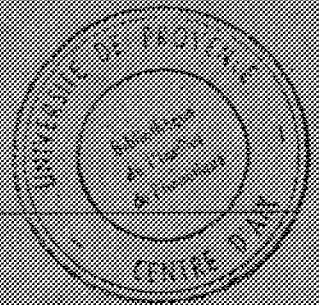
Groupement des Acousticiens de Langue Française

XIèmes JOURNÉES D'ÉTUDES SUR LA PAROLE

du Groupe

COMMUNICATION PARLÉE

Textes des exposés



STRASBOURG : 28 — 29 et 30 mai 1980

ETATS-UNIS
New-Haven

Status Report on Speech Research

Haskins Laboratory.

Groupement des Acousticiens de Langue Française

XIèmes JOURNÉES D'ÉTUDES SUR LA PAROLE

du Groupe

COMMUNICATION PARLÉE

Textes des exposés

STRASBOURG : 28 — 29 et 30 mai 1980

Les 11èmes Journées d'Etudes sur la Parole ont été organisées à l'Institut de Phonétique (Faculté des Lettres Modernes) de l'Université des Sciences Humaines de Strasbourg.
Elles ont pu se dérouler grâce à l'appui du GALF, de la Faculté des Lettres Modernes et du Conseil Scientifique de l'Université des Sciences Humaines de Strasbourg.

Thèmes de travail des XIèmes Journées :

I - PERCEPTION DE LA PAROLE
(apports de la physiologie et des modèles d'audition à la reconnaissance et à la compréhension de la parole)

Président : René CARRE

II - INTELLIGIBILITE ET QUALITE DE LA PAROLE NATURELLE, DE LA PAROLE CODEE, DE LA PAROLE DE SYNTHESE

Président : Michel CARTIER

III - VARIABILITE INTER ET INTRA LOCUTEURS

a) Observation et Analyse

Président : Max WAJSKOP
Rapporteur : Louis-Jean BOE

b) Adaptation des systèmes de reconnaissance aux locuteurs

Président : Jean-Sylvain LIENARD
Rapporteur : Camille BELLISSANT

c) Vérification et identification du locuteur

Président : Jean-Paul HATON
Rapporteur : Yves GRENIER

IV - SESSIONS AFFICHEES

THÈME I : Perception de la parole ;

Apport de la physiologie et des modèles d'audition à la reconnaissance et à la compréhension de la parole.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

QUELQUES REFLEXIONS SUR LA MODELISATION DE L'OREILLE POUR LE TRAITEMENT DU SIGNAL

CAELEN J.

Laboratoire C.E.R.F.I.A.
Université P. SABATIER
118 Route de NARBONNE
31077 TOULOUSE CEDEX

RESUME :

L'étude d'un modèle d'oreille dans une perspective d'utilisation en analyse acoustique de la parole, nous a conduit à appréhender le système auditif avec les techniques du filtrage , du codage et de modulation non linéaire. Nous proposons certaines extensions en filtrage à partir des équations de vibration de la membrane basilaire qui permettent d'extraire le squelette formantique des sons avec une bonne sélectivité temporelle, ceci grâce aux propriétés des bancs de filtres couplés. Inversement certains points physiologiques encore mal connus, peuvent être abordés avec les méthodes employées en transmission du signal, c'est notamment le cas de la modulation produite par les cellules ciliées ainsi que du codage impulsionnel dans le nerf auditif. Parmi les solutions envisagées une modulation non-linéaire (développable en série entière) retient notre attention : elle produit en effet des distorsions harmoniques du type $2f_1-f_2$. En codage impulsionnel, conformément aux observations (ANDERSON D.J. et coll 1973) nous arrivons aux conclusions dans lesquelles deux modulations sont envisageables : une modulation de fréquence jusqu'à 1000 hz (détections de f_0 et f_1 pour la parole)peu sensible aux harmoniques, et une modulation d'amplitude au-dessus de 1000 hz. (détection des formants supérieurs pour la parole). Ces modulations doivent se passer d'horloge (non observée dans l'oreille) et l'on peut envisager des codages différentiels dérivant du codage "delta".

Il est tout aussi fondamental d'étudier en détail les divers étages et constituants de l'oreille sous l'aspect de modèles, que d'aboutir à un tout cohérent. Se posent alors des questions aussi difficiles que l'adaptation des niveaux d'analyse entre eux ou les problèmes d'adaptation au signal. Nous ne donnons là que quelques règles générales pour des applications à la reconnaissance de la parole.

SUMMARY :

The study of a model of ear in order to its application in acoustic analysis for speech recognition lead us to tackle the hearing system with nonlinear filtering, coding and modulation proceeding.

In filtering, we purpose some extensions from equations of vibration of basilar membrane, which allow us to extract the formantic pattern of sounds with a full temporal selectivity, using to properties of coupled-filters bank. On the other hand, some physiological problems which still remain, can be taken up with the methods used for the transmission of signal : for instance, the modulation of hair cells and pulse coding inside the auditory nerve. Among all the technics, a nonlinear modulation (power series expansion) holds our attention an it brings out nonharmonical distorsions $2f_1-f_2$. For pulse coding, just as it is reported by D.J. ANDERSON and all (1973), we draw the conclusion that a frequency modulation is best adapted to detect f_0 and f_1, when a amplitude modulation is prefered to detect high formants. These modulations may be "delta" modulation which need not timing.

It seems to us that the main point is to fit the various levels of analysis, and we give some general rules of this adaptations, and of the adaptation to the characteristics of the signals to analyse. It is clear that this problem can be solved by persevering simulations, if one seeks applications to speech recognition.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

QUELQUES REFLEXIONS SUR LA MODELISATION DE L'OREILLE POUR LE TRAITEMENT DU SIGNAL

CAELEN J. Laboratoire C.E.R.F.I.A.
Université P. SABATIER
118 Route de NARBONNE
31077 TOULOUSE CEDEX

INTRODUCTION :

Cet article propose quelques réflexions sur les apports de la modélisation de l'oreille au traitement du signal vocal, et réciproquement quelques hypothèses sur certains points mal connus de l'audition, compatibles avec les connaissances sur le filtrage, le codage ou les techniques de transmission de l'information. Cet article n'est pas une description exhaustive d'un modèle d'oreille mais il tente de montrer les raisons qui ont présidé à tel ou tel choix dans ses options. La fig. 1 donne le schéma général du modèle que nous avons retenu.

Les points non élucidés sur le plan physiologique sont néanmoins fondamentaux pour comprendre les transformations effectuées sur le signal jusqu'au cortex. De ce fait l'apport des modèles reste limité du fait même de l'incertitude dont ils restent entourés.

Pour cerner ces problèmes nous avons choisi quelques fonctions particulières de certains transducteurs auriculaires, ce sont :

1. le filtrage de la membrane basilaire pour les prolongements qu'il apporte en analyse du signal
2. la modulation des cellules ciliées
3. le codage nerveux impulsionnel
4. l'interaction des divers niveaux d'analyse entre eux (et adaptation au signal) pour les problèmes qu'ils posent notamment pour la transmission des informations fréquentielles, temporelles et énergétiques associées aux distorsions dont ils sont inhérents.

1. EXTENSION DE L'EQUATION DE VIBRATION DE LA MEMBRANE BASILAIRE

Si l'on suppose que la membrane basilaire se comporte comme un banc de filtres couplés (une portion dx est couplée par élasticité à la portion voisine) (CAELEN J. 1979) l'équation de vibration de type hyperbolique, peut se mettre sous la forme d'un système d'équations aux

différences que l'on peut généraliser à tout autre type de couplage et pour des filtres d'un ordre quelconque.

Considérons la suite des échantillons de l'entrée ($e_0, e_1, \ldots, e_n, \ldots$) et un nombre fini, M, de points de discrétisation de la membrane basilaire ($x_1, x_2, \ldots, x_M$) également répartis ou non. L'équation aux différences la plus générale peut être mise sous la forme :

$$\sum_{k=0}^{m_1} a_{ki}\, y_i^{(n-k)} + \sum_{j=0}^{m_2} \sum_{l=-p_1}^{p_2} c_{lji}\, y_{i+l}^{(n-j)} = \sum_{k=0}^{m_3} b_{ki}\, e_{n-k} \qquad i = 1,2,\ldots M \qquad n, m_1, m_2, m_3, p_1, p_2 \in \mathbb{N}$$

avec des conditions initiales et aux limites convenables, par exemple :

$$y_i^{(j)} = 0 \quad \forall i, \forall j \; j<0 \quad \text{et} \quad y_i^{(j)} = 0 \quad \forall j, \; i<1 \text{ et } i>M$$

où :

$y_i^{(j)}$ déplacement du point x_i à l'instant t_j

a_{ki} coefficients du filtre associé au point x_i

c_{lji} coefficients de couplage des filtres associés aux points x_i et x_l

b_{ki} coefficients de pondération de l'entrée (gain)

Dans la mesure où ces coefficients ne dépendent pas de $y_i^{(j)}$ nous obtenons un système linéaire. Dans le cas contraire il est possible, sous certaines conditions, de se ramener à un système linéaire par morceaux.

Lorsque $c_{l,i} = 0 \; \forall lji$ les filtres sont indépendants entre eux (vocoder par exemple)

Sinon on distingue :

1. les couplages symétriques $p_1 = p_2$, couplage par élasticité par exemple (cas de la membrane basilaire)
2. les couplages asymétriques $p_1 \neq p_2$, couplage par viscosité par exemple.

Du point de vue de la résolution numérique on distingue le système implicite, $c_{l,i} \neq 0$ pour $j = 0$ du système explicite $c_{l,i} = 0$ pour $j = 0$. Bien que stables les systèmes implicites sont d'une résolution complexe et se prêtent mal au calcul parallèle. Les systèmes explicites qui ont par contre ces avantages ne sont stables que dans des conditions très contraignantes. Un compromis doit donc être fait pour atteindre le temps réel avec des microprocesseurs :

soit utiliser un schéma implicite avec des logiques très rapides

soit utiliser un schéma explicite avec des logiques parallèles moyennement rapides et une cadence d'échantillonnage assez faible.

En utilisant la transformée en z (si le système est linéaire) il est possible d'étudier le comportement en fréquence de ces filtres.

On obtient :

$$\left(\sum_{l} c_{lj} + \sum_{k} a_{ki} z^{-k}\right) Y_i(z) + \sum_{\partial \neq 0} \sum_{l} c_{lji} z^{-\partial} Y_{i+l}(z) = \left(\sum_{k=0}^{m_3} b_{ki} z^{-k}\right) E(z)$$

qui permet de calculer la matrice de transfert.

