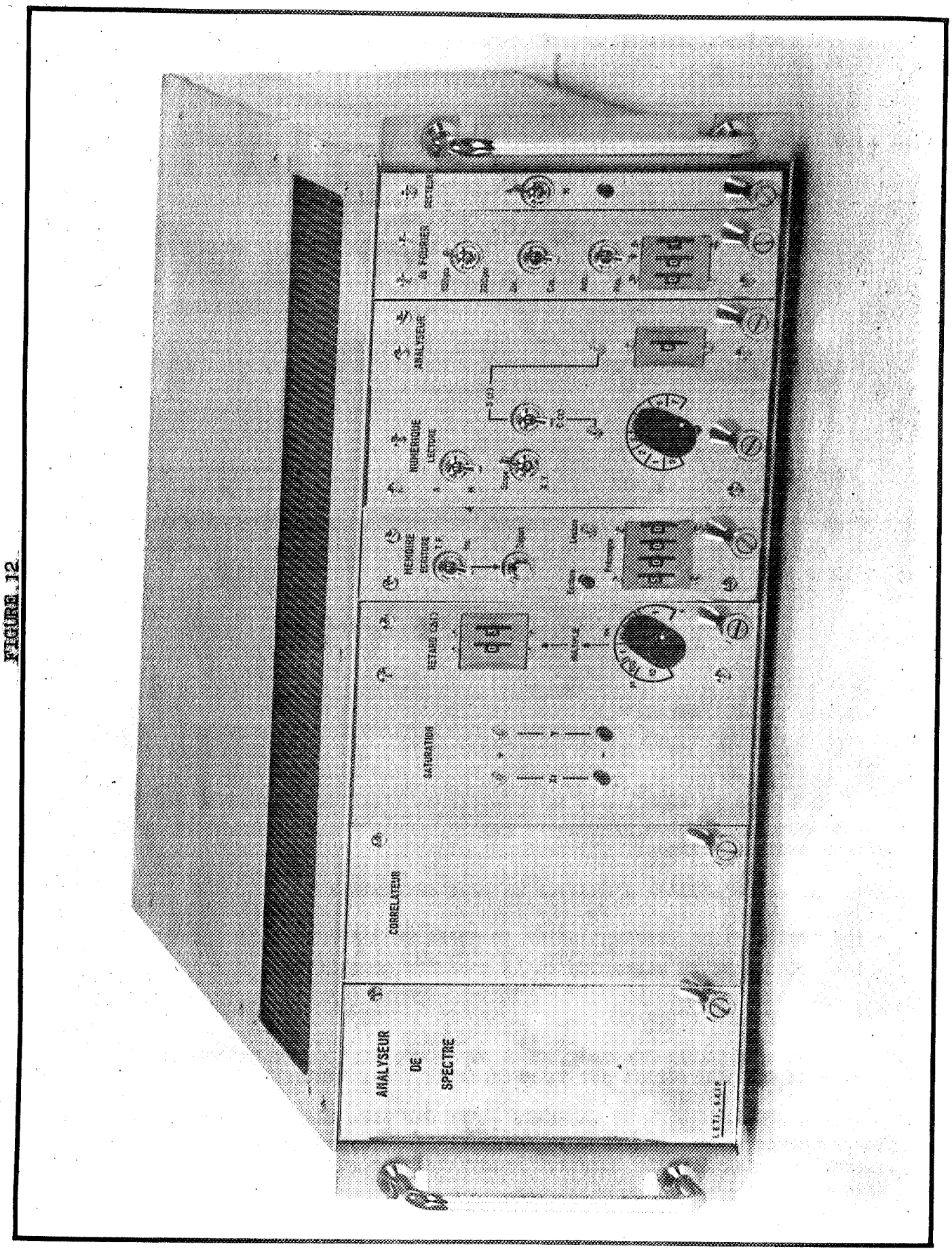
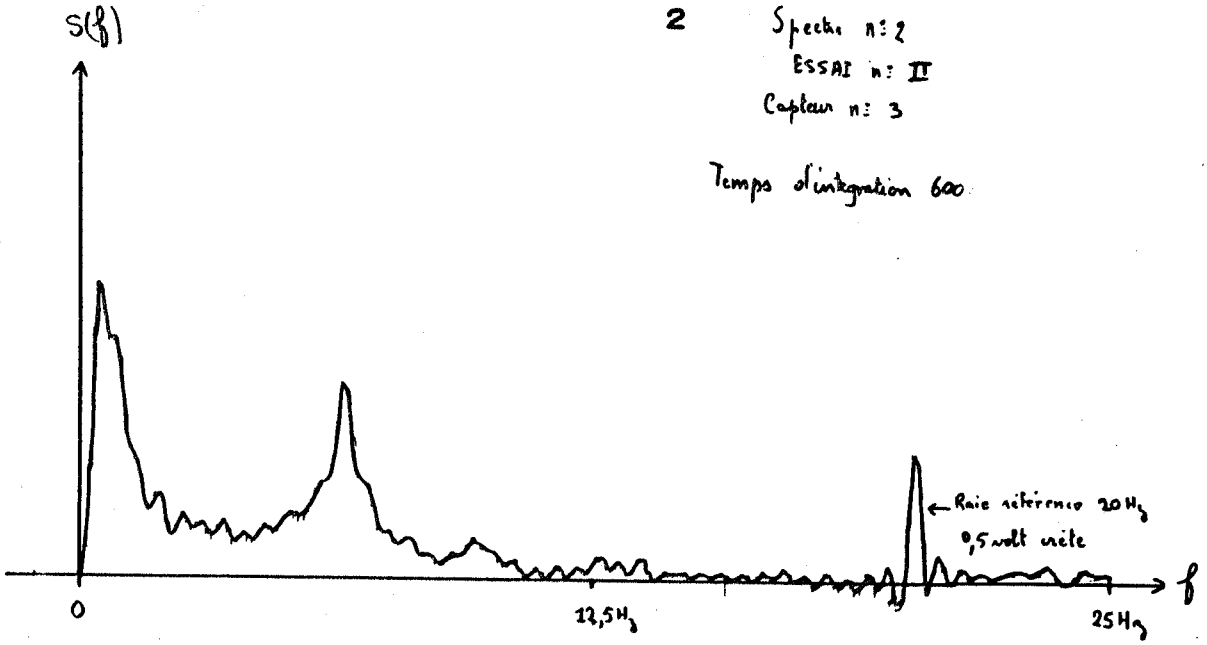

ANALYSEUR
DE
SPECTRE
CORRELATEUR
SATURATION
RETARD (ns)
MEMOIRE
ECRITURE
LECTURE
NUMERIQUE
ANALYSEUR
SECTEUR

FIGURE 12

FIGURE 13



- Spectre de Vibrations d'une conduite hydraulique.

8. EXEMPLES D'APPLICATIONS

8.1.

La figure 13 représente le résultat de l'analyse spectrale des variations de pression provoquées par un écoulement de liquide le long d'une paroi métallique.

Par cette méthode d'analyse on peut déterminer :

- les oscillations éventuelles de la masse de liquide
- les fréquences de vibration de la conduite métallique.

8.2.

Dans un deuxième exemple, nous avons appliqué cette méthode à la mesure du débit artériel par la méthode de rhéographie.

Les figures 14 et 15 montrent qu'il est plus facile d'interpréter les rhéogrammes et de mettre en évidence leurs différences sur leur représentation fréquentielle (densité spectrale) que sur leur représentation temporelle (rhéogramme classique).

8.3.

La figure 16 montre les densité spectrales d'un électroencéphalogramme à différents instants (une mesure chaque minute).

FIGURE 14 - Mesure de débit sanguin par rhéographie

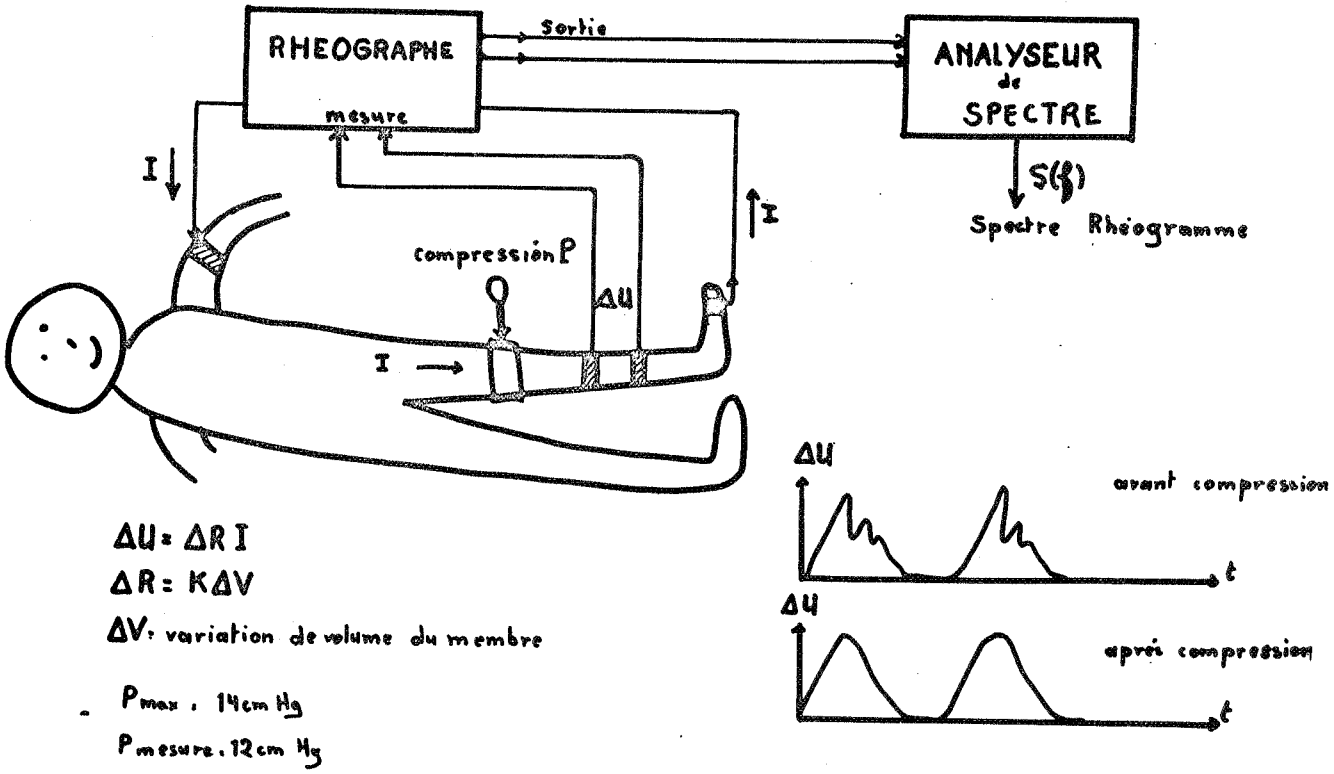


FIGURE 15 - Spectre rhéogramme

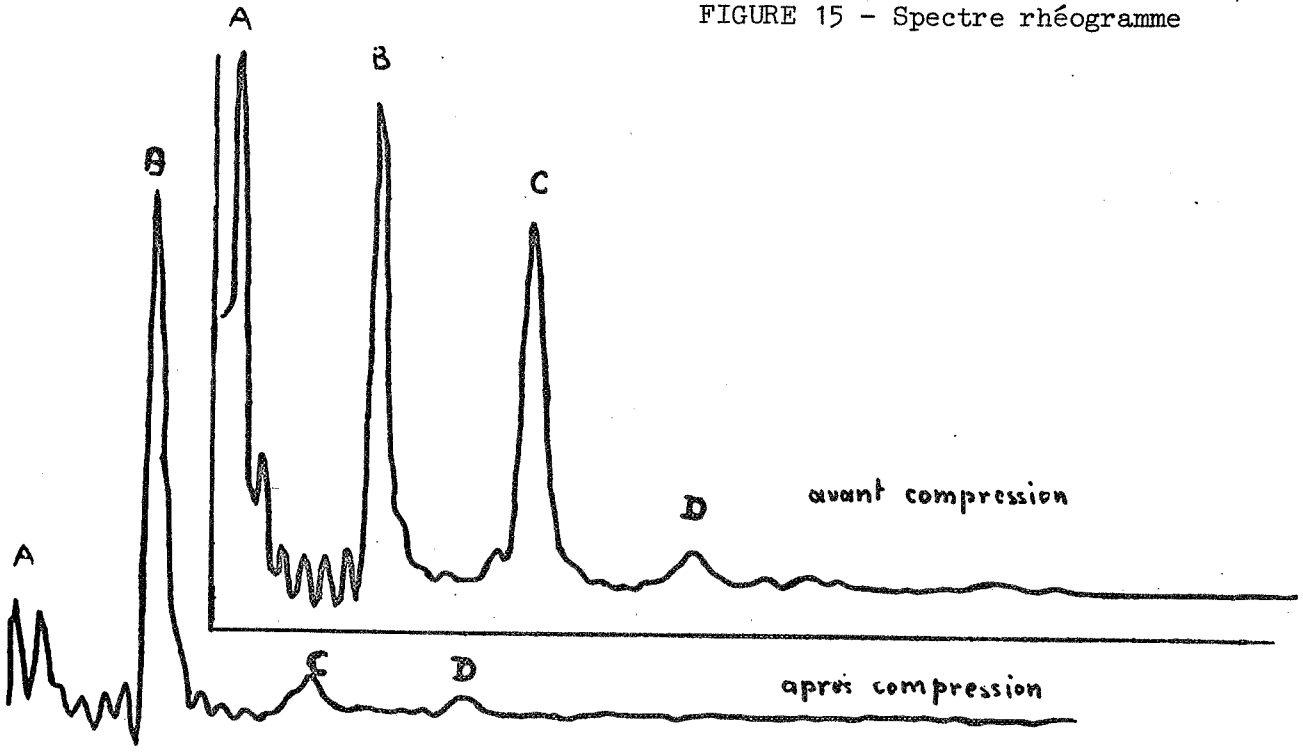
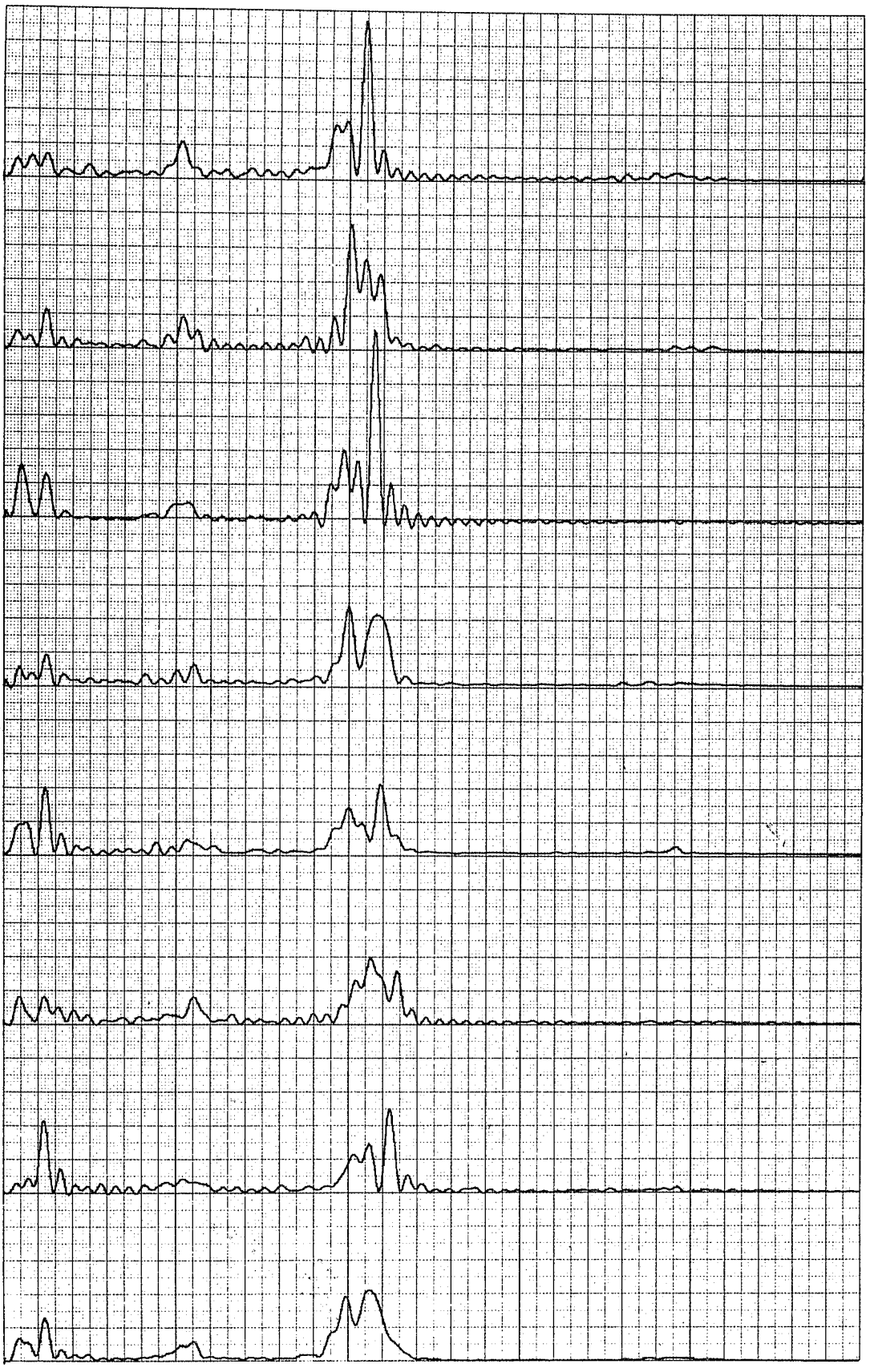


FIGURE 16



BIBLIOGRAPHIE SOMMAIRE

[1] J. MAX
Les méthodes de corrélation dans le traitement de l'Informatique. B.I.S.T. n° 96 (Septembre 1965).

[2] J. MAX et H. CHEVALIER
Analyseur statistique. Colloque sur la contribution de l'électronique aux méthodes de traitement statistique des mesures en physique. Grenoble (Avril 1966) (publié dans l'Onde Electrique, Octobre 1966).

[3] D. BERTHIER
Dispositif de calcul automatique de fonctions de corrélation pour signaux acoustiques. Colloque sur la contribution de l'électronique aux méthodes de traitement statistique des mesures en physique. Grenoble (Avril 1966) publié dans l'Onde Electrique, Octobre 1966.

[4] D. BERTHIER
Etude et réalisation d'un corrélateur automatique (multicorrélateur) fonctionnant en temps réel. Thèse de Docteur Ingénieur. Rapport C.E.A. R 3482.

[5] J. MAX
Les principales Méthodes de traitement du signal. Rapport C.E.A. R 4018.

[6] J. MAX
Statistique et cybernétique, communication au Vè congrès international de cybernétique, NAMUR (Septembre 1967).

[7] D. BERTHIER, J.M. FAUQUE, J. MAX, G. BONNET
Analyse spectrale par corrélation. Rapport C.E.A. n° R 3672.

[8] J. MAX, J.M. FAUQUE, D. BERTHIER
Les corrélateurs dans le traitement des mesures et l'analyse spectrale. Bulletin d'instrumentation scientifique et technique du C.E.A. n° 143, (Décembre 1969).

D. ABENSOUR - Quel est l'intérêt du transformateur présenté par rapport au transformateur F.F.T. ?

J. MAX - L'intérêt est principalement au niveau du rapport performances - prix.

D. ABENSOUR rappelle l'existence du système IBM.7 qui permet de faire de la F.F.T. en temps réel dans certaines conditions.

A. NEMETH - La fenêtre carrée n'introduit aucune distorsion, si sa durée est un multiple entier de la périodicité. Mais c'est un cas particulier. Habituellement, on ne connait pas à l'avance la périodicité du signal ; on la cherche même. Ainsi, la fenêtre pondérée convient mieux à l'usage général.

D.ABENSOUR donne des précisions sur le filtrage numérique.

J. MAX ne conteste pas que le filtrage numérique est une technique intéressante. Il attend la fabrication intégrée de multiplieurs numériques 8 bits.

R.A. GUEDJ - A-t-on appliqué vos méthodes au traitement du signal de parole ?

R. CARRÉ - Nous n'avons pas d'expérience sur l'application des techniques d'analyse par corrélation, mais nous travaillons sur un vocoder à autocorrélation. Cet appareil transmet des données sur la fonction d'autocorrélation du signal de parole. Nous avons l'intention d'étudier les importances relatives des paramètres transmis.

NOTION DE SPECTRE INSTANTANE DANS L ' ANALYSE ET LA SYNTHESE DES SIGNAUX

B . E S C U D I É

I.C.P.I. Laboratoire de Traitement du Signal - Lyon

INTRODUCTION

Les besoins actuels des Communications Sonores ou électromagnétiques ont conduit à l'emploi de signaux de plus en plus complexes. Ces signaux se caractérisent par une large bande spectrale et une durée grande ; leur produit BT est grand devant l'unité. Que ce soit par l'analyse spectrale classique ou celle par corrélation et transformée de Fourier, les résultats fournis traduisent mal les modulations de phase et de fréquence du signal. Ceci suggère de représenter ces signaux dans un plan "temps-fréquence" où les modulations seraient plus apparentes.

1. QUELQUES EXEMPLES D'ANALYSE SPECTRALE DE SIGNAUX MODULES OU NON

La figure 1 nous montre un signal d'aspect compliqué où l'amplitude des oscillations varie, où la phase subit des variations brusques (basculements), où la loi des passages à zéro semble assez compliquée. Sa fonction de corrélation $\Gamma(\tau)$ est cependant très proche de celle d'un signal sinusoïdal exponentiellement amorti.

La figure 2 compare les densités spectrales $\gamma_S(\nu) \rightleftharpoons \Gamma_S(\tau)$ d'un signal modulé linéairement en fréquence occupant la bande B et de durée T ($BT \ggg 1$) et un signal sinusoïdal modulé par une fonction :

$$\frac{\sin \pi Bt}{\pi t}$$

$$S_a(t) = \Pi_T(t - \frac{T}{2}) \sin 2\pi (\nu_0 + at)t$$

$$S_b(t) = \cos 2\pi \nu_0 t \, \frac{\sin \pi Bt}{\pi t}$$

Leurs densités spectrales sont apparamment du même type et pourtant $S_a(t)$ est un signal à modulation de fréquence linéaire et $S_b(t)$ un signal à modulation d'amplitude par multiplication, c'est-à-dire sans modulation de fréquence apparente.

Ces quelques exemples nous montrent comment l'analyse spectrale classique d'un emploi si commode ne peut résoudre aisément le problème de la représentation des modulations de fréquence ou de phase au cours du temps.

2. RELATIONS "TEMPS - FREQUENCE", RELATIONS D'INCERTITUDE ET DEVELOPPEMENT DE GABOR

Rappelons que la notion de fréquence est définie sur une onde sinusoïdale en régime permanent.

$$X(t) = A_0 \sin(2\pi \nu_0 t + \varphi_0)$$

Pour mesurer ν_0, il faut disposer d'au moins deux ou plusieurs périodes T_0 définies par exemple par les passages à zéro du signal X(t) (fig. 3). La notion de "fréquence" est donc un paramètre NON LOCALISE dans le temps ; il ne peut se déduire d'un seul échantillon pris à la date T_0.

Ceci nous montre que, pour "suivre" une modulation de fréquence, il faut suivre l'évolution temporelle d'un paramètre "non localisé" dans le temps.

Nous allons maintenant rappeler les relations dues à D. GABOR [1] sur l'étendue temporelle Δt et l'étude spectrale $\Delta \nu$ d'un signal S(t) d'énergie finie E_S, et muni d'une transformée de Fourier $\mathcal{S}(\nu) \rightleftharpoons S(t)$ (fig. 4). On choisit pour Δt et $\Delta \nu$ une définition énergétique d'un type analogue à celui d'un rayon de giration en mécanique :

$$\left[\begin{array}{l} (\Delta t)^2 = \dfrac{1}{E_S} \displaystyle\int_{-\infty}^{+\infty} t^2 \, |S(t)|^2 dt \\[2ex] (\Delta \nu)^2 = \dfrac{1}{E_S} \displaystyle\int_{-\infty}^{+\infty} \nu^2 \, |\mathcal{S}(\nu)|^2 dt \end{array}\right.$$

On montre alors qu'il existe une inégalité :

$$\boxed{\Delta t \cdot \Delta \nu \geq \frac{1}{4\pi}}$$

Cette relation exprime un fait bien connu des électroniciens. A tout signal S de durée T donnée, correspond une bande spectrale minimale et à bande passante B donnée pour un filtre, correspond un temps de montée minimal τ_m

$$\tau_m \simeq \frac{1}{2B} \text{ ou } \frac{1}{1{,}5\,B}$$

Cette relation est analogue aux relations d'incertitude de la mécanique quantique ; si le formalisme employé peut être le même (formalisme de Dirac utilisé par G. BONNET et G. GARAMPON [2]), toute analogie étroite est à éviter comme l'a montré R.M. LERNER [3].

Il existe un signal S(t) qui minimise l'expression de D. GABOR :

$$\Delta t \cdot \Delta \nu \geq \frac{1}{4\pi}$$

c'est le signal s(t) suivant :

$$S(t) = A_0 e^{-(t/\tau)^2} \quad \text{pour lequel} \quad \Delta t \Delta \nu = \frac{1}{4\pi}$$

Ceci n'a rien d'étonnant car l'on sait pratiquement en télécommunications, que l'impulsion gaussienne requiert la bande passante minimale à la transmission pour une durée donnée. De plus, cette fonction, exprimée sous la forme réduite : $e^{-\pi t_1^2}$ a pour transformée de Fourier :

$$e^{-\pi t_1^2} \rightleftharpoons e^{-\pi \nu_1^2}$$

Partant de ce fait, D. GABOR essaya donc de représenter les signaux dans le plan temps-fréquence. Il associa à chacune des cellules d'aire $\frac{1}{4\pi}$ un signal élémentaire gaussion :

$$\exp\left(-\pi \frac{(t-n\Delta t)^2}{2(\Delta t)^2}\right) \exp\left(-i\,2\pi \frac{kt}{\Delta t}\right)$$

et chercha les coefficients CnK tels que

$$S(t) = \sum_n \sum_k C_{nk} \exp\left(\frac{-\pi(t-n\Delta t)^2}{2(\Delta t)^2}\right) \cdot \exp\left(-i\frac{2\pi kt}{\Delta t}\right) = \sum_n \sum_k C_{nk} \Psi\left(t, \frac{k}{\Delta t}\right)$$

comme le montre la figure 5. Ce développement de S(t) sur les fonctions $\Psi_{nk}(t, \frac{k}{\Delta t})$ est d'emploi peu commode. Les fonctions Ψ_{nk} ne sont pas orthogonales et ne fournissent pas un développement unique de S. HELSTROM puis MONTGOMERY et REED montrèrent que les fonctions Ψ peuvent être du type :

$$\Psi(t,\tau,\nu) = G(t+\tau)e^{i2\pi\nu t - i(\alpha-\frac{1}{2})2\pi\nu\tau}$$

avec $\int_{-\infty}^{+\infty} G(t)dt = 1$

et occuper une cellule $\Delta\nu\,\Delta t \geq \frac{1}{4\pi}$

Alors, S(t) s'exprime par :

$$S(t) = \frac{1}{2\pi}\iint_{-\infty}^{+\infty} g(\tau,\omega)\,\Psi^{x}(t,\tau,\omega)d\tau d\omega \qquad \omega = 2\pi\nu$$

avec $g(\tau,\omega) = e^{-i\omega\tau(\frac{1}{2}-\alpha)}\int_{-\infty}^{+\infty} F(t)G^{x}(t+\tau)e^{-i\omega t}dt$

G(t) peut être la réponse impulsionnelle d'un filtre. Dans ce cas, le développement peut être obtenu avec un grille de filtres de gain $g(\nu) \rightleftarrows G(t)$ espacés en fréquence de la quantité $\frac{k}{\Delta\nu}$. Pratiquement, l'analyse effectuée en général dans l'appareil appelé sonogramme consiste à mesurer la puissance issue de chaque filtre, c'est à mesurer :

$$|Cnk|^2 \text{ ou } |g(\tau,\omega)|^2_{\tau_o,\nu_o}$$

Remarquons qu'alors toute notion de phase est abandonnée.

3. REPRESENTATION [t,ν] ET SPECTRE INSTANTANE DE PUISSANCE

Le développement de D. GABOR s'intéresse aux coefficients Cnk permettant de représenter S(t) à l'aide de signaux élémentaires $\psi_{nk}(t,\nu)$. Une autre méthode d'étude s'est manifestée chez divers auteurs. Elle consiste à déterminer la répartition énergétique (ou de puissance) dans le plan [t,ν]. Parmi ces divers travaux FANO, PAGE, LAMPARD... , nous retiendrons le travail de J. VILLE [4] qui fut si fécond en permettant par la suite à P.M. WOODWARD de définir la fonction d'Ambiguïté [5]. J. VILLE essaie de définir la notion de FREQUENCE INSTANTANEE qui, malgré l'apparente contradiction des termes, possède un sens physique assez riche, ainsi que le montre A. BLANC LAPIERRE et B. PINCINBONO [6]. J. VILLE utilise la notion introduite par A. VANDER POL [7] :

$$\nu(t) = \frac{1}{2\pi}\frac{d\Phi}{dt} \quad \text{avec } S(t) = \cos\Phi(t)$$

puis il associe à cette notion de fréquence un opérateur $\frac{i}{2\pi}\frac{d}{dt}$ agissant sur $\Phi(t)$ et il cherche l'expression de $\rho(t,\nu)$ spectre instantané qui devrait être la densité de probabilité associée à la distribution énergétique dans le plan (t,ν) :

$$\int_{-\infty}^{+\infty} \rho(t,\nu)\, dt\, d\nu = E_S$$

cf figure 6

$$\int_{-\infty}^{+\infty} \rho(t,\nu)\, d\nu = |S(t)|^2$$

$$\int_{-\infty}^{+\infty} \rho(t,\nu)\, dt = |\mathcal{S}(\nu)|^2$$

Ces trois relations permettent de définir $\rho(t,\nu)$ comme une distribution énergétique conjointe dans le plan $[t,\nu]$ J. VILLE calcule alors la fonction caractéristique associée à $\rho(t,\nu)$ c'est-à-dire la quantité

$$F(u,v) \overrightarrow{\overleftarrow{(u,v)}} \rho(t,\nu)$$

$$F(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{i[ut+uv]}\rho(t,\nu)\, dt\, d\nu$$

Dans ce calcul J. VILLE utilisait la notion de signal analytique associé à un signal $S(t)$. Si cette notion a quelque intérêt pour des signaux à bande étroite, il n'en est rien pour les signaux à large bande qui sont de plus en plus utilisés en communication. Le résultat final est que $F(u,v)$ s'exprime avec une fonction que WOODWARD appela par la suite fonction d'ambiguïté.

G. BONNET et G. GARAMPON montrèrent en 1967 que le résultat de VILLE pouvait être obtenu d'une manière élégante et condensée en utilisant pour l'analyse harmonique des signaux le langage ou formalisme de DIRAC utilisé en Mécanique Quantique [8], [9]. Le principe est le suivant : On définit deux opérateurs t et ν qui agissant sur le "vecteur" signal S fournissent

$$< t,\ S > = S(t)$$

$$< \nu,\ S > = S(\nu) \quad \text{avec } S(t) \rightleftharpoons s(\nu)$$

Ces deux opérateurs possèdent un commutateur [10] :

$$[t,\ \nu] = \frac{i}{2\Pi}$$

Calculant la valeur moyenne de l'opérateur $Q = e^{i(ut+v\nu)}$ qui n'est autre que la fonction caractéristique de la répartition conjointe (t,ν), calculée par J. VILLE, G. BONNET et G. GARAMPON obtiennent :

$$F(u,v) = \frac{e^{iuv/4\Pi}}{E_S} \int_{-\infty}^{+\infty} S^x(t)\ S(t + \frac{v}{2\Pi})\ e^{iut}\ dt$$

Quantité qui est exprimée en fonction de $S(t)$ et non pas en fonction du signal analytique associé à $S(t)$. Si $S(t)$ est REEL (cas des signaux expérimentaux !) :

$$F(u,v) = \frac{e^{\frac{iuv}{4\Pi}}}{E_S} \int_{-\infty}^{+\infty} S(t)\ S(t+\frac{v}{2\Pi})\ e^{iut}\ dt$$

posons $\frac{v}{2\Pi} = -\tau$ et $u = -2\Pi\phi$ il vient :

$$F(-\frac{\phi}{2\Pi}) - 2\Pi\tau) = \frac{e^{\frac{i}{4\Pi}\phi\tau}}{E_S} \int_{-\infty}^{+\infty} S(t)\ S(t-\tau)\ e^{-2\Pi i\phi t}\ dt$$

soit : $$F(-\frac{\phi}{2\Pi}\ ,\ -2\Pi\tau) = \frac{1}{E_S}\ e^{\frac{i}{4\Pi}\phi\tau}\ \chi_S\ (\tau,\ \phi)$$

où $\chi_S(\tau,\ \phi)$ est la fonction d'ambiguïté de P.M. WOODWARD dans le cas du signal $S(t)$ quelconque. Or cette dernière quantité est d'une importance fondamentale dans la théorie des communications. Cela chiffre le pouvoir de résolution en retard et en déplacement fréquentiel associé à un signal $S(t)$ traité par son filtre adapté [9]. Cette quantité permet de chiffrer automatiquement les performances d'un signal utilisé en RADAR, SONAR, ou en communications.

Prenons la transformée de Fourier sur u et v pour calculer $\rho(t,\nu)$:

$$\rho(t,\nu) = \int_{-\infty}^{+\infty} S^x(t - \tfrac{\tau}{2})\ S(t + \tfrac{\tau}{2})\ e^{-2\Pi i\ \nu\tau}\ d$$

$$\rho(t,\nu) = \int_{-\infty}^{+\infty} \mathcal{S}^x(\nu + \tfrac{f}{2})\ \mathcal{S}(\nu - \tfrac{f}{2})\ e^{-2\Pi ift}\ df$$

où l'on peut remarquer que S(t) est REEL dans la pratique. Connaitre $\rho(t,\nu)$ revient donc à connaitre $\chi_S(\tau, \phi)$ à un facteur près.

Remarquons que $\rho(t,\nu)$ n'est pas une quantité positive comme le désirait VILLE qui l'avait définie comme une densité de probabilité. En fait, $\rho(t,\nu)$ dt dν est un accroissement infiniment petit énergétique dans le plan temps fréquence. $P(t) = |S|^2$ est la puissance et ici on s'intéresse aux variations de cette quantité sur les domaines dt, dν, et il n'y a aucune raison que cette quantité soit essentiellement positive.

En résumé, nous avons obtenu une quantité $\rho(t,\nu)$ ou sa transformée de Fourier qui sont aptes à traduire les modulations de fréquence ou de phase en localisant l'énergie dans le plan [t,ν]. Ainsi dans le cas d'un signal modulé linéairement en fréquence, la surface d'ambiguïté traduit la pente de modulation par la direction de l'arête principale dans le plan (τ, ϕ).

4. DENSITE ENERGETIQUE DANS LE PLAN (t,ν) : THEORIE DE A.W.RIHACZEK

Nous exposerons les résultats de A.W. RIHACZEK [10] en utilisant des signaux REELS correspondant aux signaux utilisés en pratique. Ayant défini la fonction d'intercorrélation de deux signaux X(t) et Y(t) :

$$\Gamma_{XY}(\tau) = \int_{-\infty}^{+\infty} X(t)\ \ Y^{*}(t-\tau)\ dt \rightleftharpoons x(\nu)\ y^{*}(\nu)$$

nous définirons l'ENERGIE d'INTERACTION de X et Y par la quantité

$$E_{XY} = \Gamma_{XY}(0) = \int_{-\infty}^{+\infty} X(t)\ Y^{*}(t)\ dt = \int_{-\infty}^{+\infty} X(t)\ Y(t)\ dt$$

Nous allons maintenant déterminer Y(t) comme la filtrée de X(t) à travers le filtre $F(\nu, \Delta\nu)$ qui isole la bande Δν autour de la fréquence ν et nous calculons E_{XY} :

$$E_{XY} = \int_{-\infty}^{+\infty} X(t)\ \ F_{\nu\,\Delta\nu}[X(t)]\ dt$$

Cette quantité doit représenter l'interaction entre la bande spectrale $(\nu \pm \Delta\nu)$ et l'ensemble du signal, et doit pouvoir traduire les modulations de fréquence et de phase du signal S(t). En calculant la quantité :

$E_{\delta T \delta \nu}$ associée à la durée δT autour de t et à la bande spectrale $\delta\nu$ autour de ν . RIHACZEK montre que l'on peut définir une quantité $d(t,\nu)$ <u>densité énergétique dans le plan (t,ν)</u> telle que :

$$d(t,\nu) = 2\mathrm{Re}\left[S^{x}(\nu) e^{-2\Pi i \nu t} . S(t) \right] = 2S(t) \int_{-\infty}^{+\infty} S(t-\tau) \cos 2\Pi \nu \tau \, d\tau$$

S REEL

$$\int_{-\infty}^{+\infty} d(t,\nu)\, dt = 2|\mathcal{S}(\nu)|^2$$

$$\int_{-\infty}^{+\infty} d(t,\nu)\, d\nu = 2|S(t)|^2$$

Il est alors possible de montrer qu'il y a une relation avec $\rho(t,\nu)$ et qu'on plus $d(t,\nu)$ a une transformée de Fourier à deux dimensions sur t et ν telle que :

$$d(t,\nu) \rightleftharpoons J(\tau, \phi) = \chi_S(\tau, -\phi) + \chi_S(-\tau, -\phi)$$

La connaissance de $\chi_S(\tau, \phi)$ permet donc, ici encore, de déterminer la rperésentation (t,ν).

Cette nouvelle définition $d(t,\nu)$ a un intérêt dans le cas des signaux à large bande ou à modulation forte du type :

$$S(t) = F(t) \cos \Phi(t) \qquad \text{(cf figure 7)}$$

où $F(t)$ est une enveloppe définie sur un support borné 0, T et où $\Phi(t)$ est une phase modulée au cours du temps.

Si le produit BT, B bande spectrale occupée, T durée du signal est grand devant l'unité il est possible de montrer que la quantité E_{BT} se concentre dans le plan temps fréquence autour d'une courbe donnant la loi de modulation.

$$S(t) = F(t) \cos \Phi(t) \rightleftharpoons |\mathcal{S}(\nu)|\, e^{i\varphi(\nu)}$$

La concentration se fait dans le plan (t,ν) autour des points :

$$\nu = \nu_{inst}(t) = \frac{1}{2\Pi} \frac{d}{dt} (\Phi(t))$$

cf figure 8

$$t = \tau g(\nu) = -\frac{1}{2\Pi} \frac{d}{d\nu} (\varphi(\nu))$$

Exprimons maintenant le taux de concentration de E_{BT} autour de cette courbe. Cet écart est donné par l'erreur entre la loi de phase réelle et la loi de phase linéaire.

$$\Phi(t) = A + B(t-to) + C(t-to)^2 + \ldots.$$

si le terme d'écart est pris à $\frac{\Pi}{4}$ rd on obtient : $C = \left[\frac{\partial^2}{\partial t^2}\Phi(t)\right]_{t=to}$

on appelle $T_r = \frac{1}{\sqrt{\left|\frac{d\,i}{dt}\right|}}$ temps de relaxation

de même pour $\varphi(\nu)$ d'où :

$$Bd = \frac{1}{\sqrt{\left|\frac{d\,ig}{d\nu}\right|}} \quad \text{et} \quad Bd\ Tr = 1$$

avec Bd bande dynamique

Ce dernier résultat justifie ce calcul, car il est en accord avec les relations d'incertitudes du plan [t,ν]. Nous avons tracé dans le plan (t,ν) les cellules (Bd, Tr) en nous rappelant que Bd est la bande dynamique (liée à la pente de modulation) et Tr le temps de relaxation. Le cas étudié est celui d'une modulation de fréquence hyperbolique (figure 9).

$$\nu_i(t) = \frac{\nu_2}{1+\alpha(t-to)} \qquad \nu_1 = \frac{\nu_2}{1+\alpha \cdot T}$$

5. AUTRES DEFINITIONS DU SPECTRE INSTANTANE DE PUISSANCE

D'autres auteurs ont tenté de définir le spectre instantané de puissance indépendamment des notions de représentation temps fréquence. C.H. PAGE, pour sa part, remarque que dans le signal S(t) c'est la valeur à l'instant t qui importe du point de vue acoustique ; en utilisant une analogie avec les effets acoustiques sur l'oreille il définit donc le spectre instantané par : [10] [11]

$$\rho(t,\nu) = \frac{\partial}{\partial t}\left[|\mathcal{S}_t(\nu)|^2\right]$$

avec $\mathcal{S}_t(\nu) \rightleftharpoons S_t(\theta)$ $\begin{cases} = S(\theta) & \theta < t \\ = 0 & \theta > t \end{cases}$

et avec $S_t(\theta)$

Ceci revient à faire agir un opérateur coupure sur le signal $S(\theta)$ et à calculer $\Gamma_S(t,\tau)$ appelée fonction d'auto-corrélation mobile [11]

$$\gamma_S(t,\nu) = |\beta(t,\nu)|^2 \quad , \quad \rho(t,\nu) = \frac{\partial}{\partial t}\left[\gamma_S\ (t,\nu)\right]$$

$\rho(t,\nu)$ représente l'accroissement spectral sur la bande spectrale entre les instants t et t+dt. Pour des raisons identiques à celles invoquées dans le paragraphe 3, $\rho(t,\nu)$ n'est pas défini positif. Dans le cas général il peut être négatif. On montre d'ailleurs que $\rho(t,\nu)$ peut s'exprimer sous une forme analogue à $d(t,\nu)$:

$$\rho(t,\nu) = 2\int_0^{+\infty} X(t)\ X(t-\tau)\ \cos 2\pi\nu\tau\, d\tau$$

$$\rho(t,\nu) = 2\ S(t)\ \mathrm{Re}\left\{\beta^x\ (t,\nu)\ e^{-2\pi i\nu t}\right\}$$

Calculons maintenant la Transformée de Fourier à deux dimensions de $\rho(t,\nu)$. Cette quantité s'exprime en fonction de $\chi_S(\tau,\phi)$ par la relation suivante :

$$\chi_S(|\tau|\ ,\ -\phi) \rightleftharpoons \rho(t,\nu)$$

Nous constatons une fois de plus que la Transformée de Fourier à deux dimensions de cette nouvelle définition du spectre instantané est reliée à la fonction d'ambiguité. Cette dernière quantité prend donc de l'importance, et le calcul rapide et automatique de χ_S devient un des problèmes à résoudre pour l'analyse spectrale des signaux à modulations fortes.

Il existe d'autres définitions avec l'opérateur coupure temporel agissant sur le futur du signal. On consultera à ce sujet l'étude critique de ces notions faites par A. BLANC-LAPIERRE et B. PICINBONO [6].

6. SENS PHYSIQUE ET MOYENS DE MESURE DU SPECTRE INSTANTANE DE PUISSANCE

a/- Sens physique

Une étude critique détaillée des notions existantes en 1955 fut faite par A. BLANC-LAPIERRE et B. PICINBONO. Il en ressort que cette notion de spectre instantané est une notion fortement relative, comme le traduit la multiplicité des définitions existantes. Ceci a pour origine les définitions mêmes des grandeurs TEMPS et FREQUENCE. En effet pour définir théoriquement la fréquence il faut disposer d'une onde monochromatique plane de durée infinie [67]. Cependant les effets acoustiques sur l'oreille, tels celui d'une sirène ou d'un moteur à régime variable, ou les modulations de fréquence ou de phase réalisables en radiocommunications suggèrent la notion "d'évolution de la fréquence".

$$X(t) = A_o \cos 2\pi (\nu_o + a \sin 2\pi \nu_1 t)\, t$$

Si nous appelons $\varphi(t)$ la phase ou argument du cosinus :

$$X(t) = A_o \cos \varphi(t)$$

et par analogie avec le cas classique à fréquence fixe on pose :

$$X(t) = A_o \cos (2\pi\, \nu_o t + 2\pi\, a \sin 2\pi\, \nu_1 t)$$

avec $\nu(t)$ "fréquence instantanée" telle que :

$$\begin{cases} X(t) = A_o \cos \varphi(t) & \nu_i = \dfrac{1}{2\pi}\dfrac{d\varphi}{dt} \\ \nu_i(t) = \nu_o + a\nu_1 \cos 2\pi\, \nu_1 t \end{cases}$$

Dans le cas d'un signal sans porteuse du type Y(t) :

$$Y(t) = \Pi_T\,(t - T/2) \sin 2\pi\, (\nu_o + at)\, t$$

on peut soit définir Y(t) comme ci-dessus, ou à partir de la loi temporelle des passages à zéros t_k :

$$\nu_k = \frac{1}{\Delta t_k} = \frac{1}{t_k - t_{k-1}}$$

Ces diverses définitions montrent ce caractère relatif dû à la nature "complémentaire" des deux paramètres t et ν liés par la relation d'incertitude. Cependant le sens physique de $\rho(t,\nu)$ quelleque soit sa définition peut se dégager comme l'apport spectral et temporel observé à la date $(t, t + dt)$ dans la bande $(\nu, \nu + d\nu)$. Dans le cas d'un signal à modulation de fréquence ou de phase où le facteur BT $>>> 1$ on montre que l'on peut "suivre l'évolution" de $\nu(t)$ au cours du temps.

Remarquons maintenant que les travaux récents ont dégagé des notions reliées à la fonction d'Ambiguité de WOODWARD, si utile en traitement du signal. Cette fonction exprime d'ailleurs au terme E_S près la distance quadratique entre $S(t)$ et $S(t-\tau)_\tau e^{-2\pi i \phi t}$ si S est réel. Elle permet de traduire par les paramètres τ et ϕ la loi de modulation en fréquence du signal, mieux que ne le fait $\gamma_S(\nu)$ qui ne porte que sur le seul paramètre ν. Enfin signalons que dans d'autres domaines telle la mécanique quantique, on considère des quantités analogues avec un formalisme assez proche, mais l'on doit se garder d'un parallèle trop étroit.

Les travaux de G. BONNET et G. GARAMPON-JOURDAIN ont permis d'éclairer la notion de spectre instantané et de représentation temps-fréquence en utilisant un formalisme particulièrement puissant et en fournissant surtout une relation simple avec la fonction d'Ambiguité de WOODWARD, sans aucune hypothèse sur la nature spectrale du signal $S(t)$: [2]

$$\rho(t,\nu) \rightleftharpoons F(u,v) = \frac{e^{\frac{iuv}{4\pi}}}{E_S} \int_{-\infty}^{+\infty} S^x(t)\, S(t+\tfrac{v}{2\pi})\, e^{iut}\, du$$

soit :

$$\rho(t,\nu) \rightleftharpoons F(-\frac{\phi}{2\pi}, -2\pi\tau) = \frac{e^{\frac{i\phi\tau}{4\pi}}}{E_S} \chi_S(\tau,\phi)$$

$\chi_S(\tau,\phi)$ en plus de son intérêt propre quant à l'évaluation des performances d'un signal traité par son filtre adapté nous fournit un moyen de représentation des modulations de phase et de fréquence du signal. Ceci explique son importance croissante dans l'analyse et la synthèse des signaux.

b/- Mesure des spectres instantanés

Nous avons vu lors de l'étude de la définition de GABOR que l'on doit mesurer les coefficients Cn_k attachés au filtre de rang k, fréquence centrale $\frac{k}{\Delta t}$, et observé en sortie à l'instant Δt. Pratiquement, on réalise une grille de filtres de largeur de bande b jointifs et échantillonnés à la cadence $\frac{1}{2b}$. Ayant mesuré la puissance en détectant puis en intégrant sur la durée $\frac{1}{2b}$ à la sortie de chaque filtre on mesure ainsi $|Cn_k|^2$, comme le montre la figure 10.

On ne dispose donc dans l'appareil appelé sonographe qui réalise cette opération, que des modules carrés des coefficients Cn_k. De plus la représentation habituelle par noircissement d'un papier est des plus imprécise et ne permet pas de mesurer correctement ces paramètres. Les techniques actuelles de conversion analogique digitale et l'emploi des sélecteurs multicanaux devraient permettre d'améliorer les résultats classiques pour l'analyse et la synthèse des signaux.

Si l'on désire mesurer $d(t,\nu)$ au sens de A.W. RIHAZECK il faut mesurer l'énergie d'interaction entre le signal $S(t)$ et de signal filtré dans un filtre isolant $\Delta\nu$. Ceci revient à mesurer $E_{\Delta\nu,\Delta T}$ c'est-à-dire l'intégrale de $d(t,\nu)$ sur un petit domaine dans le temps fréquence. Si le signal est à modulation forte ($BT \gg 1$) RIHAZECK a montré que cette grandeur se concentre autour de la courbe $\nu_i(t)$ qui traduit la modulation. Le dispositif mesurant l'énergie d'interaction de S et de sa filtrée est représenté figure 11 :

$$E_{S_1 S_2} = \int_{-\infty}^{+\infty} S_1(t)\, S^{x}_{2}(t)\, dt \qquad S_1 \text{ et } S_2 \text{ REELS}$$

Enfin pour le cas du spectre instantané au sens de C.H. PAGE, l'estimation de cette grandeur passe par celle de $I_S(t,\tau)$ que l'on peut estimer à l'aide d'un circuit porte et d'un multicorrélateur en temps réel suivi de transformateur de Fourier qui fournit :

$$\gamma_S(t,\nu) = |\mathcal{S}(t,\gamma)|^2$$

Des différentes valeurs $\gamma_S(t_k,\nu)$, il est possible de déduire $\rho(t_k,\nu)$. Le schéma de l'opération est représenté figure 12 et deux résultats d'analyse forment la figure 13.a et b.

c/- Mesure de la fonction d'Ambiguité

Nous devons déterminer la fonction d'ambiguité :

$$\chi_S(\tau,\phi) = \int_{-\infty}^{+\infty} S(t)\ S(t-\tau)\ e^{-2\pi i\phi t}\ dt\ ,\ \underline{S\ Réel}$$

Remarquons que l'on peut écrire :

$$\begin{cases} \chi_S(\tau,\phi) = R_S(\tau,\phi) + i\ I_S(\tau,\phi) \\ R_S(\tau,\phi) = \int_{-\infty}^{+\infty} S(t)\ S(t-\tau)\ \cos 2\pi\ \phi t\ dt \\ I_S(\tau,\phi) = \int_{-\infty}^{+\infty} S(t)\ S(t-\tau)\ \sin 2\pi\ \phi t\ dt \end{cases}$$

Ces quantités R_S et I_S sont les fonctions d'intercorrélation entre $S(t)\cos 2\pi\ \phi t$ ou $S(t)\sin 2\pi\ \phi t$ et $S(t-\tau)$. Pour les estimer il suffit de disposer :

- d'un générateur fournissant $\cos 2\pi\ \phi t$ ou $\sin 2\pi\ \phi t$ déclenché par $S(t)$
- d'un multiplieur formant $S(t)\cos 2\pi\ \phi t$ ou $S(t)\sin 2\pi\ \phi t$
- d'un multicorrélateur

comme le montre la figure 13.

Divers essais entrepris actuellement ont montré que cette technique était parfaitement viable avec l'apparition des multiplieurs rapides (bande 1à10MHz ou plus) et les multicorrélateurs en temps réel. On peut ainsi calculer $R_S(\tau,\phi)$ et $I_S(\tau,\phi)$ et donc $|\chi_S(\tau,\phi)|$.

7 - EXTENSION DE LA NOTION DE SPECTRE INSTANTANE DE PUISSANCE AUX SIGNAUX ALEATOIRES LOCALEMENT STATIONNAIRES

Les signaux aléatoires rencontrés en pratique sont des signaux LOCALEMENT stationnaires. Ces divers types de non stationnarité peuvent se ramener à 3 types fondamentaux simples. Nous indiquerons rapidement ces trois types d'évolution sur la fonction de corrélation :

a) évolution "lente" de la puissance $\Gamma_S(t,0)$ c'est-à-dire de la valeur à l'origine de Γ_S. Ceci est un cas courant d'évolution lente de la puissance calculée sur la durée de stationnarité T_{st}.

b) évolution par dérive lente dans le temps des zéros successifs de $\Gamma_S(\tau)$, cas constaté dans certains types de bruit à bande relativement étroite dont la fréquence centrale évolue lentement : c'est le cas des vibrations induites par un moteur à régime variable, ou encore c'est la dérive en fréquence de certains types "d'ondes encéphalographiques"

c) évolution complexe résultant de la superposition des deux effets a) et b).

Ces types d'évolution se rencontrent dans le cas particulier de la réverbération acoustique sous-marine.

Dans une majorité de processus LOCALEMENT stationnaires l'évolution des propriétés spectrales avec t est lente devant la durée ou intervalle de stationnarité T_{st}. Dans ce cas on peut écrire : $\Gamma_S(t,\tau) = f(t) \cdot \Gamma_S(\tau)$

dans le cas d'une évolution lente de la puissance du processus ou encore :

$$\Gamma_S(t,\tau) = \Gamma_S(\tau, \tau_o) \quad \text{avec } \tau_o = \tau_o(t)$$

évolution lente d'un des paramètres caractéristiques de $\Gamma_S(\tau)$. Dans tous ces cas l'estimation du spectre "évolutif" $\gamma_S(\nu,t)$ se fera par estimation sur la durée T_{st} à l'aide d'un multi-corrélateur en temps réel du type proposé par J. MAX et D. BERTHIER.

Il existe un autre type de non stationnarité assez courant en télécommunications. C'est le milieu à paramètres lentement variables devant la bande B utile du milieu. Cette évolution est en général aléatoire et supposée stationnaire [12]. Dans ce cas on définit des spectres "évolutifs" à l'aide des paramètres suivants :

$H(t,\tau)$ réponse percussionnelle du dispositif

où $H(t,\tau)$ est le résultat observé à la date t dû à une impulsion de DIRAC apparue à l'instant $t-\tau$.

$h(t,\nu) \overset{\nu \; \tau}{\rightleftharpoons} H(t,\tau)$ gain complexe relié à $H(t,\tau)$ par transformation de Fourier.

Ces paramètres servent à définir les grandeurs suivantes :

$$\begin{cases} \Gamma_H(\tau, t_1) = E\left\{H(\tau,t)\; H(\tau, t+t_1)\right\} \\ \gamma_H(f,t_1) \underset{(\tau)}{\rightleftharpoons} \Gamma_H(\tau,t_1) \\ \sigma(\tau,\nu) \underset{(t_1)}{\rightleftharpoons} \Gamma_H(\tau,t_1) \end{cases}$$

La quantité $\sigma(\tau,\nu)$ est appelée fonction de diffusion et est directement reliée à la fonction d'Ambiguité du signal $X(t)$ se propageant dans le milieu. En effet si on traite le signal reçu par son filtre adapté on obtient une grandeur dépendant de :

$$\sigma(\tau,\nu) \quad \text{et de } \chi_X(\tau, \phi)$$

8 - SYNTHESE D'UN SIGNAL S(t) A PARTIR DE $\rho(t,\nu)$ OU DE $\chi_S(\tau,\phi)$

Si l'on adopte le point de vue de D. GABOR la représentation (t,ν) à l'aide des fonctions élémentaires $\psi_{km}(t,\nu)$ devrait permettre la synthèse d'un signal quelconque si l'on connaît les coefficients C_{km} qui le caractérise. Il suffit alors d'envoyer des impulsions pondérées par les coefficients C_{km} dans les différents filtres de rang m aux instants k convenables pour reconstituer S(t). Chaque filtre a pour réponse impulsionnelle $\psi(t - t_k, \nu - \nu_m)$.

Malheureusement la mesure sonographique ne fournit que $|C_{km}|^2$ ou $|C_{km}|$. Toute information de phase a été perdue. Ceci explique aussi certaines difficultés de synthèse des signaux acoustiques telle que la parole.

Les autres méthodes de mesure fournissent $\rho(t,\nu)$ ou $\chi_S(\tau,\phi)$ sa transformée de Fourier. Diverses méthodes de synthèse sont alors possibles. R.M. LERNER indique que l'on peut soit développer S(t) sur une base orthogonalisée de "signaux élémentaires" $v_{mn}(t,\nu)$, soit utiliser, pour des signaux de durée θ des fonctions $v_k(t)$ qui seront des translatées fréquentielles de $v_0(t)$:

$$v_k(t) = v_o(t)\; e^{2\pi i \frac{kt}{\theta}} \rightleftharpoons \mathcal{V}_0\left(\nu + \frac{k}{\theta}\right)$$

Les coefficients du développement de S(t) sur les fonctions $v_k(t)$ sont :

$$S(t) = \sum_k \beta_k v_k(t) \quad , \quad v_k(t) = e^{-at} e^{2\pi i \frac{mt}{\theta}}$$

$$\beta_k = \frac{1}{\theta} \int_0^{\theta} e^{at} S(t) e^{- i 2\pi \frac{kt}{\theta}} dt$$

Cette méthode a permis de calculer les v_k et de déterminer le signal S(t) correspondant à une fonction d'Ambiguité ou un spectre instantané donné. Pratiquement on choisit les fonctions $v_k(t)$ de manière à les réaliser commodément à l'aide d'une grille de filtres sélectifs.

9 - CONCLUSION

Il apparaît donc que toute expression "spectre instantané" contient une CONTRADICTION interne, ou les définitions des variables TEMPS et FREQUENCE. Cependant si cette définition a un caractère relatif qui oblige à parler de SPECTRE EVOLUTIF, elle n'en a pas moins un sens physique indéniable mais relatif. Cet aspect est nettement mis en relief par les nombreuses définitions de cette grandeur. Il faudra donc lier la définition à un type d'ANALYSE bien déterminé.

Nous devons remarquer que toutes ces définitions proposées ont toutes pour Transformée de Fourier à 2 dimensions une grandeur égale ou dépendant étroitement de la fonction d'Ambiguité de WOODWARD. L'intérêt des 2 notions de spectre instantané et de fonction d'Ambiguité est très net lors de la description des signaux à modulation forte. De plus les définitions proposées permettent des expressions simples pour déduire le spectre ou la fonction d'Ambiguité de la filtrée d'un signal, à partir de ces deux grandeurs calculées pour ce signal.

Enfin l'emploi du formalisme de DIRAC a permis à G. BONNET et G. GARAMPON une formulation commode du spectre instantané. Le parallèle formel que l'on peut établir entre la théorie des communications et la mécanique quantique révèle en fait les profondes différences d'interprétation de formules identiques par ces deux disciplines [9].

FIGURE 1

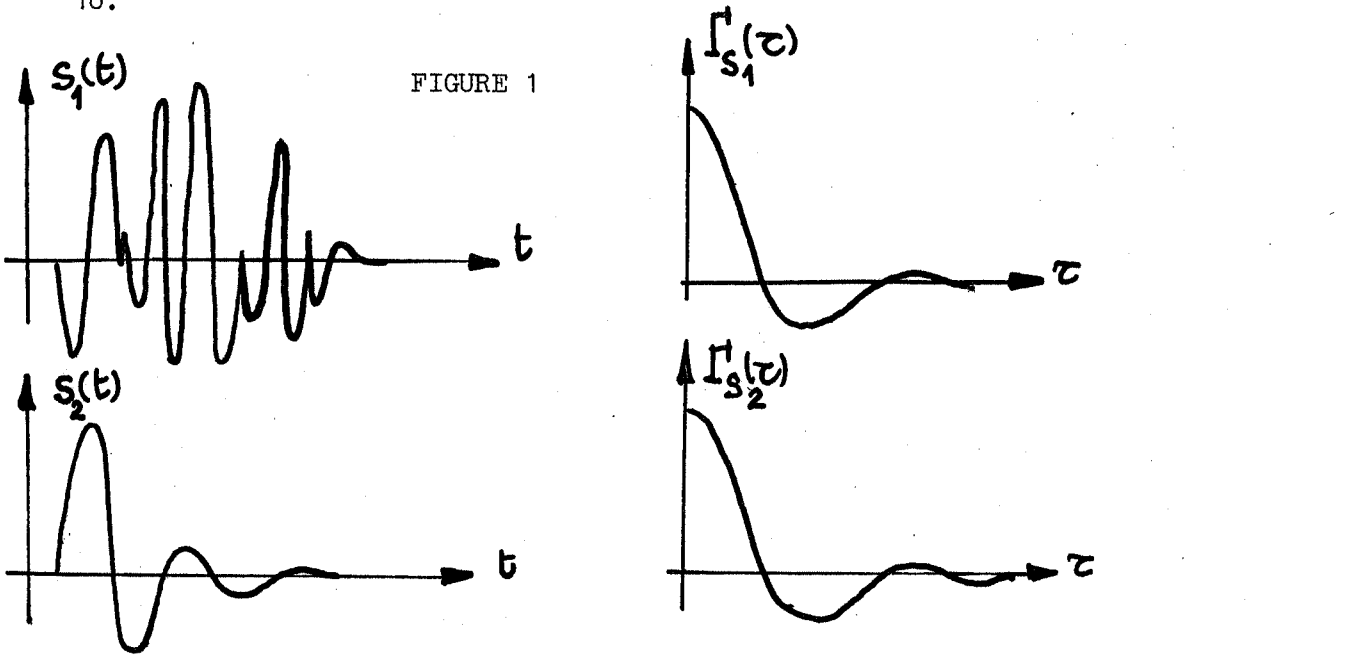


FIGURE 2

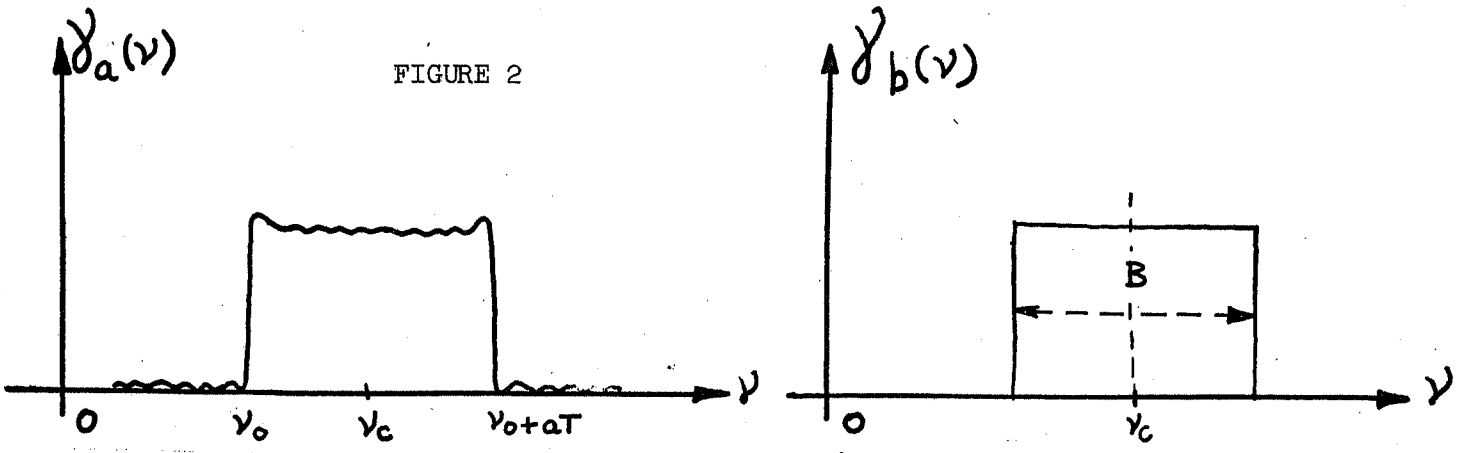
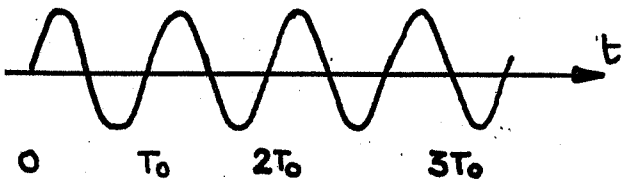
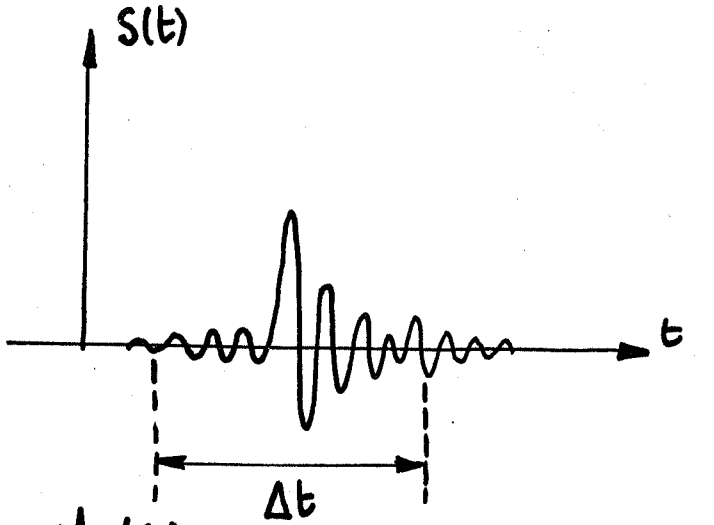


FIGURE 3





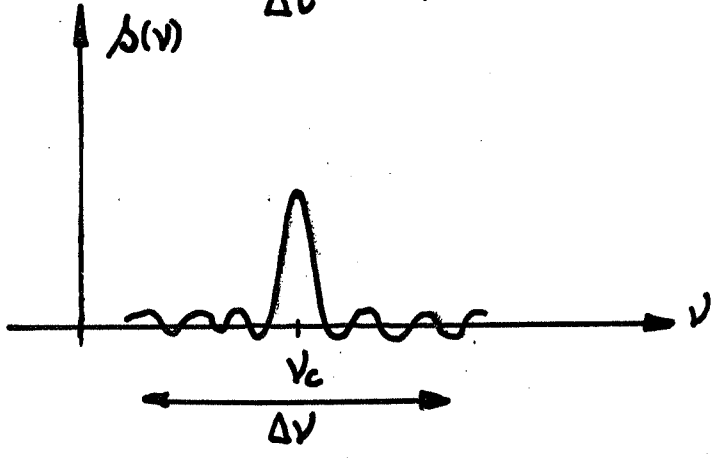


FIGURE 4

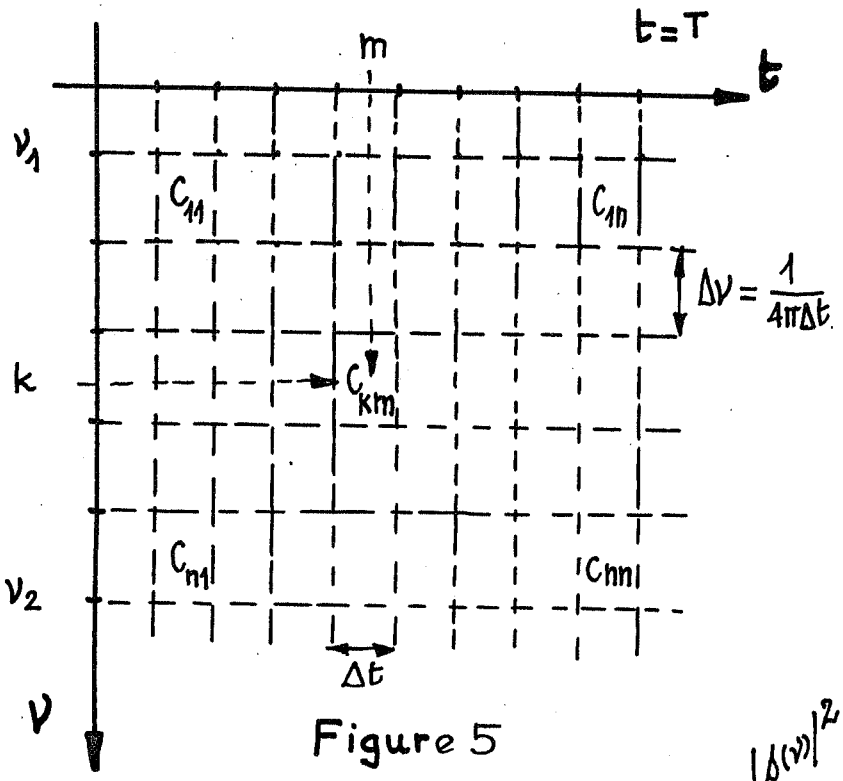


Figure 5

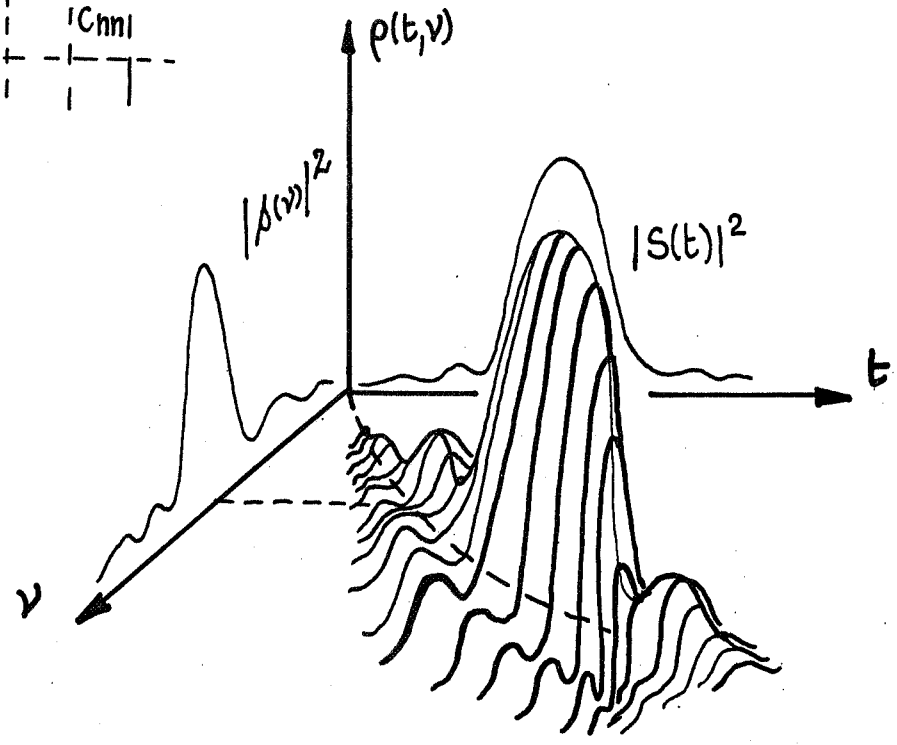


Figure 6

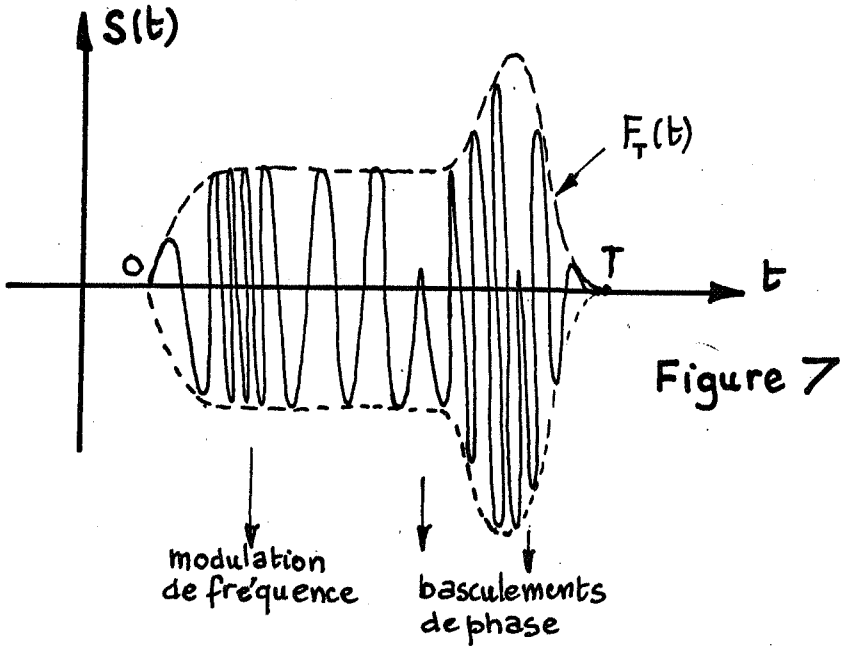


Figure 7

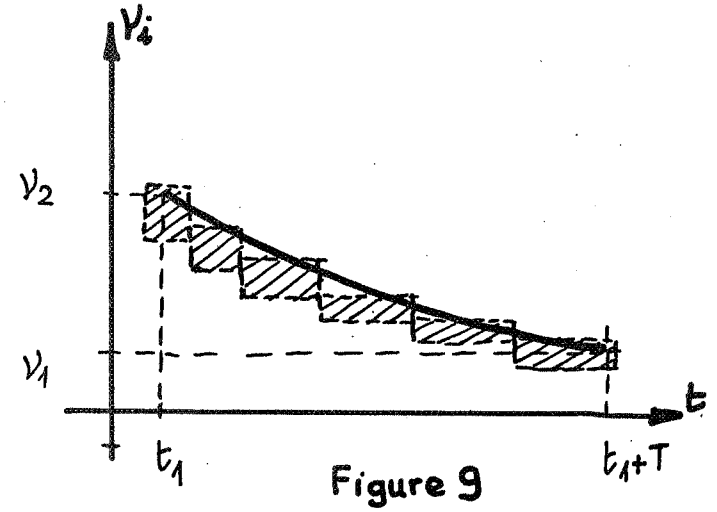


Figure 9

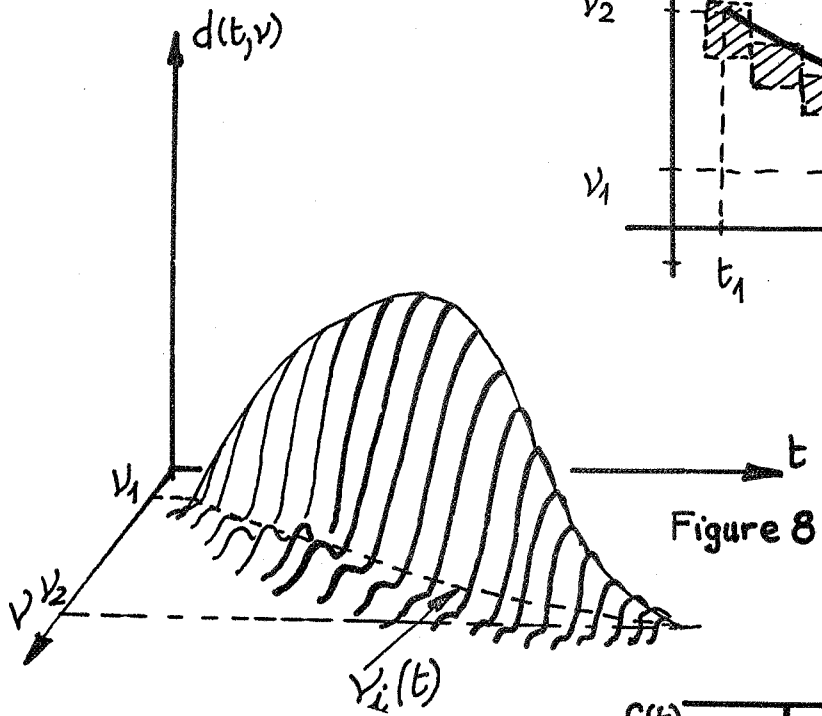


Figure 8

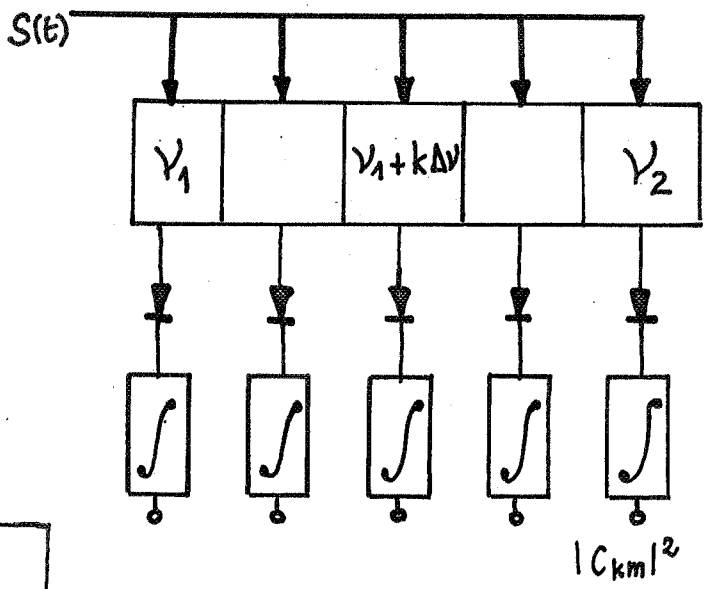


Figure 10

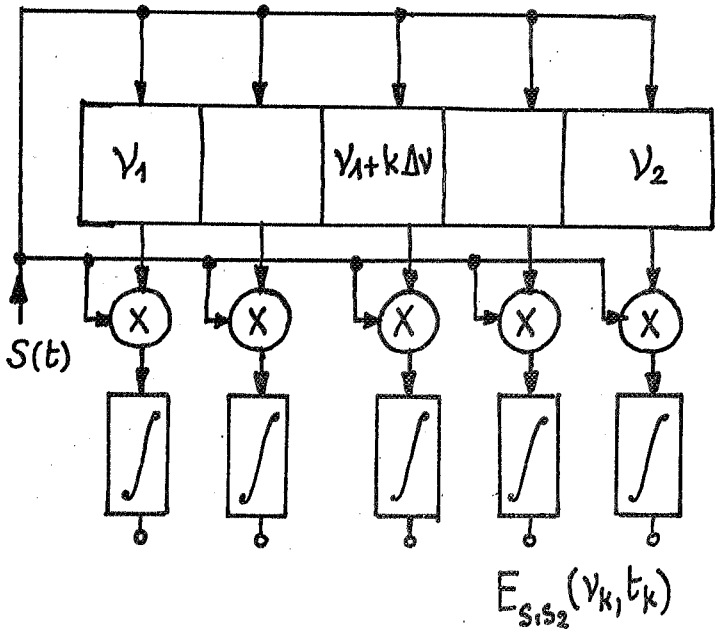


Figure 11

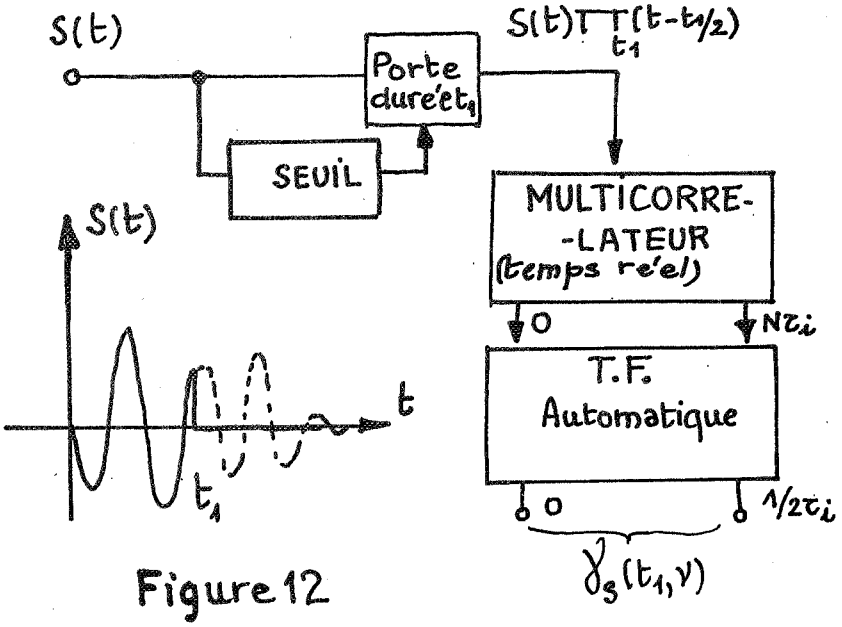


Figure 12

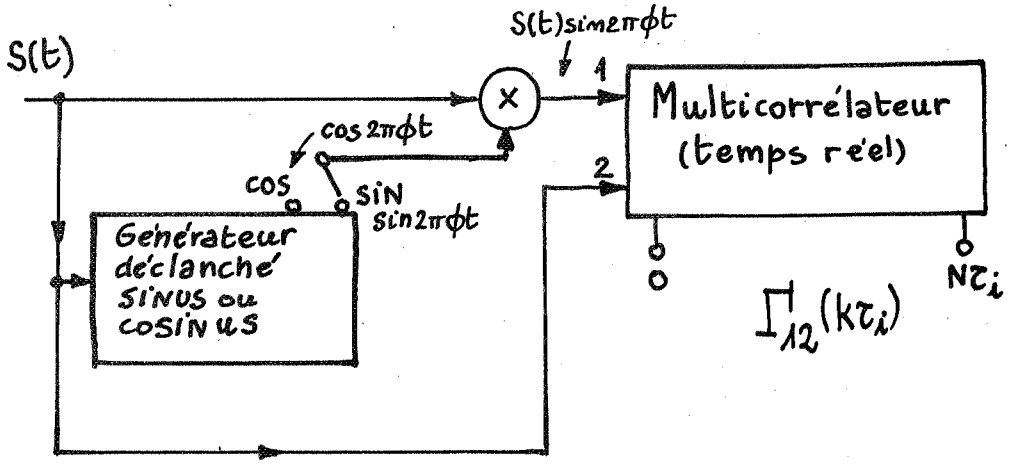
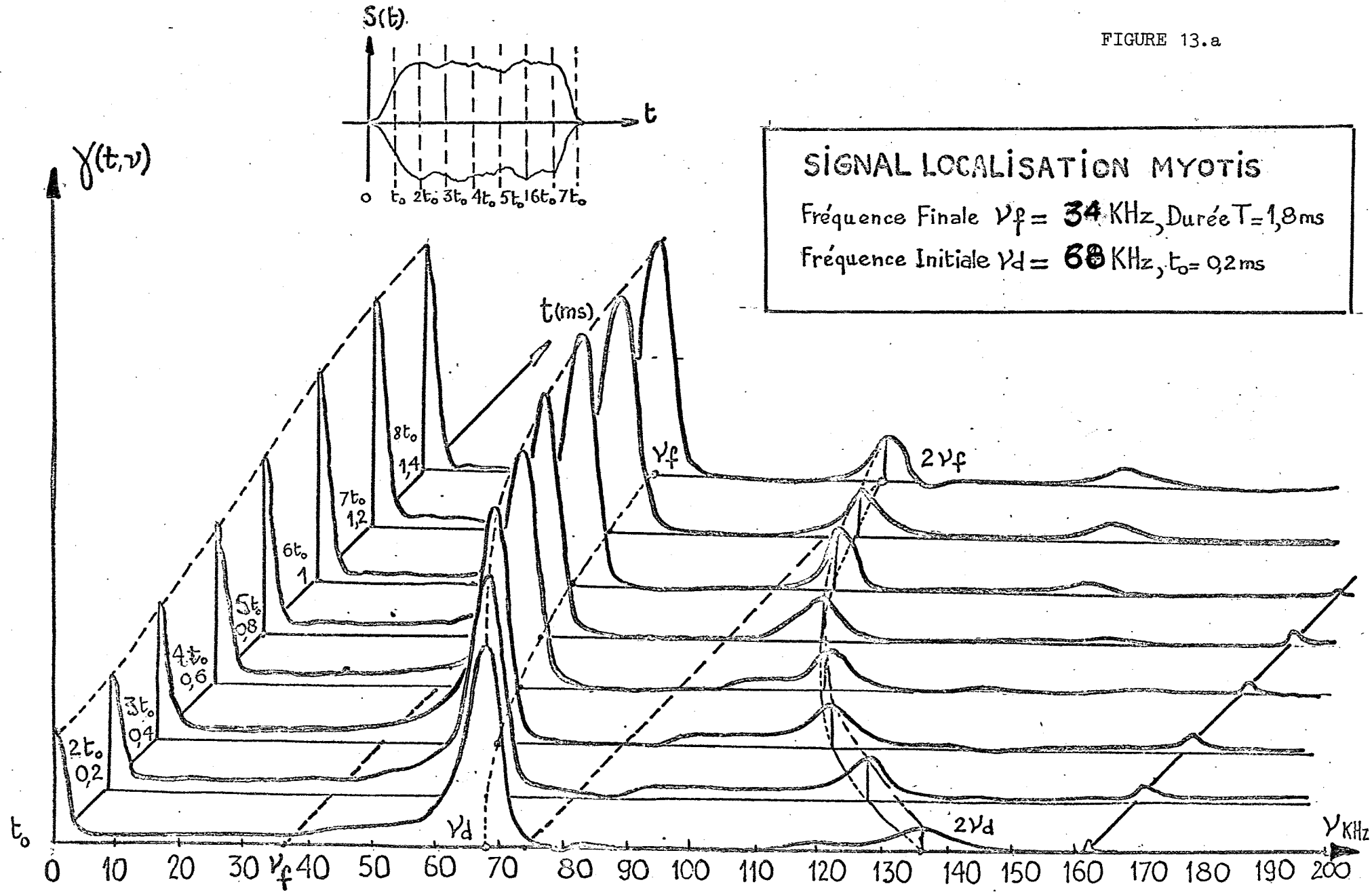


Figure 13

FIGURE 13.a


S(t)
t
0 t_o $2t_o$ $3t_o$ $4t_o$ $5t_o$ $6t_o$ $7t_o$
$\gamma(t,\nu)$
SIGNAL LOCALISATION MYOTIS
Fréquence Finale ν_f = 34 KHz, Durée T = 1,8 ms
Fréquence Initiale ν_d = 68 KHz, t_o = 0,2 ms
t(ms)
$8t_o$ 1,4
$7t_o$ 1,2
$6t_o$ 1
$5t_o$ 0,8
$4t_o$ 0,6
$3t_o$ 0,4
$2t_o$ 0,2
t_o
ν_f
$2\nu_f$
ν_d
$2\nu_d$
ν KHz
0 10 20 30 ν_f 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200

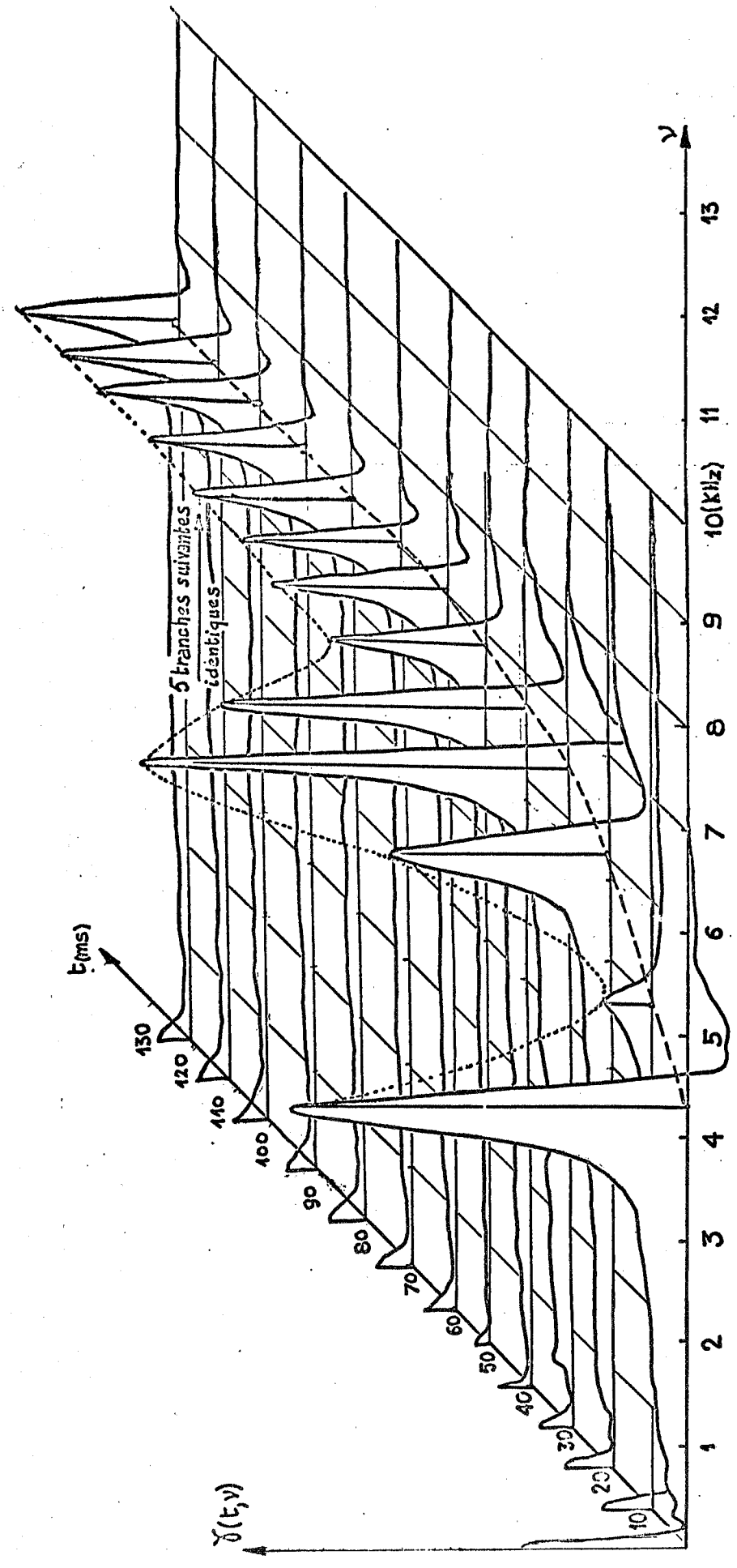


FIGURE 13.b - Cri synthétique du Troglodyte - t_o = 10 ms ; durée 0,13 s ; (A)

BIBLIOGRAPHIE

[1] D. GABOR
Jour. Inst. Elect. Eng. part III (G.B) p 429 1946

[2] G. BONNET, G. GARAMPON
Congrès Nice ASM 1967

[3] R.M. LERNER
Lectur. Communic. Syst. Th. Badhdady Ed. 1961

[4] J. VILLE
Câbles et Transmissions n° 1 1948

[5] P.M. WOODWARD
Probability and Inform. Th. Pergamon P. 1953

[6] A. BLANC-LAPIERRE et B. PICINBONO
Publ. Sc. Univ. Alger t1 1955

[7] A. VAN DER POL
Proc. Rad. Eng. vol. 18 n° 7 Juillet 1930

[8] A. MESSIAH
Mécanique Quantique

[9] G. GARAMPON
Thèse Ing. Doct. Univ. Grenoble Mai 1970

[10] A.W. RIHACZECK
IEEE Trans. Inf. Th. 1968

[11] R.M. FANO
J.A.S.A. 22,5 p. 546 1950

[12] C.H. PAGE
J. Appl. Phys. USA 23, 1 p. 103 1952

[13] T. KALIZEWSKI
The Radio Electron. Eng. Oct. p. 199 1967

D. ABENSOUR souligne l'équivalence temps-fréquence pour la représentation d'un signal sans perte d'informations.

B. ESCUDIÉ - Mais la transformée de *FOURIER* n'est pas causale. Elle est effectuée de $-\infty$ à $+\infty$. La transformée de *LAPLACE* lui est supérieure sur ce plan.

P. DEMAN - Les représentations temps-fréquence sont indispensables pour une raison pragmatique : il est nécessaire de trouver une représentation dont la perte d'informations en traitement soit minimale, compte tenu de la nécessité de respecter le principe de causalité finie dans la transmission : la décision de choix du message le plus vraisemblable doit être prise après un temps fini après la décision de son émission. Le retard en reconnaissance vocale ne doit pas dépasser quelques dizaines à quelques centaines de millisecondes, ce qui oblige à utiliser une représentation temps-fréquence à l'exclusion d'une transformation de *FOURIER* sur un signal "long".

Toutes les représentations temps-fréquence aboutissent à une représentation du signal par une somme de fonctions non orthogonales (au moins au voisinage). Ceci conduit à utiliser un quadrillage de densité supérieure au théorème d'échantillonnage pour limiter la perte d'information. Ceci est fait dans l'analyseur spectral instantané utilisé à la T.H.-C.S.F. :
$\Delta t = 4$ ms, $Q = 3$, 50 filtres pour la bande 300-3400 Hz, soit $\Delta t \,.\, 4\, F \simeq 0{,}04$.

R.A. GUEDJ - Etes-vous au courant des travaux du Professeur *CULLER* à l'Université de Santa Barbara ? Il semble qu'il utilise également une méthode de superposition de fonctions gaussiennes. Y a-t-il une analogie avec ce que vous faites ?

B. ESCUDIÉ - Oui ; il y a beaucoup de points communs. De même, des travaux ont été effectués en vue d'une généralisation du développement de *GABOR* à partir des fonctions gaussiennes.

R. CARRÉ - Dans l'analyse de la parole, on recherche essentiellement à déterminer la fonction de transfert du conduit vocal, à déterminer les fréquences de résonance (fréquences des formants). Le sonagraphe est souvent peu précis pour mesurer ces fréquences. Par ailleurs, dans le cas des voix de femmes, le nombre des composantes du spectre est réduit et ne permet pas d'imaginer l'enveloppe du spectre correspondant à la courbe de réponse du conduit vocal. Sachant que l'on ne connaît pas le signal de source, est-il possible, avec ces nouvelles méthodes d'analyse, de déterminer dans tous les cas et avec précision les fréquences des formants ?

B. ESCUDIÉ - On peut peut-être essayer, à partir de la fonction d'ambiguïté et après déconvolution, de déterminer l'évolution des paramètres. C'est de l'identification de processus à paramètres variables par détermination de la fonction de dispersion.

COMMUNICATIONS PRESENTEES AU COURS DE L'APRES-MIDI DU 2 AVRIL 1971

METHODES D'EVALUATION DE L'INTELLIGIBILITE ET DE LA QUALITE DE LA PAROLE

J. C. R I S S E T

C.N.E.T. - Institut d'Electronique fondamentale - ORSAY

Introduction : Intelligibilité et qualité, paramètres subjectifs

Cet article tente de donner un aperçu d'ensemble des méthodes développées pour évaluer l'intelligibilité et la qualité de la parole. Il ne s'agit que d'un survol : le lecteur devra pour plus de détail se reporter à la bibliographie. Mais cette présentation rapide l'aidera peut-être à situer, parmi l'ensemble des travaux effectués dans le domaine de l'évaluation de la parole, les thèmes abordés aux Journées d'Etudes d'Aix.

La parole dont on cherche à évaluer l'intelligibilité où la qualité est issue le plus souvent de canaux techniques de transmission : parole humaine transmise à distance ou parole artificielle produite par des appareils de synthèse. Mais, d'où qu'elle émane, la parole à évaluer est pres que toujours destinée à des auditeurs, et, qu'on le veuille ou non, le juge doit être en dernier ressort un sujet humain : la sanction finale de l'intelligibilité ou de la qualité ne peut résulter que d'épreuves et de tests subjectifs.

Les tests subjectifs sont d'administration délicate et souvent incommode, et l'on redoute -à juste titre- qu'ils ne donnent des résultats incertains et fluctuants. Aussi l'on tente souvent de se débarasser de ce maillon "ondoyant et divers", de ce bout de chaîne gênant, l'auditeur : on cherche à remplacer les tests subjectifs par des critères objectifs, portant sur des paramètres physiques susceptibles de mesure. Mais il faut bien voir qu'un critère objectif d'intelligibilité ou de qualité de parole ne se justifie que dans la mesure où sa validité a été mise à l'épreuve de tests subjectifs.

C'est dire qu'on ne peut éluder le problème de mettre au point des tests subjectifs fiables et significatifs-problème difficile : il faut se garder des artefacts psychologiques et choisir des épreuves adaptées à chaque circonstance. Or il existe dans ce domaine un savoir-faire souvent méconnu, car trop de spécialistes des techniques et des sciences "exactes" pensent que des données subjectives ne peuvent donner prise à une analyse scientifique. Dès maintenant, certains protocoles expérimentaux, certaines méthodes de traitement des données tirent le meilleur parti de jugements subjectifs.

1. EVALUATION DE L'INTELLIGIBILITE DE LA PAROLE

On évalue souvent l'intelligibilité en mesurant le pourcentage d'éléments phonétiques correctement perçus : on obtient ainsi la netteté (articulation score) pour les éléments phonétiques considérés (Egan 1944, 1948, Beranek 1949, Chavasse 1951, Lehmann 1962). (Les éléments présentés doivent être phonétiquement équilibrés, c'est-à-dire qu'il faut tenir compte de leur distribution statistique dans la langue considérée). Les pourcentages obtenus dépendent de la nature des éléments phonétiques considérés : un même signal de parole donnera lieu à un pourcentage plus élevé pour des mots que pour des syllabes. On a pu établir, dans le cas d'une parole humaine entachée de bruit, la relation entre les pourcentages obtenus pour différents éléments phonétiques : cela permet, à partir d'une mesure de netteté pour un élément phonétique bien déterminé, d'évaluer assez bien l'intelligibilité du signal de parole étudié.(Cette technique est cependant sujette à caution pour évaluer une parole artificielle où des éléments comme les syllables peuvent être intelligibles individuellement mais mal coordonnés, ce qui peut rendre les phrases difficiles à comprendre). M. CARTIER donne ci-joint plus de détails sur l'utilisation d'éléments phonétiques monosyllabiques appelés logatomes.

L'administration de tests de netteté est lente et fastidieuse : il faut disposer de nombreuses données pour que les tests soient statistiquement significatifs. Aussi a-t-on mis au point d'autres méthodes pour évaluer l'intelligibilité.

La méthode de l'indice d'articulation (articulation index) (French & Steinberg, 1947 ; Beranek, 1949; 1954 ; Kryter, 1962 ; Lehmann 1962, 1968) conduit au calcul d'un indice qui permet de prévoir approximativement la netteté (paramètre subjectif) à partir de caractéristiques physiques objectives du canal de transmission de la parole. (Ainsi à un indice de 0,5 correspond une netteté de 70 % pour les mots et de 98 % pour les phrases). Cette méthode est valide si la distorsion apportée par le canal consiste surtout en distorsion linéaire et addition de bruit gaussien. Pour la mettre en oeuvre on a divisé le spectre audible en bandes de fréquences (bandes d'octave), puis on a évalué la contribution de chaque bande à l'intelligibilité, ce qui permet de pondérer chaque bande de fréquence et le niveau de bruit qu'elle contient. Cette méthode ne peut s'appliquer à l'évaluation de l'intelligibilité de la parole de synthèse.

La méthode du test de rime (rhyme test) (Fairbanks 1958, Voiers 1965) teste plus spécialement la confusion entre consonnes par les sujets. Elle est rapide à apprendre et à administrer, les résultats peuvent être traités par ordinateur (Voiers 1965, House et al., 1963, Preusse, 1967). M. Voiers a beaucoup

étudié l'optimisation de cette méthode : il restreint les choix possibles à des couples de consonnes ne différant que par un seul trait distinctif phonologique (Jakobson, Fant & Halle) et les versions qu'il a mises au point permettent non seulement d'évaluer l'intelligibilité, mais de donner des diagnostics sur les défauts du système. MM. Rossi et Peckels ont adapté ce test au français en y apportant des contributions nouvelles : ils ont déjà des résultats portant sur des échantillons de parole d'origines diverses.

II. TESTS DE QUALITE

L'intelligibilité n'est pas le seul critère de qualité de la parole : on admet cependant que c'est généralement le plus important, et qu'il ne devient significatif de parler d'autres composantes de la qualité qu'à partir du moment où l'intelligibilité est correcte.

Dans certaines circonstances, des éléments divers peuvent rentrer en ligne de compte pour l'appréciation de la qualité de la parole, en plus de l'intelligibilité et de l'agrément, du caractère "naturel" en "esthétique" de la parole. Ainsi le problème de l'identification correcte du locuteur, qui n'a pas de sens pour une parole de synthèse, peut être d'une importance vitale, par exemple dans les télécommunications militaires (cf Pollack et al., 1954; Voiers 1966).

Les travaux d'un groupe d'experts de l'I.E.E.E. (I.E.E.E. Recommended Practice for Speech Quality measurements, 1969) n'ont pu aboutir à la recommandation d'une méthode unique pour évaluer la qualité de la parole. L'article consignant les conclusions du groupe distingue deux types de méthodes : les méthodes "utilitaires" et les méthodes "analytiques". Cette distinction est commode, bien que sans doute provisoire. Les méthodes utilitaires visent à caractériser la qualité par un résultat scalaire, par une note. Les méthodes analytiques cherchent à distinguer les différentes composantes, les attributs psychologiques qui concourent à l'appréciation de la qualité : il s'agit d'analyses multidimensionnelles de jugements subjectifs.

II.1 Méthodes utilitaires

On trouvera dans l'article mentionné ci-dessus (I.E.E.E. Recommended Practice...,1969) la description détaillée de trois méthodes d'administration de tests de qualité.

Dans la méthode d'isopréférence (Munson et Karlin 1962 ; Rothauser 1968), on demande aux sujets de comparer le signal test à divers signaux déduits d'un signal de référnnce en le dégradant par addition de bruit (ou dans certaines variantes, par multiplication par des signaux aléatoires convenablement choisis : cf Rothauser, Schroeder 1968). La comparaison est faite par paires : on demande aux sujets de noter lequel des signaux de chaque paire ils préfèrent. Si la méthode est mise en oeuvre de façon complète pour un système de transmission présentant un intérêt particulier, elle conduit à des courbes d'isopréférence, sortes de courbes de niveau sur lesquelles se situent les échantillons équivalents du point de vue des jugements de préférence auxquels ils donnent lieu, et à un indice de préférence, différent pour les différentes courbes d'isopréférence : cela peut conduire, pour un système donné, à une méthode objective d'évaluation de la qualité.

La méthode des préférences relatives (Hecker et Williams, 1966) utilise comme signaux de référence non plus des signaux d'une même famille dérivés d'un même signal par différents degrés de distorsion, mais des signaux de parole ayant subi des types différents de distorsion et pouvant néanmoins être rangés par ordre de qualité. Pour cette méthode, le sujet ne peut savoir, dans chaque paire présentée, lequel est le signal à étudier et lequel est le signal de référence : ce protocole évite certains artefacts auxquels peut donner lieu la méthode d'isopréférence (pour laquelle le sujet est influencé, dans ses jugements de préférence par paires, par la comparaison qu'il fait involontairement entre les dégradations du signal de référence dans différentes paires).

Enfin, dans la méthode des jugements par catégorie, on demande au sujet d'attribuer au système à tester l'un de plusieurs qualificatifs (par exemple inacceptable, mauvais, passable, bon, excellent). Cette méthode suppose que les sujets subissent préalablement une phase d'apprentissage, au cours de laquelle on leur présente un certain nombre d'échantillons de référence correspondant aux différents qualificatifs. Richards et Swaffield (1959) ont étudié des méthodes de jugement par catégorie dans le cas où l'auditeur ne se contente pas d'écouter, comme à la radio, mais participe activement à la communication, comme dans une conversation téléphonique. S'il faut évaluer un système de transmission de parole à deux voies, il faut tenir compte des possibilités des interlocuteurs à s'adapter au système et à en tirer le meilleur parti, spécialement quand les conditions d'intelligibilité sont mauvaises. Par exemple le locuteur n° 2 pourra demander au locuteur n° 1 de répéter certaines parties du message et le locuteur n° 1 pourra modifier sa façon de parler en fonction de ce qu'il en infère sur la difficulté qu'a le locuteur n° 2 à le comprendre à travers le système. L'effort qu'exige la communication devient alors un paramètre subjectif significatif du système.

II.2 Méthodes analytiques

Les méthodes analytiques visent à extraire les composantes subjectives, les critères psychologiques importants qui interviennent dans les jugements. Elles partent donc des résultats de tests subjectifs, et elles leur appliquent des techniques mathématiques de traitement de données pour mettre en évidence les dimensions importantes qui sous-tendent ces résultats.

Le problème se pose dans d'autres domaines que l'évaluation de la parole. (Si l'on veut par exemple noter significativement l'intelligence, il faut bien y distinguer plusieurs composantes). Des techniques mathématiques ont été développées qui s'appliquent au traitement de données de nature très variée (matrices de confusion entre différents stimuli , classement de ces stimulis deux à deux ou évaluation par une note des différences entre les stimuli). Le but poursuivi est de distinguer les composantes qui contribuent à ces jugements et d'évaluer leurs poids respectifs dans les jugements ; on peut aussi chercher à corréler ces composantes psychologiques et la composition physique des stimuli : dans certains cas on arrive à isoler un paramètre ou une propriété physique qui correspond étroitement à une composante psychologique du jugement. Un intérêt toujours croissant s'attache à ces méthodes, d'autant plus que les ordinateurs permettent la mise en oeuvre de techniques puissantes et de plus en plus raffinées de traitement des données (cf Shepard ; Kruskal).

Le traitement mathématique - l'analyse multidimensionnelle des données - vise à une représentation spatiale traduisant les relations géométriques entre ces données, de façon que par exemple les distances entre points représentatifs de stimuli soient d'autant plus petites que ces stimuli soient jugés plus semblables. On peut alors extraire des indications de cette représentation géométrique. Ainsi l'analyse factorielle (Harman, 1960), détermine des facteurs en fonction desquels on peut exprimer de manière condensée les informations : les nombreuses données sont représentées dans un espace à un grand nombre de dimensions où elles forment un nuage de points. On ajuste à ce nuage un sous-espace en déterminant successivement les principales directions d'allongement du nuage, auxquelles on associe un facteur. Cela revient à déterminer les axes et les moments d'inertie du nuage et du point de vue du calcul à diagonaliser des matrices.

Dans la méthode de la différentielle sémantique (Osgood et al, 1957) on demande aux sujets de situer les échantillons présentés sur des échelles dont les points extrêmes sont définis par des adjectifs de sens contraires (par exemple grave-aigu, doux-intense, naturel-artificiel...) : pour chaque échantillon on obtient ainsi un nombre pour chaque échelle (convenant par exemple que 0 correspond au pôle grave et 1 au pôle aigu). Le traitement mathématique vise d'abord à réduire la redondance entre les échelles pour lesquelles les résultats sont fortement corrélés ; puis l'on procède à une analyse factorielle des résultats pour les échelles restantes. Si l'on a présenté N échantillons à classer sur n échelles, les résultats fournissent un nuage de Nxn points dans un espace à n dimensions ; l'analyse factorielle réduit cette dimensionnalité. Si un axe obtenu par analyse factorielle fait un angle faible avec une échelle proposée, c'est que la dimension correspondante est assez bien décrite par la dichotomie correspondant à l'échelle). Mc Gee (1964) a appliqué la méthode de la différentielle sémantique à l'étude de la qualité de la parole filtrée. Ses résultats indiquent que deux dimensions contribuent principalement à la qualité : l'une correspondant à l'intelligibilité et l'autre interprétée comme le naturel. Voiers (1964) a appliqué la différentielle sémantique pour déterminer les dimensions principales de la variation de la qualité d'un locuteur à un autre.

Mais cette méthode encourt un reproche : ses résultats peuvent être sensibles aux échelles proposées, aux présupposés de l'expérimentateur (voire des sujets). On tend maintenant, dans la mesure du possible, à utiliser des protocoles expérimentaux dans lesquels on demande aux sujets les jugements les plus simples possibles : des résultats des tests on ne pourra peut-être pas tirer autant de renseignements, mais les renseignements obtenus auront plus de valeur intrinsèque. De ce point de vue, le processus le plus fiable consiste à partir de données de confusion et à en faire l'analyse factorielle : cela n'est pas toujours applicable. On peut aussi appliquer la méthode d'analyse des proximités, qui se prête à des jugements de préférences par paires donnant simplement une série d'inégalités. Si l'on dispose d'un nombre suffisant de données, on peut à partir de ces seules données de comparaison construire une représentation géométrique assez précise de l'espace des jugements, impliquant même une métrique déduite de seules données ordinales (Shepard 1962, 1966, Benzecri 1964, 1965). Une méthode plus facile à appliquer consiste à demander aux sujets de noter de 0 à 10

la similarité entre échantillons de chaque paire et à faire une analyse factorielle des données obtenues. C'est par cette dernière méthode que Mc Gree (1965) a pu montrer que l'appréciation de la parole filtrée se faisait différemment suivant l'absence ou la présence des premières harmoniques.

Barbara McDermott (1969) a utilisé des techniques semblables pour analyser des jugements de similarité et de préférence relatifs à une parole humaine transmise par différents canaux techniques. A partir de jugements de similarité, elle a pu extraire des dimensions semblant correspondre aux attributs suivants : clarté, distinction entre distorsion du signal et bruit de fond, intensité. Les jugements de préférence ont conduit essentiellement aux mêmes dimensions, mais à part la clarté, les différents sujets n'étaient pas en accord sur l'importance de ces dimensions du point de vue de leur préférence pour tel ou tel échantillon.

Voiers et ses collaborateurs (1965) ont étudié la qualité de la parole transmise par différents vocoders, en tenant compte de la variance des jugements pour tenter d'établir une échelle de qualité. Ils ont également pu isoler des dimensions bien définies dans l'espace des jugements, mais dont les correspondants physiques ne sont pas évidents.

On peut dire que les méthodes analytiques ont déjà apporté des renseignements utiles, mais qu'elles promettent bien davantage encore. Lespremiers résultats indiquent clairement que plusieurs dimensions distinctes sous-tendent les jugements de qualité. Les méthodes analytiques se prêtent à des diagnostics, elles aident à déterminer en quoi la qualité de telle parole étudiée est insuffisante. D'autre part la mise au point d'une méthode utilitaire sans reproche, fournissant un paramètre global d'évaluation de la qualité, devrait passer par des études analytiques, permettant d'évaluer les facteurs du jugement et leur importance respective; le poids de chaque facteur peut varier suivant les types de communication auxquels on destine la parole à étudier. Il est symptomatique que le comité spécialisé sur la parole de l'Acoustical Society of America ait organisé une session sur l'analyse multidimensionnelle au cours de la réunion d'avril 1971 de cette société.

CONCLUSIONS

L'évaluation de l'intelligibilité et de la qualité est une tâche délicate : il est souvent difficile de séparer, dans la perception de la parole, l'appréciation des signaux acoustiques des inférences faites à partir du contexte. Cependant, une batterie de techniques est déjà disponible, d'autres sont en cours d'étude. Il s'agit d'un domaine qui revêt une importance toute particulière, à un moment où l'on assiste au développement du traitement de la parole et particulièrement à l'apparition de différents procédés de synthèse automatique : la comparaison des résultats pose avec acuité le problème de l'évaluation de la parole.

- BIBLIOGRAPHIE -

AFNOR
Evaluation des bruits du point de vue de leur influence sur l'intelligibilité de la parole - Nov. 1970.
Commission de l'Acoustique CF/TC 43 - doc. 155.

A.D. ARKHIPOVA & M.A. SAPOZHKOV
The quality of vocoder speech. Soviet Physics - Acoustics, vol. 16 n° 3 (1971) pp. 292-298.

D. CAXTER & B. KEISER
Speech channel evaluation divorced from talker listener influence.
IEEE Trans. Commun. Technol 14 (1966) 101.

J.P. BENZECRI
Analyse factorielle des proximités. Public. de l'Institut de Statist. de l'Université de Paris.
Part I 13 (1964) ; Part. II 14 (1965)

J.P. BENZECRI
Leçons sur l'analyse des données multidimensionnelles. Laboratoire de Statistique Mathématique, Faculté des Sciences de Paris.

L.L. BERANEK
Acoustic measurements -J. Wiley 1949.

L.L. BERANEK
Acoustics-Mc Graw Hill 1954.

F. CHAVASSE
L'application des moyens d'analyse de la qualité des transmissions téléphoniques.
In La cybernétique (de Broglie) Ed. de la Revue d'Optique Paris 1951.

R.W. DONALSON & R.J. DOUVILLE
Analysis, subjective evaluation, Optimization, and comparison of the perfomance capabilities of PCM, A M, AM and PM Voice communication systems.
IEEE Trans. on Comm. Technology COM 17 (1969) 421-431.

R.L. EBEL
Estimation of reliability of ratings - Psychometrika.
16 (1951) 407-424.

J.P. EGAN
Articulation testing methods. Laryngoscope 58 (1948) 955-991.

J.P. EGAN
Articulation testing methods. OSRD Report n° 3802 Nov. 1944 (US Dpt. of Commerce Report PB 22 848).

H. EISLER
Measurement of perceived acoustic quality of sound - reproducing systems by means of factor analysis. JASA 39 (1966) 484-492.