Discussion :

L'étude d'un modèle d'oreille et plus précisément l'étude des vibrations de la membrane basilaire conduit à une généralisation des équations de filtrage qui permettent d'obtenir des filtres aux propriétés particulières :

- dissymétrie des pentes de coupure dans les aigus et dans les graves, en général plus raide dans les aigus que pour des filtres indépendants et moins raide dans les graves à surtension égale (conforme aux observations de RHODE W.S 1974) (fig 2 à 4)
- effet de masque spatial (donc fréquentiel) qui :
  - lisse le spectre dans les hautes fréquences
  - modifie les extrema du spectre par absorption ou rapprochement des crêtes. (ce qui provoque une meilleure répartition des voyelles dans le plan f_1-f_2 par exemple)
  - filtre le spectre harmonique de la voix
- nombre de pôles dépendant non seulement de l'ordre de chaque filtre mais de leur nombre M et des coefficients de couplage
- sélectivité dépendant des couplages et des coefficients de surtension

L'utilisation de tels filtres en analyse de la parole permet d'obtenir des spectres où apparait nettement le squelette formantique des phonèmes même dans les transitions rapides. (fig. 5)

2. LA MODULATION DES CELLULES CILIEES

De nombreuses hypothèses ont été avancées au sujet de la transduction des cellules ciliées (HALL J.L 1975, PFEIFFER R.R 1970, RUSSEL R. et coll 1975). Leur rôle apparait important tant au point de vue de la transmission des informations que des distorsions ou du filtrage. Au niveau des modèles des solutions ont été proposées (DE BOER 1975, DOLMAZON J.M. et coll 1977) qui font intervenir des transducteurs linéaires interactifs ou non. Il semble que des transducteurs non linéaires soient envisageables en raison du type des distorsions de la forme f_1-$2f_2$ pour des sons purs (DUIFHUIS H. 1976). Nous nous proposons d'étudier sommairement ici, quelques modulations classiques pour tenter de circonscrire ce problème dans l'optique "traitement du signal".

La cellule attachée au point x_i de la membrane basilaire reçoit en entrée le message $m_i(t)$ signal filtré au point x_i, et excite les fibres nerveuses à sa sortie, $s_i(t)$.

De façon générale nous avons $s_i(t) = F_i(m_i(t), s_i(t))$ si la cellule i est indépendante de ses voisines.

Examinons quelques modulations classiques F_i

2.1 modulation linéaire d'amplitude

$$s_i(t) = (a + m_i(t)).\cos(\omega_0 t + \varphi_0)$$

si $M(\omega)$ est le spectre du message $m_i(t)$, celui de la sortie peut s'exprimer par :

$$S(\omega) = \pi a\left[e^{j\varphi_0}\delta(\omega-\omega_0) + e^{-j\varphi_0}\delta(\omega+\omega_0)\right] + \frac{1}{2}\left[e^{j\varphi_0}M(\omega-\omega_0) + e^{-j\varphi_0}M(\omega+\omega_0)\right]$$

Notons ω_m la plus haute fréquence du message. La largeur de bande occupée par le signal $s_i(t)$ est égale à $2\omega_m$ et a une distribution symétrique par rapport à la fréquence porteuse ω_0. Pour la démodulation on peut procéder en

- une démodulation d'enveloppe, sensible au bruit
- une démodulation de produit avec un filtre de fréquence centrale ω_0 et de bande passante inférieure à $2\omega_m$

On pourrait alors considérer le réseau nerveux placé en aval comme un circuit de démodulation (filtres sélectifs notamment).

Considérons alors que la fibre qui reçoit le signal $s_i(t)$ soit un système linéaire de fonction de transfert H_f^i. La réponse de la fibre $r_i(t)$ peut se mettre sous la forme $r_i(t) = n(t)\exp(j\omega_0 t)$ qui peut être interprétée comme une modulation du message $n(t)$ par la porteuse de fréquence $\omega_0/2\pi$, ce message étant obtenu par filtrage de $m_i(t)$ par le filtre H_f décalé de ω_0 vers les basses fréquences. Le type des distorsions produites est donc :

- distorsions harmoniques de $k\omega_0$ $\quad k = 1, 2, \ldots$
- distorsions d'amplitude

Il ne semble pas que l'on ait découvert de "porteuse" dans l'oreille ni que les distorsions décrites ci-dessus conviennent.

2.2 modulation exponentielle

$s_i(t) = E.\exp(j(\omega_0 t + m_i(t))$ la porteuse est $E.\exp(j\omega_0 t)$

le spectre $S(\omega)$ de $s_i(t)$ peut se développer en série et si de plus $m(t) = \cos\Omega t$ alors :

$$S(\omega) = 2\pi E \sum_n J_n\left(\frac{\Delta\omega}{\Omega}\right)\delta(\omega-\omega_0-n\Omega)$$

J_n fonction de Bessel d'ordre n

$\Delta\omega$ excursion maximale de fréquence

Ce spectre présente (par rapport à celui du message) des distorsions non-harmoniques du type $\omega_0 \pm n\Omega$ (interférences entre le message et la porteuse). Si nous considérons encore le réseau nerveux comme un circuit de démodulation et chaque fibre comme un filtre de fonction de transfert H_f, le discriminateur idéal pour reconstituer le message $m_i(t)$ sans distorsion est $H_f(j\omega) = (\alpha\omega + \beta).\exp(j\varphi)$ α, β, φ coefficients convenables. Ce discrimateur donne en sortie un signal proportionnel à la fréquence instantanée du signal appliqué à son entrée, on peut le réaliser avec un circuit limiteur (compression de dynamique) placé en amont d'un filtre

oscillant désaccordé. La fréquence centrale f_r et le coefficient de surtension Q_r d'un tel filtre doivent vérifier la relation

$$\frac{2(f_r - f_o)}{f_r} Q_r = \frac{1}{\sqrt{2}} \qquad f_o = \omega_o / 2\pi \text{ fréquence de la porteuse}$$

Cette modulation semble plus proche de la réalité auditive que la précédente. En effet on note

- distorsions non harmoniques de type $\omega_o \pm n\Omega$
- nécessité d'un compresseur de dynamique et d'un filtre sélectif pour la démodulation, ce qui semble bien exister au niveau des fibres nerveuses.

Le réseau nerveux pourrait encore ici jouer le rôle de circuit de démodulation mais l'existence de la porteuse est un frein dans cette hypothèse.

2.3 modulation non-linéaire

Choisissons une modulation plus générale de la forme

$s_i(t) = \sum_{i=0}^{N} a_i m^i(t)$ polynome de degré N en m(t)

lorsque m(t) est une somme de tons purs de fréquences f_1 et f_2 le spectre de s(t) peut se développer en série. La contribution des termes se répartit alors comme suit :

termes du 1er ordre sur f_1, f_2

termes du 2me ordre sur $2f_1, f_1-f_2, f_2-f_1, 2f_2$

termes du 3me ordre sur $3f_1, f_1-2f_2, f_1+2f_2, f_2-2f_1, f_2+2f_1, 3f_2$

etc ...

La démodulation peut se faire par des circuits résonants qui peuvent sous certaines conditions éliminer certaines oscillations parasites et donc améliorer le rapport sigal/bruit.

Cette hypothèse pour la modulation des cellules semble donc la mieux adaptée, en effet :

- on note la présence de distorsions non-harmoniques en particulier $n_1 f_1 \pm n_2 f_2$ (HALL J.L 1975)
- la modulation ne passe plus pas la présence d'une porteuse
- si l'on considère le réseau nerveux comme un circuit de démodulation il doit être constitué de filtres sélectifs (ce qui est le cas des fibres nerveuses) et d'un compresseur de dynamique probablement présent au niveau du système nerveux également.

Conclusion

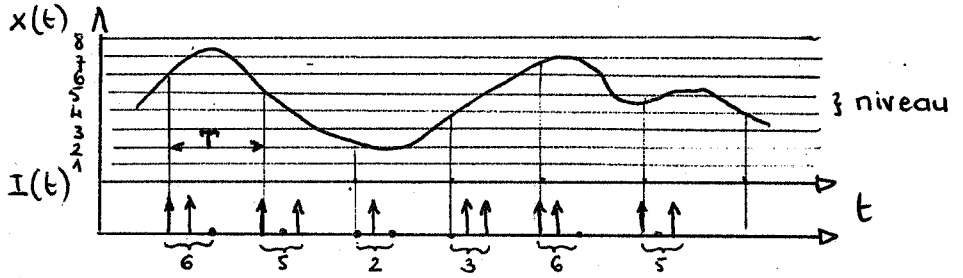
A ce stade de l'investigation, compte tenu des quelques observations sur les cellules ciliées et les fibres (encore insuffisantes) il semble qu'un modèle non-linéaire d'ordre 3 (degré du polynome) convienne pour modéliser cette partie de l'oreille. Il reste bien entendu que des observations physiologiques doivent confirmer ces résultats.

3. LE CODAGE IMPULSIONNEL

Les signaux nerveux sont des signaux impulsionnels d'amplitude et de durée constantes. Dans l'oreille cette durée est d'environ 0,7 ms ce qui rend théoriquement impossible la transmission de signaux de fréquence supérieure à 1430 hz. Or il est bien évident que la perception de sons de fréquences plus élevée ne pose aucun problème. Pour lever ce paradoxe apparent plusieurs théories ont été proposées parmi lesquelles la théorie "de la volée" est une des plus anciennes. On ne trouve pas une importante bibliographie sur ce sujet et les questions essentielles restent posées. Certes on connait les débits et la forme des impulsions (KIANG NY.S. et coll 1974, CHARLET DE SAUVAGE R. et coll 1977) mais la composition et la génération des codes demeurent enigmatiques. On peut supposer que pour transmettre les informations de durée, intensité et fréquence des codes résistants à la propagation d'erreurs et au bruit soient nécessaires dans l'oreille. La dynamique des fibres ne dépasse pas 30 dB ce qui limite les choix possibles.

3.1 Codage en modulation d'amplitude

Le signal impulsionnel de sortie est proportionnel (ou fonction monotone) au signal continu x(t) de l'entrée. On peut admettre alors que cette sortie I(t) permet de coder le spectre instantané au niveau des cellules ciliées. Il est peu probable que les signaux eux-mêmes soient codés (problème de la période réfractaire) . Par contre il semble logique de transmettre les informations du spectre d'amplitude. Dans ces conditions cela nécessite la présence dans les réseaux nerveux d'un intégrateur (neurone sommateur). Avec ces hypothèses nous aurions le schéma suivant :



où x(t) est l'enveloppe du signal de sortie d'une cellule ciliée.

Un mot de code est transmis avec la période T et pour former N mots de code il faut n bits (impulsions) tel que $2^n \geqslant N$. Cela donne 4 bits pour une dynamique de 30 dB. La période T minimum pour trans mettre un échantillon est donc n.0,7 ms soit environ 3 ms ce qui correspond au seuil de durée minimum perçue. La largeur de bande occupée par ce signal est B = n/T = 1300 hz.

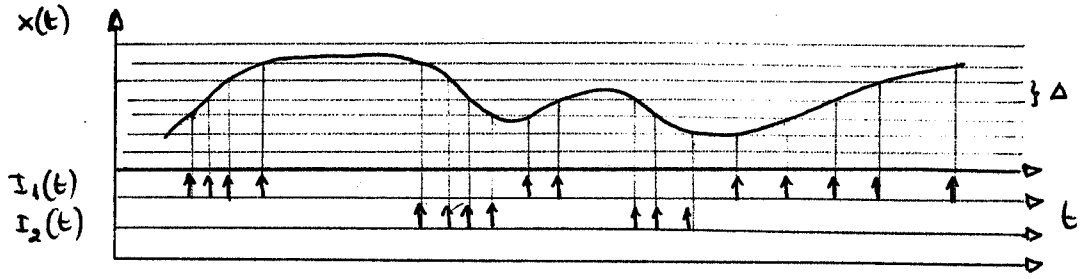
Ce codage est :

- résistant aux perturbations
- de rapport signal/bruit convenable surtout pour une largeur de bande large (il croit exponentiellement avec B)
- introduit des effets de masque temporel puisqu'il interdit la perception de durées inférieures à T
- nécessite l'existence d'une horloge

3.2 Codage en modulation delta

C'est un codage d'amplitude différentiel.Les propriétés de ce code sont identiques au précédent au plan des qualités de transmission. Il nécessite par contre deux "lignes" l'une pour coder les pentes positives et l'autre pour coder les pentes négatives. Il présente en outre l'avantage de ne pas utiliser d'horloge dont l'existence n'a pas été mise en évidence dans l'oreille. Le grand nombre de fibres pourrait alors s'expliquer en considérant qu'un mot de code est transporté par plusieurs fibres.

Le schéma ci-après montre un codage en modulation delta de l'enveloppe sur deux fibres.



$I_1(t) = 1$ si $x'(t) > \Delta$, 0 sinon (pentes positives)
$I_2(t) = 1$ si $x'(t) < -\Delta$, 0 sinon (pentes négatives)

Une transmission sans perte est obtenue seulement pour les signaux à variations bornées. Avec les valeurs précédentes on obtient un masque temporel du même ordre, c'est à dire environ 3 ms.

3.3 Codage en fréquence : les passages par zéro

En basse fréquence on a pu constater que les impulsions sont synchrones au signal d'entrée. Cela suggère donc d'examiner les codes par passage par zéro bien connus en traitement du signal. Il est évident qu'une telle technique ne peut convenir qu'en basse fréquence (précision de la détection des impulsions, période réfractaire etc...). Ce codage peut opérer directement sur le signal plutôt que sur son enveloppe où il apporterait moins d'informations (détections des extrema seulement). Sur le signal il peut contribuer à l'affinement des mesures en fréquence. La fig 6 montre un tel codage qui permet à la détection (démodulation par construction d'histogrammes - PST en américain) d'affiner la précision fréquentielle. Les impulsions sont symbolisées à la droite du signal (caractère i) pour les fibres émettrices seulement. Un simple comptage permet de dresser l'histogramme (en haut à gauche).

Ce codage est :

- fiable en basse fréquence (pour les formants F_0 et F_1 par exemple) surtout après un préfiltrage tel que celui de la membrane basilaire.
- peu sensible aux harmoniques du fondamental F_0 (gain de précision pour F_1)
- sujet aux distorsions non-harmoniques du type $n_1 f_1 \pm n_2 f_2$ pour deux sons purs, distorsion venant renforcer celles qui sont introduites par les cellules ciliées (voir 2.3).

Conclusion

A ce stade des recherches on peut imaginer que le

codage nerveux pourrait être :

- un codage par passage par zéro du signal de sortie des cellules ciliées dans les basses fréquences (jusqu'à 1400 hz par exemple)
- un codage en modulation delta de l'enveloppe du signal de sortie des cellules ciliées dans les hautes fréquences.

Dans l'état actuel des connaissances physiologiques de telles hypothèses sont acceptables. il est bien évident qu'elles doivent être confirmées.

4. PROBLEMES POSES PAR L'INTERACTION DE TOUS LES ETAGES

On peut noter tout de suite l'importance de ces interactions, esquissées quelque peu dans les paragraphes précédents au sujet des distorsions. Quel est par exemple la contribution des cellules ou celle des fibres dans l'apparition de sons de fréquence f_1-2f_2 pour des tons purs ? Il est difficile d'y répondre.

D'autres interactions interviennent directement sur le filtrage, le codage, les modulations, dont il est difficile de tenir compte. Cela limite donc l'utilisation d'un modèle dans la description fidèle de la réalité. Par contre il reste possible de simuler certaines adaptations indirectes au signal telles que :

- mouvements de la tête
- réflexes stapédiens (commande de sélectivité et de gain)
- phénomènes non-linéaires de saturation
- inhibition latérale
- masques temporels et fréquentiels
- etc . . .

qui permettent d'améliorer le rapport signal/bruit, la détection fréquentielle, les sélectivités etc ... et pour lesquelles des techniques comparables à celles qui sont mises en oeuvre dans la commande des systèmes, peuvent être utilisées. (CAELEN J. 1979)

CONCLUSION

Bien qu'imparfait un modèle peut apporter sa contribution en analyse de signaux et en particulier en analyse de la parole. C'est d'autant plus nécessaire que le signal est complexe ce qui est le cas de la parole. Inspiré par l'oreille on peut rechercher des filtrages encore plus performants (extension décrite en 1.) des détections plus précises de fréquences. Inversement les techniques du traitement du signal peuvent permettre une meilleure compréhension des phénomènes auditifs. Dans un cas comme dans l'autre l'efficacité maximum doit être trouvée à travers un modèle global qui doit faire la part entre les phénomènes principaux et les phénomènes secondaires.

BIBLIOGRAPHIE

ANDERSON D.J, TODR J.R, HIND J.E, BRUGGE J.F 1973
Temporal position of discharges in single auditory nerve fibers within the cycle of auditory nerve discharges
J.A.S.A. n° 51

CAELEN J. 1979
Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique. Thèse d'état, TOULOUSE
CHARLET DE SAUVAGE R, CAZALS Y, ARAN J.M 1977
Modèles d'oreille interne et analyse du potentiel d'action du nerf auditif. Revue d'Acoustique vol 10 n° 42
DALLOS P. 1973
The auditory periphery . Acad. Press NEW YORK
DUIFHUIS H. 1976
cochlear non-linearity and second filter : possible mechanism and implications. J.A.S.A. n° 59
DE BOER E. 1975
Synthetic whole nerve action potentials for the cat. J.A.S.A. vol 58 n° 5
DOLMAZON J.M, BASTET L. 1977
Un modèle fonctionnel du système auditif périphérique. Revue d'Acoustique vol 10 n° 42
EVANS E.F 1974
Auditory frequency selectivity and the cochlear nerve Springer Verlag BERLIN
KIANG N.Y.S, MOXON E.C 1974
Tails of tuning curves of auditory nerve fibers. J.A.S.A. vol 54 n° 6
HALL J.L 1975
Nonmonotonic behavior of distorsion product $2f_1-f_2$, psychological observations. J.A.S.A. vol 58 n° 5
MØLLER A.R 1976
Dynamic properties of primary auditory fibers compared with cells in the cochlear nucleus. Acta Phys. Scand. n° 98
PFEIFFER R.R 1970
A model for two-tone inhibition of single cochlear nerve fibers. J.A.S.A. vol 48 n° 6 part 2
RHODE W.S 1974
An investigation of cochlear mechanism using the Mössbauer technic. Acad. Press. NEW YORK
RUSSELL R., PFEIFFER R.R, KIM D.O 1975
Cochlear nerve fibers responses : distribution along the cochlear partition. J.A.S.A. vol 58 n° 4
SPATARU A. 1970
Théorie de la transmission de l'information. (Traduction) Masson et Cie PARIS
SPOENDLIN H. 1971
The organization of the cochlear receptors. Bibl O.R.L. n° 13
ZWICKER E, THERHARDT E 1974
Facts and models in hearing. Springer Verlag BERLIN