C. FAIRBANIKS
Test of phonemic differentiation : the rhyme test J.A.S.A. 30 (1958) 596-600.

H. FLETCHER & J.C. STEINBERG
Articulation testing methods. Bell Syst. Techn. J. 8 (1929) 806-854.

N.R. FRENCH & J.C. STEINBERG
Factors governing the intelligibility of speech sounds.
JASA 19 (1947) 90-119.

H.H. HARMAN
Modern factor analysis. The Univ. of Chicago Press, Chicago 1960.

M.H.L. HECKER C.E. WILLIAMS
Choice of reference conditions for speech preference tests. J. Acoustic. Soc. Am. 39 (1966) 946-952.

A.S. HOUSE, C.E. WILLIAMS, M.H.L. HECKER & K.D. KRYTER
Psychoacoustic speech tests : a modified rhyme test. Decision Sciences Lab. Electronic Syst. Div. Air Force Systems Command, Rept. ESD TDR 63-403 June 1963.

A.S. HOUSE
Articulation testing methods : consonantal differentiation with a closed reponse set. JASA 37 158-166 (1965).

I.E.E.E.
Recommended practice for speech quality measurements. IEE Trans. on Audio Electroac. AU 17 n° 3 (Sept. 1969) 225-246.

R. JAKOBSON, C.G. M. FANT & M. HALLE
Preliminaries te speech analysis. The distinctive features and their correlates. M.I.T. Acoustics Lab. Techn. Rep. n° 13 (1953).

K.D. KRYTER
Method for the calculation and use of the articulation index.
JASA 34 (1962) 1689-1697 (cf. erratum : JASA 36 (1964) 1393).

K.D. KRYTER & E.C. WHITMAN
Some comparisons between rhyme and P.B. word intelligibility test. JASA 37 (1965) 1146.

R. LEHMANN
Etude psychophysique de l'intelligibilité du langage (Thèse) Editions de la Revue d'Optique théorique et instrumentale, Paris 1962.

R. LEHMANN
Effet de masque et indice de netteté - Revue d'acoustique n° 3 (1968) 167-170.

H. LEVITT & L.R. RABINER
Use of a sequential strategy in intelligibility testing. J. Acoust. Soc. Am. 42 (1967) 609-612.

B.J. Mc DERMOTT
Multidimensional analyses of circuit quality jedments. JASA 45 (1969) 774-781.

V.E. Mc GEE
Semantic components of the quality of processed speech. J. Speech Hearing Res. 7 (1964) 310-323.

V.E. Mc GEE
Determining perceptional spaces for the quality of Filtered speech. J. speech Hearing Res. 8 (1965) 23-38.

W.A. MUNSON & J.E. KARLIN
Isopreference method for evaluating speech transmission circuits. J. Acoust. Soc. Am. 34 (1962) 762-774.

L.H. NAKATANI
Measuring the ease of comprehending speech. Proc. 7th Int. Congr. Acoust., Budapest, Hungary, 1971.

C.E. OSGOOD G.C. SUGI & P.H. TANENBAUM
The measurement of meaning - Psychol. Bul., 53 (1957) 470-538. Illinois Press, Ubana, Ill. 1957.

I. POLLACK, J.M. PICKETT & W.H. SUMBY
On the identification of speakers by voice. JASA 26 (1954) 403-406.

I. POLLACK, H. RUBINSTEIN & L. DECKER
Intelligibility of known and unknown message sets. J. Acoust. Soc. Am. 31 (1959) 273-279.

J.W. PREUSSE
Semi automatic speech intelligibility measurements. IEEE. Trans. Audio & Electroac. AU 15 (Déc. 1967) pp. 188 - 191.

D.L. RICHARDS & J. SWAFFIELD
Assessment of speech communication links. Proc. I.E.E. (London) 106 (1959) pp. 77-92

E.H. ROTHAUSER & G.E. URBANEK
New reference signal for speech quality measurements. J. Acoust. Soc. Am. 38 (1965) 940 (abstract).

E.H. ROTHAUSER, G.E. URBANEK, & W.P. PACHL
A comparison of preference measurement methods. J. Acoust. Soc. Am. 49 (1971) 1297-1308.

M.R. SCHROEDER
Reference signal for signal quality studies JASA 44 (1968) 1735.

R.N. SHEPARD
Analysis of proximities I. Psychometrika 27 125-140 (1962)
II. Psychometrika 27 219-246 (1962)

R.N. SHEPARD
Metric structures in ordinal data. J. Mathemat. Psychology 3 (1966) 287-315.

R.N. SHEPARD & J.D. CAROOLL
Parametric representation of non linear data structures. In multivariate Analysis Academic Press, NY 1966, P. 561.

W.H. TEDFORD, J.R. & T.V. FRAZIER
Further study of the Isopreference method of circuit evaluation. JASA 39 (1966) 645-649.

W.S. TORGERSON
Theory of methods of scaling. J. Wiley and Sons, Inc. NY 1958.

W.D. VOIERS, M.F. COHEN & J. MICHUNAS
Evaluation of speech processing devices. I : intelligibility, quality speaker recognisability, AFCRL 65-826, 1965.

J. SWAFFIELD & D.L. RICHARDS
Rating of speech links and performance of telephone networks. Proc. I.E.E. (London) 106 (1959) pp. 65-76.

W.D. VOIERS
Perceptual bases of speaker identity. JASA 36 (1964) 1065.

W.D. VOIERS
Perfomance evaluation of speech processing devices. II : the role of individual differences. Rept. AFCRL 66-24, Air force Cambridge Reseach Lab. Office of Aerospace Research, Badford, Mass. 1966.

W.D. VOIERS, M. COHEN & J. MICKUNAS
Effects of stimulus presentation rate upon diagnostic intelligibility scores. Sperry Rand Res. Center, Final Repst. Contract AF 19 (628) 4195, July 31 1965.

C.E. WILLIAMS & M.H.L. HECKER. Relation between intelligibility scores for four test methods and three types of speech distorsion. J. Acoust. Soc. Am. 44.

L E S L O G A T O M E S

M. C A R T I E R
C. N. E. T. - LANNION

La qualité essentielle d'une parole de synthèse, comme d'une transmission, est d'être intelligible. La notion d'intelligibilité est délicate à saisir ; elle échappe à la mesure. Mais la netteté qui, par définition, se mesure, est directement liée à la façon dont l'auditeur comprend le message. On se contente donc de mesures de netteté (pourcentage des éléments phonétiques correctement perçus).

La méthode la plus naturelle consiste à faire écouter des phrases ou des mots. Cependant, le vocabulaire et l'intelligence des sujets faussent les résultats (prévisibilité : cf. E. LEIPP [1]). C'est pourquoi STEVENSON a proposé, en 1930, l'utilisation de mots dénués de sens : les logatomes. Les logatomes sont des *associations inséparables de sons prononçables d'une seule émission de voix* (CHAVASSE [2]) du type CVC. Les consonnes sont parfois remplacées par des groupes de consonnes.

DEFINITION DES LOGATOMES

La définition des sons qui composent les logatomes est le premier élément à discuter : CHAVASSE, entre autres, recommande, sur le plan national, l'emploi de sons de la langue du pays et le respect de leur distribution statistique. E. LEIPP s'est élevé contre les sonorités étranges des logatomes habituellement employés en France. D'autre part, nous sommes membres du G.A.L.F., et des logatomes français surprendraient moins nos oreilles. En effet, les listes de logatomes issues des recommandations du C.C.I.T.T. [3] découlent de l'espéranto pour avoir une valeur internationale. Il est logique que le C.N.E.T. utilise ces listes puisque le réseau téléphonique est ouvert aux communications internationales. Pour juger la parole de synthèse française, les mots aléatoires de Mle CASTELLENGO [4] rendent sans doute mieux compte de l'intelligibilité et demandent peut-être moins d'entraînement de la part des auditeurs.

DEROULEMENT D'UN ESSAI

Les mesures sont effectuées par une équipe de cinq opérateurs. Chaque opérateur prend la parole tour à tour, les quatre autres notent ce qu'ils entendent. Un essai comporte, en général, 500 logatomes émis. Le résultat est présenté sous deux formes : netteté aux sons, netteté aux logatomes.

Précisons dans quelles conditions se font ces essais : le locuteur règle le niveau en prononçant devant un équipement normalisé la phrase des téléphonistes *PARIS BORDEAUX LE MANS ST LO LEON LOUDUN* . Les conditions standard d'écoute donnent au cours de cette phrase un niveau de 60 dBA environ sur l'oreille. Les logatomes sont ensuite émis non pas au même niveau, ce qui serait impossible à définir, mais *avec le même effort* . Ceci correspond à 72 dBA sur les voyelles des logatomes.

Le signal arrive sur un écouteur, les auditeurs sont soumis sur l'autre oreille à un bruit ambiant de 60 dBA. La façon d'appliquer la parole et le bruit n'est pas indifférente ; le niveau doit être bien défini car on est proche de la zone critique de rapport signal à bruit (cf. LEHMANN [5]).

Le procédé de réglage du niveau par une phrase se retrouve chez HOUSE [6] qui a étudié l'importance du locuteur et mis en évidence l'influence des rapports d'énergie entre voyelles et consonnes. On considère que la participation de cinq locuteurs à un essai fournit une valeur moyenne satisfaisante.

DISCUSSION

Le principal inconvénient de la méthode des logatomes est d'imposer une équipe entraînée. De plus, un essai prend beaucoup de temps et le dépouillement en est difficile.

Par contre, ce test exigeant est suffisamment sensible pour donner des résultats significatifs, même sur des systèmes de bonne qualité où la netteté aux mots serait quasi-parfaite. L'expérience acquise et la stabilité des équipes permettent de situer les performances du système étudié par rapport à des équipements connus ; (par exemple, une conversation téléphonique devient difficile lorsque la netteté aux logatomes n'atteint pas 70 % et une communication est possible à partir de 50 % à 55 %).

Une question intéressante est de connaître la correspondance entre la netteté aux logatomes et d'autres performances. Il a été démontré que les divers tests n'offrent pas la même sensibilité aux différents types de distorsion. L'indice d'articulation part de l'hypothèse contraire. KRYTER [7] admet une correspondance donnant :

Syllabes (logatomes)	50 %	70	90
Mots isolés	80	90	100
Phrases	97	99	100

Ces pourcentages ont été établis pour étudier l'influence du bruit sur l'intelligibilité. Les correspondances ne sont pas nécessairement valables pour la parole de synthèse.

Nous avons trouvé les résultats suivants pour trois vocoders à canaux différents : la netteté aux sons et la netteté aux logatomes se correspondent comme les mots isolés et syllabes de KLYTER. Par contre, les résultats aux mots dissyllabiques sont meilleurs que pour les sons.

CONCLUSION

Les questions posées restent nombreuses ; nous soulignerons les points suivants :

- Si deux systèmes de transmission ou le traitement de la parole sont de qualités comparables en présence de parole "haute fidélité", une dégradation de la parole introduite peut se traduire par des baisses de netteté différentes.
- Des signaux d'origines différentes peuvent être plus ou moins affectés par un même bruit superposé à l'écoute.
- Comme l'a fait remarquer J.C. RISSET, des règles d'assemblage défectueuses peuvent conduire à dégrader l'intelligibilité d'une parole de synthèse qui donnerait de bons résultats sur des éléments courts.

Les logatomes constituent un outil bien connu et au point qui permet, malgré tout, de situer les performances d'un système et d'en suivre les progrès d'ensemble. Son emploi est assez lourd et délicat, mais la présence au CNET d'une équipe spécialisée permet de remédier en partie à cet inconvénient.

LISTE DE LOGATOMES :

MED	NIST	DRUC	RIC	VAB
VLEK	GLOT	ĝEFT	PRUG	DOĈ
TIS	SLAR	RUG	FURS	TRUM
KLUFT	ĈIL	LOR	SUM	GAC
ŝTOF	PSAŜ	ŝURS	SAN	BLIM
SPIT	BOP	POĈ	JAST	ZED
VENG	STRUV	ŝRAN	TENG	COP
STAS	FLAB	BRIF	KREŜ	BIN
NOR	GRIL	ŝLEZ	MES	LAV
ŝOL	KEK	PLUV	FROT	HIŜ

N.B. - Toutes les lettres forment des sons individuels et se prononcent (les consonnes finales, en particulier), comme si elles étaient suivies d'un e muet. La prononciation est celle des sons du français, sauf s qui est toujours doux (comme dans son), ŝ qui se prononce ch (char), c qui se prononce ts (tsar), ĉ qui se prononce tch (match), u et e qui se prononcent respectivement ou et é, ĝ équivaut à dj et j à l'y du mot « yeux »).

BIBLIOGRAPHIE

E. LEIPP
Le problème de l'intelligibilité de la parole
Revue d'Acoustique n° 12 1970 p. 343

P. CHAVASSE
De la notion d'intelligibilité
L'audioprothésiste français n° 3 fev. 1962

Comité Consultatif International Télégraphique et Téléphonique
Qualité de transmission téléphonique - Livre rouge, tome V
publié par l'Union Internationale des Télécommunications 1962

Mle CASTELLENGO, J.S. LIENARD
Bulletin du Groupe d'Acoustique musicale n° 53 janv. 1971

R. LEHMANN
Etude psychophysique de l'intelligibilité du langage (thèse)
Ed. Revue d'Optique théorique et instrumentale 1962

HOUSE, WILLIAMS, HECKER, KRYTER
Articulation testing methods : consonantal differentiation with a closed response set
J.A.S.A, 37 n° 1 janv. 1965

KRYTER
Methods for the calculation and use of the articulation index
J.A.S.A. 34 1962 p. 1689

DIAGNOSTIC APPROACH TO THE EVALUATION OF SPEECH INTELLIGIBILITY

William D. VOIERS
TRACOR, Inc. - AUSTIN, Texas, U.S.A.

INTRODUCTION

It is a matter of common observation that speech communication--more specifically, a listener's apprehension of a speaker's linguistic intent--is essentially a dual process. One aspect of this process involves discrimination by the listener of various acoustical manifestations of the speaker's intent with respect to the identity of an elementary speech unit. The other aspect involves inferences based on contextual or extra-stimulus information, i.e., on information from sources extrinsic to the immediate acoustical manifestation of the speaker's linguistic intent. Among the more important sources of contextual information which may contribute to the process of speech communication are the listener's knowledge of the structure of the language involved;[1, 2] his knowledge of the circumstances occasioning and the purposes motivating the communication;[3] and his familiarity with dialectal and idiolectal characteristics of the speaker.[4] Depending upon the circumstances, still other sources of contextual information, e.g., his familiarity with intelligibility test material, or with other arbitrary vocabulary constraints, may significantly reduce the listener's uncertainty as to the speaker's intent.[5] Both the discriminative and the inferential aspects of the speech apprehension process are of course subjects of legitimate interest to the student of human communication. It is conceivable, moreover, that some scientific purposes can be served by experimental circumstances in which the contributions of the two aspects are not explicitly controlled. Such control would seem obligatory, however, in experiments or tests conducted to evaluate the intrinsic potential of a transmission channel, speaker, or listener from the standpoint of speech intelligibility. To the extent that a listener's responses in the testing situation are dependent in unknown degree upon contextual information, his performance necessarily provides an imperfect reflection of the characteristics of the channel or other entity under test. Cognizance of this issue is implicit, where not explicit, in the designs of most of the intelligibility tests in use today, but few, if any, provide altogether satisfactory resolution of the issue.

Although control of various contextual influences can be obtained in a number of ways (e.g., by the use of nonsense syllable tests) the objection may be raised that many of these tend to sacrifice the element of realism or face validity. In response to this objection, however, one has only to note the degree to which communications situations vary with respect to the amount and kind of contextual information available to the listener. The investigator who places a premium on realism may reasonably be asked "which reality (i.e., how much contextual information) is to be embodied in the testing situation--that of the school girl telephone conversation or that of the infantry platoon commander under combat conditions?" Ultimately, the issue must be resolved in a somewhat arbitrary manner, but it does not follow that testing procedures which provide the greatest amount of contextual information (e.g., "sentence recognition" and "word recognition" tests) are generally superior on either practical or theoretical grounds. It does follow, however, that where the relative contributions of stimulus factors and contextual factors to test results are subject to uncontrolled variation, both the quantitative and the qualitative implications of test results will be ambiguous.

Perhaps, the most notorious source of uncontrolled contextual information in the intelligibility testing situation is the listener's familiarity with the speech materials used, particularly as such knowledge almost inevitably varies as a function of the listener's history of exposure to the test materials. Various methods of controlling this factor have been devised, but their efficacy remains open to question on several counts. In the case of the venerable PB test of word intelligibility, for example, the prescribed procedure for controlling the effects of familiarity involves an extensive regimen of training, terminated on evidence that the effects of familiarity have reached an assymtotic state.[6] Necessarily, to require such procedures is to constrict rather severely the range of circumstances in which use of the test is practical; but practical constraints do not constitute the only basis for objection.

On one hand, familiarization training serves, in effect, to alter the general level of difficulty of the listener's task and, in turn, to obscure any relationship between the "real world" and the testing situation that might be claimed on the basis of absolute level of difficulty. On the other hand, it may conceivably alter quite drastically the qualitative implications of the listener's responses to the test materials. A crucial consideration in this connection is the fact that the various discriminations required of the listener in the course of recognizing a phoneme or other elementary speech unit are not of intrinsically equal difficulty, as shown by Miller and Nicely, among others.[7] Some of these discriminations are evidently accomplished with virtually perfect reliability, even under conditions of extreme signal impoverishment. Others are made with significantly less than perfect reliability under the best of conditions and may become prohibitively difficult under conditions of signal impoverishment. Given these circumstances, it would be truly remarkable if the facilitative effects of familiarization training were in fact exerted equally on all aspects of the speech discrimination task. It would seem to be a more tenable hypothesis that familiarization training serves to facilitate listener performance primarily in those aspects of the speech discrimination task that are intrinsically most difficult. It is conceivable, therefore, that familiarization training serves, effectively, to desensitize the test to deficiencies with respect to those features of the speech signal which are most crucial to the speech communication process and perhaps most vulnerable to common forms of signal impoverishment.

Finally, there is the intrinsic limitation of the PB test and related tests of "word intelligibility" that qualitative differences in erroneous responses to a given stimulus word are not amenable to unambiguous interpretation. In such tests, the effects of stimulus factors and of interphonemic constraints of the language are all but hopelessly confounded.

Superior control of contextual influences is evidently provided by testing procedures in which stimulus uncertainty is limited to a single phoneme (as in the Fairbanks Rhyme Test[8]), particularly where the listener's response options are explicitly specified (as in the Modified Rhyme Test[9] and the Phonemically Balanced Rhyme Test.[10]) With such procedures the effects of listener experience in the testing situation, and of apperceptive factors in general, appear to be substantially reduced.

However, the arbitrary restriction of a listener's response options may complicate the situation in other respects. This becomes evident when one considers the fact that the discriminations required of a listener in recognizing a speech sound are ultimately determined not by the acoustical characteristics of the stimulus as regarded in isolation but, rather, by the characteristics that distinguish the actual stimulus from what the listener conceives to be the set of possible stimuli in a given situation. Thus, the listener's response to a given speech event is criterial of the discriminability of specific acoustical speech features, depending upon the response options available to him. To constrain his response options arbitrarily is possibly to deny him opportunity to indicate the indiscriminability of one or more acoustical speech features and, conceivably to desensitize the test with respect to one or more crucial acoustical speech features.

It would seem essential, therefore, that some discretion be exercised in selecting the options available to the listener for responding to a given stimulus word or other speech unit. In addition, it would seem desirable that the differences between the correct response and the set of permissible alternatives be minimal in some sense or another and, thus, that the significance of a particular incorrect response be essentially univocal as, for example, in ensemble:

bee pea vee dee me

where each of the permissible erroneous responses differs from the stimulus word, "bee," by a single "distinctive feature." However, this approach may also pose some practical problems. One problem is that individual items cannot be "scrambled" without qualitatively altering the implications of the listener's response. If, for example, "pea" is the stimulus word in the above ensemble, only one of the response options, "bee," differs minimally and unidimensionally from the stimulus word. All remaining options differ by two or more distinctive features and, in general, the implications of a given response will vary qualitatively and quantitatively, depending upon the identity of the stimulus. Alternatively, one might choose to design the test such that only one member of a given ensemble of response options ever serves as the stimulus, but obvious, practical considerations argue rather strongly against this solution to the problem.

In principle, at least, some of the major problems associated with the multiple choice approach might be circumvented by recourse to nonsense syllable tests (for example, a rhyme test of initial consonant apprehensibility in which the set of permissible response options to each stimulus item is the complete set of consonant phonemes) but practical considerations also militate against this approach. For example, the denotation of some phonemes requires the use of unfamiliar symbols or other special instructions which may be condusive to various response biases in untrained listeners. A close examination of Wicklegren's data,[11] for example, reveals trends consistent with this suggestion.

The qualitative or diagnostic evaluation of errors may also pose formidable problems of implementation in multiple choice (more than two) tests, and although these problems are relatively easy to deal with where computer scoring facilities are available, they are not amenable to easy solution under circumstances requiring the use of manual scoring techniques. Thus, while it is possible to conceive of special circumstances in which multiple choice (greater than two) tests could find application,

both practical and theoretical considerations rather severely limit their usefulness and validity, as becomes evident when one contemplates specific details of test design, administration, scoring and interpretation. Many of the limitations of multiple choice procedures in general can be obviated, however, by recourse to the special case of two-choice testing procedures and, in particular, procedures where the response options permitted the listener in each instance differ in some minimal fashion.

The two-choice approach to the evaluation of speech apprehensibility is, among other things, consistent with the conception of speech apprehension as a multidimensional decision process, each dimension of which corresponds to a cluster of highly, if imperfectly, correlated acoustical speech features.[12] And although the various dimensions of this process have yet to find optimal characterization in the articulatory, acoustical and perceptual domains, a sufficiently precise characterization for present purposes is provided by a system of phonemic attributes or "distinctive features" similar to that of Jacobson, Fant and Halle[13] and Miller and Nicely.[14] In particular, such a system provides a basis for the construction of a two-choice test where the correctness of the listener's response to a given item is criterial--depending on the design and purposes of the investigation--of the effective fidelity with which a speaker articulates, a system transmits or the listener, himself, can discriminate the states of a particular <u>cluster</u> of intercorrelated, information-bearing, acoustical features. In addition, therefore, to providing rigorous control of extra-stimulus factors, a two-choice test can provide valuable insights concerning the specific deficiencies of systems or individuals under test. Even though data on listener performance in apprehending the speaker's intent with respect to a given phonemic attribute will not in general be perfectly criterial of the discriminability of a specific acoustical speech feature, they can often serve to delineate, rather sharply, the possible sources of deficiency or malfunction

in the speaker, channel or listener under test. Supplemented with other information, moreover, such data may in fact permit relatively precise evaluation of specific deficiencies of the entity under evaluation.

The Diagnostic Rhyme Test (DRT) was designed on the basis of the foregoing considerations. Accordingly, it is a two-choice test in which each item involves two rhyming words, the initial consonants of which differ by a single elementary consonant attribute. The listener's task is simply to indicate which of the two words has been spoken and, in effect, to indicate that he has or has not apprehended the speaker's intent with respect to a particular phonemic attribute. Among its more important theoretical advantages is the fact that erroneous responses can be unambiguously attributed to characteristics of the speaker, listener or system under test as opposed to characteristics of the context or testing situation. Among its more important practical features are: (1.) economy of administration in that the use of "minimally contrasting" word pairs results in the exclusion of excessively easy and, hence, effectively non-functional items; (2.) minimal requirements with regard to listener selection and training (previous experience with the test materials can serve to facilitate listener performance only with respect to a particular randomization of the test materials); (3.) adaptability to both manual and computer scoring schemes.

Table 1 presents the phonemic taxonomy used as a guide to the design of the DRT, in which the six dimensions: voicing, nasality, sustention, sibilation, graveness, and compactness are explicitly represented. A comprehensive discussion of the phonological and acoustical correlates of the phonemic attributes represented in the taxonomy is not feasible here but relevant papers on the subject can be found in the recent literature.[15, 16, 17]

TABLE 1

CONSONANT TAXONOMY USED IN THE CONSTRUCTION OF THE DRT (FORM III)

	/m/	/n/	/v/	/ð/	/z/	/ʒ/	/ʤ/	/b/	/d/	/g/	/w/	/r/	/l/	/j/	/f/	/θ/	/s/	/ʃ/	/ʧ/	/p/	/t/	/k/	/h/
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Nasality	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sustention	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+
Sibilation	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Graveness	+	-	+	-	-	o	o	+	-	o	+	-	o	o	+	-	-	o	o	+	-	o	o
Compactness	-	-	-	-	-	+	+	-	-	+	-	-	o	+	-	-	-	+	+	-	-	+	+
Vowel-like*	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-

* The DRT does not test for the apprehensibility of the opposition, vowel-like – nonvowel-like. However, test words are chosen so as not to confound this attribute with the six attributes for which discriminability is tested.

In all major respects, the taxonomy presented in Table 1 follows the system of distinctive features formulated by Jacobson, Fant and Halle.[1e] Such minor differences as may be observed stem primarily from attempts to derive a taxonomy most nearly in accord with the facts of phonemic perception. Although it is clear that the question of the optimal phonemic taxonomy is yet to be resolved, the one proposed here represents a somewhat arbitrary but hopefully adequate compromise with regard to a number of theoretical and practical considerations. In any case the design of the DRT provides some amount of latitude in scoring options and may thus prove adaptable to various future developments in relation to the issue of the optimal phonemic taxonomy.

THE DIAGNOSTIC RHYME TEST (DRT)

Structure of the DRT

The Diagnostic Rhyme Test (DRT) is more properly described in terms of a set of principles for item construction and selection than in terms of a specific corpus of test materials. Thus, the corpus of 96 rhyming word pairs shown in Table 2 constitutes only one realization of such principles, but one which also takes into account the results of various experimental investigations involving the use of earlier versions of the DRT. The gross structure of the test is evident in the table, where the items in each block of seven are arranged according as they are designed to test for the apprehensibility of a particular attribute of the initial consonant phoneme. The order is as follows:

1. Voicing
2. Nasality
3. Sustention
4. Sibilation
5. Graveness
6. Compactness
7. Filler item (to be used for research purposes, etc).

The positive state (e.g., grave) of each attribute is represented in the left member of each pair; the negative state (e.g., acute) is represented in the right member of each pair.

TABLE 2

SPEECH MATERIALS USED IN FORM IV OF THE DIAGNOSTIC RHYME TEST

99.*	VEAL-FEEL	43.	BEAN-PEEN	50.	ZOO-SUE	106.	DUNE-TUNE
107.	MEAT-BEAT	51.	NEED-DEED	2.	MOOT-BOOT	58.	NEWS-DUES
59.	VEE-BEE	3.	SHEET-CHEAT	66.	FOO-POOH	10.	SHOES-CHOOSE
67.	ZEE-THEE	11.	CHEEP-KEEP	74.	JUICE-GOOSE	18.	CHEW-COO
19.	WEED-REED	75.	PEAK-TEAK	82.	MOON-NOON	26.	POOL-TOOL
27.	YIELD-WIELD	83.	KEY-TEA	34.	COOP-POOP	90.	YOU-RUE
35.**	-----	91.**	-----	98.**	----	42.**	-----
71.	GIN-CHIN	15.	DINT-TINT	22.	VOLE-FOAL	78.	GOAT-COAT
79.	MITT-BIT	23.	NIP-DIP	30.	MOAN-BONE	86.	NOTE-DOTE
31.	VILL-BILL	87.	THICK-TICK	38.	THOSE-DOZE	94.	THOUGH-DOUGH
95.	JILT-GILT	39.	SING-THING	46.	JOE-GO	102.	SOLE-THOLE
47.	BID-DID	103.	FIN-THIN	110.	BOWL-DOLE	54.	FORE-THOR
55.	HIT-FIT	111.	GILL-DILL	6.	GHOST-BOAST	62.	SHOW-SO
7.**	-----	63.**	-----	70.**	-----	14.**	-----
8.	ZED-SAID	64.	DENSE-TENSE	57.	VAULT-FAULT	1.	DAUNT-TAUNT
72.	MEND-BEND	16.	NECK-DECK	65.	MOSS-BOSS	9.	GNAW-DAW
80.	THEN-DEN	24.	FENCE-PENCE	17.	THONG-TONG	73.	SHAW-CHAW
32.	JEST-GUEST	88.	CHAIR-CARE	81.	JAWS-GAUZE	25.	SAW-THAW
40.	MET-NET	96.	PENT-TENT	33.	FOUGHT-THOUGHT	89.	BONG-DONG
104.	KEG-PEG	48.	YEN-WREN	97.	YAWL-WALL	41.	CAUGHT-TAUGHT
56.**	-----	112.**	-----	49.**	-----	105.**	-----
36.	VAST-FAST	92.	GAFF-CALF	85.	JOCK-CHOCK	29.	BOND-POND
44.	MAD-BAD	100.	NAB-DAB	93.	MOM-BOMB	37.	KNOCK-DOCK
52.	THAN-DAN	108.	SHAD-CHAD	101.	VON-BON	45.	VOX-BOX
4.	JAB-GAB	60.	SANK-THANK	109.	JOT-GOT	53.	CHOP-COP
12.	BANK-DANK	68.	FAD-THAD	61.	WAD-ROD	5.	POT-TOT
76.	GAT-BAT	20.	SHAG-SAG	69.	HOP-FOP	13.	GOT-DOT
84.**	-----	28.**	-----	21.**	-----	77.**	-----

* Numbers to the left of each pair indicate the position of the item in each block of 112 items on the listeners answer sheet.

** Filler items. The manner in which these spaces are filled is at the option of the experimenter. Among other things they may be used for testing experimental items.

The apprehensibility of each attribute is tested in each of eight vowel contexts, two representing each "quadrant" of the vowel articulation diagram. Thus the four upper left blocks of Table 2 involve high, front vowels, whereas those in the four upper right blocks involve high, back vowels. The low, front vowels are represented in the four lower left blocks, while the low, back vowels are represented in the lower right blocks. No central vowels are used in the DRT.

There are two, formally equivalent items (e.g., bean-peen and veal-feel) designed to test for the apprehensibility of each attribute in each vowel context. This redundancy serves, among other things, to facilitate various tests of the reliability or consistency of listener performance over the course of a testing session. Either member of each pair may be chosen as the stimulus word in a given instance without changing the function of the item qualitatively. Choice of stimulus word affects only the polarity of the test provided by the item.

It is perhaps apparent from the table that insufficient latitude exists to permit any degree of selectivity on the basis of frequency of word occurrence in speech or printed matter. However, results such as those of Pollack, Rubinstein and Decker[19] suggest that frequency-of-use influences the perceptibility of complex stimuli primarily, if not only, as it provides a basis for the listener's expectation concerning the occurrence of the stimulus. Where other, more explicit, bases for expectation are available--as they are in the case of the DRT--frequency-of-use may reasonably be expected to have little or no influence on listener response, particularly, where the listener is required, in effect, simply to discriminate a specific aspect of the total stimulus event, rather than to "recognize" the stimulus.

It may also be noted by reference to Tables 1 and 2 that there are some minor exceptions to the rule of "unidimensional difference" between members of each word pair. This

results from the fact that all compact items are here classified indifferently with respect to graveness. Thus, while the phonemes comprising the pairs /k-p/, /g-b/, /k-t/, /g-d/, etc. differ primarily with respect to compactness, they might thus be considered to differ secondarily in terms of graveness in that the first member of each pair has a neutral or indeterminate status with respect to the latter attribute while the second member of each pair has a positive or negative status. However, various data on phonemic confusability suggest that Table 1 tends to conform most nearly with the facts of phonemic perception. Some experimental justification is also provided by results to the effect that the apprehensibility of compactness, as measured by such items, is quite differently affected by various forms of signal impoverishment than is graveness.

In recognition of experimental evidence that the acoustical correlates of the state of a given attribute may not be equally apprehensible in every instance of its manifestation, nor equally vulnerable to all forms of signal impoverishment, a number of constraints were imposed in assembling the corpus shown in Table 2.

For example, half the items designed to test for the apprehensibility of sustention lie in the voiced plane, half in the unvoiced, which symmetry is preserved within each quadrant of the vowel space. Half the items designed to test for the apprehensibility of nasality lie in the grave plane, half in the acute. Half the items designed to test for apprehensibility of compactness involve the back-front opposition while half involve the back-middle opposition. Still other symmetries may be observed on close examination of Table 2. With minor exceptions, therefore, the two halves of the vowel space, partitioned horizontally or vertically, involve identical phoneme pairs for testing the apprehensibility of each attribute.

Scoring the Diagnostic Rhyme Test

It is perhaps evident from the foregoing discussion that DRT response data can be scored in a diversity of ways, depending upon the interests of the investigator. Generally, however, greatest interest will attach to the six major "diagnostic" scores, each constituting an indicant of the gross apprehensibility of the speaker's intent with respect to a given attribute. It is possible, in addition, to fractionate each of the major diagnostic scores into various components (e.g., to obtain separate scores for the apprehensibility of sustention in the voiced and unvoiced planes) in accordance with the structure of the test as described above.

Separate scores for the apprehensibility of each state of each attribute are likely to be of interest, in that some experimental variables may affect the apprehensibility of the two states of some attributes in an asymmetrical manner. The resulting discrepancy between listener scores for the two states of an attribute is termed, bias. It is measured simply as the difference between the percent (adjusted for chance) of the time listeners correctly apprehend the positive state (e.g., voiced) of an attribute and the percent of the time they correctly apprehend the negative state (e.g., unvoiced).

Adjustment for Contextual Constraints

The format of the DRT serves to fix, quite rigidly, the amount of stimulus information which the listener must extract from each test word. A consequence of this constraint, however, is that chance alone may suffice to result in correct responses to fifty percent of the items. It is appropriate, therefore, that some adjustment be made for the effects of chance or guessing. This is accomplished by means of the familiar formula:

$$S = \frac{100(R-W)}{(T)}$$

where S is the estimated true percent-correct responses, R is the observed number of correct responses, W is the observed number of incorrect responses and $T = R + W$ is the total number of items involved in deriving a particular score. This correction is applied to all Diagnostic Rhyme Test scores.

SOME RESULTS

Time does not permit a comprehensive treatment, here, of the issue of the validity of the DRT in various applications, but one major issue may be touched upon. This is the issue of the potential value of diagnostic score patterns.

Clearly there would be little point to the DRT if it failed to yield qualitatively different results for different types of speech signal impoverishment. The accumulated results of some seven years of experience with the DRT, in its present form and in earlier forms, has more than adequately demonstrated its validity in this respect. Different types of speech impoverishment yield different patterns of diagnostic scores, and these patterns tend strongly to conform with predictions that would be made on the basis of general principles of acoustic phonetics.

Figures 1, 2, and 3 present DRT results for some representative forms of speech impoverishment. Although some time might be spent in discussing various aspects of these results, their primary purpose is simply to illustrate the diversity of patterns typical for individual types of speech impoverishment. Each of these cases will be treated in detail by forthcoming publications.

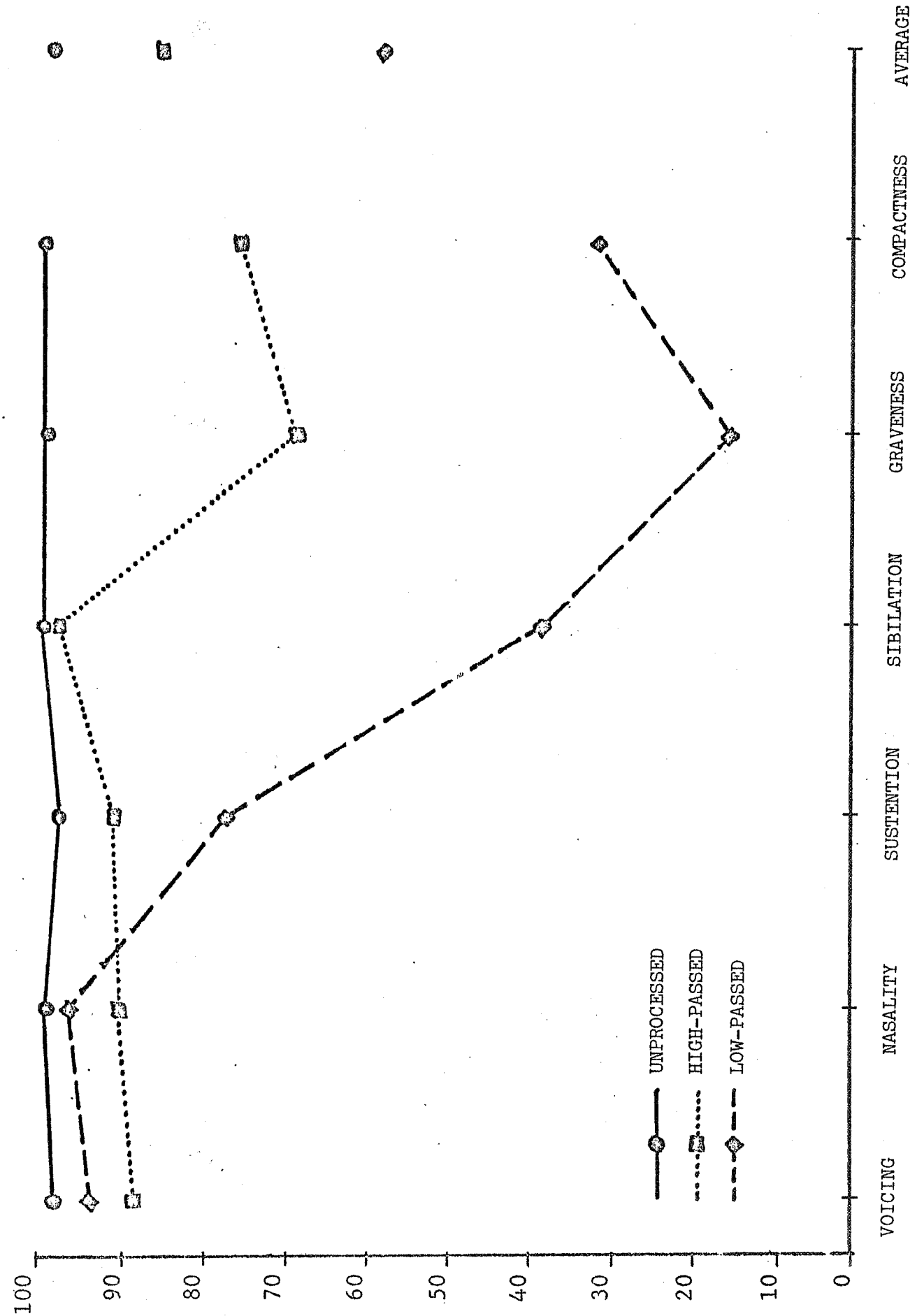


FIGURE 1. DIAGNOSTIC SCORES FOR THREE CONDITIONS OF FREQUENCY DISTORTION
(AVERAGES FOR 12 SPEAKERS)

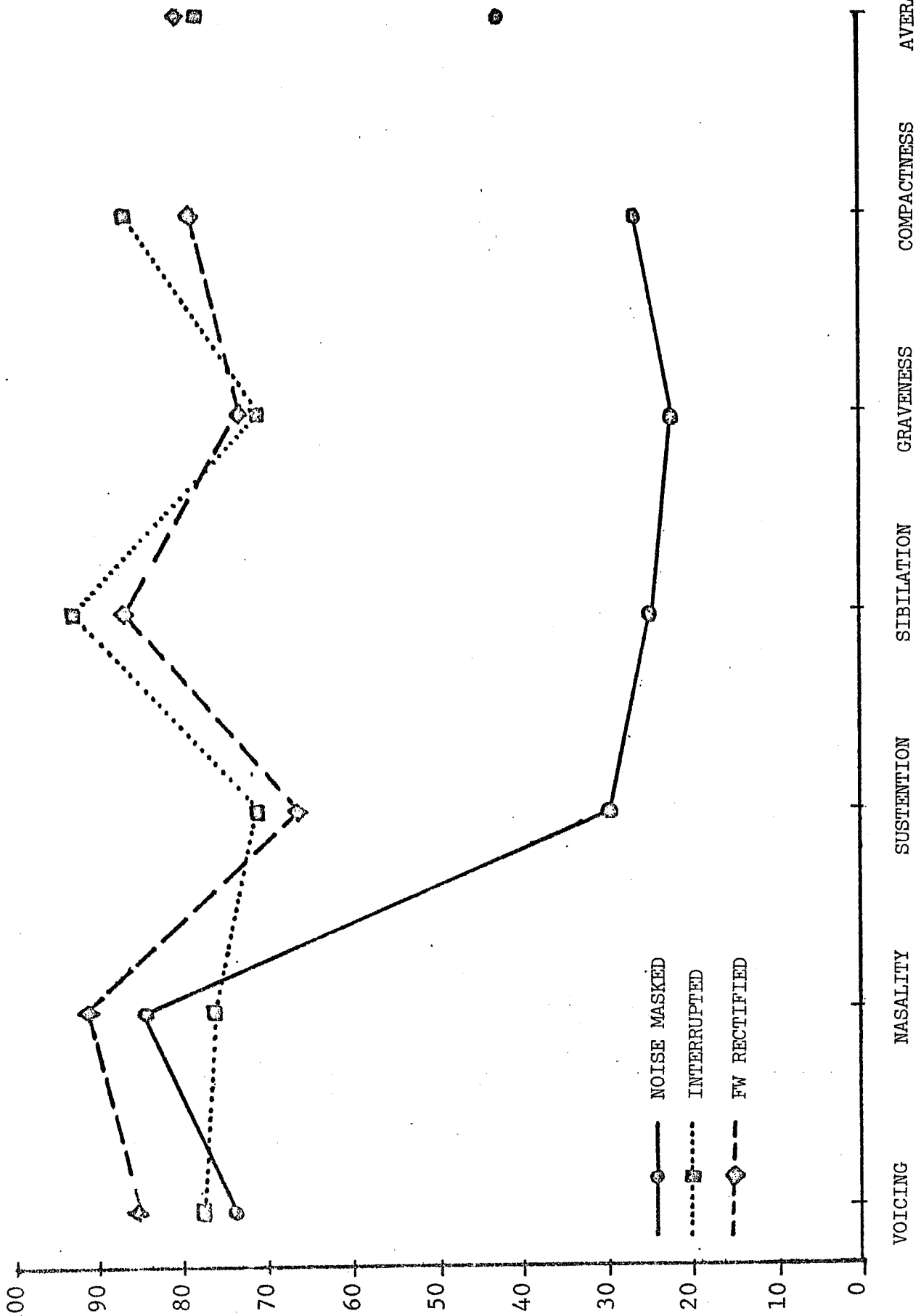


FIGURE 2. DIAGNOSTIC SCORES FOR VARIOUS FORMS OF SIGNAL IMPOVERISHMENT
(AVERAGES FOR 12 SPEAKERS)

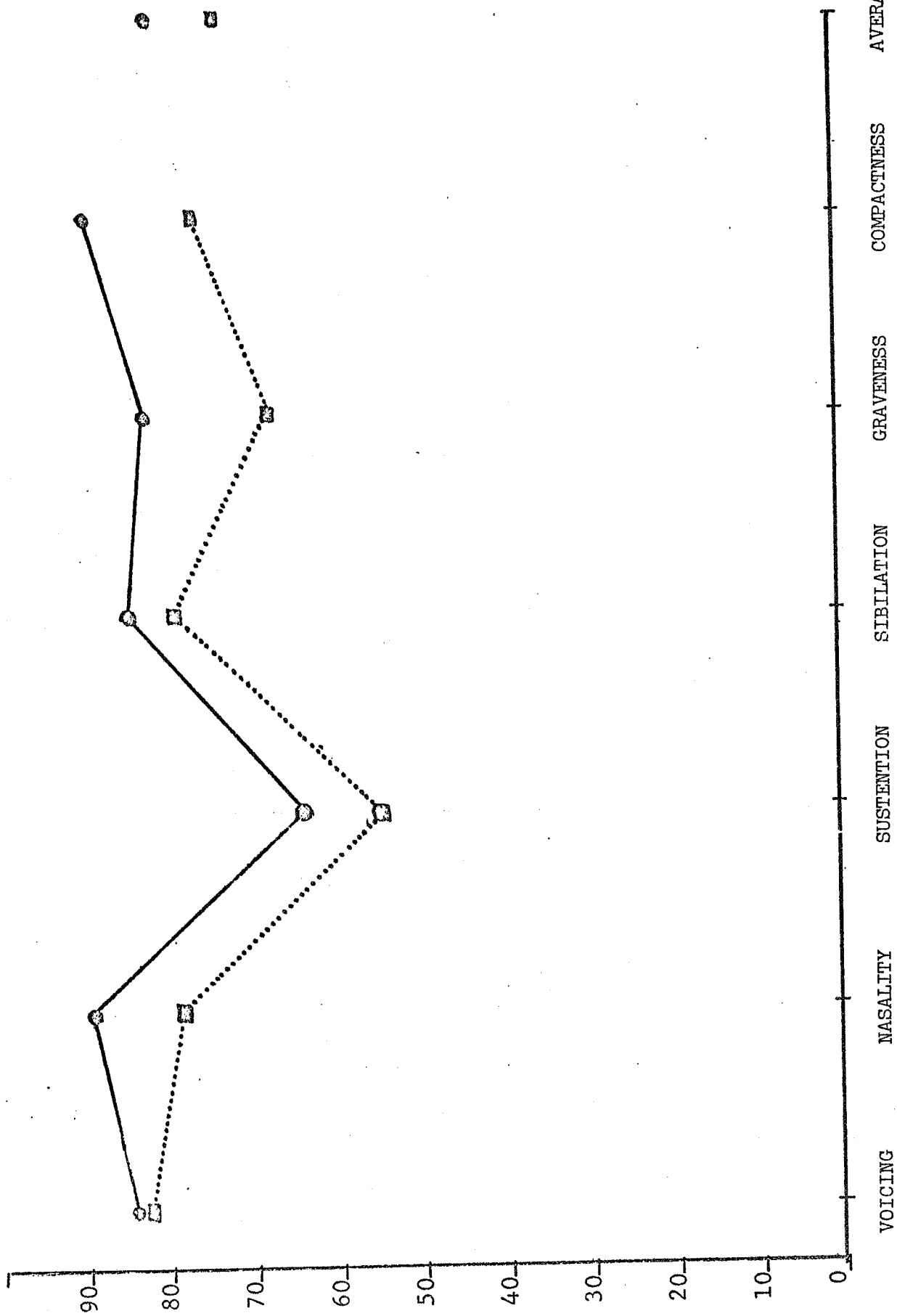


FIGURE 3. DIAGNOSTIC SCORES FOR TWO EXPERIMENTAL VOCODERS
(AVERAGES FOR 12 SPEAKERS)

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

Adaptation au français du
DIAGNOSTIC RHYME TEST
de *W.D. VOIERS*

J. P. PECKELS
Cie IBM France - Centre d'Etudes et Recherches - LA GAUDE

M. ROSSI
Institut de Phonétique - Faculté des Lettres - AIX-en-PROVENCE

PREFACE

Le "Diagnostic Rhyme Test", méthode rapide et puissante pour évaluer la netteté d'un vocodeur, a été proposée il y a quelques années par le Dr. W. D. Voiers [1], et depuis, connaît un grand succès aux Etats-Unis.[2,3,4]

Sur l'initiative du Comité de Recherche en Informatique, le Dr. Voiers a été invité à venir exposer sa méthode de test à ces journées d'Etudes sur la Parole, et le lecteur pourra se reporter au texte de sa conférence.

L'idée de l'adaptation au Français de cette méthode a pris naissance voici un an au sein du bureau du groupe "Communication Parlée". Un travail d'équipe de plusieurs membres de ce bureau a permis, au cours de l'année, non seulement de réaliser cette adaptation mais aussi de la mettre en oeuvre en effectuant des essais sur plusieurs vocodeurs réalisés en France.

Pour ce faire, il fallait une étroite coopération entre un phonéticien, des techniciens qui connaissent et utilisent les vocodeurs, ainsi que des spécialistes en mesures de netteté et d'intelligibilité.

Le groupe de travail comprenait Messieurs Cartier et Lorand du CNET (Centre National d'Etudes des Télécommunications), Monsieur Carré de l'ENSERG (Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble), ainsi que les auteurs. C'est au nom de ces personnes que les auteurs présentent l'adaptation au Français du "Diagnostic Rhyme Test" ou Test de Diagnostic par Paires Minimales.

L'exposé est divisé en plusieurs parties, à savoir :

B/d. Introduction (J-P. Peckels)

B/e. L'Analyse taxonomique de Jakobson, utilisée par Voiers. Critique et adaptation au Français (M. Rossi).

B/f. La technique du test par paires minimales et son exploitation pratique (J-P. Peckels).

B/g. Analyse des résultats en fonction de la nature des traits phonétiques et de la fréquence d'apparition des unités phoniques dans la langue (M. Rossi).

B/h. Interprétation des résultats en vue d'établir un diagnostic (J. P. Peckels).

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

première partie

I N T R O D U C T I O N

J. P. P E C K E L S

Cie IBM France - Centre d'Etudes et Recherches - LA GAUDE

Le but de nos travaux est d'adapter à la langue française un outil de test des vocodeurs ou de la parole synthétique, mais qui pourrait - à notre avis - apporter également une aide précieuse dans d'autres domaines de traitement et de transmission de la parole.

Des études approfondies et une collaboration étroite avec des spécialistes en transmission téléphonique de la parole - sous forme analogique ou numérique - ainsi qu'avec des médecins s'occupant des troubles de l'audition et de la phonation, seront toutefois nécessaires pour l'application du test dans ces domaines respectifs.

Cette méthode permet d'évaluer la netteté d'un vocodeur ou la qualité d'une parole synthétique et en même temps - après analyse détaillée des résultats obtenus - de constater un bon ou un mauvais fonctionnement du vocodeur ou du synthétiseur en question. La méthode permet donc d'établir un diagnostic contrairement à la plupart des méthodes d'évaluation utilisées jusqu'à présent qui ne permettent que de donner un chiffre de netteté global. Par ailleurs, ce test est facilement utilisable en ce sens qu'il ne fait pas nécessairement appel à des auditeurs spécialisés, contrairement au test par logatomes par exemple. En plus, il ne demande que peu de temps : une séance de test dure environ vingt minutes.

Un programme de dépouillement par ordinateur a été écrit de façon suffisamment souple pour permettre des modifications et des extensions éventuelles.

Le test de diagnostic par paires minimales, outre le fait qu'il permet de comparer des vocodeurs ou synthétiseurs de parole existante, est un outil très efficace lors du développement de ces appareils. En effet, grâce à lui on peut caractériser l'influence d'une modification d'une partie des circuits du vocodeur en développement ou en construction. Le test

contribue ainsi à optimiser les rapports coût/performance et complexité/ performance d'un vocodeur ou d'un synthétiseur de parole, et évite de figer des solutions onéreuses ou nuisibles à la netteté et l'intelligibilité, et donc à la qualité du produit final.

Le test permet également de vérifier les réglages du vocodeur une fois construit et de déterminer l'influence d'une variation de certains paramètres affectant ces réglages.

Nous donnerons maintenant une description détaillée du test par paires minimales ; tout d'abord M. Rossi va exposer la taxonomie française qu'il a définie.

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

deuxième partie

LA MATRICE PHONOLOGIQUE DE JAKOBSON , FANT ET HALLE

M . ROSSI

Institut de Phonétique - Faculté des Lettres - AIX-en-PROVENCE

VOIERS affirme, à juste titre, que le cadre proposé par MILLER et NICELY* pour tester l'intelligibilité des indices acoustiques de la parole n'est pas adéquat. Ces auteurs proposent en effet la liste des traits suivants : voisement, nasalité, affrication, durée, lieu d'articulation. Certaines de ces catégories comme le lieu d'articulation sont complexes et recouvrent en fait plusieurs traits.

VOIERS préfère à ce cadre le système taxinomique de JAKOBSON qui est plus économique et perceptuellement plus adéquat**.

La version simplifiée de ce système, adoptée par VOIERS, comprend sept types d'oppositions :

voisé	/ non voisé	=	*b* / *p*
nasal	/ non nasal	=	*m* / *b*
interrompu	/ non interrompu	=	*t* / *s*
strident	/ mat	=	*s* / θ
grave	/ aigu	=	*p* / *t*
compact	/ diffus	=	*k* / *t*
vocalique	/ non vocalique	=	*j* / *ʒ* .

La matrice taxinomique de JAKOBSON est en principe une matrice phonologique, en ce sens que, seuls, les traits distinctifs définissent les consonnes : les traits redondants sont éliminés et marqués par un 0. Exemple :

/m/ est + nasale par opposition à /b/ qui ne l'est pas,
" - non compacte " /n/ qui est compacte,
" + grave " /n/ qui est aiguë.

Cette consonne est, d'autre part, une consonne voisée, mais ce trait est redondant pour /m/ car toutes les consonnes nasales sont voisées. Le trait voisé est donc marqué par un 0 pour /m/ .

La matrice de JAKOBSON permet de classer de façon simple les consonnes du français ; comme, d'autre part, chaque trait est défini au niveau acoustique, le système taxinomique de JAKOBSON peut être, semble-t-il, utilisé de façon adéquate dans un test d'intelligibilité de parole.

Nous l'avons donc nous-même adopté. Mais un problème se pose : devons-nous construire le test en fonction d'une interprétation phonologique de cette matrice ?

1. L'analyse structurale proposée par JAKOBSON, FANT et HALLE pour les consonnes de l'anglais ou par JAKOBSON pour les consonnes du français, ne correspond pas forcément à un modèle perceptif. D'ailleurs, les analyses structurales diffèrent souvent d'un auteur à l'autre selon les critères de départ.

2. Si on part d'une matrice interprétée phonologiquement, la présence des 0, qui indiquent les traits redondants, pose des problèmes d'interprétation. En effet :

* An analysis of perceptual confusions among some english consonants J.A.S.A., 27, 2, pp. 338-352.

** Performance evaluation of speech processing devices III Diagnostic evaluation of speech intelligibility, A.F.C.R.L., 67-0101, p. 8.

a) Les 0 ne sont pas homogènes et n'indiquent pas le même type de redondance :

exemples : (1) Pour /k/ : aigu = 0
(2) Pour /j/ : aigu = 0
(3) Pour /m/ : vocalique = 0
(4) Pour /ʒ/ : aigu = 0

Dans (1), le 0 est un indice de variation contextuelle, /k/ étant aigu devant voyelle aiguë et grave devant voyelle grave ; dans (2), le 0 procède d'un choix que certains peuvent considérer comme arbitraire : on pourrait aussi bien considérer le trait compact comme redondant et le trait aigu comme distinctif ; dans (3), le trait vocalique est marqué 0 parce qu'il est impliqué par le trait nasal ; dans (4), le trait aigu est redondant parce qu'il n'existe pas dans le système de consonne constrictive à la fois grave et compacte.

La valeur des 0 n'est donc pas homogène.

b) Or, si l'on choisit une paire minimale à partir d'une opposition contenant un certain nombre de zéros, par exemple :

	grave	compact			grave	compact
p =	+	−	ou :	*ʒ* =	0	+
k =	0	+		*v* =	+	−

on ne teste pas un seul trait mais un complexe de traits, car le 0 représente un trait redondant qui peut jouer un rôle dans la perception et qui est manifesté dans l'onde acoustique.

c) Si d'autre part on veut tenir compte des zéros, étant donné que la matrice phonologique n'en donne pas la valeur, on ne sait pas exactement ce qu'on teste.

d) Pour les mêmes raisons on risque de faire une erreur dans le choix du contexte : on peut par exemple choisir /p/ et /k/ dans un contexte de voyelle aiguë (pille/quille), or dans ce cas on teste deux traits et non plus un seul car /p/ reste grave, mais /k/ est aiguë dans ce contexte. Si le trait aigu est redondant pour /k/ il n'en joue pas moins un rôle dans la reconnaissance de cette consonne.

e) Enfin, dans la mesure où la matrice phonologique conditionne le choix des paires minimales, on est amené à délaisser certaines oppositions qui pourraient nous fournir des renseignements intéressants. Ainsi, SMITH C.P.* ne prend pas en compte les oppositions *n/l* et *l/r*.

Sur la matrice qu'il utilise aucun trait ne distingue *l* de *r*, et /l/ possède deux zéros (grave et compact) dans sa définition : effectivement, on ne voit pas quel trait il aurait pu tester.

En ce qui concerne *n/l*, trois traits opposent ces deux consonnes : nasal / non nasal
non continu / continu
non vocalique / vocalique . On comprend qu'il n'ait pas choisi cette opposition complexe. Mais lorsqu'on sait que, dans le bruit par exemple ou au téléphone, /l/ et /n/ sont plus facilement confondus que /n/ et /d/, on peut se demander si, au niveau phonétique, les indices qui distinguent ces deux consonnes sont aussi complexes et nombreux que le laisse supposer la matrice phonologique utilisée par SMITH.

* Perception of Vocoder speech processed by pattern matching
J.A.S.A., 46, 6 (1969), pp. 1562-1571

En fin de compte à partir d'une matrice de ce genre, il peut arriver qu'on teste un faisceau de traits lorsqu'on croit en tester un et on omet de tester des oppositions qui sont complexes sur la matrice mais qui sont peut-être plus simples qu'il n'y paraît.

Pour toutes ces raisons, nous n'utiliserons pas une matrice structurale, mais une matrice phonétique où tous les traits sont explicités.

Ce que nous cherchons d'ailleurs à préciser, dans un test d'intelligibilité, c'est la fonction respective de chaque trait : il ne semble donc pas logique de se fonder au départ sur une interprétation struturale a priori.

1. LA MATRICE PHONETIQUE (Figure 1)

1.1. Le trait vocalique ne recouvre pas exactement la catégorie définie par JAKOBSON, FANT et HALLE*. Notre interprétation de la notion de vocalité découle directement des derniers travaux de P. DELATTRE** ; nous appelons consonnes vocaliques les consonnes dénuées de bruit et accompagnées de transitions renversées qui impliquent la présence de joints bas ; la classe des consonnes vocaliques comprend ainsi : *j*, *w*, *l* et *R* et les nasales.

1.2. Nous considérons l'opposition */l ~ n/* comme une opposition minimale, comme la seule opposition véritable de nasalité ; deux indices contribuent essentiellement à assurer l'identité phonétique de ces deux consonnes : le tempo des transitions et le degré de continuité des joints. Par contre */m ~ b/* et */n ~ d/*, que nous présentons toutefois dans notre liste, sont des oppositions plus complexes : 6 indices acoustiques et trois traits opposent ces consonnes sur le plan phonétique.

Nous verrons d'ailleurs que le taux d'intelligibilité pour ce dernier type d'opposition est très élevé alors qu'il est faible pour */n ~ l/*.

1.3. Certaines cases comportent à la fois le signe + et le signe - ; c'est le cas pour les consonnes *k*, *g* et *R*. En effet :

1.3.1. *k* et *g* sont aiguës (+) devant les voyelles aiguës et graves ou non aiguës (-) devant les voyelles graves. Pour cette raison, afin de tester de façon correcte l'opposition / compacte ~ non compacte /, nous avons opposé (*k*,*g*) à (*p*,*b*) devant les voyelles graves et à (*t*,*d*) devant les voyelles aiguës.

De même pour l'opposition / interrompu ~ non interrompu /, nous présentons (*k*,*g*) opposés à (*ʃ*,*ʒ*) dans un contexte de voyelles aiguës, car ces dernières consonnes sont toujours aiguës ;

1.3.2. *R* étant une consonne polymorphe, la plupart des traits susceptibles de la définir sont représentés nécessairement par des signes contradictoires. Ainsi dans la variété de français où elle est consonne constrictive sans battements, *R* est non vocalique, non interrompue et continue : elle constitue, dans ce cas, avec *g* une opposition minimale ; mais si elle reste suffisamment ouverte pour conserver son trait vocalique, elle constitue avec *ʒ* cette fois une opposition minimale.

* Preliminaries to speech analysis
M.I.T. Press, Cambridge (1963)

** Des indices acoustiques aux traits pertinents
Proceedings of the sixth International Congress of Phonetic Sciences
Prague (1967), pp. 35-47.

FIGURE 1 - Analyse binaire : en traits acoustiques, du système consonantique du français

	v *n*	*m*	*n*	*g*	*k*	*b*	*p*	*d*	*t*	*ʒ*	*ʃ*	*v*	*f*	*z*	*s*	*l*	*R*	*j*	*w*
nasal	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
vocalique	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	+	+/−	+	+
interrompu	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	+	+/−	−	−
continu	+	+	+	−	−	−	−	−	−	+	+	+	+	+	+	+	+/−	+	+
compact	+	−	−	+	+	−	−	−	−	+	+	−	−	−	−	−	+/−	+	−
aigu	+	−	+	+/−	+/−	−	−	+	+	+	+	−	−	+	+	+	+/−	+	−
voisé	+	+	+	+	−	+	−	+	−	+	−	+	−	+	−	+	+	+	+

Dans la variété de français où elle est consonne à battements, *R* est interrompue (présence des occlusions que sont les battements), non continue (écoulement intermittent de l'onde sonore) et vocalique ; s'il s'agit de la consonne antérieure à battements, elle est non compacte : dans ce cas l'opposition avec *l* est minimale (continue ~ non continue). S'il s'agit de la consonne postérieure à battements, elle est compacte ; l'opposition avec *l* est alors complexe.

Dans la liste des paires minimales les oppositions */g ~ R/*, */ʒ ~ R/* et */R ~ l/* sont toutes représentées. Une analyse rapide permettra de savoir laquelle de ces oppositions est minimale dans la réalisation du locuteur utilisé.

2. LA LISTE DES PAIRES MINIMALES (Figures 2.a, b, c)

Chaque trait apparaît un nombre égal de fois devant les voyelles choisies ; celles-ci au nombre de 9 sont rangées en trois classes :
. les voyelles orales non arrondies : *i* , *ɛ* , *A* ,
. les voyelles orales arrondies : *y* , *u* , *O* ,
. les voyelles nasales : *ɛ̃* , *ɑ̃* , *ɔ̃* .
Chaque trait est présenté en moyenne dans six oppositions différentes ; exemple : voisé / non voisé = *b/p*, *d/t*, *g/k*, *ʒ/ʃ*, *z/s*, *v/f*.

Nous avons cherché à répartir également chacune de ces oppositions dans chaque classe de voyelles. Nous n'avons pas toujours réussi. En effet :

2.1. Certaines oppositions sont rares à l'initiale, exemple : *d/z* , *w/j* .

2.2. Certaines autres ne peuvent pas apparaître dans tous les contextes, exemple : l'opposition de compacité *k/t* doit se réaliser devant une voyelle aiguë pour être minimale.

2.3. Le lexique impose certaines contraintes de sélection.

Chaque trait est testé 8 fois devant chaque voyelle c'est-à-dire 72 fois au total. Devant */i/*, par exemple, on présente 2 fois bile, deux fois pile, deux fois dire et deux fois tire.

Au total 432 mots ont été présentés à deux groupes de 11 personnes dans un test à choix forcé : les auditeurs, après la présentation du stimulus, doivent cocher le mot entendu dans la paire minimale qui leur est proposée. Les auditeurs étaient installés dans un Laboratoire de Langues. Les stimuli étaient diffusés à partir d'un magnétophone OPELEM et reçus au niveau de l'oreille par l'intermédiaire d'un casque SOCAPEX enveloppant.

La courbe de réponse du casque a été mesurée, à l'aide d'une microsonde et à l'aide de l'oreille artificielle B et K, par Monsieur CORSAIN, Directeur du Service d'Acoustique appliquée au C.N.R.S. (courbe de réponse : 100-4000 Hz avec ± 6 dB ; fréquence de référence : 400 Hz).

Nous avons réglé le niveau d'écoute de façon à ce que l'intensité reçue au niveau de l'oreille soit de l'ordre de 64 dB (100 mV sous 50 ohms à la sortie de la prise pour casque ; fréquence de référence : 1000 Hz).

FIGURE 2.a - Liste de paires minimales établie par M. ROSSI pour les tests de diagnostic sur synthétiseur de parole

		i		ε		A	
Voisé	N°	1-19-37-55	2-20-38-56	3-21-39-57	4-22-40-58	5-23-41-59	6-24-42-60
	opposition	b/p : bile/pile	d/t: dire/tire	v/f : ver/fer	3/ʃ : gère/chère	z/s: Zabre/sabre	g/k: gaze/case
Grave	N°	73/91/109/127	74-92-110-128	75-93-111-129	76-94-112-130	77-95-113-131	78-96-114-132
	O	f/s : fils/six	p/t: pisse/tisse	b/d :bette/dette	v/z : veste/zeste	w/j: hoir/yard	m/n: mate/natte
Compact	N°	145/163/181/199	146-164-182-200	147-165-183-201	148-166-184-202	149-167-185-203	150-168-186-204
	O	g/d : guigne/digne	k/t: quille/tille	ʃ/s : schéma/sème	3/z : gèle/zèle	r/l: ratte/latte	ʃ/s: chape/sape
Interrompu	N°	217-235-253-271	218-236-254-272	219-237-255-273	220-238-256-274	221-239-257-275	222-240-258-276
	O	g/3 :guise/gisent	k/ʃ: quiche/chiche	g/3 : gaine/gêne	k/ʃ : Caire/chaire	t/s: tache/sache	b/v: bal/val
Vocalique	N°	289-307-325-343	290-308-326-344	291-309-327-345	292-310-328-346	293-311-329-347	294-312-330-348
	O	w/v :ouirent/virent	r/3: rite/gîte	w/v :ouest/veste	l/d :lèche/dèche	j/3: yard/jars	w/v: ouaille/vaille
Nasal	N°	361-379-397-415	362-380-398-416	363-381-399-417	364-382-400-418	365-383-401-419	366-384-402-420
	O	m/b : miche/biche	n/d: Nîmes/dîmes	m/b : mêle/bêle	n/l : naine/laine	n/l: nappe/lape	n/d: nard/dard

		Y		O		u	
Voisé	N°	7-25-43-61	8-26-44-62	9-27-45-63	10-28-46-64	11-29-47-65	12-30-48-66
	oppo-si-tion	b/p : bure/pure	z/ʃ : jurent/churent	v/f : vol/fol	z/s : zone/saône	g/k: gourd/court	d/t: douche/touche
grave	N°	79-97-115-133	80-98-116-134	81-99-117-135	82-100-118-136	83-101-119-137	84-102-120-138
	O	m/n : mule/nulle	f/s: furent/surent	v/z :vosne/zone	p/t : pôle/tôle	b/d: bouse/douze	f/s: four/sourd
Compact	N°	151-169-187-205	152-170-188-206	153-171-189-207	154-172-190-208	155-173-191-209	156-174-192-210
	O	ʃ/s : chutes/sûtes	r/l : rut/luth	k/p : corps/port	g/b : gosse/bosse	k/p : coule/poule	g/b : gourd/bourre
Interrompu	N°	223-241-259-277	224-242-260-278	225-243-261-279	226-244-262-280	227-245-263-281	228-246-264-282
	O	g/ʒ : gutte/jute	k/ʃ : cure/churent	b/v : bol/vol	t/s : tort/sort	g/r : goutte/route	p/f : pour/four
Vocalique	N°	295-313-331-349	296-314-332-350	297-315-333-351	298-316-334-352	299-317-335-353	300-318-336-354
	O	r/ʒ : ruche/juche	r/ʒ : rut/jute	l/d : lors/dors	l/d : lotte/dote	l/d : louve/douve	l/d : louche/DOU-che
nasal	N°	367-385-403-421	368-386-404-422	369-387-405-423	370-388-406-424	371-389-407-425	372-390-408-426
	O	m/b : mule/bulle	n/l : nugues/Lucques	n/l : note/lotte	n/d : nonne/donne	m/b : maule/boule	n/d : nouille/douille

FIGURE 2.b

B/e/7.

FIGURE 2.c

B/e/8.

		$\tilde{\varepsilon}$		$\tilde{a}$		$\tilde{o}$	
Voisé	N°	13-31-49-67	14-32-50-68	15-33-51-69	16-34-52-70	17-35-53-71	18-36-54-72
	opposition	v/f : vintes/feintes	z/s : zende/scinde (1)	ʒ/ʃ : jante/chante	g/k : gangue/cangue	d/t : dompte/tonte	b/p : bonde/ponde
grave	N°	85-103-121-139	86-104-122-140	87-105-123-141-	88-106-124-142	89-107-125-143	90-108-126-144
	O	v/z : vain/zain	p/t : peinte/teinte	m/n : mante/Nantes	f/s: fente/sente	b/d: Bombes/Dombes	p/t: ponte/tonte
Compact	N°	157-175-193-211	158-176-194-212	159-177-195-213	160-178-196-214	161-179-197-215	162-180-198-216
	o	g/d : guinde/dinde	k/t : quinte/teinte	ʒ/z : gens/zan	r/l: rente/lente	g/b: gonze/bonze	k/p: conte/ponte
Interrompu	N°	229-247-265-283	230-248-266-284	231-249-267-285	232-250-268-286	233-251-269-287	234-252-270-288
	O	t/s : teinte/sainte	p/f : peinte/feinte	b/v : bande/vende	d/z : dans/zan	p/f : ponte/fonte	g/r : gonfle/ronfle
Vocalique	N°	301-319-337-355	302-320-338-356	303-321-3394357	304-322-340-358	305-323-341-359	306-324-342-360
	O	w/v : ointes/vintes	r/ʒ : rein/geint	l/d : lance/danse	j/ʒ: iambe/jambe	j/ʒ: ion/jonc	l/d: lombes/Dombes
Nasal	N°	373-391-409-427	374-392-410-428	375-393-411-429	376-394-412-430	377-395-413-431	378-396-414-432
	O	m/b : main/bain	n/l : nimbe/limbes	n/d : Nantes/Dante	n/l: Nantes/lente	m/b: monde/bonde	n/d: nom/don

(1) On peut remplacer cette paire par la paire suivante : Zain/sein, parceque les sujets peuvent ne pas savoir que "zende" se prononce "zinde". Je pense cependant qu'on peut sans problème faire intervenir des mots rares dans les paires minimales à condition que les sujets puissent avoir connaissance de la liste avant le test.

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

troisième partie

LA TECHNIQUE DU TEST
PAR PAIRES MINIMALES
ET SON EXPLOITATION PRATIQUE

J. P. P E C K E L S

Cie IBM France - Centre d'Etudes et Recherches - LA GAUDE

1. Présentation des listes de mots

1.1. Introduction

Nous avons vu que le but du test est de vérifier si les consonnes initiales des mots présentés sont bien comprises par les auditeurs, ou bien si elles sont prises pour d'autres consonnes.

Toutefois, on guide le choix de l'auditeur en ne lui demandant pas d'écrire dans l'absolu ce qu'il entend mais en lui présentant un mot à choisir parmi une opposition représentant une paire minimale : on entend par là deux mots qui ne se distinguent que par la consonne initiale et où, en plus, ces deux consonnes représentent les deux membres d'une des oppositions définies par M. Rossi.

Etant donné que l'auditeur a devant lui les deux mots parmi lesquels il doit choisir, il faut évidemment que le vocodeur soit d'une qualité telle que -à l'exception éventuellement de la première consonne- l'auditeur reconnaisse à coup sûr le corps du mot et ne soit pas gêné par des problèmes élémentaires de netteté.

Le test se fait donc sur six caractéristiques, avec deux oppositions dans chaque cas. Ainsi, dans la caractéristique "voisée", on oppose des consonnes voisées à des consonnes non voisées ; dans la caractéristique "compacte", des consonnes compactes sont opposées à des consonnes diffuses ; etc...

1.2. Listes

M. Rossi vient de présenter les listes des stimuli ; nous allons nous y référer à nouveau (figures 2.a, 2.b, 2.c ; pages B/e/6, 7 et 8) pour expliquer de quelle manière on crée - à l'aide d'un locuteur professionnel - la bande magnétique servant à évaluer un vocodeur.

Sur ces listes il y a 18 colonnes, soit deux par voyelle. Chaque colonne représente les oppositions dans six caractéristiques à évaluer. On trouve donc en tout 18 x 6 = 108 paires de stimuli.

Au cours du test, chaque stimulus sera présenté deux fois. Ceci augmente à la fois le nombre effectif de mots testés et permet de comparer les erreurs faites lors du premier et second passage du mot en question, afin de voir s'il y a apprentissage ou accoutumance.

Pour un test, on prépare quatre séquences -soit un total de 432 mots- en sélectionnant un membre de chaque opposition. Ces quatre séquences sont enregistrées par un locuteur professionnel (Figure 1).

Deux groupes de 4 séquences distinctes ont jusqu'ici été définis. Dans l'un de ces groupes on a enregistré séquentiellement les deux colonnes par voyelle ; dans l'autre, on a alterné avec une colonne d'une autre voyelle.
A partir de ces deux groupes, plusieurs bandes originales ont été enregistrées à différents rythmes et avec des intervalles différents entre mots.

La bande test enregistrée à la sortie du vocodeur est présentée à un groupe d'auditeurs munis de casques binauraux. Le niveau de la parole est réglé à 64 dB re 2.10^{-4} µbar (pondération "A" du sonomètre, selon les recommandations du CEI).

L'auditeur a devant lui des listes représentant tous les stimuli et on lui demande de cocher, pour chaque paire, le mot qu'il a compris (Figure 2).

Une réponse doit être donnée dans chaque cas, et l'on voit ici à nouveau qu'il est important que le vocodeur ne déforme pas les mots au point que l'auditeur croie entendre un troisième mot entièrement différent de ceux qui constituent la paire minimale. Il semble que, dans certains cas, cette condition soit difficile à réaliser, ainsi que M. Rossi va l'expliquer plus loin.

FIGURE 1 -- Exemple de choix de mots à enregistrer par le locuteur

DIRE			TIRE
PISSE			TISSE
QUILLE			TILLE
QUICHE			CHICHE
RÎTES			GÎTES
NÎMES			DÎMES
GÈRE			CHÈRE
VESTE			ZESTE
GELE			ZELE
CAIRE			CHAIRE
LÈCHE			DÈCHE
NAINE			LAINE
GAZE			CASE
MATE			NATTE
CHAPE			SAPE
BAL			VAL
QUAILLE			VAILLE
NARD			DARD

JURENT			CHURENT
FURENT			SURENT
RUTH			LUTH
CURE			CHURENT
RUT			JUTE
NUQUES			LUCQUES
ZONE			SAÔNE
PÔLE			TÔLE
GOSSE			BOSSE
TORT			SORT
LOTTE			DOTE
NONNE			DONNE
DOUCHE			TOUCHE
FOUR			SOURD
COURD			BOURRE
POUR			FOUR
LOUCHE			DOUCHE
NOUILLE			DOUILLE

ZAIN			SEIN
PEINTE			TEINTE
QUINTE			TINTE
PEINTE			FEINTE
REIN			GRAIN
NIMBE			LIMBES
GANGUE			GANGUI
PLNTE			SENTE
RENTE			LENTE
DANS			ZAN
LAMBE			JAMBE
NANTES			LENTE
BONDE			LONDE
PONTE			TONTE
CONTE			PONTE
GONFLE			RONFLE
LOMBES			BOMBES
NOM			DON

FIGURE 2 - Exemple de feuille de test

DIRE			TIRE
PISSE			TISSE
QUILLE			TILLE
QUICHE			CHICHE
RÎTES			GÎTES
NÎMES			DÎMES
GÈRE			CHÈRE
VESTE			ZESTE
GELE			ZELE
CAIRE			CHAIRE
LÈCHE			DÈCHE
NAINE			LAINE
GAZE			CASE
MATE			NATTE
CHAPE			SAPE
BAL			VAL
OUAILLE			VAILLE
NARD			DARD

JURENT			CHURENT
FURENT			SURENT
RUTH			LUTH
CURE			CHURENT
RUT			JUTE
NUQUES			LUCQUES
ZONE			SAÔNE
PÔLE			TÔLE
GOSSE			BOSSE
TORT			SORT
LOTTE			DOTE
NONNE			DONNE
DOUCHE			TOUCHE
FOUR			SOURD
GOURD			BOURRE
POUR			FOUR
LOUCHE			DOUCHE
NOUILLE			DOUILLE

ZAIN			SEIN
PEINTE			TEINTE
QUINTE			TINTE
PEINTE			FEINTE
REIN			GEINT
NIMBE			LIMBES
GANGUE			CANGUE
FENTE			SENTE
RENTE			LENTE
DANS			ZAN
LAMBE			JAMBE
NANTES			LENTE
BONDE			PONDE
PONTE			TONTE
CONTE			PONTE
GONFLE			RONFLE
LOMBES			DOMBES
NOM			DON

2. Dépouillement du test

Les stimuli mal compris par les auditeurs sont localisés, par superposition de caches ou transparents.

Les résultats sont ensuite dépouillés par ordinateur. Le programme de dépouillement est extrêmement simple, grâce à une numérotation des stimuli.

Ainsi (figures 2.a, 2.b, 2.c, pages B/e/6, 7 et 8) les stimuli portant les numéros 1 à 72 inclus représentent la caractéristique voisée, ceux de 73 à 144 inclus la caractéristique grave, etc... (145 à 216, 217 à 288, 289 à 360, 361 à 432). Rappelons-nous que par caractéristique on a 18 × 2 = 36 stimuli, présentés chacun deux fois au cours du test.

Lors de la phase d'initialisation du programme, on introduit par ailleurs pour chaque stimulus le signe "+" ou "-" selon qu'il s'agit, dans la caractéristique voisée par exemple, d'une consonne voisée ou non voisée. Ainsi, dans notre exemple, le mot bile porte les numéros 1^{+} et 19^{+}, le mot pile 37^{-} et 55^{-} (Figure 2.a, page B/e/6).

Voyons le résultat du dépouillement pour un exemple pratique :

Exemple : Réponse fausse 75 (auditeur "X").

Le stimulus présenté appartient à la caractéristique "grave" "-". Il s'agit donc d'une consonne aiguë. (Le mot en question est "dette").

Le fait de donner comme réponse fausse "75" fait donc apparaître dans les tableaux des résultats les indications suivantes :

Caractéristiques 2^{-} (auditeur "X")

Question..... 75 Erreur(s)..... 1

En clair, ceci signifie que la consonne aiguë en question a été comprise comme une consonne grave. (Le mot "dette" a été compris comme "bette").

On a déjà expliqué que chaque stimulus est présenté deux fois au cours du test. Ainsi, le mot "dette" est présenté à nouveau comme stimulus 129, et il est important de comparer le nombre d'erreurs commises dans les deux cas, ceci afin de détecter une accoutumance éventuelle chez les auditeurs. Des listes détaillées d'erreurs sont fournies par le programme.

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

quatrième partie

ANALYSE DES RESULTATS EN FONCTION DE LA NATURE DES TRAITS PHONETIQUES ET DE LA FREQUENCE D'APPARITION DES UNITES PHONIQUES DANS LA LANGUE

M. ROSSI

Institut de Phonétique - Faculté des Lettres - AIX-en-PROVENCE

Les deux groupes de sujets qui ont participé à l'expérience étaient de niveau différent : le premier était un groupe naïf, le deuxième un groupe d'étudiants de phonétique qui avaient suivi quelques séances d'éducation auditive dans le cadre de l'enseignement de la phonétique. Les résultats sont évidemment différents.

1 - Pour le premier groupe, le taux d'intelligibilité est de 92,11 %.

2 - Pour le deuxième groupe, il s'élève à 95,86 % (pour la parole naturelle : 99,5 %).

Malgré cette différence, il est intéressant de remarquer que les mêmes tendances apparaissent, dans le système de fautes, pour les deux groupes.

Le tableau des erreurs (figure 1) montre :

1 - Que certaines oppositions sont plus perturbées que d'autres :

Exemples : Aigu/grave, compact/non compact.

2 - Une répartition dissymétrique des erreurs à l'intérieur de chaque opposition : l'un des membres de l'opposition est généralement plus perturbé que l'autre.

Exemples : le terme aigu (caractéristique - 2) est plus perturbé que le terme grave(caractéristique + 2).

On peut penser a priori que cette répartition dissymétrique des erreurs n'est qu'un simple artefact dû à un manque d'homogénéité dans les réponses des sujets.

Afin de vérifier si les différences dans la fréquence des erreurs entre les oppositions, ou entre certaines classes d'oppositions, étaient significatives et ne dépendaient pas des variations entre les sujets, nous avons procédé à une analyse de la variance et calculé le rapport de Snédécor.

			GROUPE 1	GROUPE 2
CARACTERISTIQUE	1-	ERREUR (S) ...	24	11
"	1+	"	13	5
"	2-	"	80	62
"	2+	"	48	19
"	3-	"	45	28
"	3+	"	68	33
"	4-	"	8	2
"	4+	"	10	1
"	5-	"	25	14
"	5+	"	15	9
"	6-	"	16	2
"	6+	"	23	11

FIGURE 1 : Fréquence des erreurs par caractéristique pour les groupes 1 et 2 (statistique globale).

Nous aboutissons aux conclusions suivantes :

1 - La fréquence des fautes est significativement différente entre les six types d'opposition :

F = 12,6 (seuil 2,41 à 0,01, v = 11, v' = 120).

2 - Si on élimine les types Aigu/grave (caractéristiques -2 et +2) et Compact/Non compact (caractéristiques -3 et +3), la fréquence des fautes n'est plus significativement différente entre les types d'opposition.

F = 1,24 (seuil 3,04 à 0,01, v = 7, v' = 80).

Les oppositions Aigu/Grave et Compact/Non compact constituent donc une source importante de variations, c'est-à-dire de confusion : les indices de lieu d'articulation sont de loin plus perturbés que les indices de mode d'articulation.

3 - Bien que le nombre d'erreurs soit plus grand pour l'opposition Aigu/Grave que pour l'opposition Compact/Diffus, cette différence n'est pas du tout significative.

F = 3,64 (seuil 4,31 à 0,01, v = 3, v' = 40).

4 - Par contre, pour chacun de ces deux types d'opposition, la fréquence des fautes est significativement différente entre les caractéristiques négative et positive, entre les traits marqué et non marqué, ce qui ne se vérifie pas pour les autres types d'opposition :

4.1. Ainsi pour l'opposition voisé/non voisé, les consonnes sourdes sont confondues plus souvent avec les consonnes sonores (24 fautes, 6 %); les consonnes sonores sont confondues moins souvent avec les consonnes sourdes (13 fautes, 3 %) : cette différence n'est pas significative.

4.2. Mais, pour les caractéristiques 2 et 3, (Aigu / Grave et Compact / diffus) la fréquence des fautes est significativement plus grande avec les consonnes aiguës qu'avec les consonnes graves (80 fautes = 20 % contre 48 fautes = 12 %) et avec les consonnes compactes qu'avec les consonnes diffuses (68 fautes = 17 % contre 45 fautes = 11 %).

Aigu / Grave : F = 6,1 (seuil 4,35 à 0,05, v = 1, v' = 20).

Compact / Diffus : F = 5,21 (seuil 4,35 à 0,05, v = 1, v' = 20).

4.3. Nos résultats diffèrent sur certains points essentiels de ceux de VOIERS et de SMITH (figure 2). Nous rendrons compte de ces différences chemin faisant.

1. L'OPPOSITION VOISE/NON VOISE.

Le taux d'intelligibilité pour cette opposition s'élève à 95,20 %. MILLER et NICELY, dans une étude sur l'intelligibilité de la parole naturelle présentée avec certaines distorsions*, avaient déjà remarqué que l'opposition Voisé / non Voisé, dans tous les cas de distorsions (masquage, filtrage), étaient une des plus résistantes.

1.1. Contrairement à ce qu'on pourrait s'attendre, connaissant les limites du Vocodeur, (impossibilité de fonctionnement simultané de la source de bruit et de la source impulsionnelle), les contrictives ne sont pas plus perturbées que les occlusives).

1.2. Parmi les consonnes sourdes, les aiguës sont plus perturbées que les graves :

19 consonnes sourdes aiguës sont entendues comme sonores, contre 5 consonnes sourdes graves seulement.

Les erreurs sont en raison inverse de la fréquence d'utilisation de ces consonnes dans la chaîne parlée (sourdes aiguës = 25,27 % , sourdes graves = 12,74 %). On ne peut donc voir là un effet de la fréquence statistique. L'explication de ce phénomène peut être la suivante : la fréquence de coupure aux environs de 3500 Hz élimine les hautes fréquences qui constituent la partie essentielle de l'énergie des consonnes aiguës. Les consonnes aiguës deviennent ainsi des consonnes faibles - or nous savons que l'opposition de sonorité est aussi une opposition de force. Il est donc normal que les sujets tendent à percevoir ces consonnes aiguës sourdes "faibles" comme des sonores.

1.3. Parmi les consonnes sonores, par contre, les erreurs sont inversées : la fréquence des erreurs est plus grande avec les consonnes graves qu'avec les aiguës. 9 consonnes sonores graves (total des fautes = 13) sont perçues comme sourdes.

Ce type de confusion, provient du fait que les consonnes graves ont une faible énergie. Selon le niveau sonore de la consonne le détecteur de F_0 peut prendre ou ne pas prendre les impulsions laryngées.

Nous n'insisterons pas sur l'opposition Interrompu / non interrompu qui est la mieux perçue de toutes : 97,75 % d'intelligibilité. Chez VOIERS et SMITH, cette opposition est la plus perturbée, puisqu'ils obtiennent 67 % d'intelligibilité seulement. La différence de ces résultats est due à la différence du nombre d'échantillonnages par seconde dans les deux expériences.

2. LES OPPOSITIONS DE MODE D'ARTICULATION.

2.1. Nous n'insisterons pas sur l'opposition Interrompu / non interrompu qui est la mieux perçue de toutes : 97,75 % d'intelligibilité. Chez VOIERS et SMITH, cette opposition est la plus perturbée (figure 2) puisqu'ils obtiennent 67 % d'intelligibilité seulement. La différence de ces résultats est due au nombre différent d'échantillonnages par seconde dans les deux expériences.

* G.A. MILLER and P.E. NICELY,
An Analysis of perceptual confusions among some english consonants ;
J.A.S.A. 27,2, 1955.

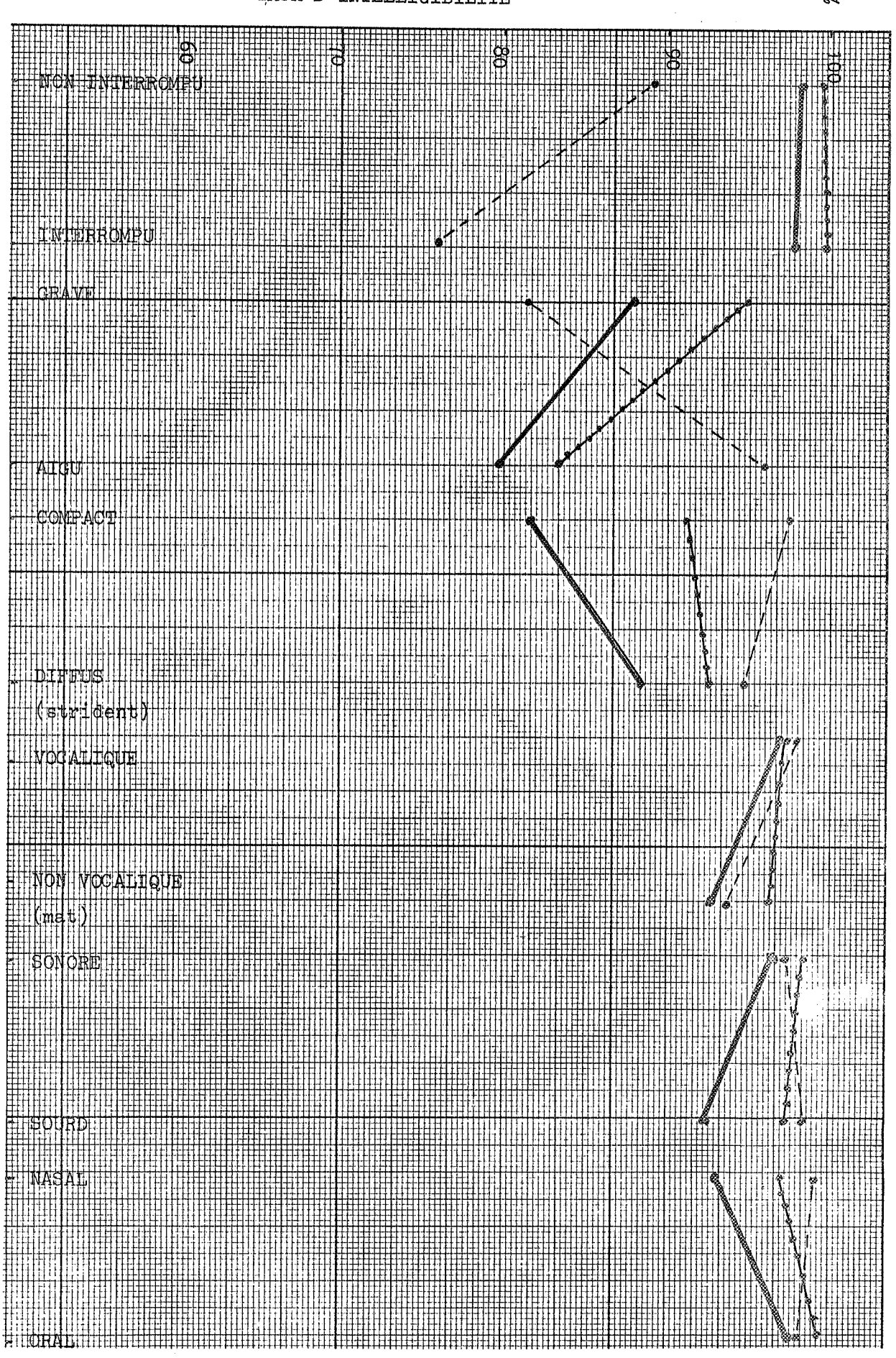
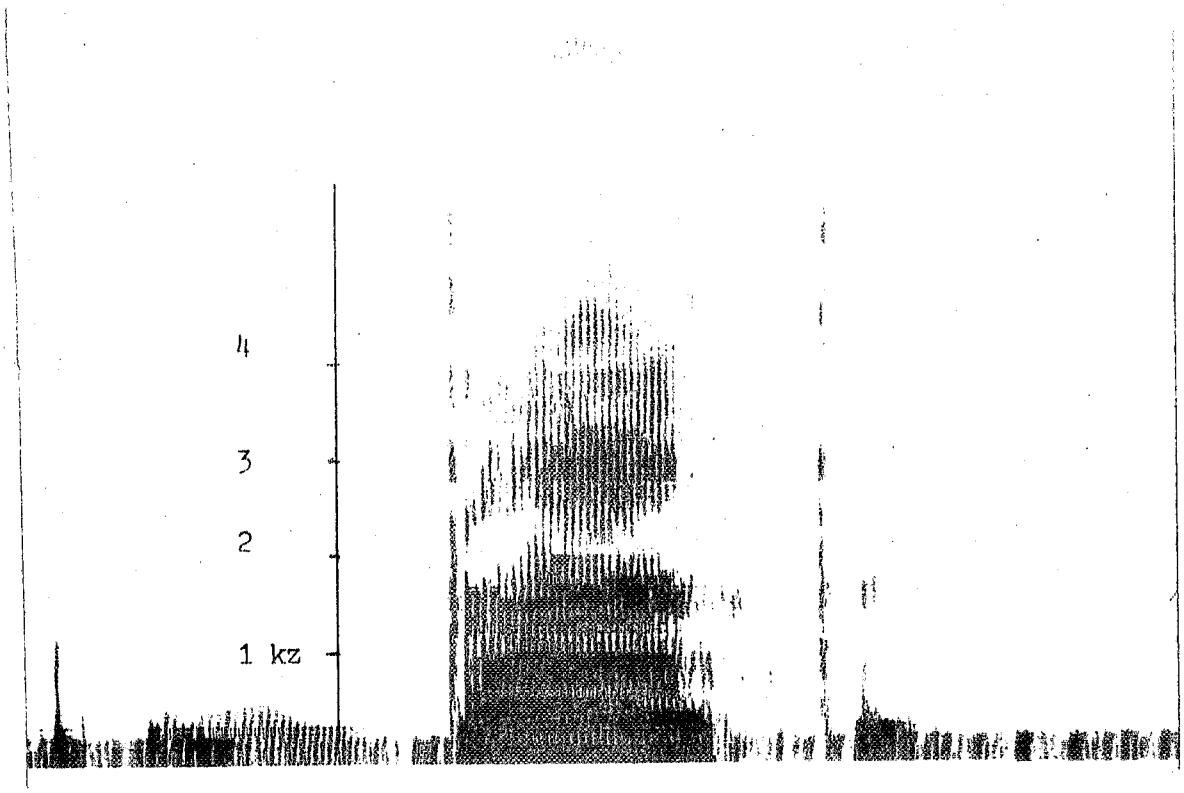


FIGURE 2 - Taux d'intelligibilité pour chaque trait acoustique : ——— Résultat de VOIERS
——— Groupe 1
+++++ Groupe 2) Résultats de PECKELS - ROSSI

FIGURE 3.a - Sonagramme de peinte : voix naturelle, locuteur IBM

O peinte 86



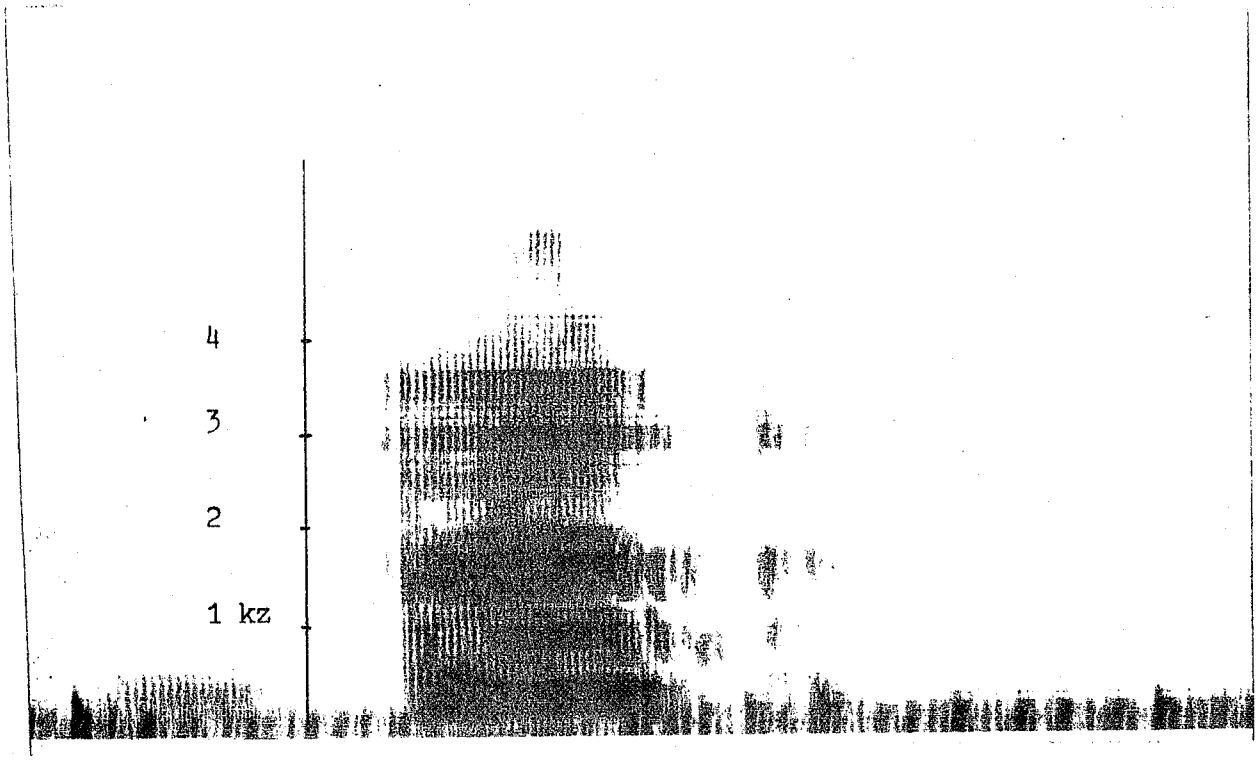
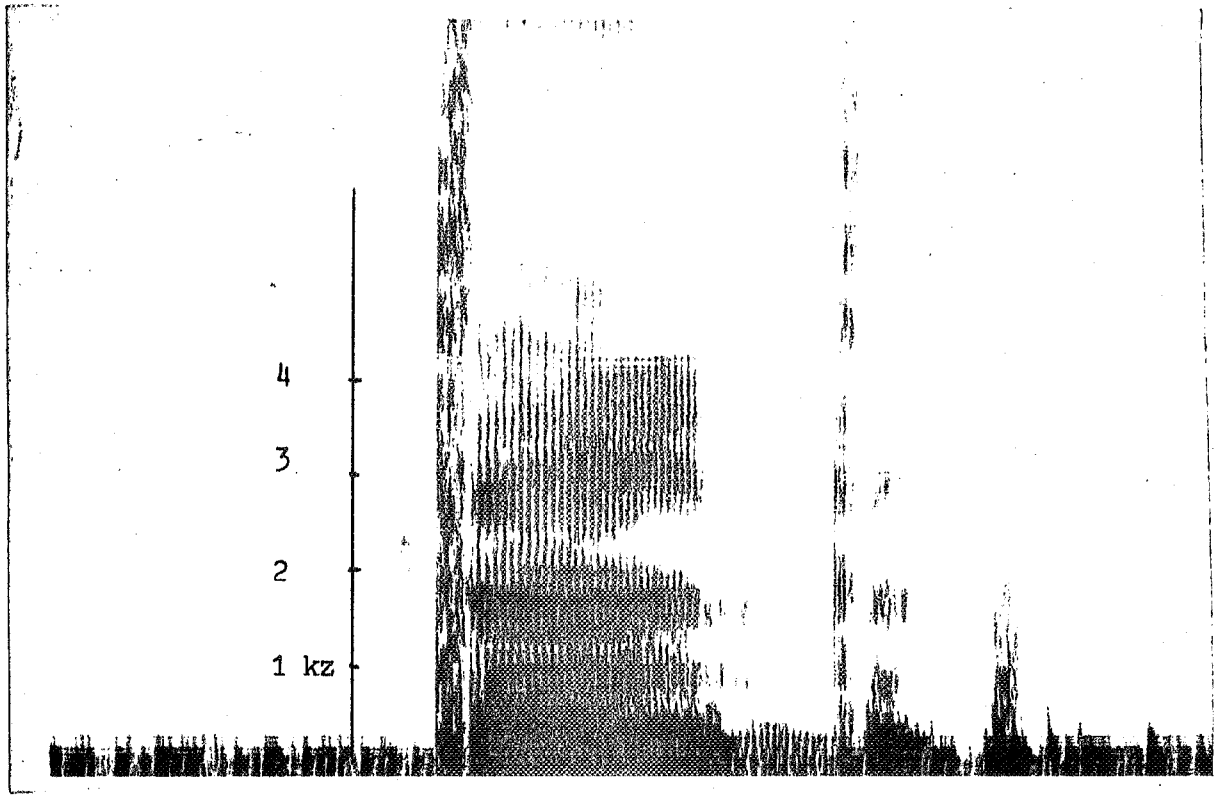


FIGURE 3.b - Sonagramme de peinte : voix reconstituée sur vocodeur IBM

V peinte 86

FIGURE 3.c - Sonagramme de teinte : voix naturelle, locuteur IBM

O teinte 140



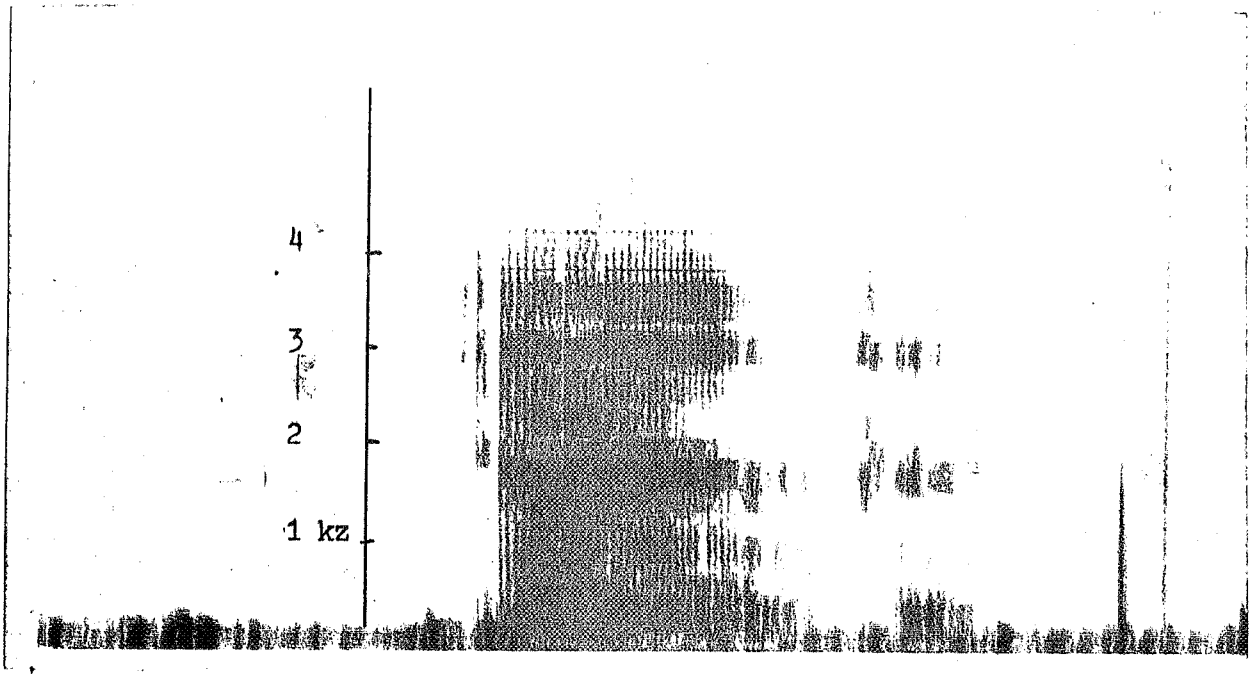


FIGURE 3.d : Sonagramme de teinte : voix reconstituée sur vocodeur IBM

V teinte 140

2.2. Vocalique / non vocalique : Cette opposition est également bien intégrée = 97,35 % d'intelligibilité. Il est cependant intéressant de noter que les erreurs ne sont pas distribuées au hasard et sont concentrées sur certains types d'opposition et même sur certains stimuli :
9 fautes sur 21 affectent le type d'opposition : /ʒ ∿ r/ ,
8 " 21 " : /ω ∿ V/ ;
le taux d'intelligibilité pour ouirent descend à 82 % .

En ce qui concerne ouirent ∿ virent , trois faits permettent d'expliquer les erreurs :
- virent, synthétisé, ne possède pas de bruits sur son spectre,
- les deux consonnes possèdent ce que DELATTRE appelle un joint bas et qui, normalement, permet la distinction ω/v,
- la quantification des spectres détruit ou masque la différence de tempo dans les transitions (la transition, en voix naturelle, est plus lente pour /ω/ que pour /v/).

2.3. Nasal / non nasal : Comme les autres oppositions de mode d'articulation, celle-ci est bien perçue ; taux d'intelligibilité : 95,08 %.

Cependant, comme pour l'opposition précédente, les fautes ne sont pas distribuées au hasard ; elles sont concentrées sur le type d'opposition /n ∿ l/ : le pourcentage de reconnaissance de /n/ dans les oppositions /n ∿ l/ descend à 84 %.

Nous avions considéré l'opposition /n - l/comme une opposition minimale ; le bien-fondé de cette analyse est corroboré par le fort pourcentage d'erreurs qui l'affecte et qui ne s'explique que par la parenté étroite qui lie ces deux consonnes.

DELATTRE identifie trois indices qui permettent la distinction de ces deux consonnes :

2.3.1. Le tempo des transitions : rapides pour /n/, lentes pour /l/.

2.3.2. La continuité des formants consonne - voyelle pour /l/, leur discontinuité pour /n/,

2.3.3. Le locus de *F* 1 à 250 Hz pour /n/, et à 400 pour /l/.

La comparaison des sonagrammes de la parole naturelle et de la parole reconstituée semble montrer que c'est la distinction au niveau des deux premiers indices qui a été altérée.

3. LES OPPOSITIONS DE LIEU D'ARTICULATION.

3.1. Aigu / Grave : Les résultats de MILLER et NICELY (op.cit.) montrent qu'en seuil masqué ou après filtrage (filtre passe-bas 200-2500 Hz). Les consonnes occlusives aiguës sont plus perturbées que les occlusives graves. Mais tandis que /t/ a tendance à être perçu comme /p/, /d/ par contre est plutôt perçu comme /g/.

SMITH et VOIERS, de leur côté, aboutissent à une conclusion diamétralement opposée : dans la parole reconstituée par Vocodeur (2400 bits/s) le trait grave est moins bien perçu - et de loin - que le trait aigu.

Les résultats - pour les consonnes du français - concordent avec ceux de MILLER et non ceux de VOIERS. Nous trouvons que la fréquence des fautes - nous l'avons déjà dit - est significativement plus grande pour les consonnes aiguës que pour les consonnes graves.

1. Les consonnes aiguës (caractéristique - 2) :

	y	s	z	d	t	ʒ	ω	f	v	b	p	total	% reconnaissance
y	22						0					22	100
s		85						3				88	96,5
z			57						9			66	86,3
d				65						23		88	73,8
t					79						31	110	71,8
ʒ						8			14			22	36,37
					POURCENTAGE MOYEN DE RECONNAISSANCE								79,80

On remarque que la fréquence des fautes pour les consonnes aiguës occlusives (*t, d*) est plus importante que pour les constrictives (*s, z*). Cette différence est significative ($F = 23$, seuil : 8,10 à 0,01, v = 1 et v' = 20). D'autre part, la fréquence des fautes, pour les occlusives, est plus grande devant les voyelles ouvertes non arrondies (*a, ɛ, ɑ̃, ɛ̃*), elle est significativement plus grande que devant les voyelles arrondies ou diffuses (*i, y, u, o, ɔ̃*).

$F = 14,3$ (seuil : 8,10 à 0,01, v = 1, v' = 20).

Ces erreurs sont concentrées sur les mots : dette, tâte, dans, teinte* pour lesquels le taux d'intelligibilité atteint à peine 52,3 %, ce qui signifie que les sujets répondent au hasard.

Ces erreurs sont spécifiques de la caractéristique aiguë (-2) : en effet, dans la caractéristique grave les mots bette, pâte, banc, peinte sont reconnus pratiquement à 100 %.

Le sens vers lequel se réalise la faute montre que le trait Aigu a perdu les indices essentiels qui permettent de le distinguer du trait grave : mais pourquoi avec les occlusives ? et pourquoi devant les voyelles ouvertes non arrondies ?

3.1.1. A cause de la coupure du filtre aux environs de 3700 Hz, le formant de bruit des dentoalvéolaires disparaît pratiquement. On a donc une consonne à faible bruit, c'est-à-dire une consonne apparentée à une labiale.

* Certains de ces mots ne figurent pas sur les listes définitives présentées ici (figure 2a, 2b, 2c, pages B/e. 6, 7 et 8) ; nous avons été amené, à la suite de ces travaux, à corriger certaines paires afin d'équilibrer la répartition des stimuli dans le test.

HOFFMAN* a pu montrer qu'un bruit d'explosion ajouté à une transition quelle qu'elle fût ne permettait jamais la perception des labiales, mais soit d'une palato-vélaire (*k*, *g*), soit d'une dento-alvéolaire (*t*, *d*).

Par contre, pour les constrictives aiguës (*s*, *z*) l'énergie en haut du spectre, qui n'est pas complètement supprimée, s'étale dans le temps. La prolongation dans le temps de cette énergie contribue à augmenter son intensité sur le plan perceptuel.

3.1.2. Pourquoi la confusion devant les voyelles ouvertes non arrondies? (figure 3a, 3b, 3c, 3d, pages B/g/5 et 6).

Il s'agit de voyelles dont *F2* a une valeur proche de celle du locus de dentalité, pour lesquelles par conséquent la transition de *F2* est à peu près neutre et se dirige vers une zone de la consonne située à 1800 Hz. Après (*p*) la transition de *F2* est à peu près neutre, et se dirige vers un locus situé à 1500 Hz environ.

La transition de *F3* est neutre après la consonne aiguë (*t*), elle est fortement négative après (*p*).

La quantification de l'énergie sur les spectres (de 5 en 5 dB) aboutit, en voix reconstituée par le vocodeur, à la suppression des transitions de *F3* pour les deux consonnes.

Or, si on se reporte aux résultats de HOFFMAN (op. cit. p. 1038), on se rend compte que, en l'absence de bruits d'explosion, avec une transition de *F3* neutre et une transition de *F2* également neutre devant la voyelle ouverte (æ), les sujets identifient (*d*) dans 50 % des cas seulement ; dans les autres cas, la consonne est identifiée comme (*b*). Si *F2* devient négatif, le pourcentage d'identification de (*b*) augmente très vite.

Ce type d'erreurs (*t* --> *p*) montre l'importance de l'énergie des hautes fréquences pour identifier correctement les consonnes aiguës diffuses devant les voyelles dont *F2* est proche du locus de dentalité.

2. Les consonnes graves, (caractéristique +2)

	w	b	p	f	v	j	d	t	s	z	total	% de reconnaissance
w	22					0					22	100
b		82					6				88	93,1
p			102					8			110	92,7
f				74					14		88	84
v					47					19	66	71,2
	POURCENTAGE MOYEN DE RECONNAISSANCE											87,40

* Study of some cues in the perception of the voiced stop consonants, J.A.S.A., 30, 11, 1958, p. 1035-1041.

Contrairement à ce qui se passe pour les consonnes aiguës, la fréquence des fautes pour les consonnes graves est significativement plus grande avec les constrictives qu'avec les occlusives.

$F = 8,44$ (seuil : 8,10 à 0,01, v = 1, v' = 20).

D'autre part, 76 % des fautes avec les constrictives ont lieu devant les voyelles arrondies. La fréquence des fautes est significativement plus grande avec les constrictives devant les voyelles arrondies (*u*,*y*) qu'avec les mêmes consonnes devant les autres voyelles.

$F = 13$ (seuil : 8,10 à 0,01, v = 1, v' = 20).

Le système des fautes est donc inversé par rapport à la caractéristique aiguë.

Les mots les plus perturbés sont furent (60 % de reconnaissance) et vous (37 % de reconnaissance) qui sont confondus respectivement avec surent et zou.

Les résultats des travaux de HEINZ et STEVENS* montrent qu'au contact des voyelles arrondies la consonne(*s*) possède un premier pôle d'énergie entre 3500 et 4000 Hz au lieu de 5000 au contact des autres voyelles ; ils montrent également qu'au premier pôle d'énergie vers 7000 Hz pour la consonne (*f*) s'ajoute une résonance aux environs de 3000 Hz qui se traduit par une augmentation de 15 dB du niveau des fréquences moyennes par rapport à celui de (*f*) devant les voyelles non arrondies. Nous relevons les mêmes caractéristiques pour nos consonnes (*f*, *v*) et (*s*, *z*).

Etant donnée la bande passante du Vocodeur, cette énergie se trouve au sommet du spectre relatif de la consonne reconstituée (figure 4, page B/g/11) ; ainsi l'énergie de (*f*) se trouve à peu près à la même place que celle de (*s*) reconstitué avec une intensité voisine. Or on sait que l'un des indices qui différencient ces deux consonnes est précisément la différence d'intensité globale qui est très faible pour (*f*).

Il reste un indice qui permettrait de distinguer la consonne grave de la consonne aiguë : c'est la présence pour (*f*, *v*) d'un formant de bruit vers 1300 Hz qui n'existe pas pour (*s*, *z*). Cet indice évite peut-être que (*f*, *v*) soient perçus dans tous les cas comme (*s*, *z*), mais il ne suffit pas à assurer l'identité des consonnes graves.