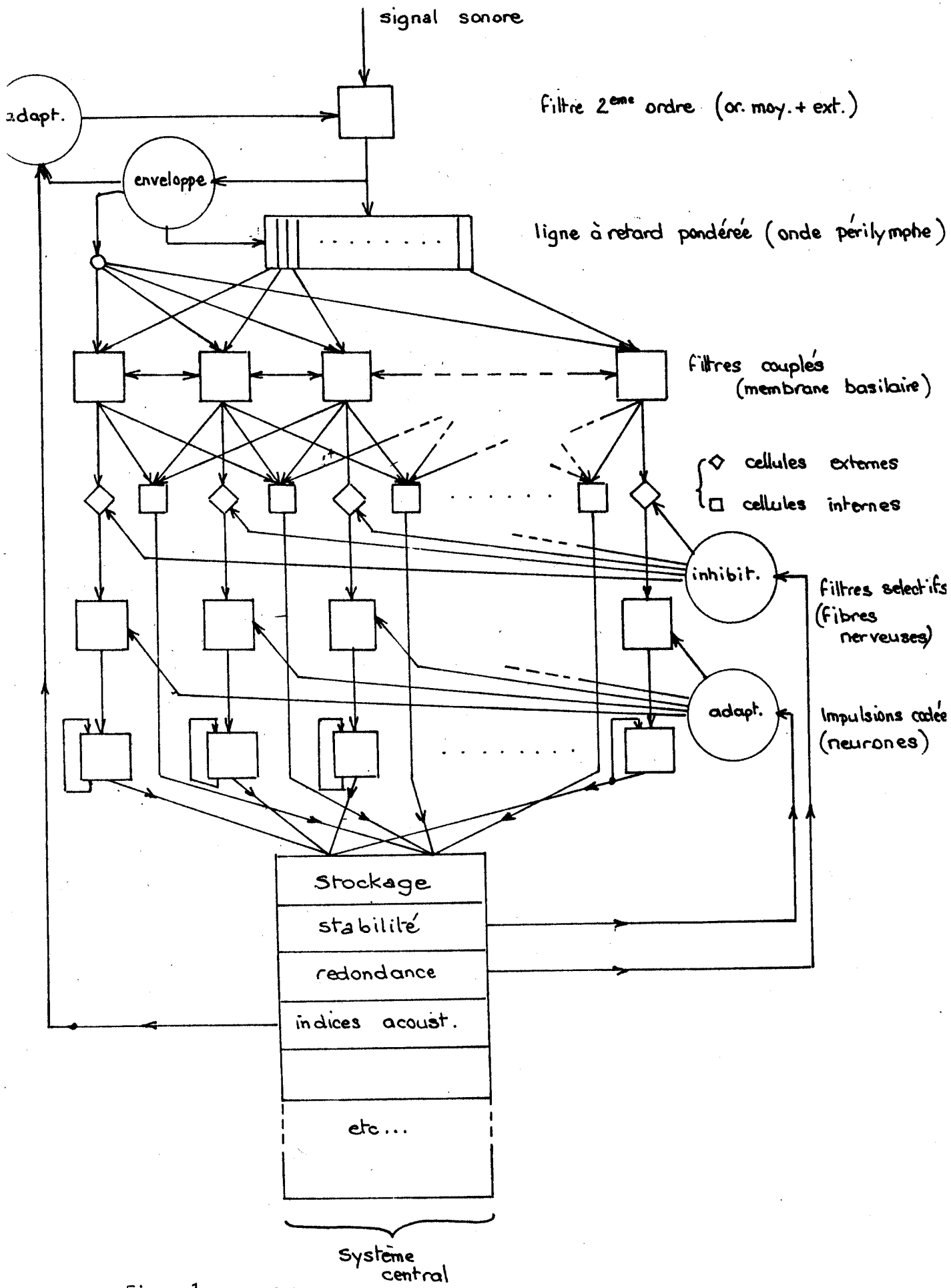


Fig. 1 : Schéma général du modèle d'oreille

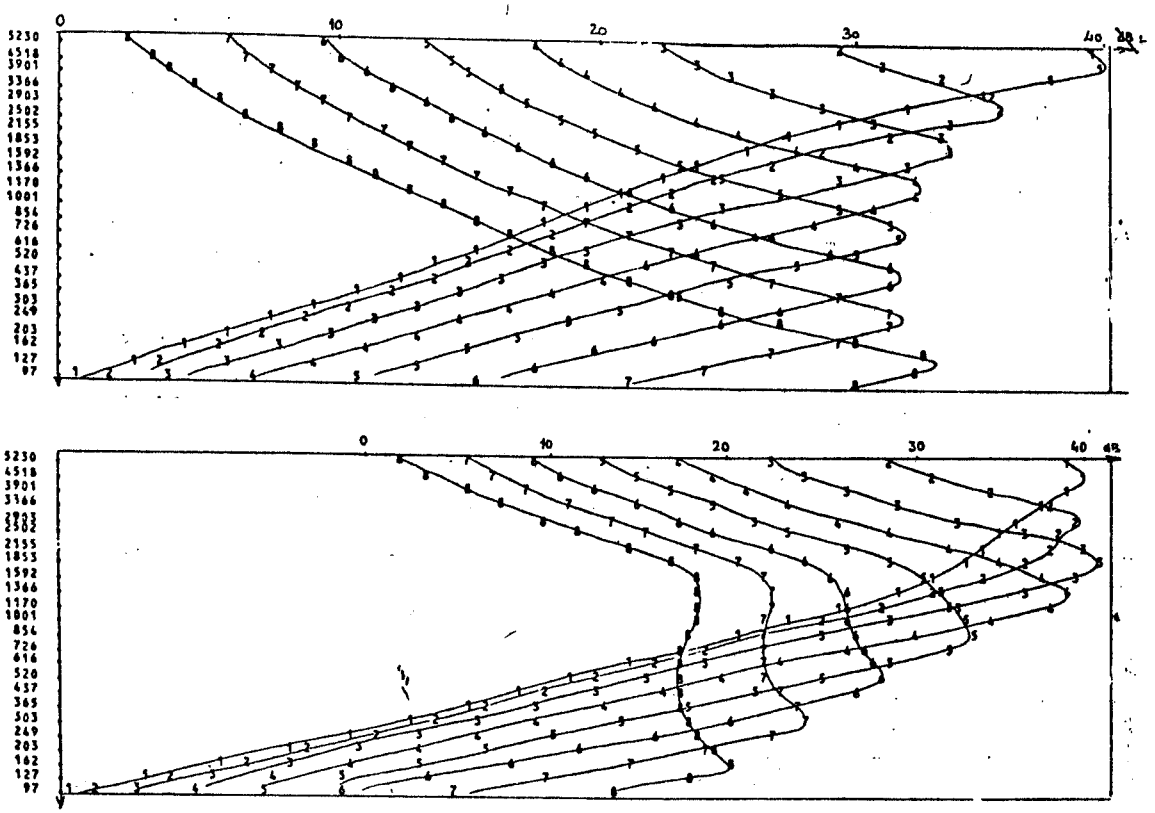


Fig. 2 : Modules de fonctions de transfert de quelques filtres du modèle. En haut, les filtres sont isolés, en bas, ils sont connectés au modèle de l'oreille externe et moyenne.