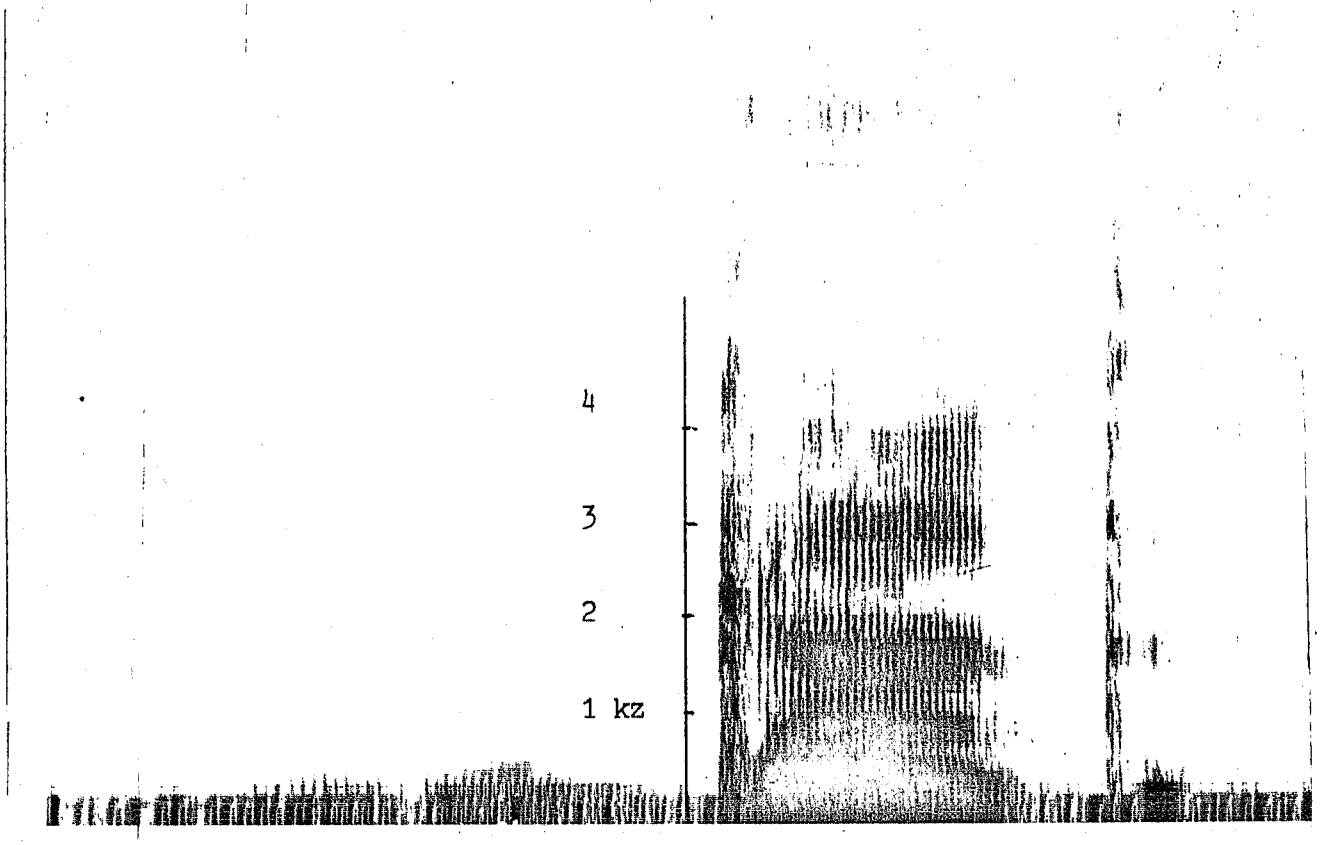
Nous avons également testé l'opposition complexe *(v ∿ z)* = / grave + diffus ∿ aigu + compact /, dans les mots vu et jus, c'est-à-dire encore dans un contexte de voyelle arrondie. Cette opposition est encore plus perturbée que l'opposition simple / aiguë + (diffus) ∿ grave + (diffus) /, car la confusion a lieu dans les deux sens avec un taux d'intelligibilité de 50 %.

Les deux consonnes reconstituées présentent le même spectre : un formant vers 1600 Hz relié au deuxième formant de la voyelle et une résonance vers 3000 Hz (résonance que l'on trouve dans tous les cas pour les consonnes compactes */ʃ, ʒ/*) reliée aux 3ème et 4ème formant de (*y*).

* On the properties of voiceless fricative consonants, J.A.S.A., 33, 5, p. 589-596.

FIGURE 4.a - Sonagramme de quinte : voix naturelle, locuteur IBM

O quinte 158



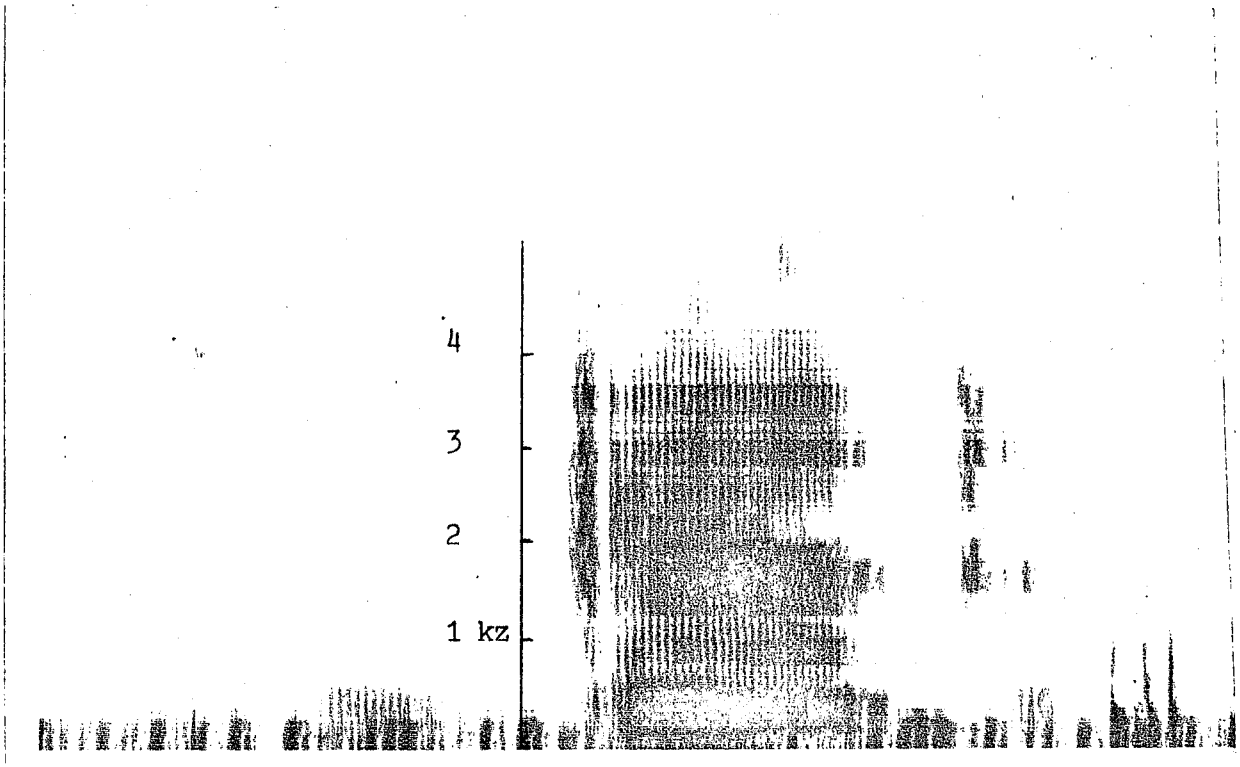
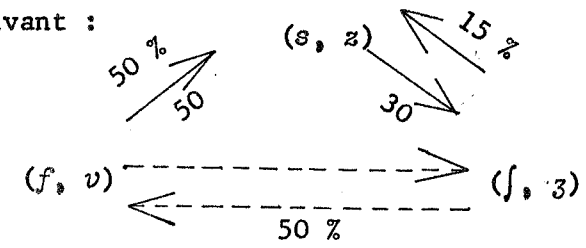


FIGURE 4.b - Sonagramme de quinte : voix reconstituée sur vocodeur IBM

V quinte 158

Il semble que l'opposition Compact / Diffus soit par ailleurs mal assurée. En effet, on note un taux d'intelligibilité relativement bas pour la consonne constrictive diffuse aiguë (*s*) qui est confondue dans 30 % des cas avec (ʃ).

On a donc le système de fautes suivant :



Il ressort de ce tableau que ce sont les consonnes graves diffuses (*f*, *v*) qui sont les plus atteintes : les consonnes graves sont particulièrement mal définies et se confondent en fait les aigues compactes dans un contexte de voyelles arrondies.

Nous avions proposé aux auditeurs un choix forcé entre deux possibilités : ils devaient choisir entre furent et surent et entre vous et zou ; mais si nous leur avions proposé un triple choix entre furent, surent et churent, combien de réponses se seraient reportées sur le mot surent ? Peut-être aucune, comme semble le prouver le caractère unilatéral des fautes pour *f* et *s* (*f* → *s*).

Le choix forcé entre deux possibilités ne semble donc pas être une méthode adéquate pour tester l'intelligibilité de certaines oppositions, notamment les oppositions de lieu d'articulation qui sont des oppositions à trois termes. Il conviendrait, dans ce cas, de proposer aux auditeurs un choix entre trois possibilités. On aurait une image plus nette du système de fautes.

3.2. Compact/diffus : La fréquence des fautes est significativement différente pour les consonnes compactes et pour les consonnes diffuses. Les consonnes compactes sont plus perturbées que les consonnes diffuses :

$F = 5{,}21$ (seuil : 4,35 à 0,05, v = 1, v' = 20).

1. Les consonnes diffuses :

	l	t	z	p	b	d	s	r	k^i	ʒ	k^o	g^o	g^i	ʃ	Total	% reconnaissance
l	65							1							66	98,4
t		62							4						66	93,9
z			41							3					44	93,1
p				60							6				66	90,9
b					56							10			66	84,8
d						36							8		44	81,8
s							31							13	44	70,4

POURCENTAGE MOYEN DE RECONNAISSANCE 88,64

Le nombre des erreurs est plus grand avec les occlusives sonores (65 %) qu'avec les occlusives sourdes. Mais la répartition des fautes entre les occlusives sonores et les occlusives sourdes n'est pas significative.

Nous avons déjà parlé des erreurs qui concernent (s) confondu avec (ʃ). Nous ajouterons simplement que (s) n'est identifié que dans un contexte de voyelle aiguë lorsque son premier pôle d'énergie est élevé (vers 4900 Hz). Dans ce cas, la partie analysée et reconstituée par le vocodeur forme une bande étroite à l'extrémité du spectre relatif. Etant donné que d'autre part, contrairement à (ʃ), (s) ne possède pas de formant de bruit au niveau du 2ème formant de la voyelle, la consonne diffuse est nettement distincte de la consonne compacte, du moins devant les voyelles antérieures non arrondies.

2. Les consonnes compactes :

	r	g°	k°	ʃ	ʒ	gⁱ	kⁱ	l	b	p	s	z	d	t	total	% de reconnaissance
r	62							4							66	93,9
g°		62							4						66	93,9
k°			59							7					66	89,3
ʃ				38							6				44	86,3
ʒ					36							8			44	81,8
gⁱ						31							13		44	70,4
kⁱ							40							26	66	66,6
	POURCENTAGE MOYEN DE RECONNAISSANCE :															82,83

Pour l'opposition compact/non compact la fréquence des fautes est significativement plus grande avec les aiguës (k^i, g^i) qu'avec les graves (k^o, g^o) : $F = 30$ (seuil 8,10 à 0,01, v = 1, v' = 20)

On aurait pu s'attendre à rencontrer davantage d'erreurs avec les compactes graves (k^o, g^o) dans un contexte de voyelles arrondies. En effet, DELATTRE a montré que seul le bruit d'explosion pouvait permettre de distinguer (g) de (b), le mouvement des transitions de la voyelle arrondie subséquente étant identique dans les deux cas. Mais étant donné la fréquence d'échantillonnage du Vocodeur (tous les 5/1000 s. pour les consonnes sourdes et à chaque impulsion laryngée pour les sonores) les spectres d'explosion de (k, g) sont bien analysés et reproduits.

La plus grande partie des erreurs a lieu avec les compactes (k, g).

Les fautes sont concentrées sur les mots : quinte, guigne, cric, confondus avec teinte, digne, trique ; le taux d'intelligibilité s'élève respectivement à 64 %, 46 % et 27 %.

Le bruit d'explosion de (k, g) se situe entre 2500 et 3500 Hz. Les deuxième et troisième formants de la voyelle adjacente convergent vers cette zone (figure 4a).

La quantification des spectres perturbe le mouvement caractéristique des transitions au contact des consonnes compactes (k, g). Le spectre de l'explosion est bien reproduit, mais les transitions disparaissent ; les formants sont droits (figure 4b). Or d'après HOFFMAN (op.cit., p. 1038), un bruit situé entre 2800 et 3400 Hz accompagné d'une voyelle dont les transitions de formants ont une pente nulle est perçu comme (d) dans 80 % des cas.

CONCLUSION.

1. Ce test, outre le diagnostic de certaines déficiences des Vocodeurs qu'il rend possible, met en lumière l'importance du contexte vocalique dans l'intelligibilité des consonnes. Il fait ressortir l'importance - selon le contexte et selon les classes de consonnes - des transitions et des formants de bruits qu'on avait eu tendance à négliger après les travaux de Groupe HASKINS.

Avec les occlusives, la direction des transitions est essentielle pour la distinction compact / diffus : (*t*, *d*) ∿ (*k*, *g*) et l'identification du trait de compacité.

Par contre, dans un contexte vocalique où le jeu des transitions est mal défini (voyelles ouvertes non arrondies), la présence d'un bruit dans le haut du spectre permet seule d'assurer l'intelligibilité des consonnes aiguës diffuses en face des consonnes graves diffuses (*t*, *d*) ∿ (*p*, *b*).

Avec les constrictives, la distribution de l'énergie sur le spectre prend le pas sur le mouvement des transitions : surtout devant les voyelles arrondies.

/*f*/ = formant de bruit à 1500 Hz + zéro acoustique entre 2000 et 6000 Hz.

/*s*/ = absence de formant de bruit à 1500 + Pôle d'énergie à partir de 4000 Hz.

/*ʃ*/ = formant de bruit à 1500 Hz + Pôle d'énergie entre 2500 et 3500 Hz.

2. On pouvait craindre que, d'autre part, le choix des auditeurs fût conditionné par la féquence d'occurrence des mots dans la chaîne parlée. En fait, nous avons remarqué que les fautes affectent dans une même proportion les mots rares et les mots fréquents.

POLLACK, RUBENSTEIN et DECKER *; comme l'a fait remarqué VOIERS, affirment que la fréquence d'occurrence du mot n'a virtuellement aucun effet sur l'intelligibilité lorsque les choix offerts à l'auditeur sont définis pour chaque stimulus.

Les erreurs caractéristiques dont nous avons rendu compte jusqu'ici (figure 5) ne sont liées ni à la fréquence d'occurrence des consonnes, ni à celle des suites CV dans la chaîne parlée.

Ainsi, dans le contexte (*y*), (*s*) a tendance à être confondu avec (*ʃ*), alors que les suites (*sy*) sont plus nombreuses et plus significatives que les suites (*ʃy*).

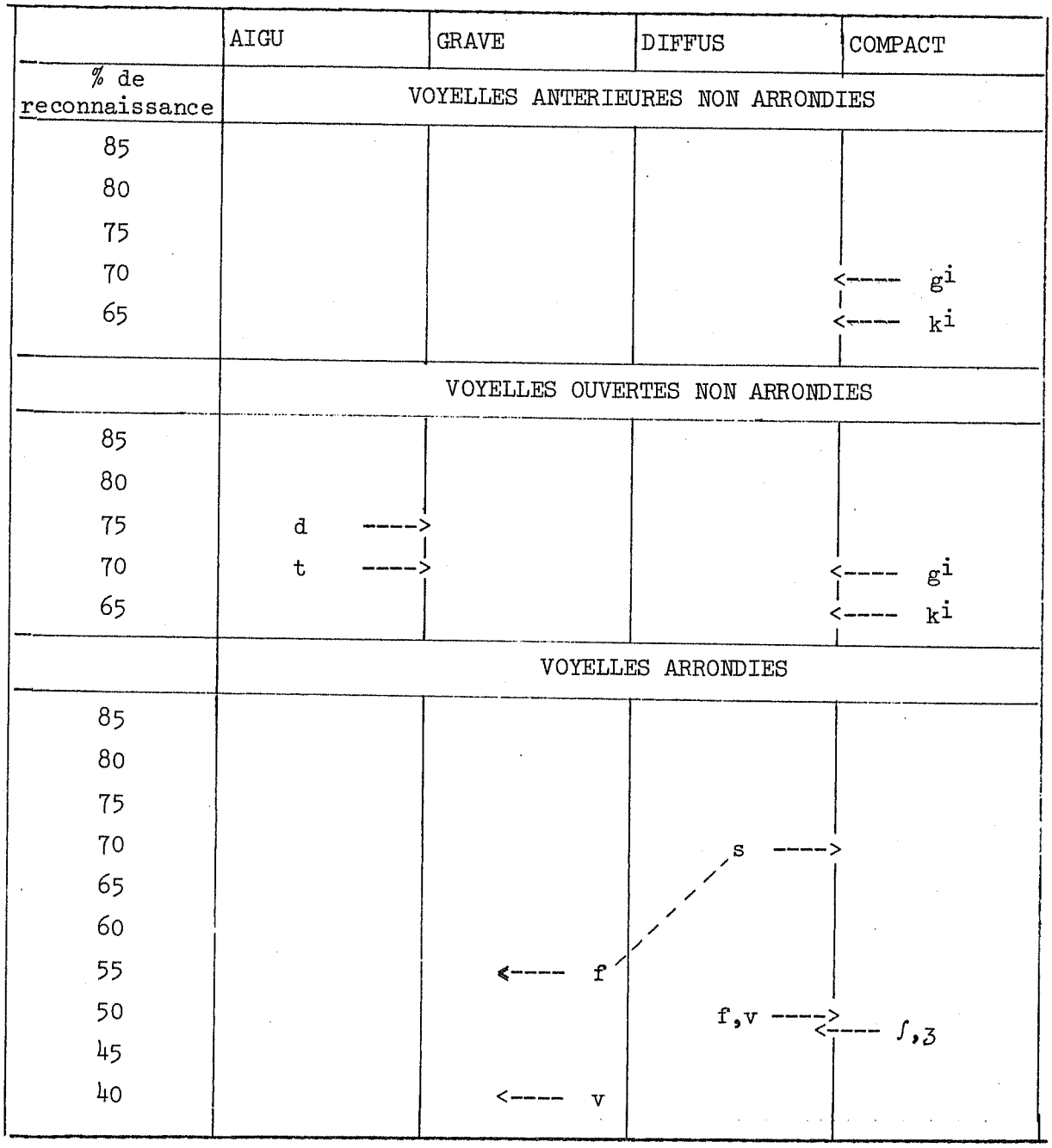
Une analyse des suites CV dans un texte de 100.000 occurrences de phonèmes a montré qu'il n'existe aucune corrélation entre la fréquence des fautes et la fréquence d'apparition des couples CV, entre la fréquence des fautes et la valeur significative d'apparition des couples CV.

3. Enfin, les résultats de cette expérience sur l'intelligibilité de la parole reconstituée par Vocoder nous incitent à des révisions déchirantes :

3.1. L'opposition /*n* ∿ *d*/, il convient de le répéter, n'est pas une opposition minimale : elle est phonétiquement complexe. Par contre /*n* ∿ *l*/ est la seule opposition minimale de nasalité, comme le prouvent par ailleurs les nombreuses confusions qui affectent ces deux consonnes sur les plans diachronique et synchronique.

* Intelligibility of Known and unknown message sets, J.A.S.A., 31, 1959, p. 273-279.

FIGURE 5 - Tableau des erreurs caractéristiques



3.2. Il conviendrait peut-être de réinterpréter les traits qui rendent compte des oppositions de lieu d'articulation :

l'opposition /v ∿ ʒ/ , par exemple, est-elle phonétiquement plus complexe que l'opposition /v ∿ z/ comme l'admettent et la taxonomie de JAKOBSON et celle que CHOMSKY et HALLE* ont récemment proposée ? D'après les derniers travaux de DELATTRE, déjà cités, il semble que v et ʒ soient acoustiquement plus proches que ne le sont v et z. Ce fait expliquerait le nombre apparemment insolite d'erreurs entre deux consonnes pourtant séparées par deux traits acoustiques.

* The sound pattern of english
Harper and Row, New York, 1968, p. 307.

LE TEST DE DIAGNOSTIC PAR PAIRES MINIMALES

cinquième partie

INTERPRETATION DES RESULTATS EN VUE D'ETABLIR UN DIAGNOSTIC

J. P. PECKELS

Cie IBM France - Centre d'Etudes et Recherches - LA GAUDE

1. Introduction

Nous avons dit, au début de cet exposé, que le test par paires minimales est un outil puissant permettant, lors du développement d'un vocodeur, de mesurer l'influence d'une modification d'une partie des circuits par exemple. Nous avons dit aussi que le test en question permet de vérifier les réglages d'un vocodeur une fois, construit, et de déterminer l'influence d'une variation de certains paramètres affectant ces réglages.

Le problème est toutefois différent selon que l'on veut, au moyen de ce test, situer un vocodeur, le classer par rapport à des résultats connus ou bien essayer de savoir comment on pourrait l'améliorer en modifiant telle ou telle partie des circuits, tout en conservant un bon rapport coût/performance.

Le premier cas est trivial ; il suffit de comparer ses résultats caractéristique par caractéristique, aux tableaux publiés montrant des résultats moyens.

Grâce aux travaux de W. D. Voiers[2,3] et de C. P. Smith,[4] ces tableaux comparatifs existent pour un certain nombre de vocodeurs (en langue américaine), et un des buts de nos travaux futurs sera de publier les équivalences en Français.

Si, par contre, à la vue des résultats que l'on obtient par rapport à ceux des autres, on s'aperçoit qu'on est dans la moyenne ou même au-dessus en ce qui concerne un certain nombre de ces caractéristiques, mais que pour d'autres le taux d'erreurs est nettement plus élevé, la question se pose :

dans quelle partie du circuit chercher, que modifier et comment ?

Comme chaque cas qui se présente dans la pratique est différent, on ne peut donner que des grandes lignes, et montrer quelles relations sont connues à ce jour entre les différentes caractéristiques et la conception du vocodeur que l'on est en train d'essayer. Par ailleurs, des relations univoques entre les résultats de test dans les différentes caractéristiques ne sont pas connues et il faudra encore un grand nombre d'essais avant de mieux pouvoir définir des remèdes pour chaque cas.

Nous allons examiner quelques constatations de W. D. Voiers et C. P. Smith, constatations que l'on peut regrouper comme suit : (il est utile de rappeler que ces auteurs ont examiné les résultats de test obtenus avec des stimuli en langue américaine).

2. Constatations de W. D. Voiers et C. P. Smith[2,3,4]

2.1. Résultats moyens sur des vocodeurs à 2400 bits/seconde

Il apparaît que ces vocodeurs -qui transmettent l'information sous forme numérique, quantifiée et échantillonnée dans le temps- rendent mal les consonnes graves (taux moyen d'erreurs : 18 %) ainsi que les consonnes interrompues (taux moyen d'erreurs : 23 %), les autres caractéristiques étant beaucoup moins affectées.

2.2. Influence du taux d'échantillonnage et de la quantification de l'énergie

D'après Voiers[2,3] les défauts notoires ainsi constatés sont dus aux taux d'échantillonnage, taux communément utilisés à cause d'impératifs techniques.

Dans un vocodeur de type analogique, dans lequel les paramètres ne sont, par définition, pas quantifiés, et où la notion de vitesse d'échantillonnage disparaît, les essais ont montré que les consonnes graves sont bien perçues, alors que les plosives continuent à être mal comprises. D'après Smith[4], la cause serait à chercher dans les filtres passe-bas utilisés lors de l'analyse, filtres qui ont pour effet une perte de résolution de l'information temporelle.

3. Conclusions

En ce qui concerne les consonnes graves et interrompues, les erreurs seraient dues -d'après ce que l'on vient de voir- au taux d'échantillonnage et, pour les plosives, également aux filtres passe-bas de l'analyseur.

En ce qui concerne les consonnes voisées, un taux d'erreur anormal nous conduit directement au détecteur de mélodie et au circuit de décision voisé/non voisé.

Le remède à trouver devient nettement plus complexe lorsqu'il s'agit des autres caractéristiques.

Smith a bien trouvé une certaine relation dans l'importance des formants les uns par rapport aux autres en ce qui concerne les caractéristiques grave, compacte et nasale, ce qui amènerait à prévoir des pas de quantification différents selon les régions formantiques caractérisées chacune par plusieurs filtres contigus. On est cependant loin d'une relation univoque entre des erreurs anormales produites par telle caractéristique et la partie physique du vocodeur à incriminer.

En plus, comme l'a montré M. Rossi, nos premiers résultats ne concordent pas avec ceux publiés et concernant la langue américaine.

Afin d'éclaircir les problèmes qui restent à la fois chez Voiers et en Français, il serait important d'avoir un vocodeur extrêmement souple, par exemple simulé sur un ordinateur sur lequel on pourrait d'une façon systématique changer des paramètres et voir l'influence sur chaque caractéristique du test. Ce n'est qu'après ces essais systématiques que l'on pourrait trouver des relations plus distinctes entre les différentes caractéristiques et les circuits physiques du vocodeur.

Ceci toutefois n'enlève rien à la puissance et la facilité d'utilisation immédiate du test de diagnostic par paires minimales, en vue de comparer entre eux différents vocodeurs et également de voir l'influence, au cours du développement, d'une modification de circuits sur les différentes caractéristiques ainsi examinées.

Le test aidera ainsi le constructeur du vocodeur à optimiser le rapport coût/performance, en jouant sur les taux d'information attribués à chaque paramètre.

B I B L I O G R A P H I E

1. W. D. Voiers, "Performance evaluation of speech processing devices, III : Diagnostic evaluation of speech intelligibility," AF Cambridge Research Labs, Final Rept., Contract AF19 (628)-4987, AFCRL-67-0101, 1967. AD 650 158.

2. W. D. Voiers, "The Present State of Digital Vocoder Technique : A Diagnostic Evaluation," IEEE Trans. Au-16, 275-279 (1968).

3. W. D. Voiers, "Diagnostic Evaluation of Speech Processing Systems," Final Rep. AFCRL-68-0568, Cont. F19628-68-C-0068, Tracor, Inc. (1968). AD 684 608.

4. C. P. Smith, Perception of Vocoder Speech by Pattern Matching. J.A.S.A. Vol. 46, N° 6 (Part 2), pp. 1562-1571, December 1969.

M. ASTIER à J.P. PECKELS

MM. PECKELS et ROSSI ont dit que le test ne pouvait être utilisé qu'avec un minimum de qualité de la transmission. Cela sous-entend-t-il que l'on ne pourra pas tracer la courbe d'intelligibilité fonction du rapport signal à bruit ? Si oui, en ce qui concerne les télécommunications sous-marines, cela limite énormément l'intérêt du test, car nous avons pu voir dans la pratique l'intérêt de comparer le rapport signal à bruit de deux systèmes de transmission pour une intelligibilité identique et égale à 50 %.

J.P. PECKELS - M. ROSSI

La qualité des vocoders est aujourd'hui telle que les tests conventionnels - par listes de mots dissyllabiques par exemple - fournissent des résultats d'intelligibilité de l'ordre de 90 %.
C'est lorsqu'on veut examiner en détail les défauts *restants* de ces vocoders que le DRT (Diagnostic Rhyme Test) révèle toute sa puissance.
Toutefois, ceci ne veut pas dire que le DRT - ainsi que sa version française - ne peuvent pas être utilisés dans un système de transmission de parole à faible rapport signal à bruit. En effet, comme W.D. VOIERS l'explique, certaines caractéristiques sont plus affectées que d'autres par la présence de bruit et par conséquent ce test permet de suivre d'une façon précise le comportement de chacune des caractéristiques en fonction du rapport signal à bruit.
L'utilisation du DRT est à déconseiller toutefois si la qualité de transmission est telle - dû à la dégradation de tous les sons, y compris les voyelles - que les auditeurs sont dans l'impossibilité de faire un choix valable parmi les deux mots qui leur sont proposés à l'écoute de chaque stimulus. C'est dans ce sens-là que la remarque du conférencier doit être interprètée.

W.D. VOIERS

I am unaware of the basis on which MM. PECKELS and ROSSI make the statement that the DRT can be used only if the quality of the transmission is above a minimum level. (At least, I am unaware that the DRT is uniquely limited in this respect). I cannot, therefore, respond in detail to their statement. I would be inclined, however, to suggest that the DRT can prove especially useful under circumstances of extreme signal impoverishment since the various diagnostic scores tend not to be equally affected by noise, frequency distortion, ...

In any case, I see no reason why the DRT cannot be used to study the effects of S/N ratio as exemplified by the enclosed figure. It might be noted, however, that conventional techniques of evaluating S/N ratio (VU peaks) may not be entirely appropriate in circumstances where primary concern is with consonants sounds. The results presented in the enclosed figures were obtained following a preliminary investigation of speaker differences under *nominally* equivalent S/N conditions (i.e., VU vowel peak/VU noise). Data from this preliminary investigation provided a basis for adjusting speech levels of individual speakers such that all six speakers involved yielded nearly identical *total* DRT scores, as averaged over a range of S/N conditions. Presumably the primary effects of these adjustments were to compensate for individual differences in voiced/unvoiced energy ratio. In any case, such normalization procedures greatly facilate examination of speaker differences in DRT score patterns, which consideration motivated the present investigation.

M. CARTIER à MM. VOIERS, PECKELS et ROSSI

Peut-on s'affranchir de la variable locuteur sans allonger exagérément la durée des tests (voir par exemple le test de PREUSSE) ?

W.D. VOIERS

We have been continuously aware of the problem posed by the effects of speaker variability upon the results of intelligibility tests in general and the Diagnostic Rhyme Test in particular. We have compared the patterns of diagnostic scores obtained from a fairly large number of speakers under a variety of experimental conditions and found systematic differences among speakers in this respect. Some of these differences are dependent, however, upon the type of speech processing involved. For example, other things equal, low pitched male speakers tend strongly to yield higher DRT scores than higher pitched speakers in cases involving pitch-excited vocoders. This tendency is not evident, however, with other types of speech degradation.

We are currently investigating the effects of the speaker's characteristic voiced/unvoiced energy ratio, his voiced/unvoiced time ratio, ..., upon his DRT score pattern unter various conditions of signal impoverishment. Results of these investigations are not yet available, however.

For the present, therefore, our solution to the practical problems of speaker selection is to examine DRT score patterns for a large number of speakers under various experimental conditions and to select one or a small number who exhibit the score patterns most typical of the group as a whole. These speakers are then used for routine testing and research purposes.

Since the DRT as conventionally employed in the past involves the use of *two* utterances of all test materials, it is quite feasible to use one speaker for one half of the test and another speaker for the other half. Our current preference is to use *three* speakers, each of whom provides one utterance of all of the DRT test words. This can be accomplished within a test period of approximately 15 minutes, as opposed to the 10 minutes previously required for a "double administration" with one speaker. Still more speakers can of course be used (at a cost in time of approximately 5 minutes per speaker) depending upon the wishes of the experimenter. For example, in the forthcoming survey of speech processing devices, which we are conducting for the 1972 AFCRL-IEEE speech conference, we utilize a total of six speakers (two of low pitch, two of high pitch and two of average pitch) which thus requires approximately 30 minutes of test time.

J.P. PECKELS - M. ROSSI

Les résultats des tests semblent montrer que, pour des voix d'hommes, le pourcentage et la répartition des fautes ne sont pas significativement différents. Cependant, il faut être prudents : les résultats sont comparables si l'on choisit des locuteurs de même âge, de même sexe et de même niveau social, appartenant à la même région linguistique. L'âge et le sexe entraînent des différences importantes de fréquence fondamentale. Or, l'on sait que pour un Fo élevé, le spectre du signal de la parole est pauvre ; s'il est vrai, comme l'a montré CARRÉ (Contribution aux études sur l'analyse et la synthèse de la parole, Grenoble - 1971, pp. 63-69), que, pour une voix de femme, l'essentiel de l'information sur la fonction de transfert est donné par la transition d'amplitude au début de la phonation, on peut s'attendre à ce qu'une voix de femme soit fortement perturbée par un vocoder.

Les variantes régionales et sociales peuvent avoir une incidence au niveau de la réalisation des traits acoustiques ; la même opposition de consonnes peut reposer sur des traits qui diffèrent d'une région à l'autre (voir ce que nous avons dit sur le polymorphisme de r) ou qui se réalisent différemment selon les régions : ainsi l'opposition de sonorité à la fin du mot en français *non* méridional (ex. t/d, dans patte/rade) est préparée par une différence de durée vocalique (voyelle brève devant la sourde, voyelle relativement longue devant la sonore). Cet indice qui joue un rôle important dans la perception de l'opposition de sonorité n'existe pratiquement pas en français méridional à cause de la réalisation de la voyelle latente finale.

M. CARTIER à MM. PECKELS et ROSSI

Le test par paires minimales réduit la difficulté de choix. Pensez-vous qu'il soit souhaitable, dans certains cas, d'augmenter sa sensibilité ? Si oui, comment (bruit à l'entrée, bruit en sortie ou autre méthode) ?

J.P. PECKELS et M. ROSSI

Tout dépend du but poursuivi. Si, comme l'indique le titre du test, l'on veut établir un diagnostic sur le fonctionnement du vocoder, on ne peut, c'est évident, apporter au signal une distorsion supplémentaire qui n'existe pas dans le vocoder.

Si, par contre, on cherche à tester l'intelligibilité des traits qui assurent l'identité des phonèmes, alors on peut, comme l'ont fait pour l'anglais *MILLER* et *NICELY* (J.A.S.A. 27.2.1955), filtrer le signal ou introduire un bruit dans telle ou telle bande de fréquence.

M. *ROSSI* pense (il y a d'ailleurs fait allusion) qu'il est souhaitable d'introduire pour les oppositions de lieu d'articulation (p/t/k) un choix multiple. Un triple choix augmenterait dans ce cas la sensibilité du test et supprimerait certaines ambiguïtés dans les résultats.

M. *ROSSI* propose en outre une méthode qui devrait nous apporter des renseignements intéressants : elle consisterait à décomposer les traits en indices acoustiques et à masquer certains de ces indices. On pourrait ainsi établir un diagnostic plus fin sur le fonctionnement des vocoders et hiérarchiser les indices qui entrent dans la définition des traits.

COMMUNICATIONS PRESENTEES AU COURS DE LA MATINEE DU 2 AVRIL 1971

ALLOCUTION D'INTRODUCTION

prononcée par Monsieur *J. A. DREYFUS-GRAF*
Président de Séance

Parmi tous les domaines de la communication parlée, la reconnaissance automatique de la parole est le terrain - j'allais dire le sable - le plus mouvant.

En effet, contrairement aux formes optiques, les formes acoustiques ne présentent pas de contours univoques et il faut d'abord les délimiter.

Comme toujours, quand il s'agit d'explorer un terrain relativement vierge, il y a les pessimistes et les optimistes. Par exemple, M. PIERCE, des Laboratoires BELL, écrivait en 1970 :
" Il est tout simplement impossible de reconnaître l'anglais phonème " par phonème ou mot par mot. "

Par contre, la Revue 0-1-Informatique rapportant, en décembre 1969, le résultat suivant d'un sondage d'opinion parmi les spécialistes :
" Les entrées vocales deviendront possibles, dans la pratique, en 1979.
" A ce moment, la carte perforée et la bande perforée auront vécu en " tant que supports de communication. "

Si nous avons fait le déplacement jusqu'à Aix-en-Provence, c'est, d'une part pour retrouver des amis, d'autre part, je pense, parce que nous appartenons au clan des optimistes les plus impénitents.

D'une manière générale, on peut distinguer sept catégories d'option, ou de piliers, sur lesquels on peut construire des machines obéissant à la parole, à savoir, successivement :

. *les transducteurs d'entrée,*

. *les présélecteurs d'informations pertinentes,*

. *les intégrateurs,*

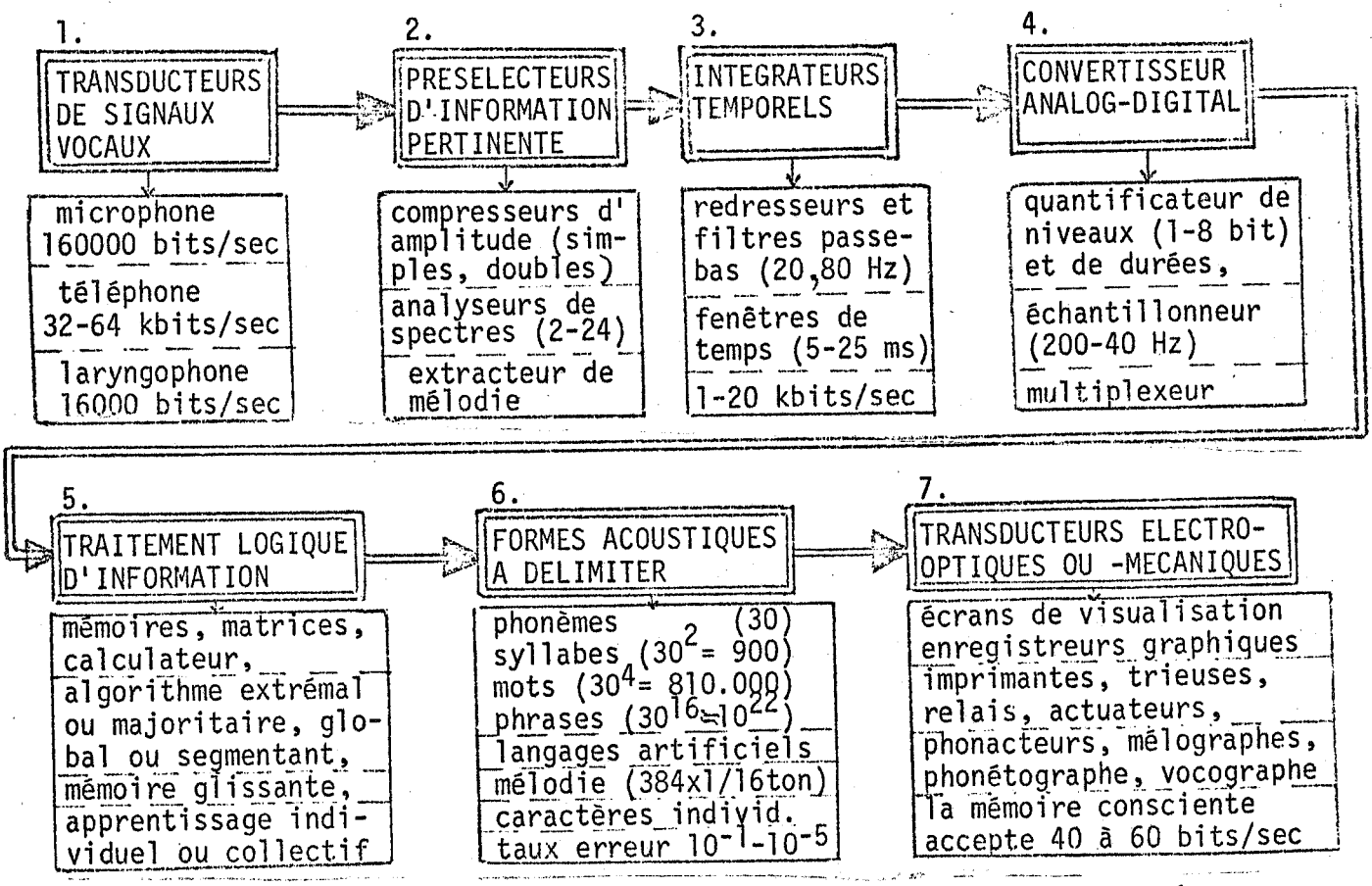
. *les convertisseurs analogues - digitaux,*

. *la logique,*

. *les formes acoustiques à délimiter,*

. *les transducteurs de sortie,*

(voir schéma général de machines obéissant à la parole, page suivante).



Les sept catégories d'options mentionnées réagissent les unes sur les autres et forment un tout. Par exemple, la logique et les formes acoustiques seront différentes selon que le transducteur de la parole sera un microphone dynamique passant 50 à 8000 Hz, une ligne téléphonique admettant 300 à 3400 Hz, ou un laryngophone transmettant 80 à 1500 Hz.

Néanmoins, chaque élément d'un tout mérite son examen séparé. En particulier, la catégorie n° 2 (Présélecteurs d'information pertinente) a suscité deux écoles, caractérisées par les conclusions suivantes :
M.P. VINCENT estime qu'il est plus important d'augmenter la puissance d'analyse syntaxique des sentences que de perfectionner les techniques de prétraitement.
Par contre, M. R.W. SCARR écrit : " Il est plus important de perfectionner l'extraction des traits pertinents, dans un système basé sur la reconnaissance des phonèmes, que d'optimiser les stratégies de décision. "

Il existe bien d'autres sujets de controverse, tels que : Faut-il chercher à reconnaître des phonèmes, des syllabes, des mots ou des phrases ? en temps réel ou en temps différé ? Faut-il prévoir des phases d'apprentissage individuel ou collectif ? par la machine seule ou par l'usager ? pour chaque mot prononcé par chaque locuteur ? pour l'ensemble des phonèmes et des prononciations normalisées ? Est-il préférable de choisir des algorithmes de décision basés sur des calculs d'extremum ou de majorités ? globaux ou segmentés ? Quels sont les taux d'erreurs admissibles ? 10 %, 1 %, voire 10^{-4} ? Quelles sont les mesures de protections correspondantes à adopter ? ...

En ce qui me concerne, je pense que la catégorie d'option la plus importante est la PRESELECTION DES INFORMATIONS PERTINENTES et qu'il faut placer d'emblée la machine dans les conditions physiologiques de l'audition où presque rien n'est linéaire.

Je vais passer maintenant la parole aux divers conférenciers. Nous allons chercher ensemble à reconnaître dans leurs paroles ce qui peut contribuer à la reconnaissance de celles-ci. Peut-être que, d'ici quelques années, les magnétophones qui nous enregistrent en ce moment seront complétés par des machines à écrire automatiques.

VOCODER NUMERIQUE

M. LAVANANT
Sté Lannionnaise d'Electronique—LANNION

I. Introduction .

Cet exposé est une présentation rapide des études entreprises à LANNION concernant le vocoder numérique.

Les raisons d'une telle étude sont multiples. A l'origine on trouve le projet A.S.P.I.C. (Analyseur synthétiseur de Parole avec Informations Codées) qui a été étudié et développé au CNET-LANNION (Département ETA). Ce vocoder est utilisé dans des études de reconnaissance de la parole et il a trouvé une application industrielle dans le projet DECLAM de la CIT. Il s'agit là d'un projet de diffusion des renseignements météo. Les autres raisons sont des conséquences de l'expérience acquise dans la technique du filtrage numérique à partir de 1967. Des programmes d'analyse et de synthèse de filtres ont été écrits par M. EL MALLAWANY au centre de calcul du CNET-LANNION. De plus, un programme appelé OSIRIS et qui est l'équivalent du BLODIB américain, permet la simulation de traitements numériques de signaux. De son côté, la S.L.E. (Société Lannionnaise d'Electronique, filiale C.G.E.) a étudié et réalisé un récepteur numérique de fréquences pour postes à clavier dans le cadre du projet PLATON. Ces différentes études nous ont amené à envisager, pour un vocoder, une solution entièrement numérique.

II. Etude d'un vocoder numérique à canaux.

L'étude s'est faite par simulation sur calculateur au moyen de 3 programmes principaux :

- analyse et détection de pitch (FFT)
- simulation d'un vocoder à bande de base.
- simulation d'un vocoder à canaux et excitation par pitch-bruit.

Les programmes ont été utilisés sur calculateur CII 90-80 puis sur CII 10070. De plus un convertisseur A/D et D/A est raccordé au calculateur Ramsès du CNET.

La détection du pitch se fait par le procédé "cepstrum" qui paraît être le mieux adapté au cas où le signal de parole est limité à la bande téléphonique 300 - 3400 Hz.

L'intérêt du vocoder à bande de base (ou à excitation vocale) est assez limité. Le taux de compression atteint est faible - de l'ordre de 10 - et correspond à un débit de 8000 bits par seconde.

Par contre, dans le cas du vocoder à canaux et excitation par pitch et bruit, le débit est de 2400 à 3200 bits par seconde et se prête beaucoup mieux aux applications de synthèse de la parole.

III. Conclusions de l'étude et réalisations envisagées.

Les programmes de simulation ont prouvé la possibilité de réaliser d'une manière numérique tous les traitements que l'on trouve dans un vocoder : filtrage passe-bande, filtrage passe-bas , génération d'impulsions périodiques et de bruit blanc. Un vocoder entièrement numérique devient ainsi particulièrement adapté aux réseaux téléphoniques tels que PLATON où la transmission et la commutation se font en MIC (Modulation par Impulsions Codées).

Différentes réalisations sont envisagées :

a) un analyseur FFT traitant un signal de parole échantillonné à 16 kHz et ayant 4 sorties :
   - le spectre à bandes étroites
   - le cepstrum $c(\tau)$
   - le spectre à larges bandes
   - la transformée inverse de ce spectre pour application à la convolution,

b) un analyseur de vocoder à canaux avec possibilité de modifier le nombre de canaux et leurs caractéristiques,

c) une unité à réponse vocale (synthétiseur) réalisant la synthèse de plusieurs voies temporelles (multiplex).

Dans tous ces projets, nous nous efforçons d'utiliser des cartes standards. L'unité de base des réseaux de calcul est une carte réalisant l'opération $A.X + B.Y$ en moins de 500 ns ; elle utilise des circuits TTL de la série 74N. Les parties de mémoires rencontrées en filtrage numérique ou en analyse FFT se présentent sous forme de registres à décalage ou de registres adressables. La technologie MOS est très intéressante grâce à son grand degré d'intégration et à sa faible consommation et il est très simple de réaliser en registres à décalage MOS une ligne à retard de T période d'échantillonnage du signal traité qui correspond à l'opérateur z^{-1} du filtrage numérique.

IV. Conclusion.

En conclusion, on peut affirmer qu'il existe actuellement des solutions numériques au problème du traitement du signal de parole. Les avantages du numérique sont essentiellement la stabilité de fonctionnement et la souplesse d'utilisation qui permet de modifier les caractéristiques de l'appareil. La constante diminution du coût des circuits intégrés devrait normalement favoriser le développement du traitement numérique des signaux.

PRETRAITEMENT ET RECONNAISSANCE DE LA PAROLE
SIMULATION ET REALISATIONS PRATIQUES

J.P. HATON , M. LAMOTTE
Laboratoire d'Electricité et d'Automatique - NANCY

Le groupe de recherches sur la parole du L. E. A. de Nancy s'intéresse à la reconnaissance de la parole et à ses problèmes annexes depuis 1967.

I - LES OPTIONS.

Rappelons les grandes options prises dès le départ, dictées pour la plupart par l'idée de réaliser un système hardware de reconnaissance. [1] [2]

a) Analyse fréquentielle de la parole par un analyseur hardware comportant un banc de 24 filtres dont les fréquences centrales s'échelonnent de 100 Hz à 7000 Hz. La fréquence d'échantillonnage est de 100 Hz. L'analyseur recherche les "pentes" de la courbe amplitude-fréquence (il travaille donc en amplitude relative) par comparaison des sorties de deux filtres consécutifs (notées 1 ou 0 selon qu'il y a augmentation ou diminution). Les données sont perforées sur ruban en temps réel, puis utilisées sur ordinateur, en attendant de connecter l'analyseur directement au système de reconnaissance.

Chaque échantillon est caractérisé par 30 paramètres, les 24 données de l'analyse plus 6 données élaborées à partir des 24 premiers.

b) Reconnaissance phonémique, du moins au stade préliminaire du traitement, pour ne pas restreindre la généralité. Nous avons choisi 30 phonèmes pour représenter la langue française.

c) Reconnaissance par matrice d'apprentissage, du type de Steinbuch, mais plus générale en ce sens que les connexions possèdent des poids variables, ajustés par apprentissage. La segmentation parole/silence est traitée à part.

d) Intégration des données fournies par l'analyse pour tenir compte de la structure continue de la parole. Les intégrateurs utilisés sont en plus munis de fuites de façon à oublier progressivement les informations passées.

II - RESULTATS ACQUIS.

La simulation sur ordinateur 10070 du système de reconnaissance a permis de faire une étude des différents paramètres de réglage et d'obtenir les "poids" de la matrice d'apprentissage.

Le pourcentage de reconnaissance de phonèmes atteint au maximun 70 % pour plusieurs locuteurs.

Nos données consistent en listes de mots assez longs, prononcés et enregistrés sans aucune précaution sur un magnétophone courant. Nous utilisons au total environ 60 mots.

La réalisation physique du système est en cours et ne pose plus de grosses difficultés. Les poids sont représentés par des résistances, cette méthode étant la plus économique.

La partie la plus délicate consistait en un détecteur de maximum permettant de sélectionner en temps réel la réponse de la matrice d'apprentissage parmi les 30 tensions de sortie. Cet appareil fonctionne, il permet d'effectuer l'opération précédente en 5 millisecondes [3].

Nous disposerons ainsi bientôt d'un appareil capable de transcrire en temps réel en leurs phonèmes constitutifs des mots prononcés par plusieurs locuteurs ; de plus, pour des locuteurs de voix très différentes (de sexes différents par exemple), l'appareil est facilement adaptable en ce sens qu'il suffit de changer les quelques cartes portant les poids de la matrice.

III - RECHERCHES ACTUELLES ET FUTURES.

Les recherches actuelles portent sur trois points principaux, tous plus ou moins liés :

a) amélioration du système de reconnaissance.

b) utilisation de contraintes autres qu'acoustiques dans la reconnaissance.

c) Méthodes d'analyse du signal vocal et recherche systématique de paramètres pour la reconnaissance.

a) L'amélioration principale consiste à effectuer une classification grossière des phonèmes en grands groupes : voyelles - occlusives - fricatives ..... avant de faire travailler la matrice d'apprentissage. Cette pré-classification, qui doit être très sûre - une erreur à ce niveau donnant une erreur finale grossière - est assurée par une série de tests portant sur les paramètres de reconnaissance. Elle assure simultanément une augmentation du pourcentage de reconnaissance et un gain de temps non négligeable.

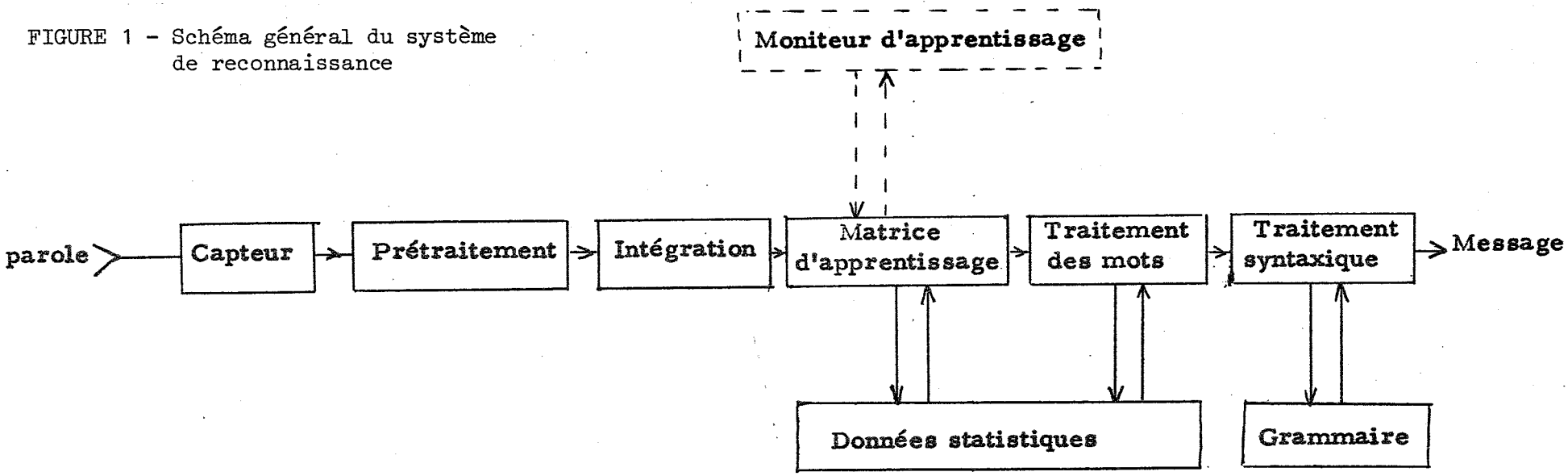
b) Dans le cadre d'une liaison directe en temps réel de l'appareil de reconnaissance avec un ordinateur, nous envisageons d'utiliser des contraintes plus évoluées : linguistiques, syntaxiques pour améliorer la reconnaissance phonémique. Avec les performances actuelles, le taux de reconnaissance au niveau de la phrase atteindrait presque 100 %. Une étude statistique du français parlé a été menée dans ce but [5] . On a étudié en particulier l'amélioration de la reconnaissance par l'utilisation des fréquences d'occurrence des diphonèmes. D'autres données statistiques vont être également utilisées. De plus, des listes de mots formés des diphonèmes les plus courants ont été dressées pour tester notre système.

Nous envisageons l'entrée orale d'un programme en ordinateur dans un avenir proche.

Grossièrement, le schéma fonctionnel du système complet peut se représenter comme sur la figure 1.

c) En ce qui concerne l'étude du signal vocal, diverses méthodes d'analyse ont été étudiées de façon critique : transformées de Haar, Fourier. Nous utilisons un programme de transformée rapide de Fourier avec lequel diverses études sont prévues. En particulier une méthode mathématique de compression d'information permet la recherche systématique des paramètres intéressants pour la reconnaissance. Cette étude est en cours. La méthode nous a déjà permis de chiffrer l'importance de chacun des 24 filtres utilisés, pour différencier entre eux les phonèmes. Il s'est avéré, en particulier, que 6 filtres transmettaient à eux seuls 92 % de l'information totale. Cette méthode est générale, applicable à tout problème de reconnaissance de formes ; elle doit nous permettre de découvrir des paramètres réellement efficaces pour la reconnaissance.

FIGURE 1 - Schéma général du système de reconnaissance



Bibliographie.

1 . M. LAMOTTE, J. BREMONT, J-P. HATON, Simulation de la reconnaissance des formes vocales par apprentissage, C. R. A. S., 269, p. 286-88, août 1969.

2 . M-J. VIGNERON, J-P. HATON, M. LAMOTTE, J. BREMONT, Recherches actuelles sur l'extraction de caractéristiques et la reconnaissance de la voix parlée, Automatisme, XV, n°12, p. 646-649, 1970.

3 . R. HORNUNG, M. LAMOTTE, Sélecteur de numéro de la voie de potentiel le plus élevé, prise parmi plusieurs dizaines de voies d'entrée, Demande de Brevet à l'ANVAR déposée le 16.11.70.

4 . J-P. HATON, M. LAMOTTE, Etude statistique des phonèmes et des diphonèmes dans le français parlé, A paraître dans la Revue d'Acoustique en 1971.

PARAMETRISATION ET PROCEDURES DE RECONNAISSANCE DE LA PAROLE

C. J. GUEGUEN , A. MAISSIS , L. F. PAU
E. N. S. des Télécommunications - PARIS

La reconnaissance du signal vocal est conventionnellement décomposée en deux phases principales : le prétraitement et la reconnaissance. C'est un développement coordonné de ces deux aspects interactifs qui constitue le but de l'équipe du Laboratoire d'Automatisme de l'Ecole Nationale Supérieure des Télécommunications. Deux axes de recherche complémentaires ont donc été particulièrement étudiés. Des travaux sur l'analyse du signal vocal par un système régis par des équations aux dérivées partielles (modèle approximatif du comportement de l'oreille) ont permis d'engendrer des paramètres originaux dotés de propriétés intéressantes. En conjonction, un ensemble de procédures de reconnaissance utilisant comme support l'analyse factorielle ont été développées dans l'optique d'une gestion hiérarchisée de celles-ci. Cet exposé se propose de faire le point sur ces travaux.

I - Phase de prétraitement.

1. Généralités sur le prétraitement.

La phase du prétraitement comprend toutes les opérations nécessaires à l'élaboration d'une série de paramètres à partir du signal temporel. La figure 1 donne le schéma détaillé de cette phase qui comprend :

1 - La segmentation du signal temporel en éléments minimaux. Pour les sons sonores, on a recours à la période du fondamental ; pour les sons sourds, c'est une durée de 10 ms qui est adoptée.

2 - La détermination des trois premiers formants pour chaque segment minimal. La méthode employée est celle des passages à zéro du signal préfiltré dans les bandes : 150 - 900 Hz pour le premier formant, 900 - 2200 Hz pour le second, 2200 - 5000 Hz pour le troisième. Du fait du caractère approximatif de la méthode, il s'agit plus de fréquences dominantes dans les bandes considérées que de formants.

3 - La détermination des paramètres proprement dits en utilisant une batterie de filtres passe-bande à large bande passante modelée sur les propriétés de la membrane basilaire. C'est sur ce point que nous allons insister dans le présent exposé.

2. Paramétrisation de la parole.

Il est connu (1) qu'un modèle approximatif de la réponse d'un point de la membrane basilaire (déplacements verticaux X (t,l) à la distance l de la fenêtre ovale)à l'excitation par un signal sonore peut être simulé par un filtre à large bande caractérisé par :

i - une fréquence de résonnance dépendant du point considéré.
ii - une bande passante à 3dB proportionnelle à cette fréquence.
iii - un gain maximal caractéristique du point étudié.

La méthode de paramétrisation basée sur ces données utilise la réponse temporelle X (t,l) de 7 points de la membrane simulée et leurs dérivées spatiales définies comme suit :

i - les 7 canaux admettent pour fréquences de résonnance : 300, 470, 800, 1300, 2100, 3600, 4600 Hz de façon à recouvrir une part importante du spectre de la parole.

ii - la dérivée spatiale définie comme $\frac{\partial X}{\partial l}$ (t,l) nécessite la création d'une deuxième série de 7 canaux de fréquences : 260, 410, 720, 1200, 1900, 3350, 4200 correspondant à un ∂l de 0,5 mm.

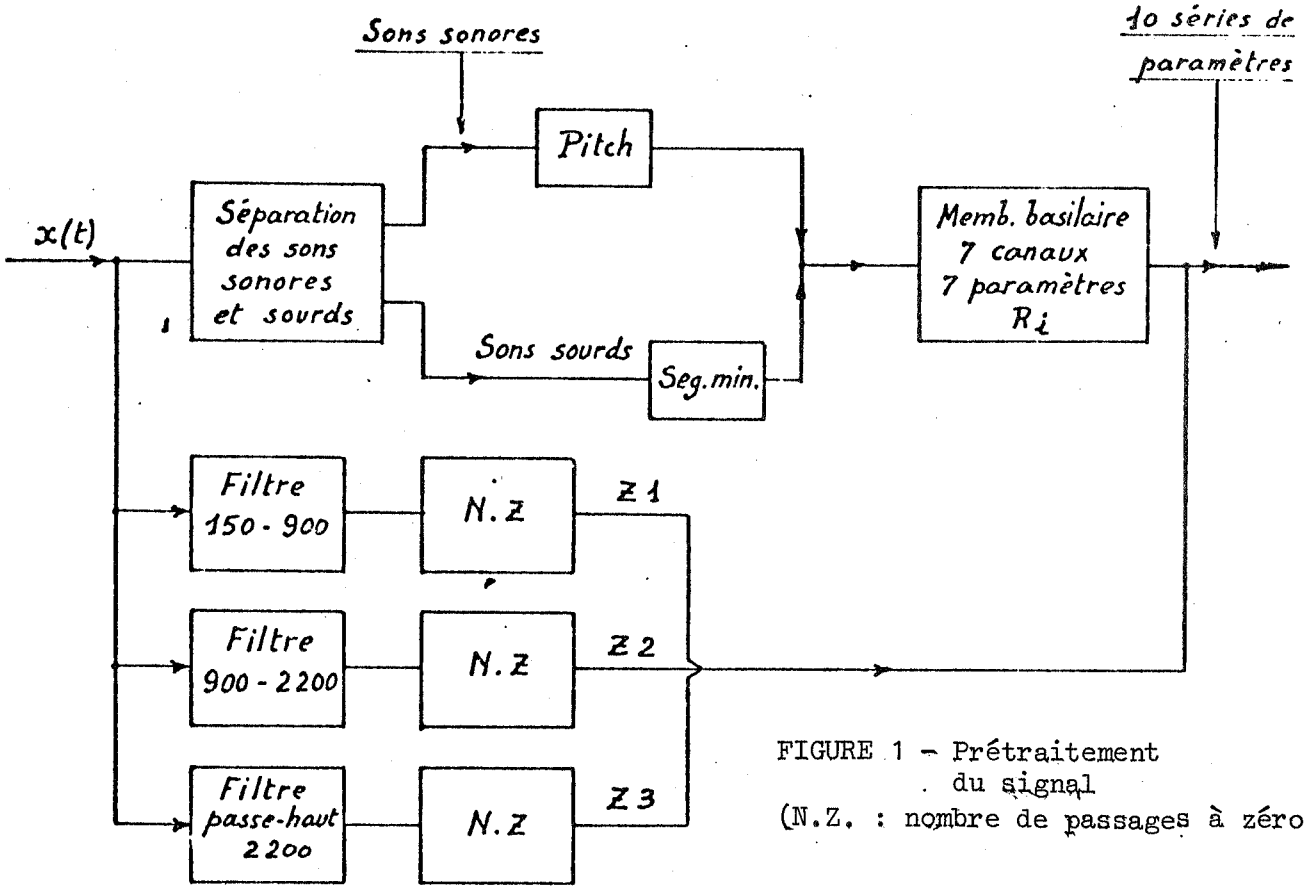
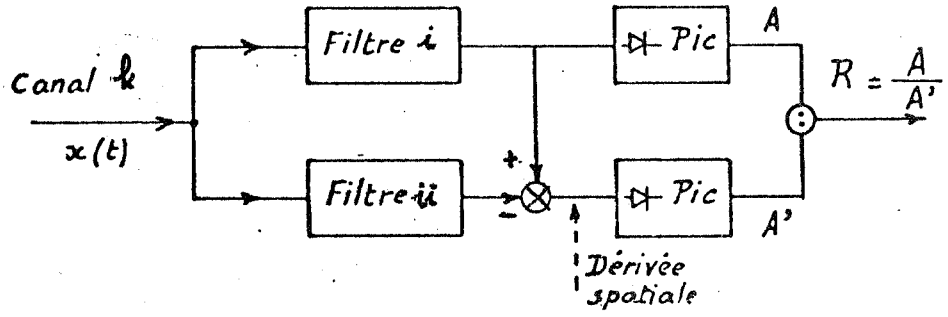


FIGURE 1 - Prétraitement du signal
(N.Z. : nombre de passages à zéro)

FIGURE 2 - Extraction des paramètres (membrane basilaire)

7 canaux - Fréquences :

I:	II:
300	260
470	410
800	720
1300	1200
2100	1900
3600	3350
4600	4200



La figure 2 illustre le mode d'extraction des paramètres relatifs au canal k . Pour un segment minimal j, on détermine le maximum du signal à la sortie du filtre de type (i) ; soit A (k,j) et le maximum de la différence des sorties des filtres (i) et (ii) correspondants ; soit A' (k,j). Le paramètre résultant est :

$$R(k,j) = \frac{A(k,j)}{A'(k,j)}$$

Pour N segments minimaux le signal temporel est ainsi représenté par une matrice R (k, j) de dimensions 7 x N.

3. Segmentation et normalisation.

L'évolution du rapport R (k, j) par rapport à j fait apparaître des régions de stabilité séparées par des régimes transitoires.

La stabilité est évidemment relative car elle dépend en particulier du contexte et du locuteur. Cependant elle met en évidence les "atomes"discernables grace à ces paramètres et permet d'attacher à chacun une valeur caractéristique. La segmentation résultante du mot est très proche du phonème.

Parallèlement, chaque canal tombe dans l'une des trois bandes pour lesqulles les formants sont définis. On peut donc associer chaque rapport R (k,j) avec l'un des trois formants (j étant associé à un phonème).

L'étude de cette représentation révèle un certain nombre de propriétés intéressantes conduisant à une normalisation par rapport à la variété des occurences possibles d'un phénomène donné.

Sur la figure 3 sont représentées un certain nombre d'expériences sur le phonème A prélevé au sein de mots prononcés par un seul locuteur. En ordonnée y sont portées les valeurs de R (k, j) (canal 2100 Hz) ; en abcisse x, la fréquence dominante (ici deuxième formant). On voit ainsi apparaître une certaine dépendance des diverses occurences approximable par une relation linéaire.

Cette propriété a été vérifiée théoriquement en assimilant la membrane basilaire à une membrane mince enroulée sur un cylindre ayant pour directrice une spirale. La relation approchée devant permettre de réaliser la normalisation sur un large domaine est du type : $1/\sqrt{f^2 - k^2}$.

A l'issue de la phase de prétraitement, on dispose donc d'une décomposition du mot en segments proches des phonèmes. Chaque phonème est caractérisé par 7 paires de paramètres susceptibles de normalisation qui permettent d'aborder la phase de reconnaissance.

II - Phase de reconnaissance.

Le problème posé ici par la reconnaissance des formes est d'effectuer un tri parmi les phonèmes inconnus caractérisés par les paramètres basiques R (k,j). Le but de ce tri est de grouper les phonèmes en classes symbolisées par le même concept abstrait. L'ensemble de ces concepts a été défini préalablement au moyen d'un échantillon d'apprentissage comprenant un nombre élevé d'occurences des divers phonèmes.

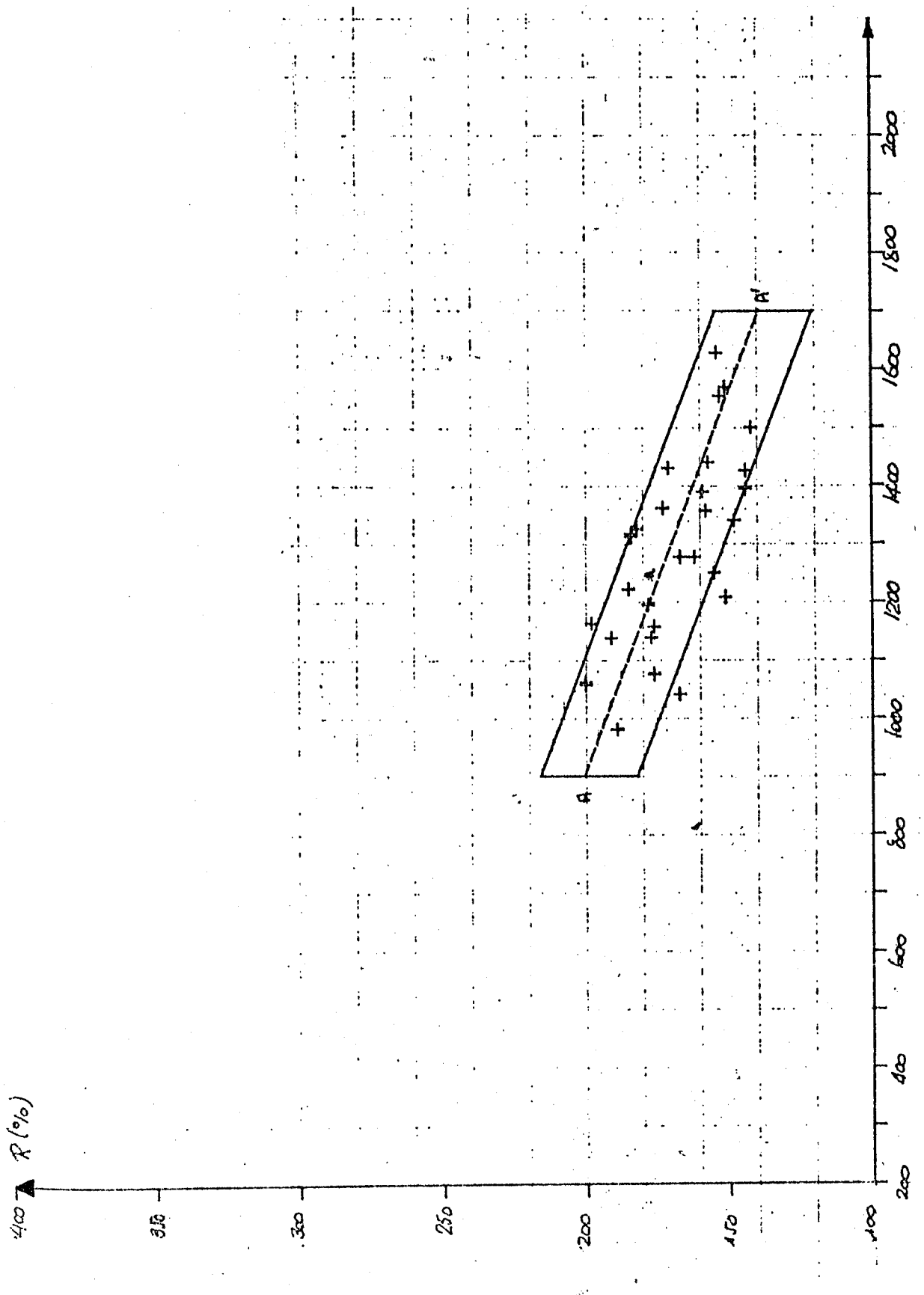
Notre approche du problème de la reconnaissance se caractérise par les aspects suivants :

- le choix des axes dans l'espace des formes d'apprentissage est réalisé au moyen de l'analyse factorielle des correspondances ; celle-ci permet de réduire la mémorisation des formes d'apprentissage à celle d'un nombre restreint de paramètres ;

- au lieu de considérer des surfaces séparatrices et des fonctions discriminantes de caractère géométrique, nous soumettons les phonèmes à reconnaître à une batterie de test statistiques dans un espace de formes de dimension réduite ;

- la décision et la non-décision résultent d'une agrégation des résultats de chacun de ces tests au moyen d'une procédure hiérarchisée en plusieurs niveaux.

FIGURE 3 - Valeurs du rapport 2100 Hz du phonème "A" pour une série de mots



1. Analyse factorielle des correspondances et apprentissage.

L'échantillon d'apprentissage est constitué par l'ensemble des paramètres normalisés, fournis après traitement d'un enregistrement sonore de la parole. L'analyse factorielle est la méthode de statistique projective utilisée pour effectuer simultanément une réduction de ces données, et une classification préalable des phonèmes (pondérés par leurs probabilités d'occurence dans l'échantillon (ou dans la langue)).

L'analyse factorielle fournit, dans un espace réduit de dimension fixée r, la meilleure approximation inertielle du nuage d'apprentissage au sens d'une métrique distributionnelle du CHI-2 sur l'ensemble des phonèmes pondérés :

$$d^2(j_1, j_2) = \sum_I \frac{[Prob(i/j_1) - Prob(i/j_2)]^2}{Prob(i)}$$

On peut en particulier obtenir une représentation plane r = 2, qui constitue une carte sur laquelle on peut visualiser les classes de phonèmes d'apprentissage. L'expérience a montré que la plupart des phonèmes vocaliques, et certains phonèmes consonnantiques (dont R et L (deux régions) formaient dans un espace r = 3 des classes géométriquement séparables (figure 4).

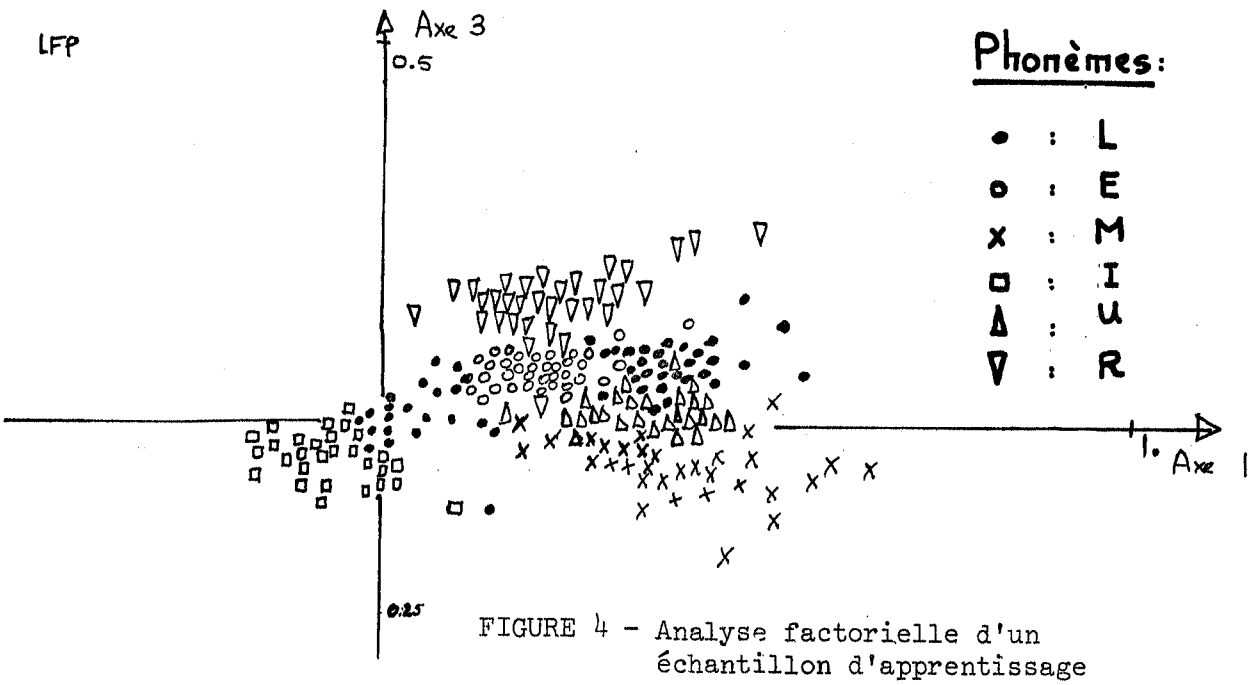
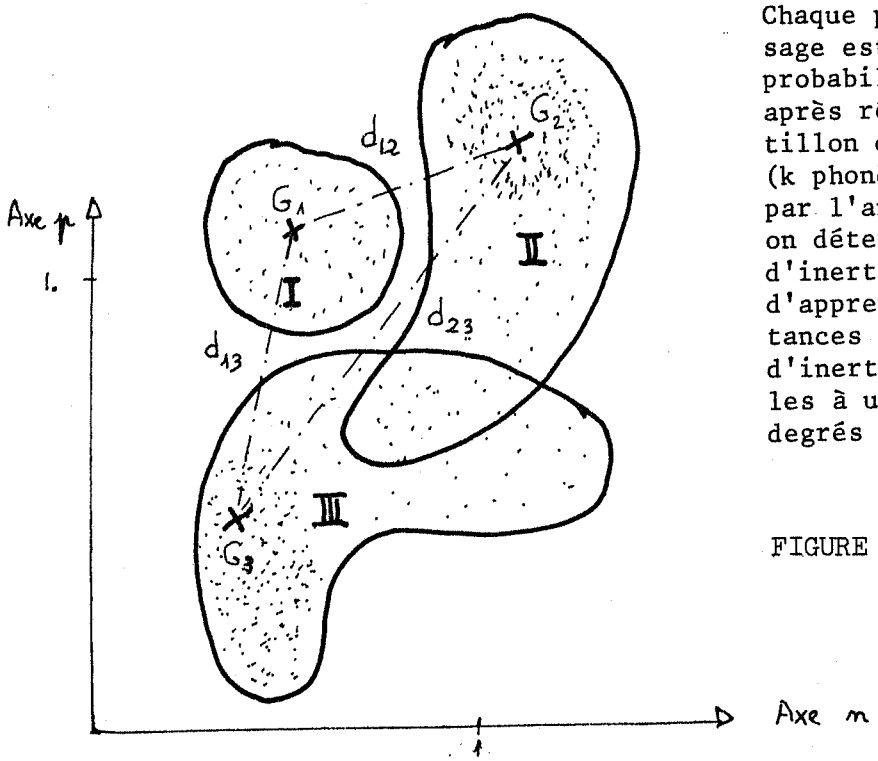


FIGURE 4 - Analyse factorielle d'un échantillon d'apprentissage

2. Tests de reconnaissance.

L'analyse factorielle des correspondances fournit par une formule linéaire les coordonnées de tout nouveau phonème observé dans l'espace réduit de dimension r. Nous pouvons tester l'appartenance de ce phonème aux différentes classes de segments minimaux au moyen d'une suite de tests statistiques de rang ou d'homogénéité, ordonnés approximativement par puissances et complexités croissantes :

2.1 - comparaison des distances du CHI-2 (ou assimilées) du nouveau phonème aux centres d'inertie de chacune des classes ; ces distances se déduisent très simplement d'une mesure des distances euclidiennes sur les cartes d'apprentissage (figure 5).



Chaque phonème d'apprentissage est pondéré par une probabilité d'occurence ; après réduction de l'échantillon d'apprentissage (k phonèmes, 1 paramètres) par l'analyse factorielle, on détermine les centres d'inertie pour chaque classe d'apprentissage ; les distances d_{ij} entre ces centres d'inertie sont proportionnelles à un χ^2 à (k-1) (1-1) degrés de liberté.

FIGURE 5 - Exemples de distances du χ^2 entre classes

2.2 - comparaison des indices de similarité entre le nouveau phonème et chacune des classes d'apprentissages ; ces indices utilisent les ultra-métriques de SHEPARD, CAROLL ou KENDALL, et nécessitent une étude théorique approfondie.

2.3 - comparaison des densités respectives de chacune des classes d'apprentissage ou voisinage du nouveau phonème ; cette comparaison se fait par une règle perfectionnée de voisinage d'ordre n_i , et l'on cherche à maximiser la quantité :

$$\frac{n_i \; P_i}{(N_i + 1) V_i}$$ (figure 6)

où : i = classe de phonèmes d'apprentissage

N_i = nombre de phonèmes d'apprentissage de la classe i

P_i = probabilité d'occurence d'un phonème de la classe i dans le dictionnaire à analyser ; cette probabilité peut être différente de la probabilité d'occurence dans l'échantillon d'apprentissage

n_i = entier positif, lié à la puissance du test

Vi = volume minimal d'un voisinage du nouveau phonème, tel que n_i phonèmes de la classe i soient intérieurs à ce voisinage.

Si le maximum trouvé est inférieur à un certain seuil, on pourra décider qu'il y a ambiguïté.

2.4 - test de rang destinés à lever le doute entre 2 classes ; ayant défini une relation d'ordre, à valeurs bivalentes L (i, j) entre points i, j de l'espace réduit de dimension r, on compare après pondération les statistiques :

$$\sum_{j} L(i,j) \qquad\qquad L(i,j) = 0,1$$

dans lesquelles i représente le nouveau phonème, et où la sommation porte sur les points j de l'une ou l'autre des deux classes considérées.

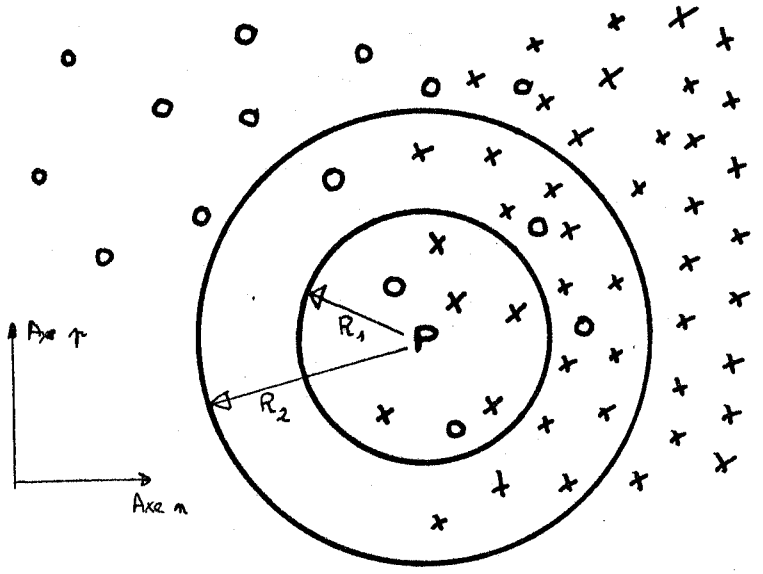


FIGURE 6 - Règle du voisin d'ordre n_i .

Exemple : 2 classes de phonèmes d'apprentissage ;

$n_i = 5$; volumes réduits à des circonférences.

Le phonème non identifié P sera alloué à la classe i (O ou X) telle que :

$$\frac{n_i P_i}{(N_i+1)\pi R_i^2} = \text{Sup}\left[\frac{n_1 P_1}{(N_1+1)\pi R_1^2}, \frac{n_2 P_2}{(N_2+1)\pi R_2^2}\right]$$

3. Décision.

Ayant évalué théoriquement les performances asymptotiques des tests précités, nous associons à chacune des statistiques correspondantes des probabilités d'appartenance et de fausse classification. Lorsque le phonème observé appartient à une classe séparable de toutes les autres, il est fréquent qu'un test rapide fournisse une probabilité d'appartenance élevée et une probabilité très faible de mouvaise classification ; une décision définitive pourra alors être prise à l'issue de calculs assez brefs. Par contre, lorsque le phonème observé se situe dans une zone de recouvrement de deux ou plusieurs classes, il faudra d'une part calculer les statistiques associées à plusieurs test, et d'autre part arbitrer à un niveau supérieur entre les classifications fournies par ces tests pris isolément. La procédure hiérarchisée régulera donc le volume des calculs, et ajustera les valeurs des paramètres incorporés dans les différents tests ; elle prendra une décision de classification ou de doute qui minimisera la probabilité de fausse classification du système de reconnaissance tout entier.

Cet exposé a décomposé l'état actuel des travaux menés à l'E.N.S.T. en deux phases distinctes. Il est bien évident que l'amélioration du processus global de reconnaissance ne peut être envisagée que par un dialogue entre analyse et reconnaissance. Les études actuelles tendent à l'aide des procédures définies à affiner les paramètres d'analyse. Les limitations ainsi révélées des paramètres conduisant à l'élaboration de procédures plus puissantes. Une de ces limitations est la nécessité d'accéder à une mesure des paramètres proche du temps réel. C'est avec ce soucis que sont développées les méthodes analyse en gardant à l'esprit la possibilité d'une génération rapide des paramètres à l'aide d'un codeur analyseur spécialement adapté à cette tache. (3)

Références :

(1) J.L. FLANAGAN : "Speech Analysis, Synthesis and Perception". New York, Springer 1965.

(2) A. MAISSIS : "Traitement des Signaux aléatoires", tome IV. Convention DGRST 7002186, Juin 1970.

(3) A. MAISSIS, F. FIEVET, PH. WALRAVE : "A new Analog-Digital, Filtering methode for real-time Speech recognition". IEEE Audio and Electr., Dec. 1970 (page 385.388).

SEGMENTATION DE LA PAROLE ET RECONNAISSANCE DES SYLLABES A L'INTERIEUR DES MOTS

G. MERCIER
C. N. E. T. - LANNION

I. - INTRODUCTION

L'expérience que nous allons décrire se place dans le cadre général du schéma de l'automate de reconnaissance présenté dans la fig. 1.

L'objet de cette étude est donc de décomposer le flux de parole sortant du vocoder à canaux sous forme numérique en éléments phonétiques, ensuite de paramétriser ces éléments (ce qui équivaut à réduire l'information sortant du vocoder et caractérisant ces éléments), de les identifier avec un certain degré de confiance et enfin de reconstituer et d'interpréter le message de départ, à partir tout à la fois du vocabulaire du langage de communication, de ses données syntaxiques et sémantiques et des données phonétiques qui viennent d'être identifiées :

Dans ce système on distingue 4 phases essentielles : une phase de segmentation, une phase de paramétrisation des phonatomes, une première phase d'identification et une phase de reconnaissance proprement dite. Dans ce qui suit nous insisterons surtout sur les 3 premières phases et en particulier sur la partie segmentation.

Remarque :-Nous partons de mots ou d'expressions ; c'est-à-dire nous supposons que la segmentation en mots est un problème résolu alors qu'en fait il ne l'est pas. On peut cependant court-circuiter cette phase en considérant comme mots certaines expressions comme "aller à", "bandes magnétiques", etc..., ou en laissant un silence assez long entre chaque mot.

- La fin de parole est déterminée soit par un long silence c'est-à-dire un niveau moyen d'énergie faible pendant un temps assez long (et sans pitch) ou par un temps maximum, fixé, de parole.

Avant de passer à la décomposition du mot en syllabe, nous séparons donc la parole du bruit en tenant compte du pitch et du niveau moyen d'énergie dans les canaux.

II. - DECOMPOSITION DU MOT OU DE L'EXPRESSION EN SYLLABES

Pour cette décomposition, on utilise une première segmentation correspondant très grossièrement à une segmentation du mot en phonèmes.

II.1 - Considérons la parole telle qu'elle sort du vocoder après la segmentation parole - bruit. A ce stade, elle est représentée par une suite d'échantillons $X_1, \ldots X_t, \ldots X_{tm}$ dans le temps, l'échantillonnage se faisant toutes les 10 ms environ, chaque X_t étant lui même un vecteur (fig 2)

$$X_t = (x_{1t}, \ldots, x_{jt}, \ldots x_{nt}, x_{n+1,t})$$

x_{jt} : énergie dans le canal j à l'instant t $\quad j = 1,n$

$x_{n+1\ t}$: valeur du pitch à l'instant t

On détermine aisément l'énergie E (t) du tème échantillon par la relation :

$$E(t) = \sum_{j=1}^{n} x_{jt} \qquad (1)$$

C'est-à-dire en faisant la somme de l'énergie de chacun des canaux.

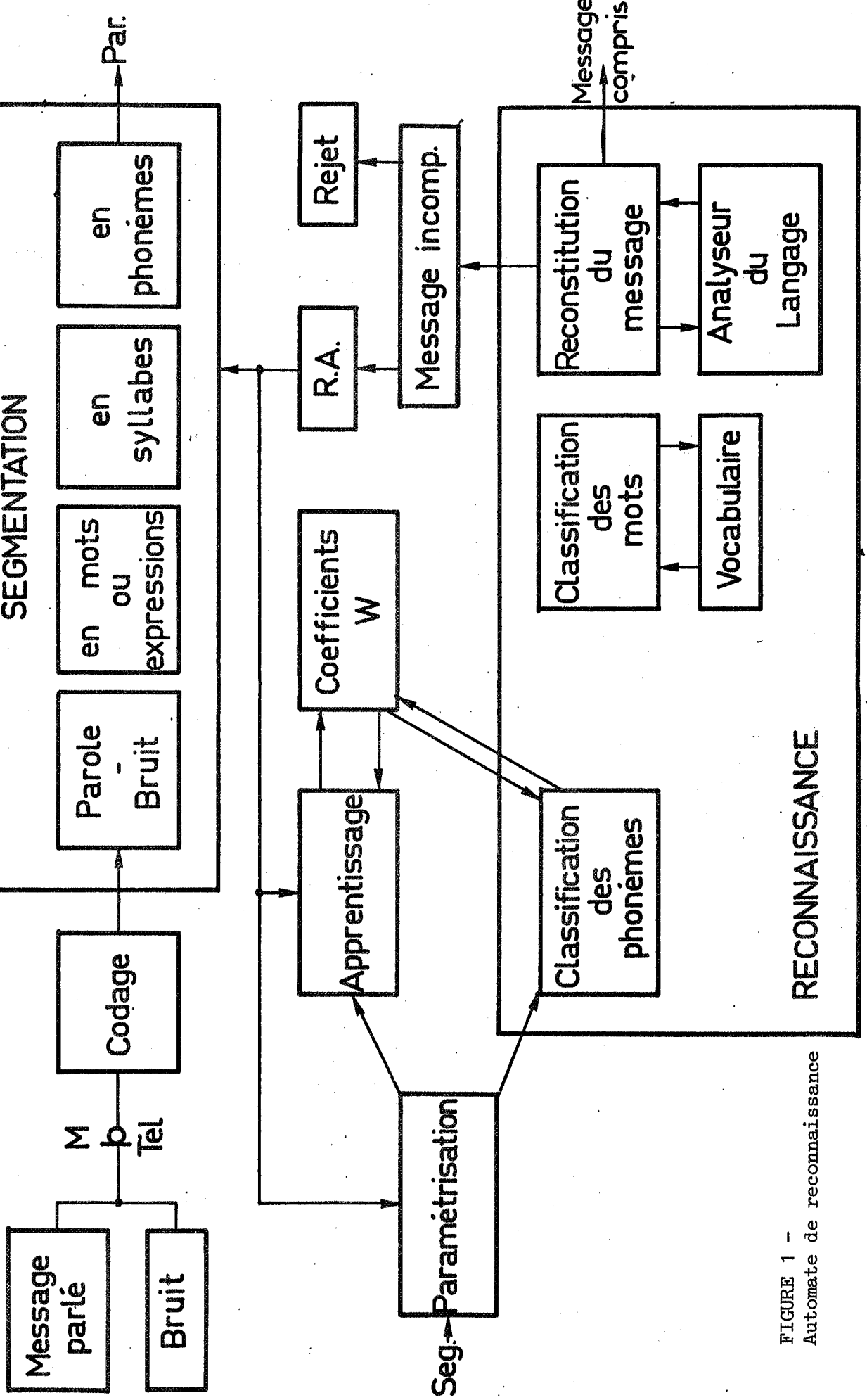


FIGURE 1 -
Automate de reconnaissance

FIGURE 2

MOT PRONONCE ZERO

CANAL 1	2	3	4	5	6	7	8	9	10	11	12	13	14	PITCH	E(L)
7	9	6	4	3	2	1	2	4	2	2	2	3	2	188	49
8	11	7	5	4	3	2	2	5	3	2	3	3	2	188	60
9	11	7	5	4	3	1	2	5	3	2	3	3	2	133	60
10	11	7	6	3	2	2	2	5	3	2	3	3	2	136	61
10	9	7	6	2	2	1	1	4	2	2	3	3	2	141	54
10	8	7	5	3	2	1	1	4	2	2	2	3	2	144	52
11	7	7	5	3	2	2	1	4	2	2	3	3	2	145	54
11	6	6	5	3	2	2	2	3	2	2	3	3	2	147	52
11	6	4	3	2	1	2	2	2	2	2	3	3	1	147	44
10	5	2	2	2	1	2	2	2	1	1	2	2	1	0	35
10	4	1	1	3	1	2	2	2	2	2	2	2	2	0	36
9	5	4	4	3	2	3	3	3	3	3	3	3	3	0	51
8	7	7	5	4	3	3	3	4	4	3	4	4	4	0	63
8	10	10	6	6	4	3	3	5	5	4	6	5	5	136	80
10	11	11	8	7	5	4	4	6	7	5	7	6	5	138	96
11	11	11	9	8	5	4	4	6	8	6	8	6	6	136	103
12	11	12	9	8	6	4	4	6	8	6	8	7	6	134	107
12	11	12	9	8	6	4	4	5	8	6	8	7	7	134	107
12	11	12	9	8	6	4	4	5	8	7	8	7	7	135	108
13	11	12	8	7	6	4	3	4	8	7	8	7	7	134	105
13	11	11	8	7	6	4	3	5	8	7	8	7	7	134	105
13	11	11	8	7	6	4	4	5	8	7	7	7	7	133	105
13	12	11	8	8	6	4	4	5	9	8	7	6	6	132	107
13	11	11	8	8	6	4	5	6	9	8	7	6	6	133	108
12	11	11	8	8	5	5	4	6	8	8	6	6	6	132	104
12	10	11	7	7	4	4	4	6	7	6	5	4	5	151	92
12	9	10	6	7	5	4	3	5	5	3	4	4	3	151	80
11	7	9	5	5	4	4	4	5	4	3	4	4	5	0	74
11	6	7	6	4	5	4	6	5	4	3	4	4	5	82	74
10	7	7	7	6	5	5	7	5	4	3	4	5	5	0	80
9	8	6	7	6	6	7	8	5	4	3	3	4	5	0	81
8	7	7	7	7	7	8	7	5	4	3	3	4	5	0	82
8	8	7	7	7	8	8	5	4	3	2	3	3	5	0	78
7	6	6	7	8	9	8	4	2	2	2	2	3	4	0	70
7	5	7	6	7	9	8	3	2	2	2	2	3	3	0	66
6	4	6	5	6	8	7	2	2	2	2	2	3	2	0	57
5	4	5	6	7	7	7	3	2	0	1	2	3	2	0	54
4	4	7	5	7	8	6	3	2	1	1	2	3	2	0	55
3	2	6	5	7	8	6	2	2	1	1	2	2	2	0	49
1	3	6	6	6	7	6	2	1	0	1	2	2	2	0	45
0	3	7	6	7	7	5	1	1	1	1	2	2	2	0	45
0	4	5	4	6	6	5	1	1	1	1	2	3	1	0	40
2	8	9	6	8	6	5	2	2	1	1	3	3	2	64	58
7	11	11	9	10	9	5	3	2	1	2	3	5	4	0	82
9	11	11	10	10	9	5	3	2	1	2	3	4	3	0	83
11	11	11	10	10	9	5	2	2	1	2	3	4	3	150	84
12	12	11	9	10	8	5	3	2	1	2	3	5	4	149	87
12	13	11	9	10	8	5	3	2	2	3	3	5	4	153	90
12	13	11	9	10	9	5	3	2	2	3	4	5	5	154	93
12	13	12	9	10	9	5	3	2	1	2	3	5	4	158	90
12	13	12	9	10	9	5	3	2	1	2	3	5	4	158	90
12	13	12	8	10	10	5	3	2	1	2	3	5	4	161	90
12	12	11	8	10	10	5	3	2	1	2	3	4	4	164	87
11	11	11	9	10	10	5	3	2	1	2	2	4	4	166	85
11	11	10	9	9	9	5	2	1	1	2	2	3	3	166	78
10	11	9	8	9	8	4	1	1	2	2	2	3	2	166	72
10	9	8	6	7	6	4	0	2	1	2	3	3	1	0	62
9	8	6	5	4	5	4	0	1	1	1	2	2	1	0	49
9	7	5	5	5	6	3	1	2	0	1	2	2	2	0	50
8	7	5	4	5	5	3	1	2	1	1	2	2	1	0	47

Par définition (fig. 3) : Un segmentsera constitué des l échantillons se trouvant entre un minimum d'énergie $E(t_1)$ et un autre minimum d'énergie $E(t_3)$, l'échantillon correspondant au maximum d'énergie $E(t_2)$ se trouvant entre les échantillons t_1 et t_3.

Un segment est donc essentiellement défini par un maximum d'énergie et deux minima d'énergie, correspondant l'un à l'énergie du premier échantillon du segment, le second à l'énergie du dernier échantillon du segment.

Remarque : On élimine certains maxima et minima secondaires en considérant un seuil s : si la différence entre le maximum et l'un des 2 minima d'énergie est inférieure à s (en général s = 6) on ne coupe pas le mot : on ne se considère pas en présence d'un segment significatif.

Ayant ainsi détecté un segment, on recherche ses paramètres caractéristiques et on les garde en mémoire.

Caractérisation du segment (fig. 4)

Les paramètres du segment sont :

- l'adresse de début P. et l'énergie E(P.) du premier échantillon du segment.
- l'adresse de fin P_3 et l'énergie $E(P_3)$ du dernier échantillon du segment.
- l'adresse du maximum P_2 et la valeur de son énergie $E(P_2)$.
- la longueur P du segment.
- le nombre de fois où on a détecté la présence de pitchs : nombre de pitchs.
- le quotient de pitch : nombre de pitchs divisé par la longueur l du segment.
- l'énergie du segment : somme de l'énergie de tous les échantillons du segment.

Ces paramètres définis, on passe à la phase de décomposition en syllabes proprement dite.

II.2 - Deuxième segmentation (fig. 5)

On détermine une syllabe à partir des segments obtenus lors de la première segmentation ; pour cela on se définit :

- 2 seuils pour l'énergie de la syllabe : MUR 1 seuil minimum $\simeq$ 500
  MUR 2 seuil maximum $\simeq$ 1000
- 1 seuil pour la différence entre l'énergie du maximum du segment et l'énergie de fin du segment MUREC $\simeq$ 40.
- 1 seuil pour la différence entre l'énergie du maximum du segment et de fin du segment précédent : MUREV.
- 1 seuil pour la durée entre les 2 maxima de 2 segments consécutifs.
- 3 seuils pour le quotient de pitch de la syllabe : ANIVA $\simeq$ 0,5
  ANIVB $\simeq$ 0,2 à 0,3
  ANIVC $\simeq$ 0,15

Supposons qu'on veuille définir la syllabe I à l'instant j (c'est-à-dire lors de l'étude du jème segment) et soit $E^j(I)$ l'énergie de la syllabe I à l'instant j et e (j) l'énergie du jème segment, on pose :

S.P. SEGMENT (1ere Segmentation)

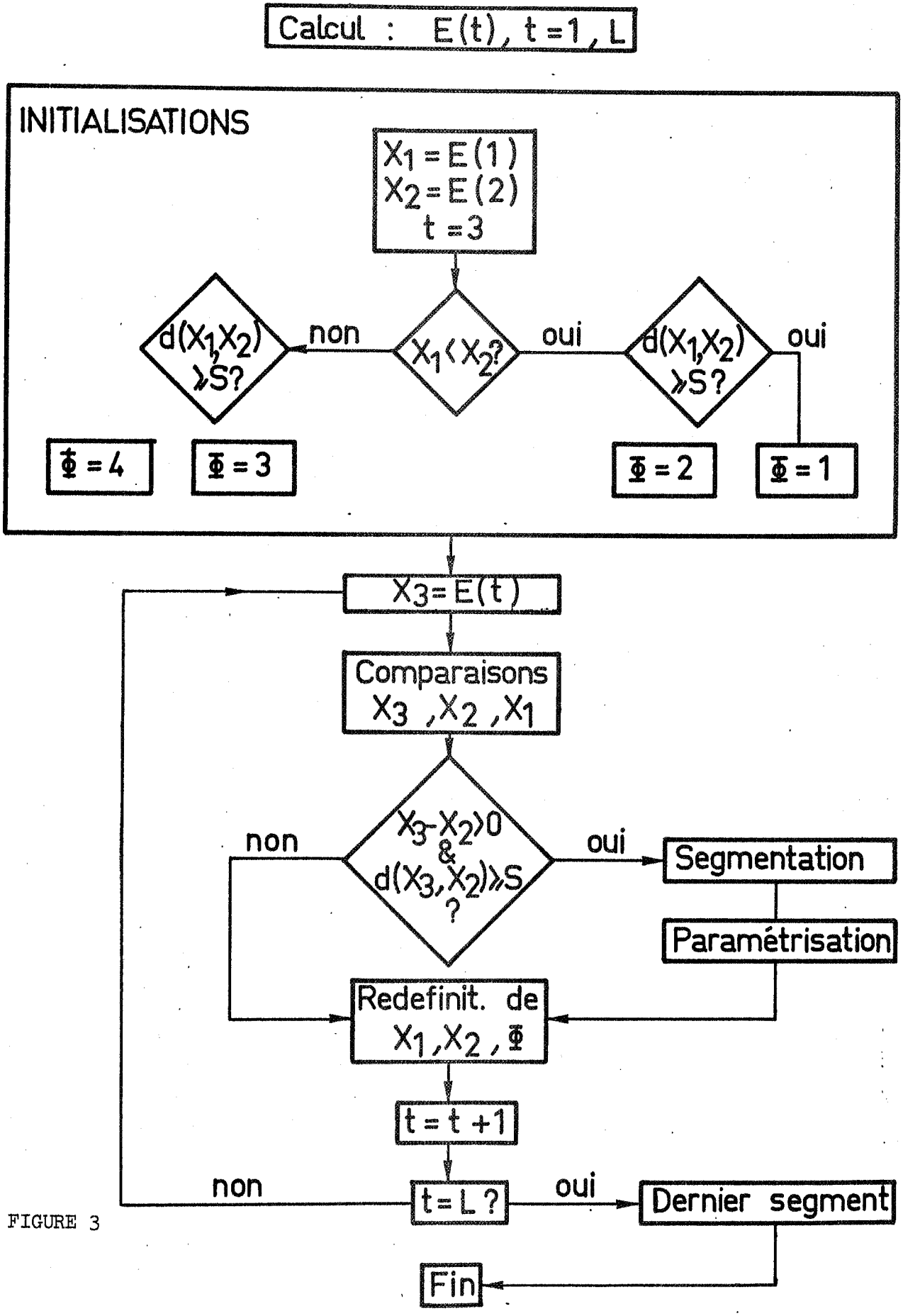


FIGURE 3

PHONEME ZIN 0000000000000000000000000000000000000000000000000000000000000000

NOMBRE D ECHANTILLONS 28

4	6	5	4	3	0	1	1	2	0	1	3	2	0	0	32
8	9	6	5	4	2	2	1	3	0	2	3	2	0	0	47
10	8	6	3	4	2	1	1	2	0	2	3	2	0	146	44
11	7	6	4	3	2	1	0	1	0	1	3	3	0	150	42
11	7	5	4	1	2	1	0	2	0	1	3	3	0	154	40
11	7	4	2	2	1	1	1	3	0	0	3	3	1	158	39
11	5	3	1	2	2	1	0	2	0	0	3	3	0	162	33
10	3	2	0	2	0	1	0	1	1	1	3	3	0	159	27
9	3	2	0	1	1	1	0	2	1	2	3	3	0	73	28
8	4	1	1	1	2	1	0	2	2	2	3	2	1	0	30
6	4	3	2	3	2	2	1	2	1	2	3	2	2	0	35
5	7	6	5	4	4	2	3	4	2	3	3	3	3	145	54
8	9	10	8	7	6	4	6	7	4	5	7	2	3	144	86
10	11	12	10	8	7	6	8	9	7	7	7	2	5	146	109
11	12	12	13	10	9	7	10	11	7	9	3	3	5	147	122
12	12	11	12	11	10	8	10	11	7	8	3	3	5	150	123
12	10	10	11	11	11	8	10	10	7	7	7	2	4	154	120
13	9	9	11	11	11	8	10	9	7	7	7	3	4	152	119
13	9	9	10	10	10	9	10	9	6	6	7	3	3	152	114
12	9	8	10	10	9	8	9	8	6	6	7	2	3	154	107
12	9	8	9	9	9	8	9	8	6	5	7	3	1	150	103
12	8	8	9	9	9	8	9	8	6	5	7	3	2	152	103
12	8	7	9	9	8	8	8	8	5	4	3	2	2	154	93
12	8	7	9	8	7	7	8	7	5	3	3	2	1	151	87
12	8	6	8	8	6	7	7	6	4	2	3	3	0	101	80
11	8	5	6	6	5	6	5	5	3	2	3	2	0	101	67
11	7	3	4	4	4	6	4	4	1	1	3	2	0	112	54
10	5	2	3	4	3	5	3	4	1	0	3	2	0	0	45

PARAMETRES DE LA SEGMENTATION

NUMERO DU SEGMENT	DEBUT DU SEGMENT	FIN DU SEGMENT	ADRESS. DU MAX	VALEUR DU MAX	VALEUR DU DERN.ECH	LONGUEUR DU SEGMENT	NOMBRE DE PITCHS	QUOTIENT DE PITCH	
1	1	8	2	47	27	8	6	.7500	304
2	9	23	16	123	45	20	16	.8000	1679

PARAMETRES DE LA DECOMPOSITION EN SYLLABE

NUMERO	AD.MAX	ENERDEB	ENERFIN	LONGUEUR	QUOTPITCH	NBPITCH	ENERGIE	ADDEBUT	ADFIN	SEGADEB	S
1	16	32	45	28	.7857	22	1983	1	28	1	

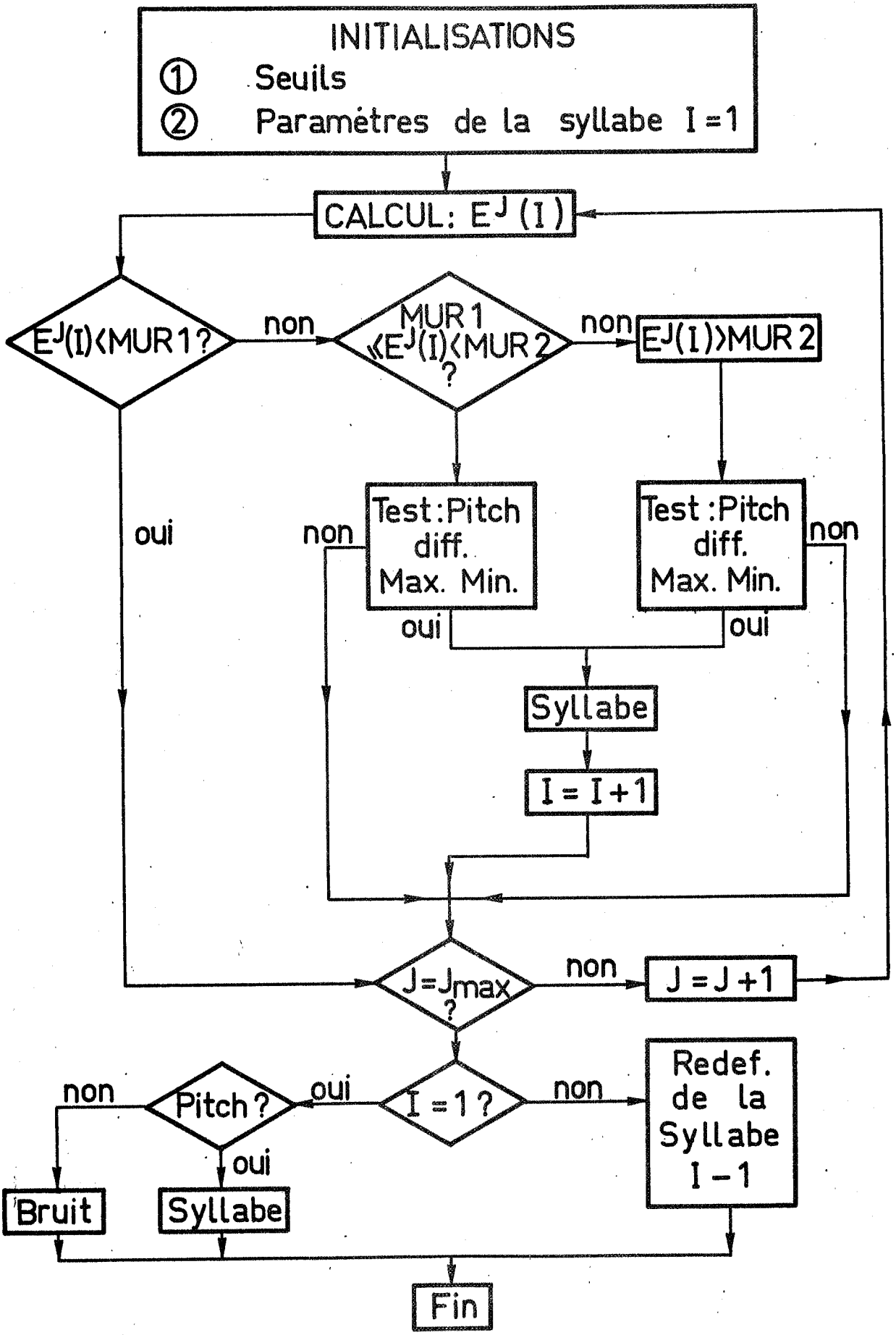
DEBUT DE VOYELLE	FIN	DEBUT DE CONSONNE	FIN
18	25	2	16

FIGURE 4

c/d/6.

FIGURE 5

S.P. SYLLABE



$$E^{o}(1) = 0$$

$$E^{j}(I) = E^{j-1}(J) + e(j) \qquad (2)$$

On compare $E^{j}(I)$ à MUR 1 et MUR 2 ; 3 cas sont possibles (fig. 4).

II.2-a - $E^{j}(I) < MUR\ 1$

Théoriquement l'énergie n'est pas suffisante pour former une syllabe ; on passe à l'étude du segment j + 1.s'il existe sinon on met le segment étudié dans la syllabe I - 1 si I ≠ 1 et on redéfinit les principales caractéristiques de la (I-1)ème syllabe.

Si I = 1 et s'il n'existe plus de segments on fait un test sur le pitch : si le quotient de pitch est suffisant, on se considère en présence de syllabe, sinon on se considère en présence de bruit.

II.2-b - $MUR\ 1 \leq E^{j}(I) < MUR\ 2$

On fait le test de la différence entre le maximum du segment et l'énergie du dernier échantillon ; si cette différence est trop petite, on considère qu'on n'a pas encore de syllabes, on passe à l'étude du segment suivant, sinon et à condition que le quotient de pitch soit suffisant on se considère en présence de syllabe.

II.2-c - $E^{j}(I) > MUR\ 2$

On se considère en présence de syllabe si le quotient de pitch est assez élevé.

Remarque : - Si la distance en temps entre 2 maxima relatifs à 2 syllabes est trop petite et si en même temps la différence d'énergie entre ces maxima et le minimum qui les sépare est trop faible on réunit les 2 syllabes en une seule.

A la fin de cette décomposition en syllabes on définit également comme pour la présegmentation les paramètres caractéristiques de la syllabe (fig. 6 et 7) : la longueur, l'adresse de début, l'adresse de fin, l'énergie du premier échantillon, du dernier échantillon, l'adresse du max et son énergie, l'énergie de la syllabe, le nombre de segments de la syllabe.

A la suite de cette décomposition en syllabe on passe à la segmentation suivante :

III. - SEGMENTATION DE LA SYLLABE EN PHONEMES EN VUE DE L'IDENTIFICATION DE CES DERNIERS

Si on définit la syllabe par le triphonème C_1VC_2 la zone du maximum de la syllabe correspondra en général au début de la voyelle de la syllabe et à la fin de la consonne C. de la syllabe. Quant au début de la consonne C_1 il est défini par la zone de début de la syllabe (en général on élimine la zone de silence entre 2 syllabes).

La fin de la voyelle V et le début de C_2 sont définis par l'échantillon v d'énergie e(v) telle que :

$$e(v) = e_{max} - \frac{e_{max} - e_{fin}}{2} \qquad (3)$$

où e_{max} est l'énergie du maximum de la syllabe et e_{fin} l'énergie du dernier échantillon de la syllabe. Ayant de cette manière décomposé la syllabe en ses éléments constitutifs on va extraire les paramètres permettant d'identifier ces éléments.

Remarque : C_1 peut correspondre à 2 ou 3 consonnes ainsi que C_2.

DE VOYELLE		DE CONSONNE	
16	22	1	13

NUMERO DU BLOC 44

PHONEME FET 000000000000000000000000000000000000000000000000000000000000

NOMBRE D ECHANTILLONS 23

3	3	3	2	2	4	2	2	3	3	4	7	2	5	0	45
3	4	4	3	2	3	2	2	3	4	4	7	3	5	0	49
3	4	6	4	3	2	3	3	3	4	4	7	3	5	0	54
4	5	5	4	4	3	3	3	3	4	5	7	2	4	0	56
4	4	5	4	4	4	3	3	4	4	5	7	3	5	0	59
5	4	5	3	4	4	3	3	4	5	5	7	3	6	0	61
5	4	5	4	4	4	3	2	3	5	5	7	2	5	0	58
6	8	8	7	5	4	4	4	5	7	6	7	3	5	0	79
8	12	11	9	7	6	4	5	6	9	8	7	2	6	0	100
10	12	12	9	8	6	5	5	7	9	9	7	3	7	138	109
12	11	12	8	8	6	5	5	6	9	8	7	2	6	141	105
12	10	11	8	7	5	5	4	5	8	7	7	2	6	142	97
13	11	11	8	6	5	5	4	5	8	7	3	2	6	146	94
13	11	10	8	6	5	5	4	5	7	6	7	2	5	146	94
13	11	10	7	6	5	5	4	5	8	6	3	2	6	144	91
13	10	10	7	6	5	4	4	5	7	6	7	3	6	145	93
13	10	10	7	6	5	4	3	4	7	5	7	3	5	148	89
13	10	9	7	6	4	4	3	4	7	5	7	2	5	150	86
13	10	8	7	5	4	4	3	4	6	5	7	2	5	149	83
12	9	7	6	5	4	4	1	3	5	4	7	2	4	149	73
12	8	7	6	4	4	3	1	2	4	3	7	3	3	149	67
12	7	7	5	4	3	3	0	1	3	3	7	3	3	149	61
11	6	5	4	3	2	2	0	1	2	2	3	3	0	0	44

PARAMETRES DE LA SEGMENTATION

NUMERO DU SEGMENT	DEBUT DU SEGMENT	FIN DU SEGMENT	ADRESS. DU MAX	VALEUR DU MAX	VALEUR DU DERN.ECH	LONGUEUR DU SEGMENT	NOMBRE DE PITCHS	QUOTIENT DE PITCH	
1	1	23	10	109	44	23	13	.5652	1747

PARAMETRES DE LA DECOMPOSITION EN SYLLABE

NUMERO	AD.MAX	ENERDEB	ENERFIN	LONGUEUR	QUOTPITCH	NBPITCH	ENERGIE	ADDEBUT	ADFIN	SEGADEB	SEGA
1	10	45	44	23	.5652	13	1747	1	23	1	

DEBUT DE VOYELLE	FIN	DEBUT DE CONSONNE	FIN
11	20	1	10

FIGURE 6

FIGURE 7

PHONEME PLATT

NOMBRE D ECHANTILLONS 15
PARAMETRES DE LA SEGMENTATION

NUMERO DU SEGMENT	DEBUT DU SEGMENT	FIN DU SEGMENT	ADRESSE DU MAX	VALEUR DU DERNIER ECH	LONGUEUR DU SEGMENT	NOMBRE DE PITCHS	QUOTIENT DE PITCH	ENERGIE DU SEGMENT
1	1	6	4	75	6	4	.6667	436
2	7	15	10	47	9	8	.8889	750

PARAMETRES DE LA DECOMPOSITION EN SYLLABE

NUMERO	AD MAX	ENERG DEB	ENERG FIN	LONGUEUR	QUOTPITCH	NBPITCH	ENERG	AD DEBUT	ADFIN	SEGADEB	SEGADFIN
1	10	44	47	15	.800	12	1186	1	15	1	2

IV. - EXTRACTION DES PARAMETRES DES PHONEMES

- Comme éléments représentatifs de la voyelle de la syllabe on prend les m_V échantillons compris entre le début et la fin de voyelle ; ce sont ces m échantillons qui seront utilisés pour l'apprentissage des coefficients w des hyperplans de séparation : fig. 8 - 9).