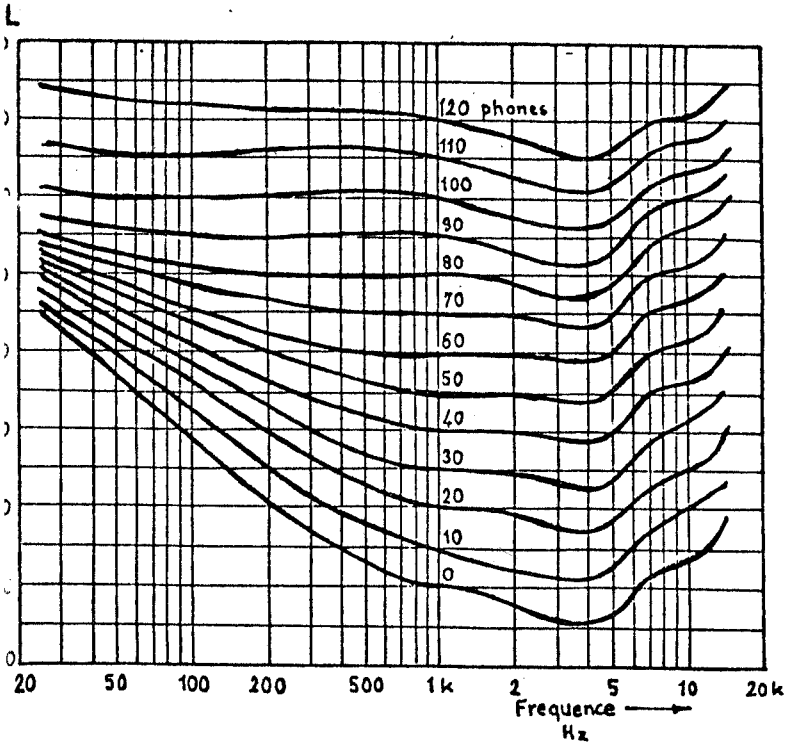


Fig. 3 : Courbes de sensibilité.
D'après FLETCHER et MUNSON

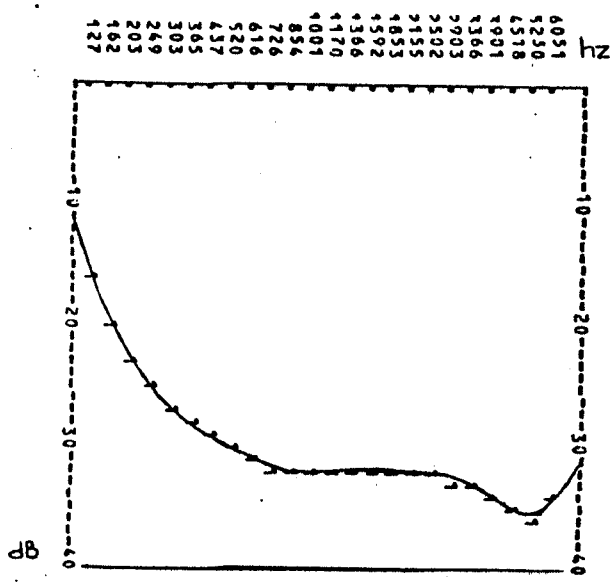


Fig. 4 : Courbe de sensibilité du modèle à 30 dB

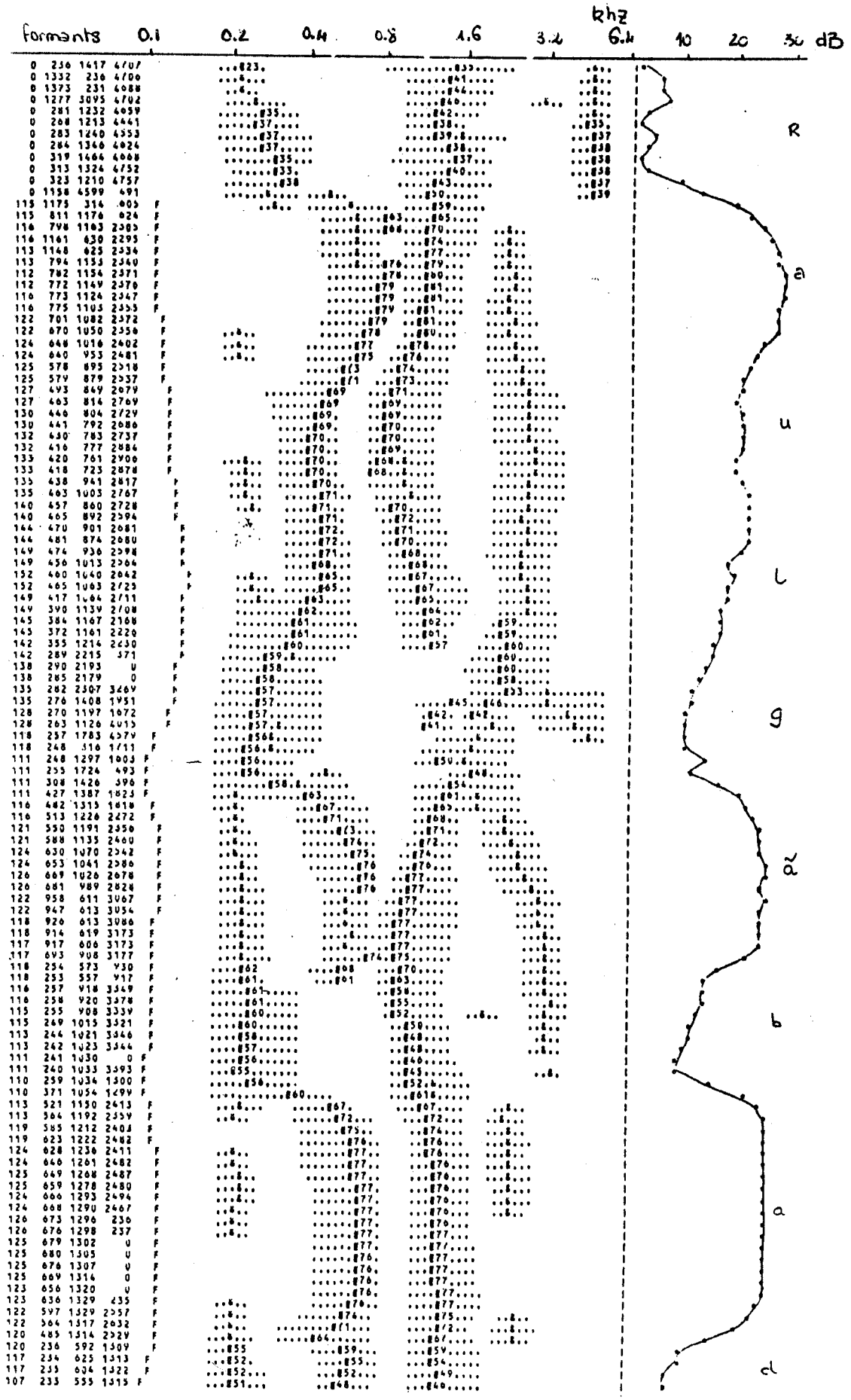


Fig. 5 : tracé du spectrogramme à la sortie de la M.B. avec les conventions : F pour fondamental... Pour bande d'énergie, pour crête principale, & pour crête secondaire. Phrase traitée : "Raoul gambade".

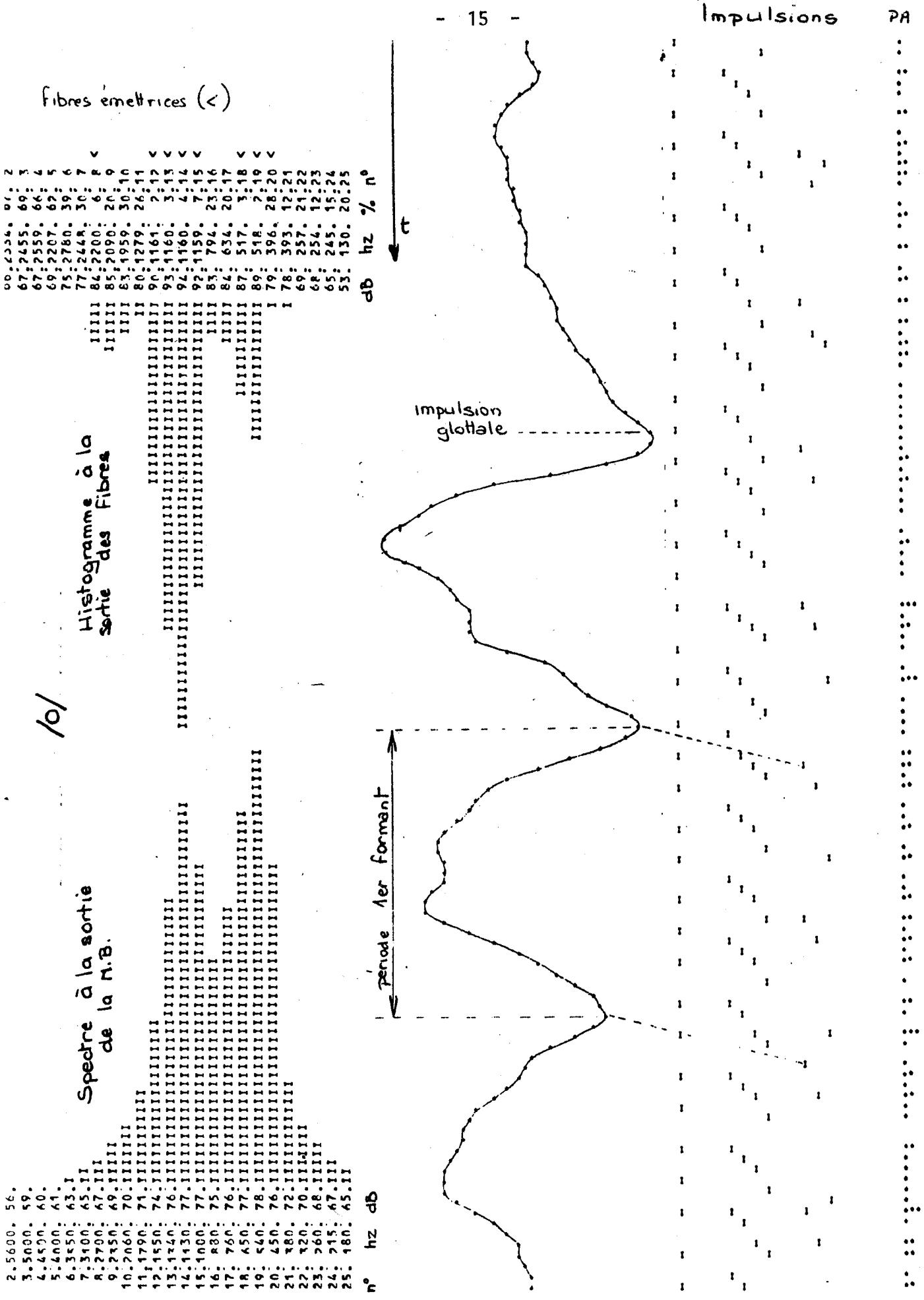


Fig. 6.12 : Impulsions délivrées par les fibres au signal /o/ avec inhibition latérale. Histogramme et coupe spectrale correspondantes.

THEME II : Intelligibilité et qualité de la parole naturelle, de la parole codée et de la parole de synthèse.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

PERCEPTION AUDITIVE DES FRICATIVES PAR LES DEFICIENTS AUDITIFS

S. BARTH, R. CHULLIAT

INSTITUT NATIONAL DE JEUNES SOURDS DE CHAMBERY

B.P. 15 73 160 COGNIN

RESUME : le but de ce travail est de rechercher le degré d'altération de la perception des fricatives consécutive à une déficience auditive et de classer les traits qui les caractérisent par ordre d'informativité.

PERCEPTION OF FRICATIVES BY DEAF PERSONS

S. BARTH, R. CHULLIAT

NATIONAL INSTITUTE FOR THE DEAF AT CHAMBERY

B.P. 15 73 160 COGNIN

SUMMARY : the aim of this work is to find to what degree deafness alters the perception of fricatives and to classify the features of these ones according to their informatory qualities.

I INTRODUCTION

Lors de la création du test par paires minimales, M. ROSSI avait suggéré l'utilisation de ce outil dans le domaine des troubles de l'audition de la parole.

Après les travaux de FONTANEZ et BERAHA (1977), BESSON et PASCAL (1975), il nous a paru utile de donner un développement à cette idée en pratiquant certaines adaptations rendues nécessaires par nos sujets déficients auditifs de naissance.

Notre choix s'est porté sur la perception des fricatives dont la difficulté est connue de tous les praticiens de la surdité. De nombreux travaux fournissent des données comparatives générales sur celle-ci : HEINZ et STEVENS (1958), HARRIS (1958), DELATTRE et al. (1964), CHAFCOULOFF et al. (1976), AUTESERRE et BOE (1976), CHAFCOULOFF et DICRISTO (1978). Pour les déficients auditifs, nous noterons ceux de BRADBERRY (1970), WALDEN et MONTGOMERY (1973), DANHAUER (1974), DANHAUER et SINGH (1975) qui utilisent une analyse en traits.

En nous basant sur la matrice phonétique de M. ROSSI présentant une analyse binaire des traits, nous avons préparé un test par choix forcé constitué de logatomes. La raison de cette démarche est que nous avons maintes fois constaté que le sourd placé dans une situation de choix forcé favorise généralement, en cas de doute, le mot qu'il connait ou qu'il utilise couramment (ex : caniche sera plus souvent "reconnu" que canisse).

II MATERIEL PHONETIQUE

Nous avons retenu les oppositions suivantes établies sur un seul trait :

- voisé - non voisé

f ~ v

ʃ ~ ʒ s ~ z

- grave - aigü

f ~ s

v ~ z

- compact - diffus

ʃ ~ s

ʒ ~ z

Chacun de ces groupes a été associé aux voyelles i, a, u. Les consonnes ont été placées en position implosive et explosive.

Les 84 logatomes obtenus, distribués au hasard en plusieurs listes, ont été enregistrés par un locuteur masculin, adulte, entendant, à raison d'un item toutes les 5 secondes (magnétophone REVOX A 77, micro SENNHEISER MD 441, bande SONY H.L), dans le chambre sourde de notre laboratoire.

III PROCEDURE D'EXPERIMENTATION

Ayant constaté que l'utilisation des prothèses auditives de nos sujets amenait une augmentation de la dispersion des résultats (voir S. BARTH, R. BENFADHEL, G. MAJO dans le même volume), nous avons préféré opérer avec un amplificateur de table INTERACOUSTICS DS 4 équipé d'un casque KHOS.

Avant chaque test, il était demandé au sujet de régler lui-même le niveau sonore de restitution de la bande, de manière à obtenir une écoute confortable. Le corpus a été présenté à 30 personnes réparties de la manière suivante :

- 10 déficients auditifs moyens (D.A.M.)
- 10 déficients auditifs sévères (D.A.S.) (15 à 16 ans)
- 10 déficients auditifs profonds (D.A.P.)

Les degrés de surdité correspondent aux normes du Bureau International d'Audiophonologie.

Dans chacun des groupes, les audiogrammes ne présentaient pas un écart supérieur à $\pm$ 5 dB par rapport à l'audiogramme moyen.

Chaque sujet disposait d'une feuille de test préimprimée.

IV RESULTATS

IV 1) relation entre le taux d'erreur global et la perte auditive moyenne

Une étude de régression appliquée aux pertes auditives moyennes (normes BIAP) et aux pourcentages d'erreurs commises conduit à la relation :

$$y = 0.29\ x - 2.64$$

où x est la perte auditive moyenne exprimée en décibels HL et y le pourcentage global d'erreurs.