- Comme éléments représentatifs de la consonne C_1 (et de même pour C_2) on prend les m_{C_1} échantillons situés entre le début et la fin de la consonne. Jusqu'à présent, pour des questions de rapidité de calcul et de simplicité, cette matrice était réduite à un seul vecteur : sa moyenne dans le temps, mais devant les résultats très moyens obtenus lors de la reconnaissance (étant donnéela variance des différents échantillons, cette moyenne n'est pas suffisamment caractéristique de la consonne), nous ajoutons actuellement à la moyenne d'autres paramètres tels que la longueur de la consonne, le quotient de pitch, la valeur de l'explosion (c'est-à-dire la pente maximum) ; ceci ne sera peut-être pas encore suffisant, et il est possible qu'il faille choisir d'autres solutions pour représenter la consonne. Il semble même, que pour améliorer sensiblement la reconnaissance d'un point de vue acoustique, l'effort doive être porté sur ce point précis de la paramétrisation de la consonne (fig. 10 - 11).

V. - RECONNAISSANCE

V.1 - Principe de la méthode : cf [2]

Soit m le nombre de classes, soit X un échantillon de la classe i, on pose :

$$(\forall i,\ i = 1, m) \text{ et } (\forall X, X \text{ élément de } i)\quad g(X) = ai$$

avec $ai = (ai1, \ldots, aim - 1)$

(4)
$$aij = \left(\frac{m}{m-1} \cdot \frac{m-j}{m-j+1}\right)^{1/2} \qquad \text{si } i = j$$

$$aij = -\frac{1}{m-j}\left(\frac{m}{m-1} \cdot \frac{m-j}{m-j+1}\right)^{1/2} \qquad \text{si } i > j$$

$$aij = 0 \qquad \text{si } i < j$$

C'est-à-dire on transforme tout échantillon de la classe i en un point d'un espace à p dimensions $p = m - 1$ et ce point correspondant au sommet d'un polyédre équilatéral à m sommets, la distance entre chaque sommet et le centre 0 étant égal à 1.

On essaie ensuite d'approcher cette application g par l'application linéaire suivante H qui à X fait correspondre $H(X) = W^t X$, application de R^{n+1} dans R^p tel que :

$$X \rightsquigarrow H(X) = W^t X, \qquad p = m - 1$$

(5)
$$H^t(X) = \left(h_1(X), \ldots, h_p(X)\right) \qquad h_j(X) = \sum_{i=1}^{n+1} w_{ij}\, x_i$$

$$X = (x1, \ldots xn, xn+1)^t$$

w matrice des vecteurs poids à $n+1$ lignes et p colonnes qui permet de définir les hyperplans séparateurs.

FIGURE 8 - Paramètres caractéristiques de voyelles

TET		8												
11	13	12	9	8	7	5	5	6	9	7	3	3	7	138
12	10	12	9	8	6	5	5	5	9	7	3	3	5	139
13	10	12	9	8	6	5	4	5	8	7	3	3	6	143
13	11	11	8	7	6	5	4	5	8	7	3	2	6	149
13	11	10	7	6	6	5	4	5	8	7	3	2	6	148
13	12	10	7	6	6	5	4	4	7	6	3	3	5	149
13	12	10	7	6	5	4	4	4	7	6	3	3	5	150
TEU		8												
12	12	12	11	9	8	7	7	9	6	7	7	3	4	149
12	12	12	11	8	8	7	7	8	6	6	7	2	4	150
12	12	12	11	8	8	7	7	8	5	6	7	3	4	151
12	11	11	10	8	7	6	6	7	4	6	7	3	4	149
12	11	11	10	8	7	6	5	7	4	4	3	3	3	153
12	11	11	10	8	7	6	6	7	4	5	3	2	2	154
12	11	11	10	8	6	6	5	6	3	6	3	3	1	153
GA O		7												
12	11	11	12	11	10	8	10	8	7	8	7	2	4	153
12	10	11	12	12	10	8	10	8	6	8	7	3	4	150
12	10	10	11	11	9	9	10	7	6	8	7	3	4	149
11	9	10	11	11	9	8	10	6	5	7	7	3	3	148
11	9	10	10	11	9	8	10	7	5	7	7	3	4	147
11	8	8	10	10	9	9	10	6	4	6	7	2	4	149
GAP		4												
11	9	10	12	12	9	7	8	9	6	7	7	2	3	153
11	9	10	11	11	9	7	8	9	6	7	7	2	3	44
11	9	10	11	10	9	7	8	8	5	7	3	3	2	147
GUE		5												
13	13	10	8	7	6	5	6	9	8	8	7	10	4	152
13	13	10	7	7	6	5	6	9	8	8	7	10	4	150
13	13	10	7	7	5	5	5	8	7	8	7	3	3	148
13	12	10	7	6	5	5	5	8	7	7	7	3	3	146
GUI		1												
7	8	6	7	6	5	4	4	5	4	5	7	3	6	0
GØ O		7												
13	12	10	8	8	8	9	5	3	3	5	3	3	3	150
13	12	11	8	8	8	9	5	3	2	4	3	3	3	148
13	12	10	8	8	9	9	5	3	2	3	3	2	2	148
13	11	10	8	7	8	9	4	2	2	4	3	2	2	149
13	11	10	7	7	8	8	3	2	1	3	3	2	2	146
13	10	9	7	7	8	8	3	2	0	2	3	3	1	151
GAU		13												
13	12	11	11	9	8	9	10	6	4	6	3	3	4	147
13	12	11	11	9	8	9	10	6	4	6	3	3	4	151
12	12	11	11	9	8	9	10	5	4	6	3	3	3	150
12	12	12	11	9	8	9	10	5	4	6	3	3	4	151
12	12	12	11	9	8	9	9	5	4	5	3	2	3	154
11	12	12	10	9	8	9	9	5	3	5	3	3	3	153
11	12	12	11	9	8	9	9	5	3	5	3	3	3	151
11	11	12	10	9	8	9	8	4	2	4	3	2	3	151
11	11	12	11	9	7	9	7	3	2	4	3	2	3	152
11	10	11	10	9	7	9	6	3	2	4	3	2	2	150
11	10	11	10	8	8	9	6	3	2	3	3	3	2	148
10	9	10	10	8	8	9	5	2	1	2	3	2	0	146
GU O		2												

FIGURE 9 - Paramètres caractéristiques de voyelles

11	8	8	7	6	5	5	4	5	7	7	7	3	3	145
GBN 7														
13	12	12	10	9	10	9	5	4	3	6	3	3	4	149
13	11	10	10	9	10	9	5	3	2	5	3	2	3	148
13	10	10	10	9	10	8	4	3	3	4	3	2	3	151
13	9	10	10	10	10	8	4	4	3	3	7	3	3	153
13	9	9	9	10	8	7	3	3	2	3	7	2	3	151
13	9	9	9	9	6	7	3	3	2	2	3	2	2	150
GBU 1														
8	9	7	7	6	8	7	3	2	1	1	3	2	2	131
GAN 10														
11	10	14	12	10	8	8	10	6	5	7	7	2	5	136
11	9	12	13	10	9	9	9	6	4	7	3	2	4	140
12	9	10	11	11	9	9	9	5	4	5	3	2	4	144
12	9	8	10	10	9	10	8	5	4	4	3	3	3	148
12	9	9	10	9	10	10	7	5	4	5	3	2	4	149
12	9	9	10	10	10	10	7	5	4	4	3	3	3	149
12	8	9	9	9	10	10	6	5	4	4	3	3	3	148
12	8	8	9	9	11	10	6	4	3	3	3	2	2	148
12	8	8	9	9	11	9	6	4	3	3	3	2	2	150
GAIN 10														
12	13	12	12	9	7	6	6	7	10	9	3	3	5	147
12	9	9	11	10	8	6	6	8	9	9	3	3	5	149
12	9	9	10	9	9	7	7	9	8	8	3	3	6	149
12	8	8	9	9	9	7	7	9	7	6	7	2	5	150
12	8	8	9	8	9	7	7	9	7	6	7	2	5	150
12	8	8	9	9	9	7	7	9	6	6	7	2	4	149
12	8	8	9	8	8	7	7	9	6	5	7	2	4	151
11	8	8	8	8	8	7	7	9	5	5	7	3	3	151
11	7	7	8	8	7	7	7	8	5	4	7	3	2	148
GUEI 12														
12	12	10	9	7	6	5	5	6	9	8	7	2	5	154
12	12	11	9	7	6	5	5	6	8	8	7	3	4	151
12	11	10	9	7	6	5	4	6	8	7	7	3	4	151
12	11	10	9	7	6	5	4	6	8	7	7	3	4	151
11	12	10	9	7	6	5	4	6	8	7	7	3	4	152
11	12	11	9	7	6	5	4	5	8	7	7	3	4	156
11	12	11	9	7	5	4	4	5	8	6	7	2	4	153
11	11	11	8	7	5	4	4	5	8	6	7	2	4	155
11	11	11	8	7	5	4	3	5	7	7	7	2	4	153
11	10	10	8	6	5	4	3	4	6	5	7	3	3	154
11	9	9	8	6	5	4	3	4	6	5	7	3	3	153
GAI 8														
13	13	12	8	7	6	5	4	6	9	8	3	2	4	146
13	12	12	9	7	6	5	4	6	9	7	3	2	4	146
13	12	11	9	7	6	5	4	5	8	7	7	3	3	148
12	12	11	9	7	6	5	4	5	9	7	3	2	4	150
12	12	10	8	6	6	5	4	5	8	7	3	3	4	152
12	11	10	8	6	5	4	3	4	8	6	7	2	3	153
12	11	9	7	6	5	4	3	4	7	5	7	3	2	154
GJET 5														
13	11	8	6	5	4	3	3	4	7	6	3	3	4	153
13	11	8	6	5	4	3	2	4	6	6	7	3	4	152
13	10	8	6	5	4	3	2	3	6	5	7	2	4	152
13	10	7	5	4	3	3	2	3	6	5	7	3	4	153

FIGURE 10 - Paramètres caractéristiques de consonnes Cl

```
    11   10   11   12   10    9    6    7    8    5    5    7    3    5  152
SA 0      16
     3    4    5    2    2    3    2    2    2    2    3    7    3    4    0
     3    2    4    3    4    4    2    2    2    3    3    7    3    4    0
     3    2    3    3    3    4    3    1    3    4    3    7    3    4    0
     3    4    4    3    4    4    3    3    3    3    2    3    3    4    0
     3    4    3    4    5    3    3    3    3    3    3    7    3    4    0
     3    4    4    3    5    4    3    4    3    3    4    7    3    3    0
     3    4    3    3    3    3    4    4    4    4    4    7    2    4    0
     3    4    4    4    3    4    4    3    4    3    3    7    2    4    0
     4    4    6    3    5    5    4    4    4    4    3    7    2    4    0
     4    4    5    4    4    4    3    3    4    4    4    7    3    5    0
     5    6    6    5    4    4    4    3    4    4    5    7    3    5    0
     5    7    8    7    6    6    4    4    5    3    4    7    3    4    0
     6   11   11   11   10    8    6    8    8    5    6    7    3    5    0
     8   10   11   12   11    9    7    8    9    6    6    7    2    5  147
    10   10   11   12   11   10    7    8    8    5    5    7    2    5  150
JA 0      18
     5    7    6    4    3    2    1    2    4    4    2    7    3    0    0
     6    7    8    5    4    2    1    2    4    4    4    7    3    1    0
     7    8    9    6    5    2    2    2    5    6    6    7    2    3   26
     9    8    9    6    5    3    2    3    5    7    6    7    3    4  145
    10    9    8    6    5    3    2    2    5    5    6    7    2    5  149
    10    9    8    6    5    3    2    2    5    5    7    7    3    6  152
    10    9    7    6    4    2    2    2    4    5    8    3    3    6  153
    10    8    6    5    4    2    2    2    5    5    8    3    3    7  155
    10    8    6    4    3    3    2    2    5    5    7    3   10    7  156
     9    8    6    4    4    3    2    2    5    6    7    3   10    7  156
     8    8    6    4    4    2    2    2    4    5    7    3   10    7  154
     8    9    7    5    4    3    2    2    5    6    7    3    3    6  153
     8   10    9    7    5    5    3    3    6    5    6    3    2    5  152
    10   11   10    8    6    6    4    4    7    5    7    7    3    3  150
    11   12   11   10    8    7    5    7    8    6    7    7    3    2  154
    11   12   12   11   10    8    6    8    9    6    7    7    2    2  153
    11   11   12   12   10    9    7    9    9    6    7    7    3    2  157
CHA       15
     1    2    2    0    0    1    1    3    4    4    4    7    2    2  127
     0    2    3    0    1    0    1    4    5    6    6    7    3    4    0
     1    2    3    0    1    2    2    4    6    7    7    7    3    7    0
     2    2    4    3    1    2    2    4    6    7    7    3   10    8    0
     3    2    3    3    1    2    2    4    6    7    8    3   10    7    0
     3    3    3    3    2    2    2    4    7    8    8    3   10    8    0
     3    3    1    2    2    2    2    4    6    8    8    3   11    8    0
     3    4    3    3    3    2    2    4    6    7    8    3   11    8    0
     3    3    3    2    3    3    2    4    6    7    8    3   10    9    0
     3    4    3    3    2    3    3    3    6    8    8    3   10    9    0
     4    3    4    2    3    3    3    3    6    8    9    3   10    9    0
     5    4    5    4    3    4    3    3    6    7    9    3   10    9  150
     5    4    5    5    5    4    3    3    6    7    8    3   11    9  150
     6    5    6    5    5    5    3    3    6    6    6    7    3    6    0
BA 0      4
     6   10   10    9    7    6    5    8    6    4    6    7    2    4    0
     8   12   12   13   10    8    7   11    7    5    8    7    3    5    0
    10   11   12   13   11    9    7   10    7    5    8    7    3    5  153
PA 0      3
```

FIGURE 11 - Paramètres caractéristiques de consonnes Cl

	3	9	9	9	8	7	7	7	5	5	5	7	2	5	0
	7	11	12	13	11	10	9	10	7	6	8	7	2	6	0
DA	0	5													
	5	5	6	5	5	5	3	4	4	4	4	7	3	1	0
	7	9	10	9	7	6	5	6	7	7	5	7	2	4	0
	8	10	12	12	9	7	5	6	9	7	6	7	3	5	0
	9	10	11	11	10	8	6	7	9	6	6	7	2	4	152
TA	0	3													
	1	5	7	8	7	7	5	4	5	6	5	7	2	5	0
	5	10	10	11	10	9	6	8	9	7	6	7	3	6	0
GA	0	12													
	6	9	6	3	3	2	0	1	1	2	3	3	2	0	0
	7	10	6	3	3	2	1	1	2	2	4	3	2	0	123
	8	10	6	3	4	1	1	0	1	1	3	3	2	0	123
	9	9	6	3	4	1	1	0	0	0	1	3	2	0	0
	9	7	5	2	3	0	1	0	0	0	0	3	2	0	140
	10	7	4	2	1	0	1	0	0	0	0	3	2	0	140
	10	6	3	1	0	0	1	0	0	0	0	3	2	0	0
	10	7	7	5	5	6	3	5	6	7	4	7	3	5	0
	10	11	8	7	6	7	4	4	6	7	6	7	2	5	0
	10	11	10	9	6	6	5	5	7	7	7	7	3	6	0
	11	10	12	11	8	7	6	7	9	7	7	7	3	5	145
KA	0	4													
	0	1	3	4	4	4	4	7	7	5	5	7	3	4	0
	2	5	5	6	5	6	4	8	8	6	7	7	3	7	0
	3	5	5	4	4	3	4	4	6	3	4	3	3	5	0
MA	0	8													
	10	7	5	3	2	1	2	4	4	2	3	3	2	1	152
	10	7	5	3	2	1	2	4	4	3	3	3	3	2	151
	11	7	5	3	3	1	2	4	3	2	2	3	2	1	153
	11	8	6	3	3	1	3	4	4	2	2	3	2	0	159
	11	8	6	3	3	1	3	4	4	1	2	3	2	0	157
	11	9	7	6	6	3	5	6	4	1	2	3	2	0	158
	11	11	10	11	10	9	7	10	6	5	4	7	3	3	158
NA	0	9													
	6	7	5	3	4	1	1	1	4	2	2	7	3	0	0
	7	8	6	5	4	3	1	1	5	2	3	7	2	0	0
	9	7	6	5	4	4	1	2	4	3	3	7	3	0	145
	10	7	6	5	4	3	1	1	4	2	3	7	3	0	145
	10	7	6	5	5	3	1	1	5	2	3	7	3	0	148
	11	7	6	5	5	2	1	2	5	3	2	3	3	0	150
	11	8	7	7	6	5	3	5	7	4	2	3	2	1	148
	12	10	9	10	9	7	6	7	9	6	5	7	2	4	0
GNA		16													
	6	7	5	2	4	1	1	0	2	1	2	3	2	0	0
	7	7	5	4	4	3	1	1	3	2	2	3	3	0	141
	8	7	5	5	4	3	1	1	2	2	3	7	3	0	142
	9	7	5	5	4	3	1	1	2	2	2	7	3	0	148
	10	7	5	4	4	2	2	2	2	1	2	3	3	0	148
	11	7	6	4	4	2	1	2	2	1	1	3	3	0	150
	11	7	6	4	4	2	1	2	2	2	0	3	3	0	150
	11	8	6	4	4	2	1	1	2	1	0	3	2	0	151
	11	8	6	5	4	3	2	2	1	3	2	3	3	2	146
	11	9	8	6	5	3	2	2	2	4	2	7	2	3	146
	12	9	9	7	6	4	2	3	3	5	3	7	2	4	153

L'erreur (différence) entre l'approximation H et l'application g peut être représentée par le vecteur E ci-dessous :

(6) $E = g(X) - H(X) = ai - H(X)$

de composante $ej = aij - hj(X)$

et la matrice optimale W* correspondant aux hyperplans séparateurs optimaux est celle qui minimise l'erreur quadratique moyenne.

(7) $E_X\, e2 = E_X \left(\sum_{j=1}^{p} e_j^2 \right)$

Une fois obtenus W* par apprentissage (c'est-à-dire H*) la règle de décision utilisée pour déterminer la classe de X est la suivante :

On calcule les m produits scalaires.

(8) $Fi(X) = ai\, H(X) \qquad i = 1,\ldots,m$

et X appartient à la classe i correspondant au produit scalaire maximum.

Remarque : Calculer les m produits scalaires revient d'ailleurs à calculer m fonctions linéaires.

V.2 - Reconnaissance de la voyelle de la syllabe

Pour cette reconnaissance, on se définit 15 classes de voyelles : $\mathring{A}$, $\bar{A}$, E, I, O, U, ON, OU, AN, IN, EI (faire, fer), AI (mais), ET (thé), EÛ (beurre, neuf).

Soit X_t l'un des échantillons de la voyelle $\mathbf{z}$ étudiée

$\mathbf{z} = (X_{t1},\ldots X_t,\ldots X_{tm})^t$ et $X_t = (x_{1t},\ldots x_{nt})$

Ayant au préalable calculé les coefficients W* par apprentissage, pour chaque X_t, on calcule les 15 produits scalaires précédemment définis :

$F_1(X_t)\ldots, F_{15}(X_t)$

puis les 15 fonctions suivantes :

(9) $f_i(\mathbf{z}) = \sum_{t=t1}^{tm} Fi(X_t) \qquad i = 1,15$ (9)

et pour classe probable de $\mathbf{z}$, on prend celle qui correspond au maximum des 15 valeurs $f_i(\mathbf{z})$ $i = 1,15$.

On range d'autre part ces 15 valeurs par ordre de grandeur décroissante ce qui permet également de prendre en considération les classes correspondant aux valeurs voisines du maximum et ainsi de donner plusieurs réponses possibles.

V.3 - Reconnaissance de la consonne C1 de la syllabe

Pour cette reconnaissance on utilise actuellement 20 classes de consonnes :

v, f, z, s, j, ch, b, p, d, t, g, k, m, n, gn, y, w, r, l, h.

Le "h" correspondant à l'absence de consonnes.

On se définit d'autre part 5 voyelles de base A, E, I, O, U.

Soit alors $\bar{X}$ le vecteur représentatif de la consonne (sa moyenne, ou sa moyenne plus d'autres composantes telles que pitch , longueur, explosion, etc...) ; soit d'autre part V_b la voyelle de base dont le produit scalaire se rapproche le plus du produit scalaire maximum calculé lors de la reconnaissance de la voyelle ; connaissant les coefficients $W^*_{CV_b}$ déterminés par apprentissage, on calcule les 20 produits scalaires $F_1(\bar{X}),\ldots,F_{20}(\bar{X})$ définis en V_1, on les range par ordre de grandeur décroissante et on choisit les réponses ayant la plus forte probabilité de s'avérer justes.

Remarque : - On pourrait également procéder comme pour la reconnaissance de la voyelle, en calculant les produits scalaires échantillon par échantillon.

- On procéderait de même pour la reconnaissance de la consonne C_2 de la syllabe.

- Le problème crucial est celui de la bonne paramétrisation de la consonne.

V.4 - Reconnaissance du mot

Utilisant la reconnaissance de la consonne C_1 et de la voyelle de chacune des syllabes prononcées, les programmes donnent actuellement la réponse la plus probable d'un point de vue phonétique ; notre but actuel est d'utiliser les différentes réponses possibles et à l'aide d'un dictionnaire de rechercher la signification exacte du message de départ.

VI. - APPRENTISSAGE

Avant de parler des expériences et des quelques résultats obtenus, nous allons parler de l'apprentissage et décrire très brièvement la méthode utilisée. Pour une étude plus approfondie cf [4] , [9] , [10] .

VI.1 - Principe général

Cet apprentissage permet de calculer la matrice optimale W^* en utilisant la méthode itérative suivante :

Soit $W = (W_1,\ldots,W_j,\ldots,W_p)$ W_j étant le $j^{ème}$ vecteur colonne $p = m - 1$

on obtient à la $(k+1)^{ème}$ itération les p vecteurs $W_j(k+1)$ en posant

(10)
$$W_j(0) = \text{un vecteur quelconque} \qquad j = 1,p$$
$$W_j(k+1) = W_j(k) - \varepsilon \nabla_W e_j^2(k) \qquad j = 1,p$$

ε : constante positive très petite permettant de contrôler la convergence de la suite.

∇_W gradient par rapport à W.

On montre que lorsque $k \rightsquigarrow \infty$ $W_j(k) \rightsquigarrow W_j^*$ $\forall j$ et donc que $W(k) \rightsquigarrow W^*$ lorsque k tend vers l'infini

cf [4 - 10]

VI.2 - Apprentissage particulier des coefficients des voyelles et des consonnes

En fait on calcule 6 matrices optimales différentes :

- une première matrice W^*_V pour la séparation de la classe des voyelles
- une deuxième matrice $W^*_{C_1A}$ pour la séparation de la classe des consonnes C accompagnées de la voyelle de base A.
- 3ème matrice $W^*_{C_1E}$ pour la séparation de la classe des consonnes C_1 accompagnées de la voyelle de base E.
- 4ème matrice $W^*_{C_1I}$ pour la séparation de la classe des consonnes C_1 accompagnées de la voyelle de base I.
- 5ème matrice $W^*_{C_1O}$ pour la séparation de la classe des consonnes C_1 accompagnées de la voyelle de base O.
- 6ème matrice $W^*_{C_1U}$ pour la séparation de la classe des consonnes C_1 accompagnées de la voyelle de base U.

Pour ce calcul, le locuteur prononce un certain nombre de fois chaque voyelle dans un ordre prédéterminé (ou chaque syllabe) et ensuite le programme réévalue W à chaque itération, c'est-à-dire à chaque prononciation à l'aide de la relation (10) pour obtenir les 6 matrices optimales W* utilisées lors de la reconnaissance des phonèmes.

Remarque : Actuellement on refait un apprentissage pour chaque locuteur.

VII. - EXPERIENCES - RESULTATS

Deux types de programmes ont été mis au point :

1°) Des programmes en temps semi-réel sur Ramsès 1 L (calculateur maison, cycle de base 4-5 µs) qui permettent d'appliquer les algorithmes dans des conditions de fonctionnement normal et de tester la validité de ces algorithmes (fig. 12).

Pour un locuteur donné, il reconnaît 15 voyelles avec un pourcentage de reconnaissance de 85 % environ, ces voyelles étant reconnues à l'intérieur de n'importe quelle syllabe et de n'importe quel mot. Si en plus on tient compte de plusieurs réponses possibles le pourcentage atteint facilement 95 %.

Malheureusement, pour les consonnes les résultats sont moins bons : ils varient de 50 à 65 % pour les 20 consonnes pour une syllabe donnée : le pourcentage s'améliore malgré tout si on tient compte de plusieurs réponses possibles. Une grosse partie de ce mauvais pourcentage est certainement due à la trop simple paramétrisation et nous espérons améliorer ces résultats (fig. 13 - 14 - 15 - 16).

Au point de vue segmentation en syllabes, le pourcentage d'erreur serait de l'ordre de 10 % mais il est très difficile de donner des chiffres car finalement tout dépend de la manière dont on prononce les mots.

Le temps de réponse non optimisé est de l'ordre de 5 à 10 s pour une syllabe, sur le calculateur Ramsès 1 L.

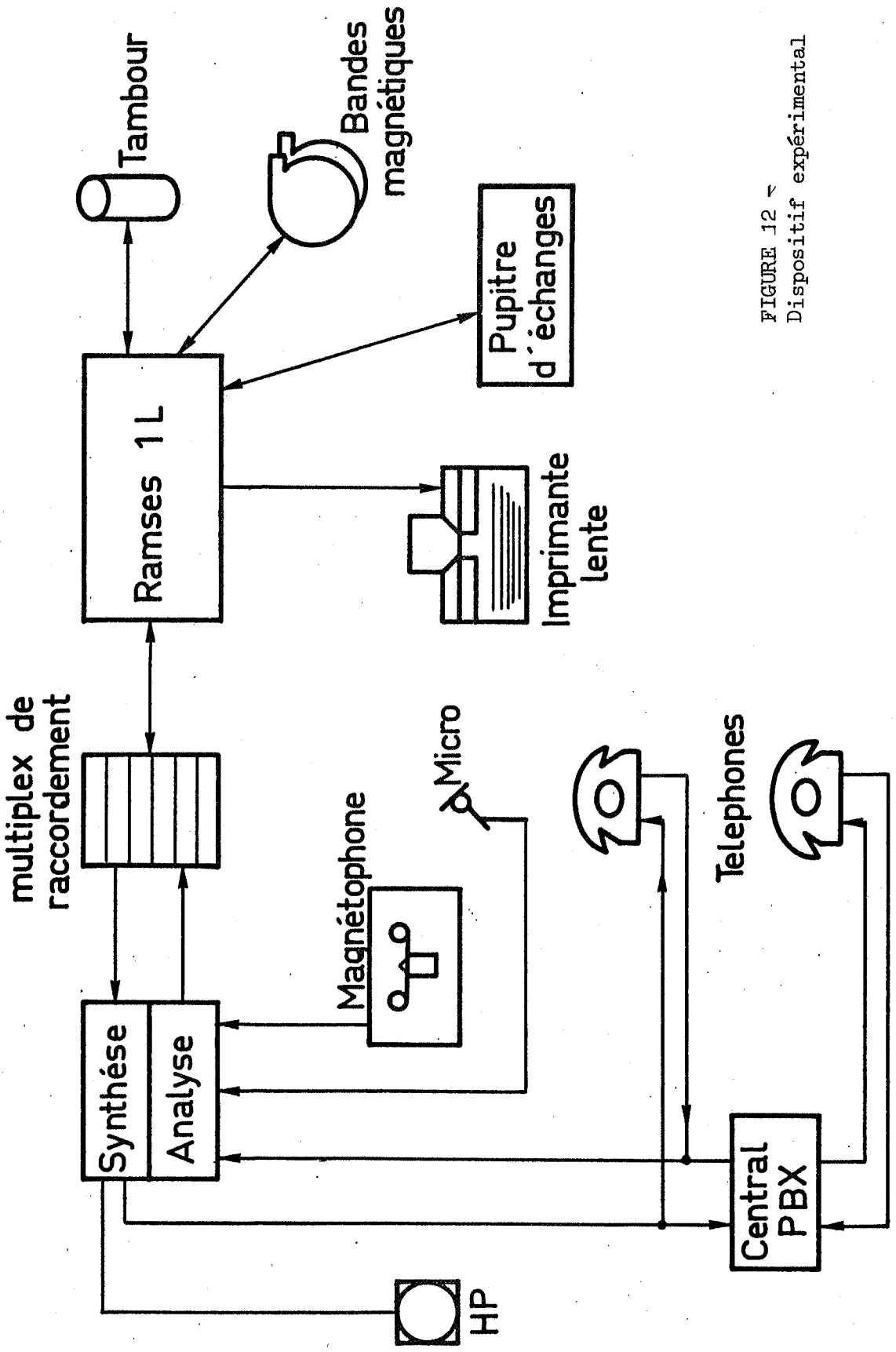


FIGURE 12 -
Dispositif expérimental

FIGURE 13

Reconnaissance des syllabes dans les mots

MOT PRONONCE :ZERO

SYLLABE NO	CONSONNES Cl	VOYELLES
1	L Z	AI U ET EI I
2	H R P	OU O

PREMIERE REPONSE:LAI HOU

MOT PRONONCE: ZERO

SYLLABE NO	CONSONNES Cl	VOYELLES
1	L Z N	AI ET I EI
2	H R P	O OU

PREMIERE REPONSE: LAI HO

MOT PRONONCE: UN

SYLLABE NO	CONSONNES Cl	VOYELLES
1	H N K	UN A AN

PREMIERE REPONSE: HUN

MOT PRONONCE: DEUX

SYLLABE NO	CONSONNES Cl	VOYELLES
1	P H B	E EU

PREMIERE REPONSE: PE

MOT PRONONCE: >TROIS

SYLLABE NO	CONSONNES Cl	VOYELLES
1	W	A AN AU

PREMIERE REPONSE:WA

FIGURE 14

Reconnaissance des syllabes dans les mots

MOT PRONONCE: QUATRE

SYLLABE NO	CONSONNES Cl	VOYELLES
1	K T D P	A UN ON EU
2	R W	AU E O

PREMIERE REPONSE: KA RAU

MOT PRONONCE: CINQ

SYLLABE NO	CONSONNES Cl	VOYELLES
1	D P T Z Y	UN A

PREMIERE REPONSE: DUN

MOT PRONONCE: PLUS

SYLLABE NO	CONSONNES Cl	VOYELLES
1	L	U I

PREMIERE REPONSE: LU

MOT PRONONCE: MOINS

SYLLABE NO	CONSONNES Cl	VOYELLES
1	W N	UN AN A A

PREMIERE REPONSE: WUN

MOT PRONONCE: ≠?'∞EGAL

SYLLABE NO	CONSONNES Cl	VOYELLES
1	P T	I U ET
2	N D K W T	A UN ON

PREMIERE REPONSE: πPI NA

FIGURE 15

Reconnaissance des syllabes dans les mots

MOT PRONONCE: BANQUE DE DONNEES

SYLLABE NO	CONSONNES Cl	VOYELLES
1	P B	AN ON A
2	K T P B	E EU
3	N Z Y GN	E ON EU
4	GN	AU
5	M	ET I EI

PREMIERE REPONSE: PAN KE NE GNAU MET

MOT PRONONCE: APPRENTISSAGE

SYLLABE NO	CONSONNES Cl	VOYELLES
1	H	A ON AN
2	N	AN
3	T	I
4	S	ON A A UN EU
5	J	E EU ON

PREMIERE REPONSE: HA NAN TI SON JE

MOT PRONONCE: AUXILIAIRE

SYLLABENO	CONSONNES Cl	VOYELLES
1	P T K	AU O
2	Z S H G YN	I
3	CH Y	EI AI U
4	R T	OU AU E

PREMIEREREPONSE:PAU ZI CH EI RROU

FIGURE 16

Reconnaissance des syllabes dans les mots

MOT PRONONCE RECONNAISSANCE

SYLLABE NO	CONSONNES Cl	VOYELLES
1	R W	E AU
2	K P B	AU
3	Z L Y H	U EI ON AI I
4	W P K D Z S L T B	AN A AU ON

PREMIERE REPONSE RE KAU ZU WAN

MOT PRONONCE APPOSITION

SYLLABE NO	CONSONNES Cl	VOYELLES
1	H B P	A ON
2	T P H K D	AU O
3	W Z Y N L GN B V	I U
4	Y M Z	AN ON AU OU

PREMIERE REPONSE HA TAU WI YAN

MOT PRONONCE COMPLEMENT

SYLLABE NO	CONSONNES Cl	VOYELLES
1	T K P WG	ON OU AN
2	P H T B K	E ON
3	H L Y Z N GN	U I ET AI EI
4	N M Z	AN OU

PREMIERE REPONSE TON PE HU NAN

MOT PRONONCE AU REVOIR

SYLLABE NO	CONSONNES Cl	VOYELLES
1	H R	O OU
2	R W	AU E O OU
3	W R B V	A ON UN

PREMIERE REPONSE HO RAU WA

2°) Des programmes en temps différé sur le calculateur CII 10070

- Actuellement le programme d'apprentissage des coefficients est au point et va donc permettre de faire un apprentissage plus long et meilleur. D'autre part un programme d'accélération de la convergence est également en préparation. Ce programme permettra ainsi de calculer les coefficients W utilisables ensuite par le programme de reconnaissance en temps réel sur Ramsès.

D'autre part le programme de reconnaissance des syllabes fonctionne également en temps différé, ce qui nous permet de modifier la paramétrisation des consonnes et d'autre part d'utiliser un petit dictionnaire et de voir ainsi le vocabulaire que peut reconnaître le programme de reconnaissance.

CONCLUSION

Les résultats actuels ne sont sans doute pas suffisants pour permettre une conclusion définitive. La segmentation n'est qu'une méthode très simple parmi d'autres ; elle permet cependant d'obtenir certains résultats malgré sa simplicité ; quant à la reconnaissance, il faut attendre une amélioration de la paramétrisation des consonnes et également une amélioration de l'algorithme d'apprentissage auquel on doit ajouter un processus d'accélération de la convergence. Il serait également indispensable d'approfondir les études sur l'adaptation des coefficients à plusieurs locuteurs et surtout les études de syntaxe et de sémantique du langage de communication.

BIBLIOGRAPHIE

(1) Keiichi Abe, Kazuo Hatano et Teruo Fukumara : Performance evaluation of word recognition system with dictionary Electronics and Communications in Japan Vol 52-6, n° 6 - 1969

(2) CHAPLIN (W.G.), LEVADI (V.S.) A generalization of the linear threshold decision algorithm to multiple classes. Computer and Information Sciences II, Tou (A.P.), New-York, London (1967)

(3) DREYFUS-GRAF (J.A.) : La parole humaine et l'informatique : Automatismes. Dunod - n° 9 septembre 70

(4) GLADYSHEV (E.G.) On stochastic approximation. Theory of probability and its applications - U.S.A. (1963) 10 pp. 275-278

(5) GRESSER (J.Y.) Reconnaissance automatique de mots et de langages parlés, dans le compte rendu des journées d'études sur la parole, du GALF et de l'AFCET, édité par l'ENSERG, février 70

(6) GRESSER (J.Y.), MERCIER (G.) Exemple de reconnaissance automatique de la parole. Commutation et électronique, n° 32 - janvier 71 pp 48 à 64.

(7) MALMBERG B. Structural linguistics and human communication Springer - Verlag - Berlin - Göttingen - Heidelberg 1963

(8) MENON (K.M.N.), JENSEN (Paul, J.) and DEW (Donald) Acoustic properties of certain VCC utterances The journal of the acoustical society of America Vol 46 - n° 2 (1969) pp 449 - 457

(9) MERCIER (G.) Reconnaissance des formes. Approximation des fonctions de décision et application à la reconnaissance des phonèmes - Thèse 3ème cycle - Faculté des Sciences, Université de Rennes (1969)

(10) MERCIER (G.) Approximation stochastique et reconnaissance acoustique d'un vocabulaire limité Annales des Télécommunications - tome 25 - n°s 5-6 - mai-juin 1970

(11) Naoyuki Okada and Tuneo Tamati : Semantic information of natural language and its extraction and classification Electronics and Communications in Japan, vol. 52-C, n° 10 - 1969.

(12) SARIDIS (G.) Learning applied to successive approximation algorithms IEEE transactions on systems science and cybernetics, vol SSC. 6 n° 2 avril 70

(13) VYSOTSKIY (G. YA.) RUDNYY (B.N.), TRUBIN-DONSKOY (V.N.) et TSEMEL (G.I.) An experiment in oral control of a computer engineering cybernetics, n° 2, 1970. pp 320-327.

APPLICATION DES TECHNIQUES STATISTIQUES A LA RECONNAISSANCE DE LA PAROLE

C . BERGER - VACHON

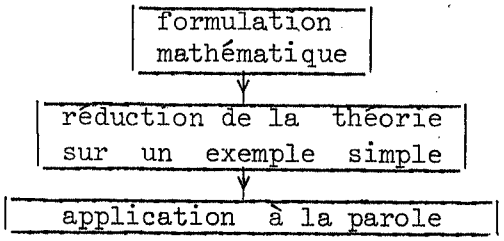
Laboratoire de Physique Electronique - Faculté des Sciences - LYON

1. BUT DE CET EXPOSE

Les techniques générales de reconnaissance des formes sont applicables au domaine acoustique. Le but de cet exposé est de montrer une formulation mathématique du problème, pris sur un exemple simple, son adaptation à une discrimination acoustique, en tenant compte de contraintes électroniques. Nous sauterons les démonstrations mathématiques qui peuvent être trouvées dans la référence [3]. Une abstraction de la technique utilisée peut être également trouvée en [1].

L'exposé développe la base de ces techniques et aujourd'hui les résultats qu'il énonce présentent un intérêt quelque peu "historique". Le schéma ci-dessous indique le plan de ce qui va suivre :

FIGURE 1 - Schéma de l'exposé



2. FORMULATION MATHEMATIQUE

On considère r populations notées $\omega_1, \omega_2, \ldots, \omega_r$ que nous nous proposons de distinguer.

De la première population, on tire n_1 échantillons ; de la deuxième, n_2 échantillons, ... , de la r^{ème}, n_r échantillons. Chacun de ces échantillons se compose de p caractères continus.

Nous pouvons noter le k^{ème} échantillon de la i^{ème} population par :
$x_i^{(k)} = (x_{i1}^{(k)} , x_{i2}^{(k)} , \ldots , x_{ip}^{(k)})$.

On suppose que les effectifs des échantillons $n_1, n_2, \ldots n_r$ sont assez grands pour que l'on puisse utiliser l'approximation gaussienne des densités de probabilité.

Le j^{ème} caractère du i^{ème} échantillon est alors distribué approximativement selon une loi normale :

$$X_i^{(j)} \rightarrow N(m_{ij}, \sigma_{ij})$$

où m_{ij} et σ_{ij} sont les estimateurs sans biais de la moyenne et de la variance :

$$m_{ij} = \frac{1}{n_i} \sum_{s=1}^{n_i} x_{ij}^{(s)} \quad \text{et :} \quad \sigma_{ij}^2 = \frac{1}{n_i - 1} \sum_{s=1}^{n_i} (x_{ij}^{(s)} - m_{in})^2$$

Abandonnons cette formulation générale pour quelque chose de plus concret. On pourrait, bien entendu, continuer à développer la théorie à l'aide d'un tel langage.

3. EXEMPLE SIMPLE

Représentons les échantillons dans un espace réduit à deux caractères ($p = 2$) : X_1 et X_2 ; supposons de plus qu'il n'y a que deux populations ($r = 2$).

Par exemple, ω_1 représente la prononciation de "ZERO" et ω_2 celle de "UN". Chaque chiffre est caractérisé par deux mesures X_1 et X_2. (Voir paragraphe 4).

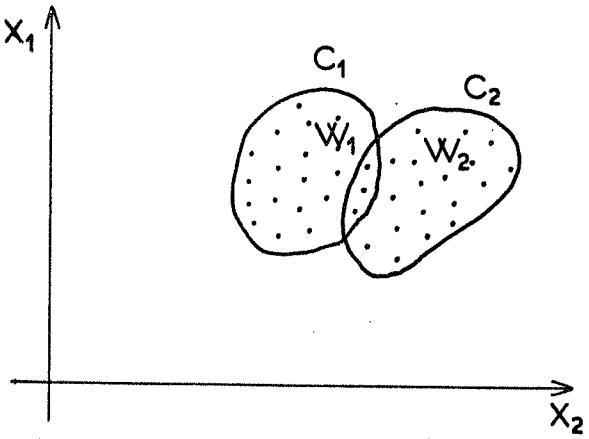


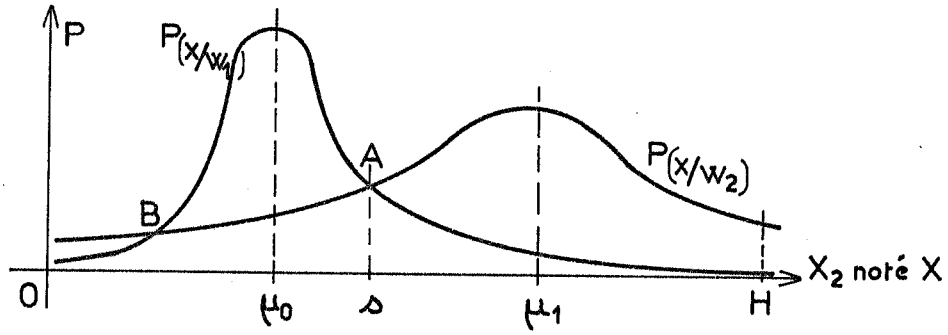
FIGURE 2 -Classes représentant les échantillons de ω_1 et ω_2 dans un espace à deux dimensions.

Les différentes observations (ou échantillons) se placent à l'intérieur des contours C_1 et C_2 (figure 2).

Il faut trouver une fonction de décision simple permettant de séparer les masses ω_1 et ω_2.

Si on représente la densité de la projection des points de ω_1 sur la dimension X_1, et celle des points de ω_2 sur la dimension X_2, on observe deux densités gaussiennes centrées en μ_0 et μ_1 (figure 3) où $P(X/\omega_i)$ est une probabilité conditionnelle.

FIGURE 3 - Répartition probabilistique de ω_1 et ω_2 sur la dimension X_2.



O et M représentent les limites de variation de X_2.

Effectuons une discrimination de type descriptif introduite par une matrice de coût du type suivant :

$$L = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Si on évalue le coût moyen minimum de misclassification on trouve les formules :

$$\boxed{R = \sum_{i=1}^{2} P(\omega_i)\,[1 - q_i]} \qquad (1)$$

où $P(\omega_i)$ est la probabilité "a priori" de la classe ω_i

et $$q_i = \int_{D_i} P(x/\omega_i)\ dx \qquad (2)$$

D_i est le domaine où on effectue la décision d_i, c'est-à-dire que si on observe $x \in D_i$, on décidera que ω_i était présent à l'entrée.

q_i est le pourcentage de bonne classification de la classe ω_i.

La théorie de minimisation montre que D_i peut être déterminé par la formule ci-dessous où $P(x,\omega_i)$ est la probabilité conjointe d'observer x et ω_i.

$$D_i = \left\{ x : \frac{P(x,\omega_i)}{P(x,\omega_j)} > 1 \right\} \qquad (3)$$

D'après le théorème de Bayes, on peut écrire que :

$$P(x,\omega_i) = P(x/\omega_i) \quad P(\omega_i) \qquad (4)$$

Par application de la formule (3) nous placerons la limite pour x = s tel que

$$P(x,\omega_1) = P(x,\omega_2) \qquad (5)$$

Soit, lorsque les classes sont équiprobables (équation 4)

$$\boxed{P(s/\omega_1) = P(s/\omega_2)} \qquad (6)$$

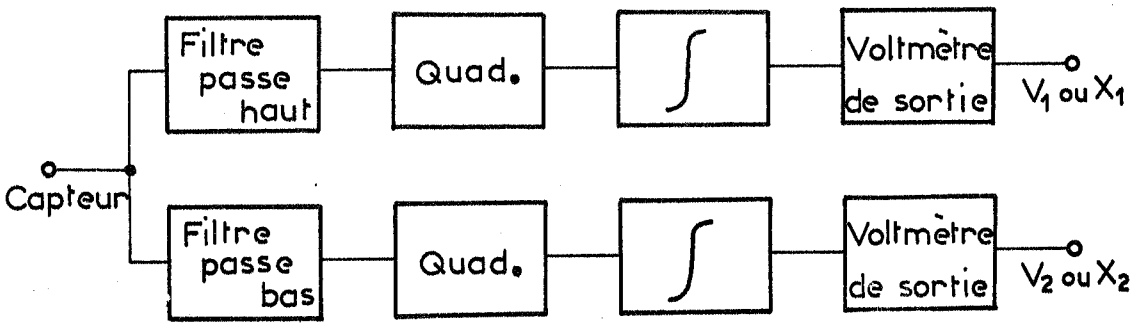
s est la position du seuil de décision. Elle constitue une approximation car il y a deux points A et B (figure 3) qui conduisent à l'équation (5).

Donc $D_0 = \{ x : P(x/\omega_1) > P(x,\omega_2)\}$, c'est-à-dire D_0 est représenté par le segment Os sur la figure 3 et de la même manière on peut voir que D_1 = sM.

4. APPLICATION

Considérons le vocoder "squelettique" de la figure 4.

FIGURE 4 - Génération des paramètres X_1 et X_2



La tension de sortie V_1 représente la valeur de l'énergie haute fréquence pendant un intervalle de temps (ex : 600 ms) lors de la prononciation d'un digit. V_2 représente l'énergie BF pendant le même temps.

La prononciation de "ZERO" conduit donc à deux valeurs qui peuvent être placées dans le plan X_1 X_2 de la figure 2. Appliquons la théorie résumée ci-dessus pour placer un seuil de décision entre "ZERO" et "UN" sur le canal BF.

On construit les distributions de probabilité des valeurs V_2 pour la prononciation des chiffres, comme cela a été indiqué sur la figure 3.

L'intersection des deux gaussiennes, située entre μ_0 et μ_1, conduit à la détermination théorique de la valeur du seuil de décision s. On peut ensuite estimer la qualité de cette discrimination à l'aide de l'équation (1) qui donne le pourcentage de reconnaissance à espérer, si on connaît les probabilités de prononciation a priori de chacun des chiffres.

Remarquons que généralement on ignore $P(\omega_1)$ et $P(\omega_2)$. La politique de moindre risque conduit à prendre $P(\omega_1) = P(\omega_2)$, comme on pourrait le voir dans la référence (2).

BIBLIOGRAPHIE

1. INABA.H et HIRAMATSU.K :Characteristic evaluation function and decision function in pattern recognition (1966) - Tokyo Electrical Engineering Collège. TOKYO - Japon.

2. ABEND.K : Compound Decision Procedures for pattern classification (Décembre 1967). Technical Report AMRL.TR.67.10. Aerospace Medical Research lab. Wright Patterson Air Force Base. OHIO.

3. Ch. BERGER-VACHON et G. MESNARD : Evaluation de l'efficacité d'un système de paramètres dans un problème de reconnaissance des formes. L'Onde Electrique (Fasc.11. p.920 décembre 1970).

IDEES GENERALES
SUR LA RECONNAISSANCE DES FORMES
APPLIQUEE A LA PAROLE

C. ROCHE

Laboratoire de Reconnaissance des Formes
Institut de Programmation - Faculté des Sciences - PARIS

I - Introduction

1. La notion d'opérateur

Le laboratoire de reconnaissance des Formes de la Faculté des Sciences de Paris travaille exclusivement avec trois classes d'outils :

- les ordinateurs
- les capteurs
- la matière grise

Nous passerons sur la troisième classe et ne décrirons que les deux premières en montrant leurs analogies. Nous laisserons au lecteur la possibilité de philosopher et de penser que ces analogies recouvrent aussi le domaine de la troisième classe.

Les capteurs que l'on utilise couramment sont :

- des capteurs optiques : camera dont la sortie est discrétisée ou flying spot comme le Calife du LCA. Ces capteurs, donnent une représentation discrète et matricielle de l'intensité sur une surface rectangulaire plane d'un champ électromagnétique à fréquence visible. Ce champ est modulé soit par une "diapositive" (Calife) soit pour la caméra par les objets vus par l'objectif.

Un capteur optique est donc un opérateur qui a un certain champ électromagnétique associe une valeur discrète ou une série de valeurs discrètes, qui sont les résultats de cet opérateur ou attributs. Ce sont ces résultats qui constituent les données introduites en ordinateurs en vue d'un traitement (de reconnaissance le cas échéant).

- des capteurs acoustiques : vocodeur à canaux délivrant un ensemble de valeurs toutes les 25 ms, ou discrétiseur à 10 KHZ. Les capteurs transforment un champ de pression mécanique en données digitales. Voilà encore des opérateurs.

Ces opérateurs sont ici des opérateurs figés cablés ayant des utilisations bien définies.

Nous disposons d'une autre classe d'opérateurs, les ordinateurs. Chacune des instructions-machine de l'ordinateur est un opérateur qui à un certain état de registre fait correspondre un autre état. La mise en marche de ces opérateurs est séquentielle dans le temps.

Un programme est un opérateur qui actionne ces opérateurs élémentaires et qui à partir de leurs résultats, donne lui-même un certain résultat qui correspondra à un certain état de l'ordinateur considéré comme un automate.

2. Identité entre la reconnaissance de forme et la notion d'opérateur.

a) La définition classique du problème général de la reconnaissance des formes est de trouver une méthode automatique qui partage un certain ensemble de données en sous-ensembles D_i ayant chacun un nom $N(D_i)$ qui correspond à celui que donnerait l'homme à l'une quelconque des données éléments de D_i.

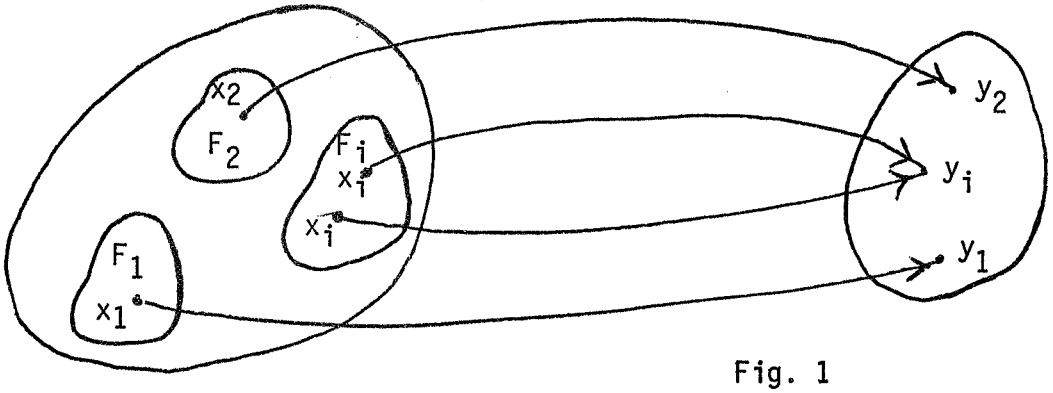


Fig. 1

La figure 1 montre d'un côté l'ensemble des données considérées partagé en sous ensembles (F_1, F_2, ..., F_i...).

De l'autre côté est représenté l'ensemble des noms $\{y_i = N(F_i)\}$.

Si l'on veut automatiser ce partage en sous-ensembles, il faut automatiser le calcul de la fonction $y_i = N(x_i)$

$$(N(x_i) = N(F_i), \forall x_i \in F_i)$$

Autrement dit créer un opérateur qui à la donnée x_i associe le résultat $y_i = N(x_i)$.

b) L'ordinateur sera donc ici considéré comme un opérateur, le plus riche et le plus souple des opérateurs réels qui existent actuellement. Nous l'utilisons donc tout naturellement pour faire de la reconnaissance de forme.

3. Notion de niveaux différents d'opérateurs

L'opérateur constitué par le capteur travaille sur l'extérieur pour fournir des données à l'ordinateur. Celui-ci à son tour travaille sur ces données pour fournir à l'homme le "nom" de la forme reçue par le système. D'où la notion de deux niveaux d'opérateurs de reconnaissance (si l'on excepte l'homme).

De la même façon, considérons un programme comme formé de différents étages d'intégration. A chacun de ces étages travaillent des opérateurs sur les résultats des opérateurs du niveau immédiatement inférieur, et dont les résultats servent d'entrée aux opérateurs du niveau immédiatement supérieur.

4. Exemples de tels opérateurs

Nous reporterons le lecteur à la bibliographie donnée en annexe et qui montre des exemples d'opérateurs :

- arithmétiques : tels que les filtres linéaires digitaux servant à extraire des données fournies par un capteur de signaux visuels, les "traits" ayant une certaine orientation [1] [8].
- syntaxiques : tels que l'opérateur [2] qui peut servir à reconnaitre des constellations de points disséminés dans un plan (recalage sur les étoiles) ou servir dans un étage supérieur pour reconnaitre des formes visuelles constituées de segments. Dans ce dernier cas, l'étage inférieur peut être un ensemble de filtres linéaires tels que ceux décrits en [1].

Un cablage ou un programme (FFT) qui "calcule" la transformée de Fourier est aussi un opérateur c'est un opérateur arithmétique, qui se place en général au niveau inférieur dans les traitements d'image, à l'étage immédiatement supérieur à celui des capteurs, à moins qu'il ne soit même avant ces capteurs comme dans les traitements optiques.

5. Qualités des opérateurs

Deux qualités sont nécessaires aux opérateurs de reconnaissance des formes : l'invariance à certaines transformations et l'information qu'ils apportent quant au but à rechercher (dite information semantique).

L'opérateur Transformée de Fourier à 2 dimensions est invariant aux translations mais non aux rotations et similitude. L'invariance aux translations lui procure certains avantages importants. Une étude qui pourrait être faite est la mesure de son information sémantique lors des cas particuliers de réalisations pratiques (en cablé, programmé, en optique...).

II - Idées théoriques de base sur la mesure de l'information des opérateurs

1. Les opérateurs O_j et O_{id}

Considérons un ensemble de données sur lesquels vont travailler un ensemble d'opérateurs O_j. Pour chaque donnée, O_j a comme résultat ou attribut, un des éléments de

$$\{a_j(k) \;/\; k \in (1, 2, \ldots, m_j)\}$$

Dans les exemples fournis par la suite ces opérateurs O_j seront effectivement des sous-programmes et $a_j(k)$ sera la valeur de la variable de sortie du sous-programme.

Sur ces données peut travailler (du moins pendant la phase d'apprentissage un opérateur O_{id} (pour "opérateur idéal") ayant comme résultat $\omega(i)$. Cet opérateur est au départ représenté par l'homme ou le "professeur" indiquant à l'ordinateur quel est le nom $\omega(i)$ de la forme qui d'après lui correspond à la donnée considérée. C'est cet opérateur qu'il faudra approcher le plus possible par un opérateur automatique en ordinateur.

2. Introduction des probabilités sur les résultats des opérateurs

Associons à chaque valeur $\omega(i)$ une probabilité à priori $p[\omega(i)]$. C'est la probabilité à priori d'avoir la forme $\omega(i)$ sans rien encore connaitre sur la donnée à analyser.