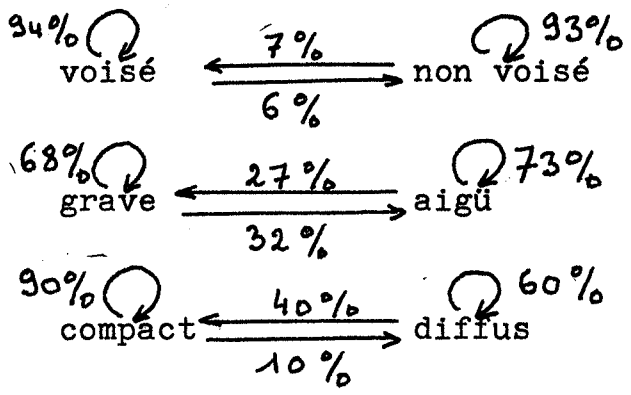
Par extrapolation de nos points expérimentaux, nous obtiendrions un taux d'erreur nul avec une perte moyenne d'environ 10 dB.

Notons, cependant, que deux DAS ont obtenu des résultats supérieurs à la moyenne des DAM.

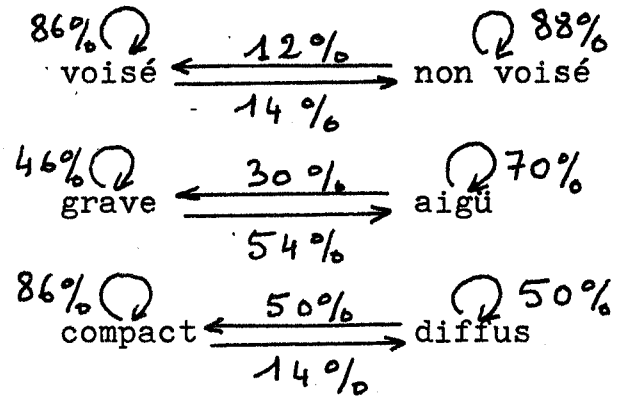
IV 2 Confusions entre traits

Tous contextes confondus, nous obtenons, pour chaque groupe de sujets, les résultats suivants (→ indique le sens de la confusion et ↻ une perception exacte) :

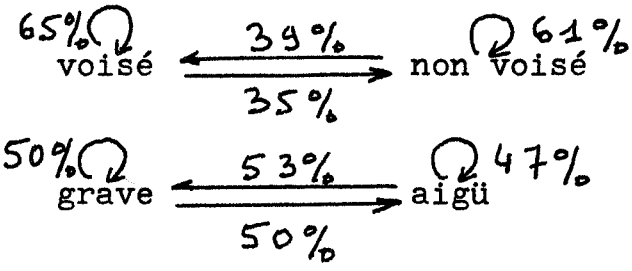
DAM

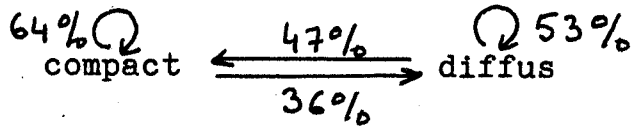


DAS



DAP





En comparant les taux de reconnaissance, on obtient pour chaque groupe une "hiérarchie" des traits

DAM	DAS	DAP
voisé (94)	non voisé (88)	voisé (65)
non voisé (53)	voisé (86)	compact (64)
compact (90)	compact (86)	diffus (53)
aigü (73)	aigü (70)	grave (53)
grave (68)	diffus (50)	non voisé (51)
diffus (60)	grave (46)	aigü (47)

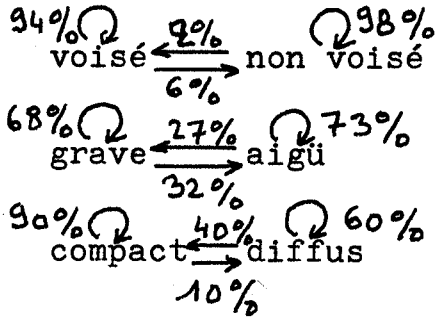
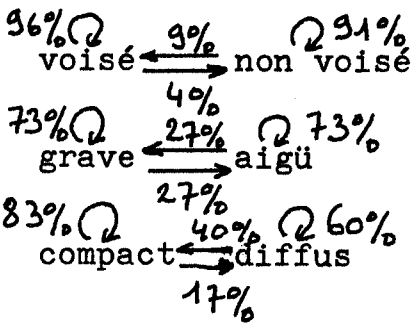
Les nombre entre parenthèses sont les taux de reconnaissance. On constate bien une augmentation des erreurs parallèle à celle du degré de surdité.

Il est possible de classer les traits en tenant compte de leurs résistances aux altérations audiologique :

voisé (81)
non voisé (77)
compact (76)
aigü (63)
grave (55)
diffus (54)

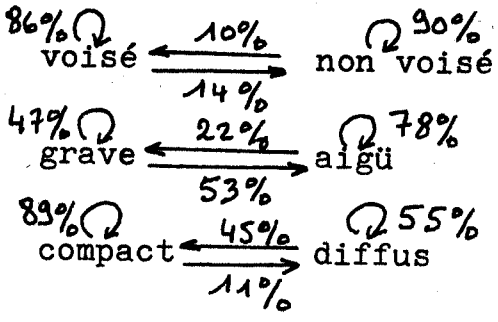
Notons que l'opposition voisé-non voisé est la plus résistante et présente une certaine symétrie au niveau des confusions. Par contre, il existe une dissymétrie de celles-ci pour les oppositions grave-aigü et compact-diffus.

IV 3 Influence de la position de la consonne

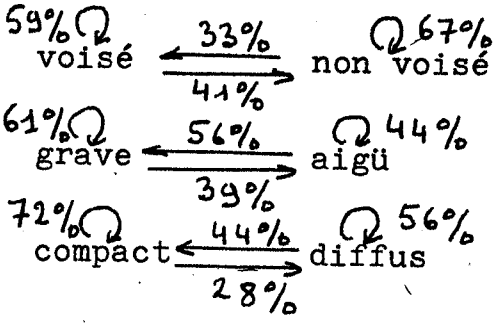
	implosive	explosive
DAM		

implosive

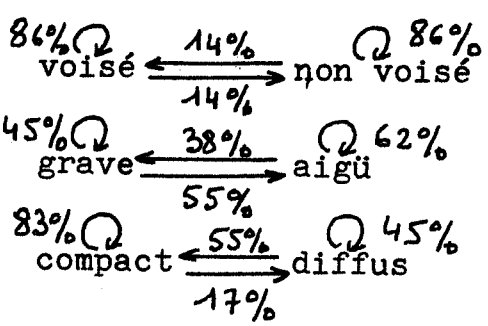
DAS

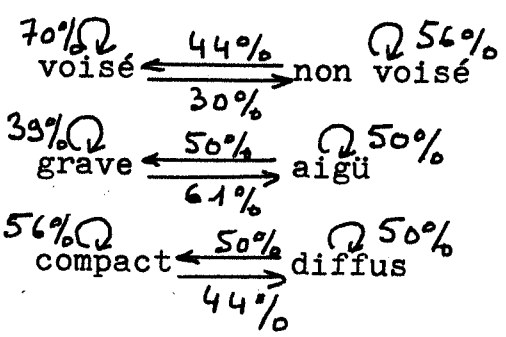


DAP



explosive





Comme en IV 3, on peut tirer une classification des traits pour chaque groupe.

implosive

DAM

non voisé (98)
compact (97)
voisé (91)
aigü (73)
grave (63)
diffus (60)

DAS

non voisé (90)
compact (89)
voisé (86)
aigü (78)
diffus (55)
grave (47)

explosive

DAM

voisé (96)
non voisé (91)
compact (83)
aigü (73)
grave (73)
diffus (60)

DAS

voisé (86)
non voisé (86)
compact (83)
aigü (62)
diffus (45)
grave (45)

DAP

compact (72)	voisé (70)
non voisé (67)	non voisé (56)
grave (61)	compact (56)
voisé (59)	diffus (50)
diffus (56)	aigü (50)
aigü (44)	grave (39)

En classant ces traits du mieux reconnu au moins bien perçu :

implosive	explosive
compact (86)	voisé (84)
non voisé (85)	non voisé (77)
voisé (78)	compact (74)
aigü (65)	aigü (61)
grave (57)	grave (52)
diffus (53)	diffus (51)

Au niveau des consonnes, on observe que :

- ʒ est mieux reconnu par nos sujets en position explosive où le voisement l'emporte sur la compacité.
- ʃ est mieux reconnu en position implosive, la compacité étant mieux perçue.
- z obtient un meilleur score en position implosive ainsi que f . Celle-ci semble favoriser l'opposition grave-aigü.
- on ne note pas de différence significative entre les deux positions pour s et v .

IV 4 Influence de la voyelle associée

D'une manière globale, les taux de reconnaissance des consonnes sont :

taux			voyelle associée
DAM 75 %	DAS 68 %	DAP 53 %	a
74 %	66 %	51 %	i
68 %	64 %	49 %	u

La reconnaissance est moins bonne lorsque v et ʒ sont associés à u . Par contre, elle s'avère meilleure lorsque f et s sont associés à a et quand z est associé à i .

V CONCLUSION

Ce travail fait partie d'un programme d'étude qui vise à cerner les problèmes d'éducation auditive que pose la scolarisation de jeunes déficients auditifs. Les premiers résultats semblent montrer que l'on peut classer les traits suivant un degré d"informativité" ou plus précisément de résistance aux altérations créées par une déficience auditive.

BIBLIOGRAPHIE

PECKELS, M. ROSSI : le test de diagnostic par paires minimales. VIIe JEP 1976.

J.P. BERAHA : recherche sur la pertinence des indices acoustiques de la parole chez le malentendant appareillé. XVe Assises de la prothèse auditive - PARIS 1977.

M.E. BRADBERRY : a distinctive feature analysis of initial consonants of preschool deaf children who received verbo-tonal therapy - unpublished thesis - university of Tennesse KNOXVILLE.

B.E. WALDEN, A.A. MONTGOMERY : dimensions of consonant perception in normal hearing and impaired listeners - Annual convention of the american speech and hearing association - DETROIT 1973.

J.L. DANHAUER : a multi dimensional analysis of hearing impaired subject's responses to sixteen consonants - unpublished doctoral dissertation - OHIO university ATHENS.

J.L. DANHAUER, SINGH : language and speech 18/1, p. 42-64 - 1975.

CHAFCOULOFF et al : TIPA 3 p. 61-113 - 1976.

D. AUTESERRE, L.J. BOE : TIPA 3 p. 9-57 - 1976.

FRANQUEVILLE : contribution à l'étude de l'audition filtrée dans des bandes de fréquences correspondant aux restes auditifs des enfants sourds - thèse école Nationale de la Santé Publique - PARIS 1962.

FRISINA : the auditory channel in the education of deaf children - American Annals of the Deaf - Nov. 1966.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

SEUILS DE DETECTION D'UN BRUIT DE TYPE "MALT" AJOUTE A LA PAROLE (VOYELLES STATIONNAIRES).

COMBESCURE Pierre

CNET LANNION
22301 LANNION

RESUME FRANCAIS

La méthode d'isopréférence, qui peut être utilisée pour juger de la qualité des codeurs ajoutant du bruit au signal de parole, cherche à donner une valeur de rapport signal sur bruit comme note de qualité d'un codeur. De nombreuses optimisations de codeur sont faites en cherchant à minimiser la puissance de l'erreur de prédiction et donc de quantification. Très peu d'études ont été faites sur la perception de ces bruits par l'oreille humaine. Cet article présente une étude des seuils de détection du bruit de quantification sur la parole, plus exactement sur les parties voisées (certaines voyelles) où ce type de bruit est le plus sensible. Une hypothèse psychoacoustique tenant compte de certaines caractéristiques de la perception par l'oreille humaine (masquage fréquentiel) permet de prévoir ces seuils avec un indice de détermination de 91 %. Quelques conséquences sur la méthode d'isopréférence et sur la façon d'optimiser les codeurs en fonction du paramètre bruit de quantification en sont déduites.

ENGLISH SUMMARY

Isopreference method, which can be used to evaluate quality of coders, gives a single value of signal to noise ratio as a measure of quality. Many optimisations of coders have been made by minimising power of the error of prediction and therefore of quantization. Very few studies have been undertaken about perception of such noise by the human ear. This paper describes an experiment on detection threshold of quantization noise. The thresholds have been measured for vowels and for twelve speakers by a transformed up and down method. Speech material (low-pass filtered at 3400 Hz) was presented to the subjects through high quality headphones. The thresholds are strongly dependent on the identity of vowels and speakers. The variation of threshold was 20 dB for the entire measurement. Female voices are more sensitive to quantization noise than male voices. A psychoacoustic hypothesis which would predict thresholds with an index of determination of 91 %, is proposed. Some useful consequences can be derived from this experiment.

- The detection thresholds of quantization noise are strongly dependent on spectral shape of vowels. The important parameters are magnitude of formants and their masking effects on adjacent frequencies.

- Optimisation of coders should take advantage of properties of the perception of speech by shaping quantization noise for minimising the intensity of perceived noise.

- The isopreference method must be used extremely carefully. When quantization noise differs from white noise, the results will be strongly dependent on spectral shape of speech, and thus on sentences and speakers used.

SEUILS DE DETECTION D'UN BRUIT DE TYPE "MALT" AJOUTE A LA PAROLE (VOYELLES STATIONNAIRES)

COMBESCURE Pierre CNET LANNION

Les codeurs de parole impliquent nécessairement une quantification du signal analogique. Les lois de quantification les plus utilisées introduisent des dégradations qui ne tiennent pas compte du comportement de l'oreille humaine d'une manière fine. La méthode d'isopréférence, qui peut être utilisée pour juger de la qualité des codeurs ajoutant du bruit au signal de parole, cherche à donner une valeur de rapport signal sur bruit comme note de qualité d'un codeur. Cette valeur peut-elle être définie sans ambiguité et dans quelles conditions ? De nombreuses optimisations de codeur sont faites en cherchant à minimiser la puissance de l'erreur de prédiction et donc de quantification. Hors, perceptuellement, une puissance d'erreur de quantification petite ne signifie pas forcément une distorsion petite du signal de parole.

Il est donc intéressant d'étudier comment est perçue la dégradation de la parole due au bruit de quantification et en particulier les seuils de détections de ces bruits. Plus précisément nous avons considéré la distorsion introduite par le M.I.C., distorsion qui peut être simulée grâce à l'addition d'un bruit blanc corrélé à la parole. Ceci est réalisé par un "MALT" qui, de plus est utilisé pour fournir l'échelle de référence en rapport signal sur bruit dans les essais d'isopréférence. Une préexpérience ayant montré que le bruit est plus sensible sur les parties voisées du signal nous nous sommes limités à l'étude d'un groupe de voyelles stationnaires. De plus nous avons considéré des codages de bonne qualité dans la bande téléphonique. On peut penser que les jugements sont alors influencés principalement par la détection du bruit sur certaines parties de la parole. Lorsque la qualité se dégrade les seuils de détection ne sont plus suffisants mais un programme de sonie des bruits détectés (type ZWICKER 1965) pourrait sans doute apporter des éléments de réponse.

MALT : Mesureur d'Agrément des Liaisons Téléphoniques.

I. PROCEDURE ET CORPUS EXPERIMENTAL

I.1. Définition du corpus

L'enregistrement des cinq voyelles stationnaires (a , ø , i , ɔ , y), a été effectué, en chambre sourde, sur magnétophone TANDBERG en utilisant un réducteur de bruit (DOLBY normal). Dix locuteurs ont participé à l'expérience (5 hommes n° 3, 4, 5, 6, 10 et 5 femmes n° 1, 2, 7, 9, 12). Il leur a été demandé de parler à une force normale, à 10 cm du microphone (Beyer Dynamic), en soutenant les voyelles durant deux secondes environ. Nous avons demandé à deux locuteurs (n° 10 et n° 7), de refaire le même enregistrement mais en forçant la voix (environ + 10 dB). Ces deux enregistrements sont considérés comme provenant des locuteurs n° 11 et 8. Les 60 voyelles, ainsi obtenues (5 x (10 + 2)) ont ensuite été numérisées après filtrage dans la bande 0-3400 Hz (48 dB/octave). Les caractéristiques de la numérisation ont été les suivantes : codeur linéaire 12 bits, durée d'enregistrement 1 seconde, fréquence d'échantillonnage 8 kHz. Une fenêtre de pondération a été appliquée au début et à la fin de l'enregistrement pour éviter les "clicks" au démarrage (problème des seuils de déclenchement automatique) et rendre plus naturelle la fin de la voyelle (durée de la fenêtre 20 ms).

I.2. Méthode automatique de poursuite de seuil à pas variable

Le niveau d'isopréférence situe en qualité un système de transmission par rapport à un signal de référence dégradé par un bruit multiplicatif.

Nous avons utilisé pour notre expérience un "MALT" dont le rapport signal sur bruit peut être commandé par calculateur. On a pris comme voyelles de référence les 60 voyelles passées à travers le MALT avec un rapport signal sur bruit de 69 dB. Il s'est avéré que cette valeur était toujours en dessous du seuil de détection du bruit multiplicatif (le rapport signal sur bruit de la chaîne d'écoute étant par ailleurs de l'ordre de 60 dB). Les voyelles étaient présentées par paire : voyelle de référence - voyelle bruitée.

La méthode utilisée pour déterminer les seuils de détection est une méthode entièrement automatisée de poursuite de seuil à pas variable (LEWITT et WETHERILL, 1965). Les caractéristiques de la procédure sont les suivantes :

- le point de convergence recherché est le point 50 % de la courbe psychométrique (c'est-à-dire le point où l'on a 50 % de réponses positives).

. une réponse positive entraîne, pour le passage suivant, une augmentation du rapport signal sur bruit du MALT, donc une diminution du niveau de bruit à détecter.

. une réponse négative entraîne pour le passage suivant une diminution du rapport signal sur bruit du MALT, donc une augmentation du niveau du bruit à détecter.

. le nombre total de basculement autour du seuil est en général de huit.

. les pas de variations du rapport signal sur bruit sont les suivants : 4 dB avant le premier franchissement du seuil, 2 dB après le premier franchissement, 1 dB après le troisième franchissement.

. la valeur du seuil de détection est calculée comme étant la moyenne arithmétique des valeurs du rapport signal sur bruit lors des basculements. Le nombre de basculements étant pair, le nombre des valeurs supérieures ou inférieures au seuil est identique réduisant ainsi les effets de biais.

. lors de la présentation des paires (A-B) la position de la référence n'est pas connue de l'auditeur, de plus elle est changée automatiquement au cours du test. Une possibilité de réécoute est prévue.

. une préexpérience sur deux auditeurs a permis de déterminer les valeurs initiales du rapport signal sur bruit pour chaque voyelle. Lors de cette expérience il nous est apparu que le nombre initial de basculements choisi (6) était trop faible (les jugements des sujets n'étant pas tous stabilisés au bout de 6 basculements).

En effet la stabilité des jugements peut facilement être observée, sur les réponses des sujets, grâce à cette méthode. Nous avons donc choisi de porter à 8 le nombre de basculements et de juger de la stabilité des résultats en faisant la différence entre la valeur "seuil" définie ci-dessus et la valeur calculée sur les quatre derniers basculements. Lorsque cette différence dépassait 1,5 dB deux basculements supplémentaires étaient rajoutées. De cette façon, seuls 3 % des jugements n'ont pas rempli notre critère de stabilité.

- la parole a été présentée aux auditeurs sur casque, en écoute binauriculaire monophonique (KOSS/747). Chaque sujet a pu régler le niveau d'écoute à la valeur qu'il jugeait la plus confortable pour le test. Les niveaux choisis ont ainsi varié entre 75 dB et 85 dB (en linéaire, recueilli sur oreille artificielle CEI/CCITT).

. 10 sujets audiologiquement normaux ont participé à l'expérience (25-35 ans).

. le test a été divisé en quatre séances d'une demi-heure environ pour chaque sujet.

. les consignes données aux sujets étaient les suivantes :

- vous allez entendre une paire de voyelles
- si vous percevez une différence appuyez sur le bouton 1
- si vous ne percevez pas de différence appuyez sur le bouton 2
- si vous désirez réécouter appuyez sur le bouton 3.

II. RESULTATS

II.1. Analyse globale et premières remarques

Les tableaux n° 1, 2, 3, 4, 5 fournissent pour chaque voyelle et pour chaque locuteur la moyenne des seuils de détection sur les 10 auditeurs ainsi que l'écart type de cette moyenne. On remarquera que les écarts types sont assez importants (2 dB), ceci étant sans doute dû au nombre d'auditeurs un peu restreint ainsi qu'au caractère de la tâche (les expériences psychoacoustiques conduisent très souvent à des écarts types (inter-sujet) importants).

Il est intéressant de noter que :

. un écart de 20 dB existe entre le seuil de détection le plus élevé et celui le plus bas, toutes voyelles et tous locuteurs mélangés (e.J.D 37, OFE(N) 58).

. pour une même voyelle prononcée par divers locuteurs l'écart peut atteindre 10 dB.

. certains locuteurs se distinguent très nettement de l'ensemble des autres ainsi les voyelles prononcées par les locuteurs n° 6, 10 (J.D. MF(N)) donnent lieu à des seuils de détection (exprimés en S/B) systématiquement plus faibles (2 à 3 dB) que la moyenne des locuteurs masculins.