Analysons maintenant la donnée avec l'opérateur O_j : s'il donne le résultat $a_j(k)$, la probabilité d'avoir la forme $\omega(i)$ devient :

$$p[\omega(i) \mid a_j(k)]$$

On peut de même considérer les probabilités suivantes :

$$p[\omega(i), a_j(k)]$$
$$p[a_j(k)]$$
$$p[a_j(k) \mid \omega(i)]$$

Nous avons entre ces nombres les relations suivantes :

$$p[\omega(i) \mid a_j(k)] = \frac{p[\omega(i), a_j(k)]}{p[a_j(k)]}$$

et

$$(1) \qquad p[\omega(i) \mid a_j(k)] = p[\omega(i)] \frac{p[a_j(k) \mid \omega(i)]}{\sum_i p[\omega(i)] . p[a_j(k) \mid \omega(i)]}$$

(relation de Bayes).

3. Estimation des opérateurs O_j

Si nous trouvons un résultat $a_j(k)$ tel que les valeurs $p[\omega(i) \mid a_j(k)] = 0$ pour toutes les valeurs de i sauf l'une d'elle i_1 pour laquelle $p[\omega(i_1) \mid a_j(k)] = 1$; la probabilité à postériori d'avoir la forme $\omega(i_1)$ est 1. L'opérateur O_j a dans ce cas extrait la forme $\omega(i_1)$ de la donnée considérée.

Si nous trouvons un résultat $a_j(k)$ tel que pour toutes les valeurs de i

$$p\,[\omega(i) \mid a_j(k)\,] = p\,[\omega(i)\,],$$

la "situation" probabiliste n'a pas changé et on dit intuitivement que l'opérateur O_j n'a apporté aucune "information"...

La théorie des questionnaires dit que le minimum du nombre moyen de questions binaires (réponse par oui ou non) à poser pour obtenir une probabilité 1 est avant l'expérience, l'entropie :

$$H\,[\Omega\,] = -\sum_i p\,[\omega(i)\,]\ \mathrm{Log}\ p\,[\omega(i)\,] \tag{2}$$

après l'expérience, l'entropie conditionnelle :

$$H\,[\Omega \mid a_j(k)\,] = -\sum_i p\,[\omega(i) \mid a_j(k)\,]\ \mathrm{Log}\ p\,[\omega(i) \mid a_j(k)\,] \tag{3}$$

(le logarithme est de base 2).

L'information apportée par la réponse est :

$$I\,[\Omega \mid a_j(k) = H\,[\Omega\,] - H\,[\Omega \mid a_j(k)\,] \tag{4}$$

On sait (FANO [3]) que $I\,[\Omega\,;\,a_j(k)\,]$ est positif dans tous les cas.

La formule (4) montre l'information apportée par la réponse $a_j(k)$ de l'opérateur O_j ; si on passe à la moyenne sur tous les $a_j(k)$, on obtient l'information mutuelle moyenne entre O_j et O_{id} :

$$I\,[A_j\,;\,\Omega\,] = \sum_{i,k} p\,[\omega(i),\,a_j(k)\,]\ \mathrm{Log}\ \frac{p\,[\omega(i),\,a_j(k)]}{p\,[\omega(i)\,].p\,[a_j(k)\,]} \tag{5}$$

Cette quantité représente l'information au sens de Shannon de l'opérateur considéré dans le cas de la reconnaissance des formes comme un canal d'information.

C'est cette quantité proposée par LEWIS [4] (voir FU [5]) que nous avons utilisée pour estimer les opérateurs dans les problèmes de reconnaissance de forme.

4. Association d'opérateur

a) Opérateurs en parallèle :

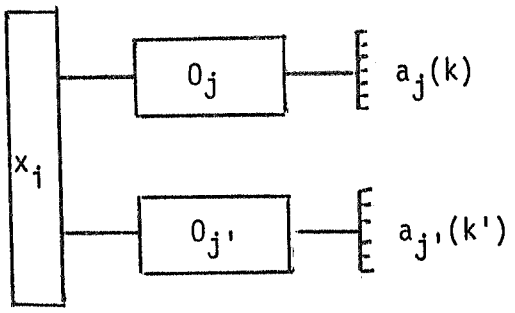
Mettre en parallèle deux opérateurs O_j et $O_{j'}$ sur une certaine donnée c'est considérer l'opérateur $O_j \otimes O_{j'}$ qui donne comme résultat $[a_j(k),\,a_{j'}(k')]$.

Quelle est l'information apportée par l'ensemble de ces résultats ? Et quelle est la moyenne de cette information ?

On démontre facilement la relation suivante :

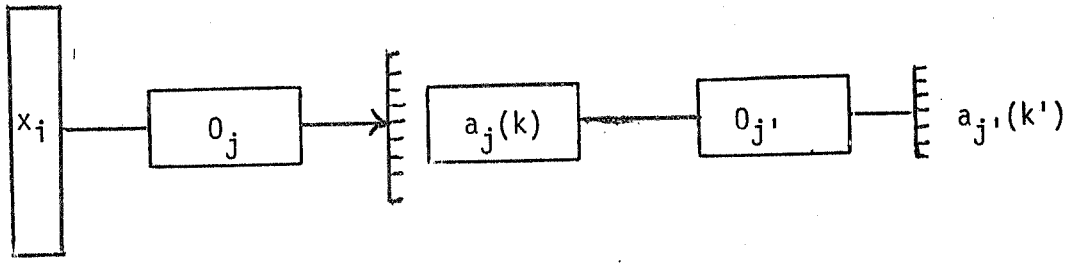
$$I(A_j ; \Omega) + I(A_{j'} ; \Omega) - I(A_j ; A_{j'})$$

$$\leqslant I(A_j \otimes A_{j'} ; \Omega) \leqslant I(A_j ; \Omega) + I(A_{j'} ; \Omega)$$



Cela signifie que l'ensemble de deux opérateurs en parallèle est d'autant plus informant que chacun d'eux l'est, mais aussi qu'ils sont d'autant plus indépendants.

b) Opérateurs en série :
Mettre en série deux opérateurs, c'est faire travailler l'un sur les résultats de l'autre.



On démontre de la même façon que dans ce cas :

$$I(A_{j'} ; \Omega) \leqslant I(A_j ; \Omega)$$

Cela signifie que toute opération sur des données diminue l'information apportée par ces données.

III - Application pratique : reconnaissance de sons particuliers à la parole

1. Remarques préliminaires

a) Les probabilités

Nous avons voulu appliquer ces méthodes dans des domaines pratiques précis.

Dans chaque cas nous avons donc calculé l'information apportée par un opérateur en utilisant la formule (5). Cette formule utilise des probabilités d'occurence. Ces dernières sont calculées pendant une période d'apprentissage. Un certain nombre de données sont présentées à l'ordinateur, chacune avec le nom de la forme (phonèmes dans le cas de la parole) correspondante établie par le "professeur", en d'autres termes avec le résultat de l'opérateur expérimental O_{id}.

A chacune de ces données on trouve les résultats $a_j(k)$ des différents opérateurs. A chaque occurence commune de $\omega(i)$ et $a_j(k)$, des cases indicées pour ces deux nombres sont incrementées et si le nombre de données N est suffisant et si l'on a ergodicité, le contenu de ces cases représente avec une certaine erreur ϵ et à la constante multiplicative N près la probabilité

$$p\,[\omega(i),\ a_j(k)\,]$$

b) Reconnaissance séquentielle

Si on a le résultat $a_j(k)$ de l'opérateur O_j, les probabilités $p\,[\omega(i)\,]$ se transforment donc en

$$p\,[\omega(i)\mid a_j(k)\,]$$

comme il a été expliqué en II - 2. Le calcul de cette dernière quantité se fait par la formule de Bayes (1).

Si maintenant on considère le résultat $a_{j'}(k')$ de l'opérateur $O_{j'}$, les nouvelles probabilités sont :

$$p\,[\omega(i)\mid a_j(k),\ a_{j'}(k')\,]$$

Dans le cas où O_j et $O_{j'}$ sont indépendants on peut appliquer la formule de Bayes pour calculer cette dernière quantité :

$$p\,[\omega(i)\mid a_j(k),\ a_{j'}(k')\,] = p\,[\omega(i)\mid a_j(k)\,]\ \frac{p\,[a_{j'}(k')\mid \omega(i)\,]}{\sum_i p\,[\omega(i)\mid a_j(k)\,]\ p\,[a_{j'}(k')]}$$

Mais dans le cas où O_j et $O_{j'}$ ne sont pas indépendants cette formule est fausse et si elle est appliquée, donnera une erreur d'autant plus grande en moyenne que $I(O_j, O_{j'})$ est grand.

D'où, dans le cas de l'utilisation de la formule de Bayes en analyse séquentielle une autre raison encore plus impérative que la première (II - 4.) pour que $I(O_j, O_{j'})$ soit le plus faible possible.

2. Lecture de mots écrits à la main (analyse séquentielle)

Un programme a été fait qui lit des mots écrits à la main sur un maillage de cellules photoelectriques. Nous reportons le lecteur à [1] et [9] qui exposent la méthode utilisant les idées précédemment décrites pour la sélection automatique des lettres à lire et des caractéristiques intéressantes de ces lettres.

3. Reconnaissance de la parole (analyse séquentielle)

Le but de ce programme a été de détecter les "phonèmes" de manière probabiliste en utilisant la même méthode dans des données expérimentales fournies par un "vocodeur". Ce vocodeur est le vocodeur à 12 canaux du CNET qui supprime beaucoup de fréquences hautes intéressantes. La figure 2 montre la sortie du vocodeurlors de la prononciation des mots "Forte pluie" ; on obtient chaque ligne toutes les 23 millisecondes, chaque ligne donne les valeurs digitalisées des filtres sélectifs et la valeur du "pitch".

a) On a fourni à l'ordinateur 17 opérateurs heuristiques tels que :

- existence d'un pitch different de zéro.
- nombre de maximum relatifs (formants).
- difference relative entre le 1er maximum et le minimum.
- difference entre le plus grand maximum et le plus petit minimum.
- variation de l'intensité moyenne entre deux lignes successives.

Le choix des opérateurs a été fait en recherchant ceux qui étaient invariants aux translations en fréquence et suivant l'axe des intensités (échelles logarithmiques).

b) Ensemble d'apprentissage. Pour l'établissement des probabilités

$$p\,[\omega(i),\ a_j(k)\,]$$

il faut présenter au professeur des données précises et qu'il soit capable de dire de quel phonème il s'agit. On utilise pour cela un programme de visualisation VISU dont on voit un exemple de listing sur la figure 2.

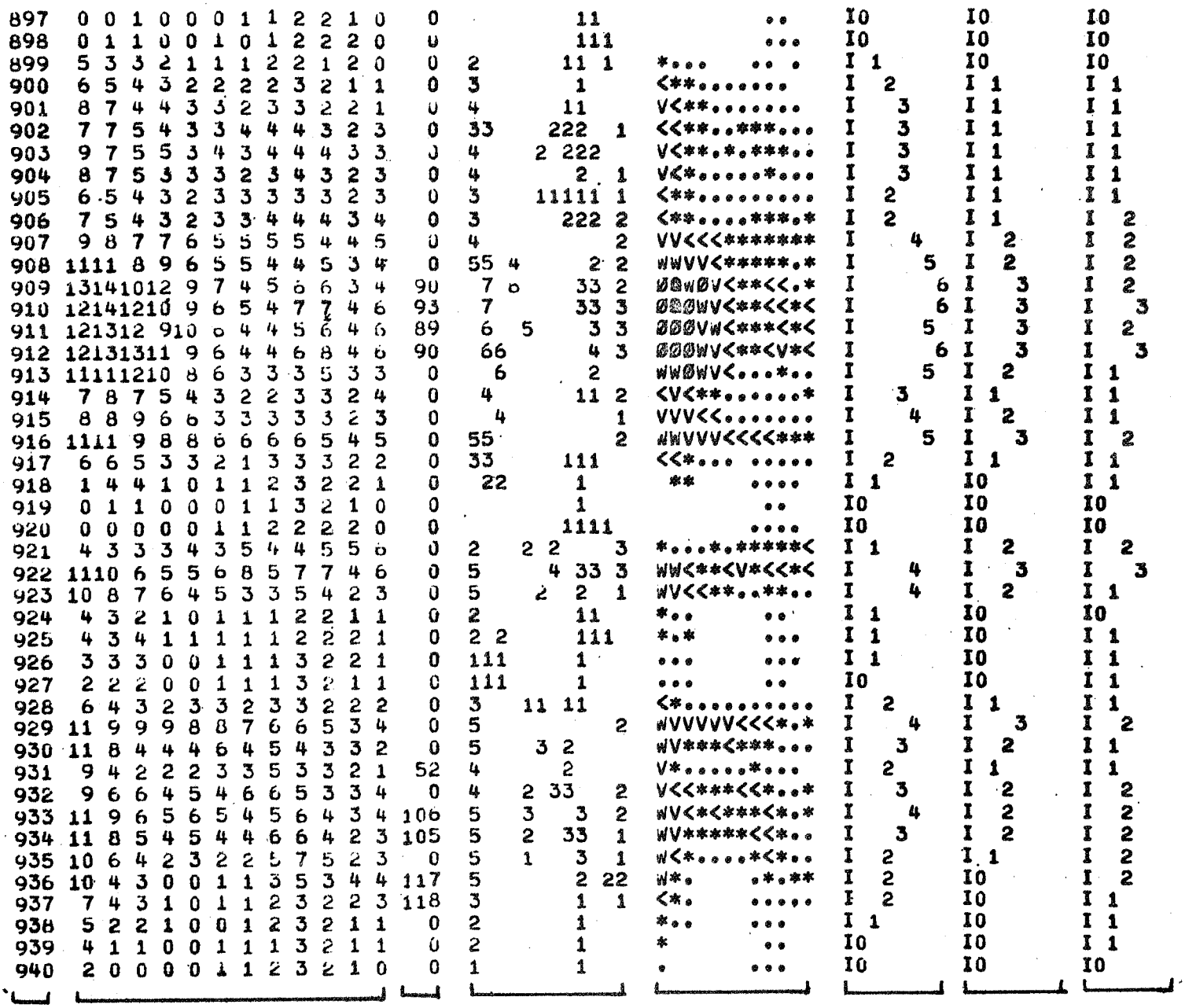
c) Le programme PROBA construit la dessus les tableaux des probabilités (Fig. 3).

d) CORREL donne tous les $\{I[A_j\ ;\ A_k\,]\ |\ j \neq k \in (1,\ 2..)\}$

e) SYNTHE travaille séquentiellement en utilisant la méthode décrite en II et déjà utilisée pour la lecture de lettres manuscrites (III - 2.) pour reconnaître des phonèmes dans un ensemble de données réelles. L'ensemble de données de test sur lesquelles travaille SYNTHE est différent de l'ensemble d'apprentissage.

La figure 4 montre les résultats du programme SYNTHE sur des données test représentant le mot STRATUS prononcé par un certain locuteur devant le vocodeur.

FORTE PLUIE



A B C D E F_1 F_2 F_3

A Line number

B Filter digitized outputs

C Pitch values

D "Formant" amplitude and frequency

E Another display of filter outputs

F_i Four filter output sum, i=1 first to four, i=2 five to eight, i=3 nine to

FIGURE 2

V I S U program giving the vocoder information, when the french words "forte pluie" are pronounced

C/f/10.

OPERATEUR NO : 1

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.075	.004	.030	.077	.080	.075	.035	.020	.011	.014	.007	.020	.009	.004	.024	.009	.022	.017	.033	.004	.671
2	.004	.004	.007	.004	.007	.004	.004	.004	.004	.007	.004	.011	.004	.028	.004	.014	.009	.004	.007	.017	.329

OPERATEUR NO : 2

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
2	.006	.003	.003	.003	.006	.003	.003	.003	.003	.003	.003	.003	.003	.009	.003	.012	.004	.003	.003	.004	.166
3	.007	.003	.003	.003	.003	.006	.006	.004	.003	.003	.003	.004	.006	.007	.006	.003	.004	.003	.003	.003	.111
4	.028	.003	.003	.003	.004	.012	.003	.004	.003	.003	.003	.007	.003	.003	.009	.003	.003	.003	.006	.004	.142
5	.011	.006	.006	.003	.011	.016	.011	.007	.004	.004	.003	.006	.003	.006	.006	.003	.003	.003	.007	.006	.145
6	.006	.003	.006	.009	.019	.014	.009	.004	.004	.006	.003	.007	.003	.003	.003	.003	.011	.004	.009	.004	.141
7	.003	.003	.007	.011	.016	.006	.004	.003	.004	.006	.004	.004	.003	.003	.003	.003	.006	.003	.006	.003	.110
8	.003	.003	.009	.034	.009	.006	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.009	.004	.003	.122

OPERATEUR NO : 3

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
2	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
3	.003	.006	.021	.021	.051	.011	.004	.007	.004	.011	.004	.012	.006	.015	.006	.011	.015	.011	.019	.007	.322
4	.003	.003	.003	.004	.004	.007	.006	.007	.003	.003	.003	.007	.003	.004	.007	.003	.004	.003	.006	.003	.112
5	.018	.003	.003	.029	.003	.034	.018	.003	.004	.003	.003	.003	.003	.006	.004	.003	.003	.003	.003	.006	.198
6	.009	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.076
7	.022	.003	.003	.003	.003	.003	.003	.003	.004	.003	.003	.003	.003	.003	.004	.003	.003	.003	.003	.003	.096
8	.006	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.071

OPERATEUR NO : 4

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.007	.002	.002	.005	.002	.002	.006	.002	.002	.002	.002	.004	.002	.004	.002	.004	.002	.002	.002	.002	.076
2	.018	.002	.002	.009	.002	.004	.011	.002	.004	.002	.002	.004	.002	.002	.004	.002	.002	.002	.006	.002	.109
3	.012	.004	.002	.004	.005	.004	.004	.002	.002	.005	.002	.005	.002	.002	.002	.002	.005	.002	.004	.004	.100
4	.006	.002	.004	.002	.004	.005	.002	.005	.004	.002	.002	.002	.002	.005	.004	.002	.002	.004	.002	.004	.096
5	.004	.004	.002	.002	.002	.026	.002	.005	.002	.002	.002	.002	.004	.002	.007	.002	.005	.004	.007	.004	.118
6	.004	.002	.002	.013	.026	.005	.004	.005	.004	.002	.004	.007	.004	.009	.004	.006	.007	.006	.006	.002	.146
7	.002	.002	.015	.015	.015	.007	.002	.002	.002	.005	.002	.004	.002	.002	.002	.005	.004	.002	.004	.002	.114
8	.002	.002	.004	.005	.004	.002	.004	.002	.002	.004	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.058
9	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.002	.061
10	.002	.002	.002	.004	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.055
11	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.055

$\|p[\omega(i), a_j(k)]\|$

For different $\&_j$

OPERATEUR NO : 5

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.007	.003	.018	.026	.008	.030	.005	.010	.007	.007	.003	.008	.003	.010	.007	.012	.010	.007	.007	.005	.256
2	.020	.007	.008	.012	.033	.012	.008	.003	.005	.005	.003	.003	.005	.007	.008	.003	.007	.007	.014	.005	.209
3	.020	.003	.003	.005	.020	.008	.007	.005	.003	.005	.005	.003	.003	.005	.008	.003	.007	.005	.007	.003	.173
4	.014	.003	.003	.012	.003	.005	.005	.005	.003	.005	.003	.005	.003	.003	.003	.005	.005	.003	.005	.003	.125
5	.005	.003	.003	.007	.003	.003	.010	.003	.003	.003	.003	.008	.003	.005	.003	.003	.003	.003	.003	.005	.102
6	.003	.003	.003	.008	.007	.010	.005	.003	.003	.003	.003	.007	.005	.005	.003	.003	.003	.003	.005	.007	.135

OPERATEUR NO : 6

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.006	.003	.003	.003	.003	.075
2	.020	.003	.003	.003	.008	.003	.012	.004	.003	.003	.003	.003	.003	.003	.006	.009	.004	.003	.003	.003	.173
3	.029	.003	.004	.008	.019	.029	.009	.009	.004	.004	.003	.008	.003	.013	.012	.003	.004	.003	.013	.006	.240
4	.006	.004	.017	.012	.025	.022	.004	.004	.003	.006	.004	.009	.006	.004	.004	.003	.015	.006	.004	.004	.213
5	.003	.004	.004	.009	.006	.004	.006	.003	.004	.006	.003	.006	.003	.004	.003	.003	.003	.004	.009	.006	.122
6	.003	.003	.004	.031	.008	.003	.003	.003	.004	.003	.003	.003	.003	.003	.003	.003	.003	.006	.004	.003	.108
7	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.070

OPERATEUR NO : 7

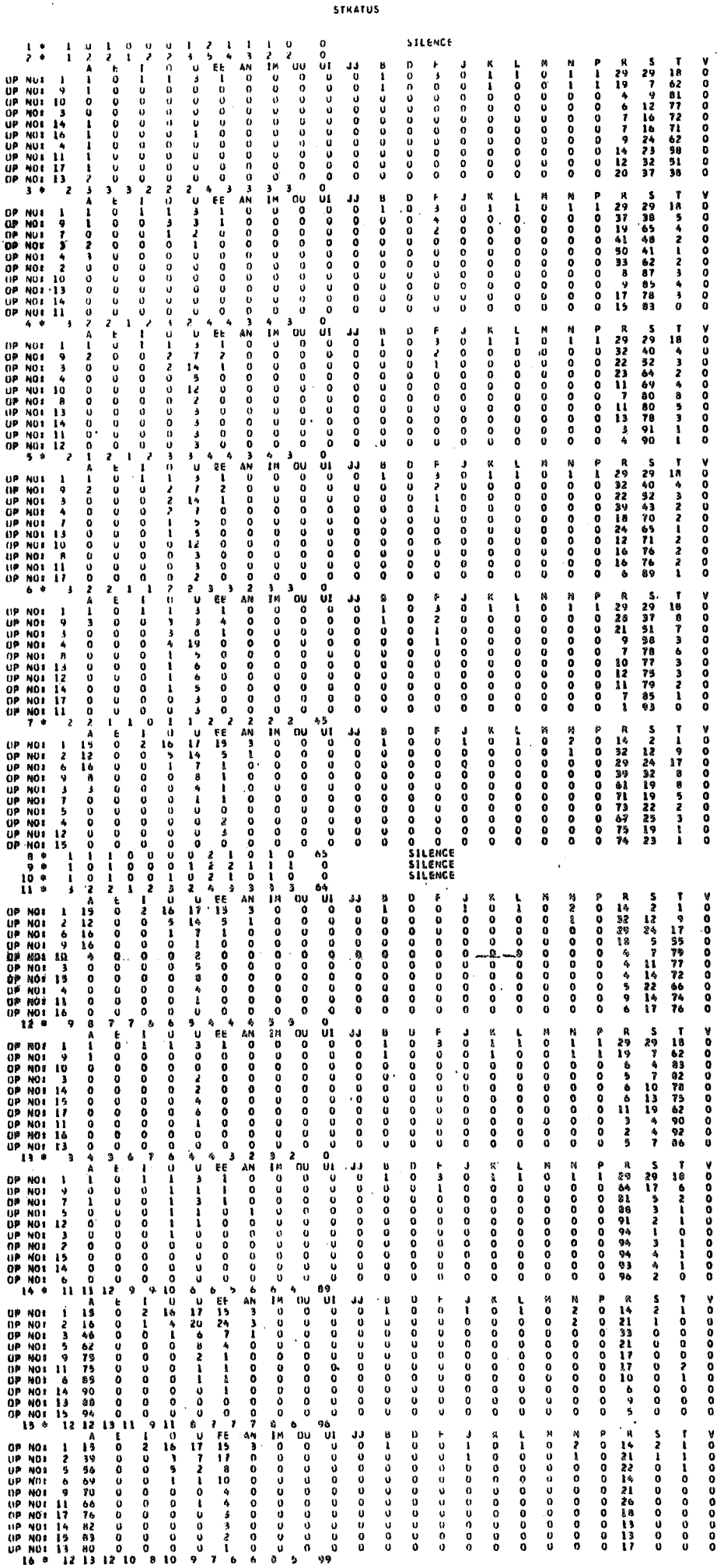
REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.003	.003	.005	.003	.005	.005	.003	.003	.003	.005	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.087
2	.008	.005	.022	.024	.020	.014	.005	.003	.007	.007	.005	.008	.007	.016	.014	.005	.007	.007	.012	.007	.302
3	.030	.005	.003	.016	.024	.027	.007	.010	.003	.005	.003	.005	.003	.003	.007	.012	.014	.008	.008	.003	.241
4	.022	.003	.003	.016	.020	.008	.018	.007	.005	.005	.003	.008	.003	.005	.003	.003	.005	.003	.008	.005	.173
5	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.071
6	.003	.003	.003	.008	.003	.010	.005	.003	.003	.003	.003	.007	.003	.005	.003	.003	.003	.003	.005	.007	.125

OPERATEUR NO : 8

REP NO :	PROBCO																			PRK
	A	E	I	O	U	EE	AN	IM	OU	UI	JJ	B	D	F	J	K	L	M	N	P
1	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.044
2	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.058
3	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.065
4	.013	.002	.002	.003	.003	.002	.007	.002	.002	.002	.002	.004	.002	.004	.002	.003	.003	.002	.002	.002	.080
5	.013	.002	.002	.004	.003	.002	.006	.002	.003	.002	.002	.002	.002	.005	.003	.003	.002	.002	.002	.004	.079
6	.005	.002	.002	.003	.002	.002	.003	.002	.002	.002	.002	.003	.002	.004	.002	.002	.002	.002	.002	.003	.067
7	.004	.002	.002	.004	.003	.008	.002	.005	.002	.003	.002	.006	.004	.003	.005	.002	.002	.002	.002	.002	.080
8	.003	.003	.002	.003	.003	.009	.002	.003	.002	.002	.002	.002	.002	.004	.002	.004	.002	.002	.007	.002	.071
9	.002	.003	.004	.003	.007	.008	.003	.004	.003	.003	.002	.002	.002	.003	.002	.003	.003	.003	.006	.002	.082
10	.002	.002	.005	.004	.018	.008	.003	.002	.003	.003	.002	.004	.002	.002	.002	.002	.006	.003	.004	.002	.083
11	.002	.002	.005	.006	.007	.004	.002	.002	.002	.004	.003	.002	.002	.002	.002	.002	.004	.003	.003	.002	.066
12	.002	.002	.005	.014	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.080
13	.002	.002	.002	.004	.002	.007	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.044
14	.002	.002	.002	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.043
15	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.042
16	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.042

FIGURE 3

FIGURE 4



Explication de la figure 4 montrant les résultats de l'analyse

Chaque ligne de données (numérotées 1, 2, 3...) est analysée à condition que l'énergie totale soit supérieure à un certain seuil (sinon, l'ordinateur imprime SILENCE). Entre chaque ligne sont indiquées par colonne les probabilités des phonèmes (A, E, I, ..., R, S, T, V) modifiées par les résultats des opérateurs choisis.

Le nom de l'opérateur est indiqué en début de la ligne des probabilités. L'opérateur choisi à chaque itération est celui qui est le plus informant en moyenne vis à vis des phonèmes à extraire au sens de la formule (5).

Dans le cas du son complexe STRATUS, l'analyse a reconnu successivement

?, S, S, S, S, R, T, T, R, A, A, A, R, R, R, R, R, T, ?, U, U, U, R, S, S, S, S, S,

Remarque : Le vocodeur considéré vocodeur du CNET à 12 canaux détecte mal les S, R et L.

Dans ce cas, le mot "stratus" aurait malgré cela été reconnu avec un autre étage de reconnaissance de mots tel que celui décrit dans [9] qui utilise la redondance d'information au niveau des lettres formant les mots.

Les erreurs dans la reconnaissance proviennent dans l'ordre d'importance :

1. Du fait que l'ensemble des opérateurs heuristiques trouvé est insuffisamment informant. En particulier, le vocodeur lui-même en tant qu'opérateur informant.
2. Du fait que les opérateurs considérés ne sont pas entièrement indépendants (voir §§ II - 4 et III - 1 - b).
3. Du fait que le nombre d'échantillons d'apprentissage a été trop faible (voir § III - 1 - a).

IV - Méthodes pour l'automatisation de la génération des opérateurs

La méthode appliquée précédemment dans deux cas différents de reconnaissance de formes : formes visuelles et acoustiques n'est en fait qu'une automatisation de la sélection des meilleurs opérateurs et de leur association en but de créer un programme unique de reconnaissance. Il faut en effet fournir "à la main" ou "heuristiquement" les opérateurs dans lesquels se fait la sélection. Comment arriver à une automatisation de la génération de ces opérateurs ? C'est dans les conditions actuelles encore dans le domaine de l'utopie.

Malgré tout un premier essai simpliste a déjà été tenté dans cette direction: voir [7]. Cette génération était une création aléatoire d'opérateurs utilisant les instructions élémentaires de l'ordinateur, suivie d'une sélection des opérateurs les plus informants : elle a réussit dans la génération d'un opérateur de prétraitement de figures de taches blanches sur fond noir ; ces opérateurs extrayaient trois formes élémentaires : noir, blanc, bord de tache.

Cette méthode est dans le cas général irréalisable pratiquement. Actuellement, nos études portent sur une génération semi automatique qui serait plutôt une aide à la génération par l'homme d'opérateurs de reconnaissance de forme.. Elles appliquent des idées très proches de celles expliquées précédemment.

Bien que ces idées soient générales aux formes visuelles et acoustiques, nous appliquons actuellement ces méthodes de génération aux données vocales.

BIBLIOGRAPHIE

[1] J. C. SIMON, A CHECROUN, C. ROCHE : "Procédés de filtrage digital d'une image de ligne". C. rendu à l'Ac. des Sc. t. 272 p. 337-340. (25 janvier 71).

[2] J. C. SIMON, A. CHECROUN, C. ROCHE : "A method of comparing two patterns independant of possible transformations and small distorsions. IEEE Symposium of November 5-7-1969 in Argonne Illinois on Feature Extraction and Selection in Pattern Recognition.

[3] FANO R. M. : "Transmission of Information". MIT Press (1963).

[4] LEWIS P. M.: "The characteristic Selection Problem in Recognition Systems". IRE Trans. Infor. Theory 8 pp. 171-178 (1962).

[5] FU K. S. : "Sequential Methods in Pattern Recognition and Machine Learning" Academic Press. (1968).

[6] SHAW A. C. : "The formal description and Parsing of Pictures" SLAC report n° 84 Stanford Linear Accelerator Center, Stanford University, Calif. (March 1968).

[7] J. C. SIMON et C. ROCHE : "Application of Questionnaire Theory to pattern recognition". International Joint Conference on Artificial Intelligence. London August 1971.

[8] KAHN E. : Thèse de doctorat 3ème cycle. Fac. des Sc. Paris, Dec.

[9] C. ROCHE et HEUDE : "Lecture rapide et reconnaissance des Formes". Public. Interne. Institut de Programmation. Juin 1970.

RECONNAISSANCE DE PHONEMES
AU MOYEN D'UNE COHLEE ARTIFICIELLE

P. ALINAT
THOMSON - C.S.F. - CAGNES-sur-MER

Le but de l'étude dont le principe va être sommairement décrit*, est de reconnaître les phonèmes de la langue française. Il est certain que dans une conversation courante entre individus, de nombreux phonèmes ne sont pas réellement prononcés : grâce à cela le débit de parole peut être plus rapide et la fatigue moins grande. Etant donné la grande redondance au niveau mots d'une part, et au niveau phrases et idées d'autre part, l'information n'est pas perturbée. Par contre, une reconnaissance se situant au niveau phonèmes le sera bien évidemment. Toutefois , dans le futur, pour des dispositifs capables de traiter une conversation, il sera nécessaire de reconnaître tous les phonèmes réellement prononcés, de façon à pouvoir, en fonction du vocabulaire et des règles mis en mémoire, reconstituer le message réellement émis. En attendant, nous imposons au locuteur une prononciation lente et bien articulée.

Les ondes sonores émises par les organes vocaux d'un individu peuvent être traduites en une tension électrique f(t) au moyen d'un microphone. Le signal f(t) peut être considéré comme stationnaire de temps à autre sur des durées de 100 ms environ. Parfois, aussi, il est purement transitoire. C'est un tel signal que l'on cherche à décomposer en une suite de phonèmes.

Comme toujours dans ce genre de problème, nous sommes conduits à faire subir au signal f(t) un prétraitement avant de résoudre le problème de classification. Le prétraitement consiste en un changement de la base au moyen de laquelle on définit le signal. Cette étape est très importante car c'est d'elle que dépend la simplicité des opérations de classification.

La détermination du prétraitement adapté à un problème particulier est en général difficile. Toutefois dans notre cas, on peut se laisser guider par notre connaissance de l'oreille interne (mécanique tout au moins).

Cela amène à construire une batterie de filtres passe-bande dont les fonctions de transfert et la répartition des fréquences centrales, sont à peu près conformes aux mesures des expérimentateurs (VON BEKESY - FLANAGAN). Le nombre des filtres est de 66.

* Etude financée par la Direction des Recherches et Moyens d'Essais PARIS

Ils sont synthétisés actuellement sous forme de filtres RC actifs, mais pourraient fort bien l'être sous forme numérique. Les fréquences centrales des filtres extrêmes sont 200 Hz et 5 KHz. La limite supérieur s'est d'ailleurs révélée trop basse car elle ne permet pas de reconnaître les consonnes fricatives S, Z et F, V.

Les sorties de filtres sont après détection-intégration, échantillonnées toutes les 4 ms. On obtient ainsi une courbe $F_t(\omega)$ qui peut être considérée comme le "spectre" du signal f(t) à l'instant t. Cette courbe approxime l'enveloppe des déformations de la membrane basilaire. Les courbes $F_t(\omega)$ sont mises sous forme numérique au moyen d'un convertisseur analogique digital.

A ce stade, il faut déterminer les formants $F_t(\omega)$, c'est-à-dire grossièrement les 2 ou 3 principaux pôles de la fonction de transfert du conduit vocal. Pour cela, après avoir fait subir des filtrages linéaires à $F_t(\omega)$, on détermine les maxima du résultat et on considère que leurs positions indiquent celles des formants cherchés.

Des relevés expérimentaux, réalisés à partir d'une vingtaine de locuteurs masculins (fondamentale à 120 Hz environ) prononçant chaque phonème dans des configurations variées, ont montré pour les phonèmes soutenus (voyelles et consonnes fricatives) que les formants se produisent dans des plages spécifiques à chaque phonème. Ces plages ont en moyenne une largeur de bande relative F Max/F Min de 1,5. Pour des voix féminines, il semble qu'il faille légèrement décaler les plages, la largeur de bande relative restant constante.

On a pu vérifier également que la diffréence entre voyelles normales et nasales, porte sur l'amplitude du premier formant par rapport au second : la nasalité diminue systèmatiquement l'importance du premier formant par rapport au second.

Un système de reconnaissance très sommaire a été construit à partir de ces constatations de façon à vérifier en temps réel sur un grand nombre d'individus, la validité des principes retenus. Ce système se limite aux phonèmes soutenus. Les critères nécessaires pour la reconnaissance de chaque phonème sont la présence de formants dans certaines zones et l'absence de formants dans d'autres zones pendant un certain temps. On ne fait pas la distinction entre sourdes et sonores car la présence du fondamentale n'est pas détectée. De plus, on ne tient pas compte des amplitudes relatives des formants : on ne peut donc pas reconnaître les voyelles nasales.

A part quelques ennuis dûs à des formants parasites crées au cours du calcul, le système donne pour des locuteurs masculins, de bons résultats pour les voyelles I, É, E (le) È, EU (leur) A, Ô (port) O (kilo) U (pur) et OU, et les consonnes fricatives CH - J. Il a été vérifié qu'il serait possible de reconnaître également S, Z et F, V, en prolongeant la batterie de filtres vers les hautes fréquences, et les voyelles nasales IN, AN et ON en tenant compte des amplitudes relatives des formants. Les consonnes explosives (P, B, M, T, D, N, K, G) et liquides (L,R) nécessitent pour être reconnues de découvrir les critères utiles et de construire un système les faisant apparaître.

Le système de reconnaissance évite tout apprentissage et la quantité d'information conservée en mémoire est relativement faible.

BIBLIOGRAPHIE

- VON BEKESY

"Experiment in hearing"
Mc Graw Hill Book

- FANT G.

"Acoustical Theory of Speech Production"
MOUTON and CO. - 1960

- LIBERMAN, INGERMAN, LISKER, DELATTRE, COOPER

"Minimal Rules for Synthetizing Speech"
J.A.S.A. Vol. 31 n° 11 - Novembre 1959.

- S.J. CAMPANELLA et D. PHYFE

"Application of a Model of the Analog Ear to Speech Signal Analysis"
IEEE Trans. Audio and Electro. Acoustics - Vol. AU 16 n° 1 - Mars 1968.

- J.FLANAGAN

"Speech Analysis, Synthesis and Perception"
NEW YORK Academic Press Inc.

- P. ALINAT

"Essai de Reconnaissance des Phonèmes au moyen d'une Cohlée Artificielle"
Troisième Colloque sur le Traitement du Signal et ses Applications
Nice du 1er au 5 Juin 1971.

RECONNAISSANCE DE LA PAROLE EN TEMPS REEL

J. CAELEN S. CASTAN G. PERENNOU

Cybernétique des Entreprises
et Reconnaissance des Formes

U.E.R. Informatique -UNIVERSITE PAUL SABATIER-TOULOUSE

Les études sur la reconnaissance automatique de la parole en temps réel, menées par le laboratoire, sont effectuées au moyen d'un Vocoder de 14 canaux du C.N.E.T., connecté à l'ordinateur I.B.M. 7044 du Centre de Calcul Universitaire de Toulouse.

1. RECONNAISSANCE PAR LES METHODES GLOBALES

Les premiers travaux ont porté sur la reconnaissance d'un vocabulaire restreint (les chiffres de 0 à 9) prononcé par un petit nombre de locuteurs.

Une méthode globale de reconnaissance n'utilise pas l'évolution temporelle du signal sous sa forme directe. Un mot est essentiellement défini par trois parties : le début et la fin, périodes transitoires courtes, le milieu plus mal défini quant à la stabilité mais de durée plus longue. Ces trois zones sont paramétrées en "moyenne" c'est-à-dire que sont effectuées des moyennes temporelles spectrales sur tout ou partie des bandes de fréquence filtrées. Chaque partie, quelle que soit sa durée est codée par le même nombre de paramètres, ici le nombre de canaux du Vocoder. Le spectre se trouve réduit uniformément pour tous les mots et ne reflète plus le caractère évolutif du signal, sa réalité physique se trouve en partie détruite. Ce type de méthode se heurte donc à certaines difficultés dès que les mots à traiter ne sont plus monosyllabiques. Par exemple le mot diviser (soit di-vi-ser) se prête mal à un tel découpage où le milieu est une succession de spectres stables et de spectres instables.

Expériences réalisées

Le vocoder délivre des "lignes" de valeurs qui représentent les spectres instantanés du signal acoustique. Un mot est quantitativement symbolisé par un ensemble de telles lignes et chacune des parties début, milieu, fin par respectivement une, deux, une lignes prises dans l'ensemble comme suit : la première ligne du mot caractérisera le début, deux lignes de moyennes caractériseront le milieu et la dernière ligne caractérisera la fin. 4 vecteurs de R^{15} soit un point de R^{60} est la description spatiale du mot.

Pour grouper ces points en catégories on utilise des discriminateurs du 3ème type (I) (2) c'est-à-dire que l'on cherche les hyperplans séparateurs au sens du meilleur facteur de tolérance. Le taux de bonne reconnaissance obtenu en utilisant cette méthode, varie entre 80% et 90% dans les conditions précisées plus haut. De très bons taux ne pourront cependant pas être atteints semble-t-il, par de telles méthodes.

2. PRINCIPES DE RECONNAISSANCE EN COURS D'ETUDE

Pour éliminer le principal défaut de la méthode précédente nous avons été amenés à segmenter un mot en fonction de sa stabilité spectrale. Les périodes instables indiquent avec une bonne précision les consonnes de sorte que les mots ainsi décrits par une succession de blocs correspondent assez bien avec la notion usuelle de syllabe. Par exemple le mot "zéro" sera décomposé en zér et ro (voir annexe II).

Comme l'expérimentation a lieu dans un bruit ambiant non négligeable il est nécessaire, avant de commencer ce découpage, d'extraire la parole du bruit (voir annexe I).

Ce problème résolu, on se trouve ramené à un problème de reconnaissance de monosyllabes (voir annexe III) pour lesquelles on peut par exemple utiliser la méthode globale précédente.

Considérons maintenant un arbre dont les sommets s'identifient à une succession de syllabes et chaque arc à une syllabe. On se déplace à mesure que les syllabes sont identifiées, à la fin du mot le sommet où l'on se trouve correspond :

- soit à un mot du vocabulaire
- soit à une succession de syllabes sans signification dans le vocabulaire autorisé

Ce deuxième cas peut se produire essentiellement pour deux raisons :

- mauvaise identification de syllabes
- perte de syllabe en début ou fin de mot.

Si l'on fait l'hypothèse que la fin du mot est perdue (respectivement le début) et si le sommet où l'on se trouve correspond à un seul mot du vocabulaire il suffit de décider par le sommet terminal successeur (respectivement de l'arbre miroir) de l'interprétation à donner à ce mot tronqué. Pour tous les mots une double analyse sera faite en parcourant les arbres directs et miroirs, la cohérence des résultats pourra être ainsi vérifiée. En conclusion :

a) deux interprétations identiques avec le mot et le mot "miroir" pour toutes les syllabes bien identifiées.
b) interprétation unique avec le mot (respectivement le mot "miroir") s'il a perdu sa fin (respectivement son début).
c) contrôle d'erreurs sinon

Ces interprétations ne sont rendues possibles que si le vocabulaire n'est pas ambigü au sens de la reconnaissance c'est-à-dire si les mots ne présentent aucune syllabe dans un même ordre de succession. Une règle que doit satisfaire un tel vocabulaire est que les mots qui le constituent doivent présenter un ordre différent dans la succession des syllabes qui les forment. Par exemple perforatrice et perforation resteraient indiscernables lors de la perte de la syllabe finale.

De tels processus d'identification plus généralement tout processus à contrôle de cheminement, de par leur structure séquentielle s'adaptent mieux à la structure également séquentielle du langage. Avec les réserves faites quant au choix du vocabulaire la probabilité de faux aiguillage ne dépend plus alors que de l'erreur d'identification sur les syllabes individuelles ce que l'amélioration constante des analyseurs de parole tend à réduire de plus en plus. A l'heure actuelle ces processus réduisent les confusions entre les mots par rapport à des méthodes globales, puisque les ramenant à des monosyllabes pour lesquelles le taux d'erreur est moindre. Des contrôles plus généraux et plus systématiques peuvent dès lors, sur un modèle similaire, être envisagés sur d'autres éléments phonétiques de façon à reconstruire et à redonner au mot sa véritable fonction qui est avant tout de signifier.

BIBLIOGRAPHIE

1- G. PERENNOU - *Contribution à l'étude des discriminateurs. Calcul et optimisation - Thèse d'Etat Juin 1968.*
2- S. CASTAN - G. PERENNOU - *On optimization Discrimination - Delf Workshop IEEE - Août 1968 -*
3- FARRENY - SENTENI - *Reconnaissance de la parole - Diplôme d'Ingénieur E.N.S.E.I.H.T. - Juin 1969 -*
4- BORTOLET - KOPP - *Segmentation et reconnaissance de la parole - Diplôme d'Ingénieur E.N.S.E.I.H.T. - Juin 1970 -*

ANNEXE I (3) SEGMENTATION : Séparation parole-bruit.

Un ensemble de tests permet de séparer grossièrement la parole du bruit. Appelons $I_i(t)$ la valeur délivrée par le canal i à l'instant t. Nous dirons qu'il y a zone de parole à l'instant t si :

$$\exists\, i,j,k \quad : \quad I_i + I_j + I_k \geqslant 10$$

ou $$\exists\, i_1 i_2 \ldots i_7 \; : \; \sum_{k=1}^{7} I_{i_k} \geqslant 6$$

ou $$\sum_{i=5,6,7} V_i \geqslant 5 \quad \text{avec} \quad V_i = I_{i+1} - I_i$$

ou $$\begin{cases} \exists\, i,j \; : \; I_i \geqslant 10 \quad I_j \geqslant 10 \\ \exists\, i_1 \ldots i_5 \; : \; I_{i_1} \geqslant 6 \ldots . \; I_{i_5} \geqslant 6 \end{cases}$$

ou Pitch $\neq 0$ et $\exists\, i,j \; : \; I_i \geqslant 10 \quad I_j \geqslant 10$

ANNEXE II (4)

Les tests de l'annexe I délimitent les zones de parole que l'on va étudier plus précisément.

Le début (1) et la fin (2) de parole présentent des variations de fortes amplitudes tel. que le montre la figure 1

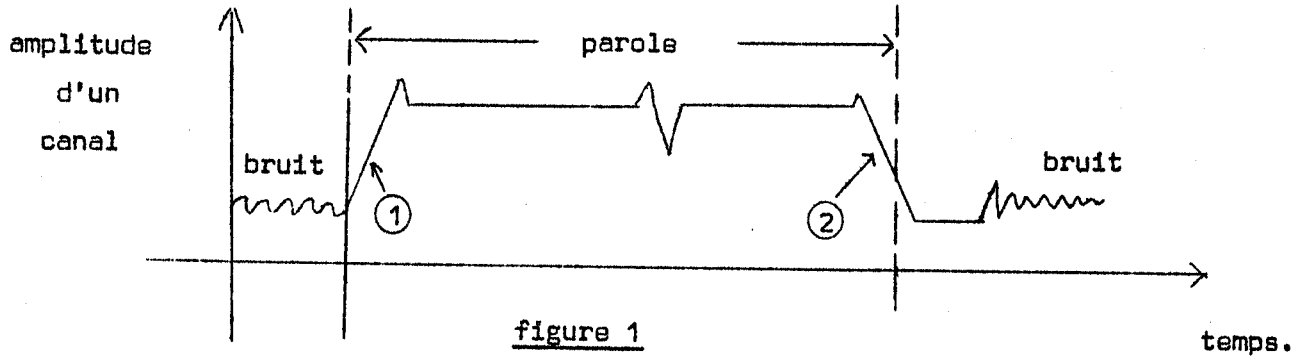


figure 1

On définit-la pente verticale élémentaire $p_i = I_i(t+1) - I_i(t)$

-la pente verticale $P = \sum_{i=1}^{14} p_i$

-l'amplitude moyenne instantanée $I_M = \frac{1}{14} \sum_{i=1}^{14} I_i(t)$

Un nouvel ensemble de tests permet de séparer plus finement la parole du bruit une fois que la séparation grossière a été faite.

1) en zone de bruit :

$\{P_M < 8\} \cap \{I_M < 5\} \Longrightarrow$ zone de bruit

sinon : zone de transition initiale bruit-parole

2) en zone de parole :

$\{\min_{+} P < -9\} \cap \{I_M(t+1) < 5 \text{ ou } I_M(t+2) < 5 \text{ ou } I_M(t+3) < 5\}$

zone de transition finale parole-bruit

sinon : zone de parole.

ANNEXE III

SEPARATION SYLLABE-SYLLABE ET PARAMETRISATION.

La transition syllabe-syllabe est caractérisée par une période instable et déterminée par la méthode précédente (Annexe II).

Nous définissons $\pi_i = I_{i+1}(t) - I_i(t)$ que nous appelons pente horizontale à l'instant t. On calcule la moyenne π_{i_M} de ces pentes sur un bloc syllabe, et on cherche les intervalles de temps Δt pour lesquels $\pi_{i_M} - \varepsilon < \pi_{i_{m,\Delta t}} < \pi_{i_M} + \varepsilon$.

$\pi_{i_{M,\Delta t}}$ est la moyenne des pentes sur l'intervalle $\Delta t \subset \{\text{syllabe}\}$.

Un choix convenable de ε détermine les Δt stables qui caractérisent les syllabes.

Le bloc syllabe étant affiné on le caractérise par la moyenne de pentes horizontales , soit un paramètre de 13 composantes dans le cas d'un vocoder à 14 canaux.

Un autre type de paramétrisation a été envisagé sur les syllabes précédemment déterminées.

On considère les valeurs $I_{i\,\Delta t}(R)$ comme valeurs d'une courbe $I_i = f(i)$ dont nous connaissons 14 valeurs en 14 points. On cherche le polynôme de degré 8 qui interpole cette courbe au sens des moindres carrés, soit $P(i)$ / $°P=8$, dont on calcule les extréma par la formule $P'(i) = 0$. La valeur de ces extréma aux points considérés constitue le vecteur paramètre de la syllabe. Pour éviter toute fausse introduction d'extréma, il a fallu "lisser" le polynôme.

Remarque : Il apparaît que la perte d'information due au vocoder nuit à l'obtention précise des formants, en effet : un examen des résultats nous fait penser que le 1er maximum correspond au 1er formant (en général peu net) et au 2ème formant, le 2ème maximum correspond au 3ème formant.

R. CARRÉ à P. LAVANANT

Comment se fait la multiplication dans votre unité spécialisée ?

P. LAVANANT

Les valeurs sont mises sous forme logarithmique. Le temps de multiplication devient alors le temps d'addition des logarithmes augmenté du temps de conversions de linéaire en logarithme et de logarithme en linéaire.

B. ESCUDIÉ à J.P. HATON

Vous avez indiqué que vous utilisiez le développement de *KARHUNEN-LOEVE*. Dans ce cas, utilisez-vous les fonctions sphéroïdales qui permettent d'avoir une base de fonctions orthogonales approchant les fonctions "concentrées" en temps et fréquence ?

J.P. HATON

La méthode de compression d'information utilisée ici consiste à minimiser l'entropie de l'information au sens de *SHANNON*. Pour cela, on effectue des transformations orthogonales de la base de départ et, en ce sens, la difficulté mentionnée n'est pas gênante.

R.A. GUEDJ à J.P. HATON

Quelle est la nature des 24 + 6 paramètres ? Y a-t-il un travail sur les règles phonologiques de la langue française, analogue à l'ouvrage de *CHOMSKY* et *HALLE* "Sound patterns of English"?

J.P. HATTON

Les paramètres de reconnaissance sont constitués des 24 sorties de filtres et de 6 paramètres acoustiques déterminés empiriquement à partir des "formes acoustiques" fournies par l'analyseur.

M. DESGOUTTE

Une analyse phonologique générative du français a été entreprise par *SANFORD M.* et *SCHANE A.* (Université de San Diego) :

. French phonology and morphology - M.I.T.Press (1968)
. Langages n° 8 - Introduction à la phonologie générative - Larousse.

J. GUIBERT à J.P. HATTON

Quelle est la durée d'intégration ?

J.P. HATON

L'intégration des paramètres de reconnaissance est effectuée en permanence et quel que soit le phonème, ce sont les paramètres qui sont spécifiquement destinés à la reconnaissance de certains groupes de phonèmes.

A. NEMETH - L'intérêt pratique de la reconnaissance consisterait à poser la question à l'orginateur par ligne téléphonique. A-t-on essayé la reconnaissance de parole passée par micro à charbon et ligne téléphonique ? A-t-on pensé à utiliser des mots spéciaux pour éviter les confusions ?

J.Y. GRESSER - Le C.N.E.T. s'est livré à de telles expériences entre les centres de Paris et Lannion. Dans l'état actuel des méthodes de reconnaissance et des transmissions (bruits de commutation, diaphonie...) lorsque l'appareil de prétraitement de la parole est situé à distance, les résultats sont mauvais.

Je pense qu'il faut utiliser, plutôt que des mots spéciaux, des langages spéciaux conçus pour la communication vocale avec les machines. Nous effectuons actuellement, au C.N.E.T., des recherches sur de tels langages. Ces langages devraient être des sous-ensembles du langage naturel, où les contraintes syntaxiques, sémantiques, etc.. devraient permettre de lever les ambiguïtés constatées à des niveaux inférieurs (phonémiques, etc..).

S. CASTAN - Jusqu'à présent, les langages de programmation sont des langages écrits. Nous pensons qu'en vue d'augmenter les pourcentages de reconnaissances, il est souhaitable d'étudier des langages de programmation qui soient des langages parlés tenant compte de certaines contraintes liées aux erreurs de reconnaissance les plus fréquentes.

J.P. HATTON à S. CASTAN

Comment faites-vous la différence entre un silence d'occlusion précédant une consonne plosive et un silence entre mots, les deux étant caractérisés par les mêmes paramètres au point de vue des pentes ?

S. CASTAN

La décision ne se fait pas sur la pente. La pente est utilisée pour segmenter l'élément phonétique et la décision se fait au fur et à mesure sur l'arbre de décision.

M. VALENCIEN à P. ALINAT

Le modèle de la membrane basilaire proposé n'est pas conforme aux descriptions anatomophysiologiques. Jusqu'à 300 Hz, la déformation est totale, au dessus, la déformation, partie de la base, atteint un maximum puis tombe brutalement. Les fibres excitées le seront toujours depuis la base jusqu'à la fréquence correspondant à la coupure.

P. ALINAT

La cochlée réalisée n'est qu'une approximation de la nature. Il est hors de question de faire une copie fidèle. Je n'ai, de plus, pas eu le temps de la décrire en détail.
Pour ce qui est des basses fréquences, il faut remarquer que, justement, la borne inférieure de la bande de fréquence traitée est de 200 Hz ; quant aux fameuses inhibitions présentées dans le système nerveux, on en tient compte au niveau du filtrage linéaire que l'on fait subir au signal $F_t(\omega)$.

M. DECHAUX

La batterie de filtres utilisée comporte des filtres à large bande suivies, après détection, d'un filtre passe-bas à 50 Hz (3 pôles) et

d'un échantillonnage à 250 Hz. Le filtrage passe-bas ne semble pas trop sévère. Cela est apparu de manière indirecte : cet analyseur sert d'organe d'entrée à un dispositif d'analyse et de synthèse des sons téléphoniques qui a donné une très bonne reproduction des consonnes plosives et ceci grâce à l'analyse fine des transitoires que permet la batterie de filtres.

J.C. RISSET - Plusieurs conférenciers ont évoqué une analogie entre le système de reconnaissance sur lequel ils travaillent et l'audition. Que pensent-ils de l'importance des différences résiduelles entre audition et système artificiel déjà mise en relief par le Docteur *VALLENCIEN* ? Ainsi le traitement réalisé par l'oreille interne ne se réduit pas à une analyse floue en fréquence, c'est plutôt une analyse temps-fréquence.

A un autre niveau, la perception ne se borne pas à reconnaître des classes définies a priori : il y a plusieurs niveaux hiérarchisés et la perception lève les ambiguïtés en passant d'un niveau à l'autre ; dans la compréhension de la parole par l'auditeur la segmentation n'est pas antérieure aux autres opérations, la décision est souvent différée. Les analogies entre systèmes artificiels ne sont-elles pas insignifiantes par rapport aux différences essentielles qui subsistent dans les modalités de reconnaissance de parole ? Et la reconnaissance automatique de la parole ne suppose-t-elle pas résolus les problèmes d'intelligence artificielle ?

R. CARRÉ - A la suite des exposés on peut constater que la plupart des travaux présentés sur la reconnaissance de la parole ne tiennent pas compte des études effectuées sur la production et la perception de la parole, des études sur le rôle et l'importance des paramètres en perception et des règles linguistiques.

Il existe deux démarches : dans l'une on traite des paramètres sans savoir ce qu'ils représentent exactement ; dans l'autre, on cherche puis on traite des paramètres réputés les plus intéressants.

La première démarche ne me séduit pas car, pour obtenir des résultats, on est souvent obligé d'opérer par retouches sans savoir ce que l'on fait ce qui ne permet pas d'étudier les problèmes fondamentaux.

J.Y. GRESSER - Les études sur la reconnaissance de la parole ont pour objet les machines et non les êtres vivants. Ces derniers fournissent des indications intéressantes sur la manière dont une machine à comprendre la parole peut fonctionner. Mais l'imitation de la nature n'est pas forcément la bonne méthode et le comportement des êtres vivants n'est pas non plus une référence pour les machines.

De plus l'oreille n'est pas le seul organe de l'audition ; le cerveau joue un rôle essentiel dans la perception auditive. Il n'existe pas de théorie exploitable de son fonctionnement. Est-il vraiment indispensable d'en attendre une ?

Nous préférons faire ce que nous pouvons avec les machines actuelles et tous les moyens même peu élégants sont bons.

B. ESCUDIÉ - Il est à regretter que toutes les études ne soient pas plus en liaison avec les études de bionique conduites sur l'oreille et la biophysique.

Secrétariat du G.A.L.F.
C. N. E. T. - Route de Trégastel - 22 . LANNION

Imprimerie E.N.S.E.R.G.
23 rue des Martyrs - 38 . G R E N O B L E