Inversement les voyelles prononcées par la locutrice n° 7 (F.E.) donnent lieu à des seuils de détection (exprimés en S/B) systématiquement plus élevés que la moyenne (3 à 4 dB). Cette locutrice possède pourtant un fondamental assez bas (175 Hz) mais par contre une pauvreté très marquée des spectres de sa voix dans les aigus (1500-2500 Hz). D'autre part les seuils de détection (exprimés en S/B) obtenus sur les voyelles prononcées par la locutrice n° 12 sont systématiquement plus faibles que la moyenne (2 à 3 dB environ). Cette dernière malgré un fondamental très élevé (250 Hz) peut au vu des résultats sur les seuils, être rattachée au groupe des hommes, ceci étant sans doute dû à la très grande richesse du spectre de sa voix dans les aigus (1500-2500 Hz).

II.2. Différences homme-femme

Le tableau n° 6 donne pour chaque voyelle les moyennes et écarts types de la moyenne pour les deux groupes hommes, femmes ainsi que la valeur des tests de signification des différences entre les moyennes des groupes.

Les aspects suivants sont à noter :

. toutes les différences entre les moyennes homme/femme sont significatives à plus de 98 % sauf pour le a.

. les voyelles prononcées par des locutrices présentent en moyenne des seuils de détection (exprimé en S/B) plus élevés que celles prononcées par des hommes, l'écart étant en moyenne de 2 à 3 dB mais pouvant atteindre 6 dB sur le i.

II.3. Différences entre voyelles

Le tableau n° 7 donne les moyennes générales et écarts types des moyennes ainsi que les valeurs du test t de signification des différences entre ces moyennes pour les cinq voyelles.

On peut en déduire le classement des voyelles par ordre de seuil de détection (exprimé en S/B) décroissant
le "o", le "u" et le "i", le "e" enfin le "a"
mais ces moyennes cachent en réalité l'aspect important suivant :

pour les voyelles prononcées par des femmes le classement (o, i, u, e, a) est significatif chaque différence entre ces moyennes étant significatives ; pour les locuteurs hommes, seul le o est significativement différent des autres voyelles.

II.4. Différences voix normale-voix forcée

Le tableau n° 8, enfin, récapitule les résultats pour les deux locuteurs auxquels nous avons demandé de refaire un enregistrement en forçant la voix.

Les résultats sont fort différents pour la locutrice (n° 7, n° 8) et le locuteur (n° 10 et n° 11).

Pour la locutrice :

- les seuils de détection ont décru, de 3 à 6 dB suivant les voyelles (sauf pour le i). Les spectres des voyelles correspondantes laissent apparaître

1 - une élévation du fondamental (175 à 220 Hz).

2 - un rehaussement du niveau relatif des fréquences aiguës (1500 à 2500 Hz) par rapport aux fréquences graves.

Ces phénomènes sont d'ailleurs classiques lorsque l'on demande à un locuteur de forcer la voix.

Pour le locuteur masculin l'effet est moins évident :

Le seuil de détection a décru de façon sensible uniquement sur le u, sur le a et le e l'effet inverse s'est produit. L'examen des spectres des voyelles laisse apparaître

1 - une faible élévation du fondamental (120 Hz à 130 Hz).

2 - un rehaussement du niveau relatif des fréquences aiguës (1500-2500 Hz) qui n'est sensible que pour le u, par contre pour le a et le e on observe une diminution de ce niveau relatif.

III. HYPOTHESE D'INTERPRETATION DES RESULTATS

Les voyelles prononcées étant stationnaires, nous avons supposé qu'une description spectrale des phénomènes pouvait servir de base à l'étude des traitements effectués par l'oreille. Nous avons donc calculé le spectre d'amplitude de chacune des 60 voyelles (en dB par rapport au maximum).

Sur ces spectres, nous avons ensuite placé le niveau du bruit multiplicatif rajouté par le "MALT" lors de la détection.

Ce niveau est obtenu en appliquant la formule n°(1) caractéristique du "MALT".

NB(X) = niveau de densité spectrale du bruit à la détection par rapport au maximum du spectre pour la voyelle x.

E (X) = énergie du signal par rapport au niveau maximum du spectre d'amplitude de la voyelle x.

$(S/B)_X$ = valeur du rapport signal sur bruit à la détection pour la voyelle x

$$(1)\ \forall X \in [1,60] \quad \boxed{NB(X) = E(X) - (S/B)_X}$$

Pour chaque voyelle on entreprend une étude du spectre en recherchant la zone de fréquence et le niveau de bruit détecté en appliquant l'algorithme suivant :

1 - rechercher le premier maximum du spectre

2 - calculer son effet de masque sur les fréquences supérieures en lui appliquant les pentes de masquage ci-dessous (TAB 1).

3 - à chaque harmonique du spectre, tester si le niveau de cette harmonique dépasse de 10 dB le niveau masquant calculé.

4 - si non, on passe à l'harmonique suivant et ainsi de suite jusqu'à 3400 Hz on retient alors le niveau masquant obtenu (Nm).

5 - si oui, on conserve le niveau de l'harmonique moins 10 dB (Nh) et on reprend l'algorithme au point 2 en utilisant le nouveau maximum du spectre. Lorsque l'on est arrivé à 3400 Hz on retient pour ND(x) la valeur la plus faible (en tenant compte du signe) des Nm et Nh.

Les valeurs des pentes de masquage approximent celles utilisées par ZWICKER (1967) dans son modèle de sonie. Le niveau de référence, auquel nous avons attaché la valeur 0 dB, correspondant aux maxima des spectres de chaque voyelle c'est-à-dire à des pressions acoustiques de l'ordre de 70 à 80 dB.

En effet le maximum de chaque spectre se trouve en général à -5, -6 dB du niveau d'énergie de la voyelle (énergie globale entre 75 et 85 dB).

Le masquage est schématisé par une pente dont les caractéristiques sont les suivantes :

Bande 1250-3400 Hz	niveau général	-20 à -30 dB	pente -35 dB/octave
	niveau fort	- 5 à -15 dB	pente -30 dB/octave
Bande 750-1250 Hz	niveau général	-10 à -20 dB	pente -30 dB/octave
	niveau fort	0 à -10 dB	pente -28 dB/octave
Bande 500- 750 Hz	niveau général	-75 à -15 dB	pente -25 dB/octave
	niveau fort	0 à 7,5	pente -20 dB/octave
Bande 0- 500 Hz	niveau général	- 5 à -15 dB	pente -20 dB/octave
	niveau fort	0 à - 5 dB	pente -18 dB/octave

Quant à la règle choisie d'un niveau de bruit perceptible à -10 dB en dessous du niveau des harmoniques dans la zone de détection, on peut la comparer aux résultats expérimentaux obtenus par ZWICKER (1954) pour la détection d'un bruit entre deux fréquences pures lorsque ces dernières sont séparées de moins d'une bande critique, ce qui est toujours le cas pour les voyelles étudiées.

Il trouve en effet que le bruit peut être détecté dès que son niveau dépasse le niveau des harmoniques diminué de 15 dB. La différence entre les deux valeurs est peut-être due à la complexité du signal des voyelles par rapport au signal composé uniquement de deux fréquences.

Les valeurs de ND(x) et NB(x) obtenues pour chaque voyelle et chaque locuteur sont données dans les tableaux n° 9, 10, 11, 12, 13 ainsi que le coefficient de corrélation entre les deux ensembles de valeurs pour chaque voyelle. Une vision globale des résultats est donnée à la figure n° 1.

La meilleure courbe d'accord pour rendre compte de ce nuage de point est une relation linéaire (NB(x) = 0,96 * ND(x) - 0,85) l'indice de détermination de la relation étant de 91 %.

On voit que la corrélation obtenue est assez bonne et on peut donc penser que les phénomènes psychoacoustiques importants pour la détection du bruit ont bien été pris en compte. On pourrait peut-être améliorer un peu la corrélation en tenant compte d'effets psychoacoustiques tels que : bande critique autour du formant, niveau exact d'émission de chaque voyelle. Mais, vu le nombre un peu restreint d'auditeurs (10), il nous a semblé inutile de rechercher une meilleure corrélation entre les valeurs calculées et les valeurs obtenues subjectivement.

IV. CONCLUSIONS ET REMARQUES

L'étude psychoacoustique précédente montre que :

. le seuil de détection du bruit de quantification des codeurs (bruit type MALT) est fortement lié à l'allure spectrale du signal de parole dans les parties voisées. Les paramètres importants, influant sur le seuil, sont l'amplitude des différents formants et leurs effets de masque sur les fréquences voisines.

. les zones de détections possibles se trouvent être les zones extérieures aux régions des formants, le plus souvent dans les hautes fréquences du fait de l'allure spectrale des voyelles.

Ces hypothèses permettent d'expliquer, par exemple, pourquoi :

. les voix de femmes sont en général plus sensibles au bruit de quantification que les voix d'hommes. En effet, en moyenne, ces dernières présentent des niveaux relatifs d'amplitude des formants d'ordre 2 et 3 plus faibles que ceux des hommes et donc des effets de masque moins importants.

. la voix forcée est moins sensible que la voix normale au bruit type MALT. En effet la voix forcée présente en général des amplitudes des formants plus fortes que la voix normale (élévation relative du niveau des hautes fréquences).

. la forme spectrale du bruit de codage a une très grande importance sur l'impression subjective ressentie. Ainsi les codeurs MICDA (CARTIER et Coll. 1977) ; dont le bruit de quantification est proche d'un bruit rose paraissent moins "bruités" que les codeurs ayant un bruit de quantification type MALT pour un même rapport S/B global.

On peut tirer parti, lors de la conception des codeurs, du masquage du bruit de quantification par les formants en donnant une forme spectrale à ce bruit qui ressemble le plus possible au signal de parole de façon à maximiser l'effet du masque de la parole. Ceci a d'ailleurs déjà été utilisé par ATAL et SCHROEDER (1979 a, b).

Enfin, l'étude précédente montre les limites de validité de la méthode d'isopréférence. Il faut en effet s'assurer que le bruit introduit par le codeur à tester est fort semblable au bruit injecté par le MALT. En effet la procédure de test semble forcer les sujets à baser leur jugement principalement sur l'intensité relative des bruits détectés sur les phrases émises. Dès que le bruit de codage s'écarte d'un tel bruit, on trouvera des effets phrases et des effets locuteurs importants. Le corpus utilisé influera alors sur les valeurs d'isopréférence et une valeur moyenne unique de signal sur bruit n'aura que peu de significations si l'on ne s'assure pas de la validité du corpus.

BIBLIOGRAPHIE

ATAL B, SCHOEDER M, (1979 a) Voir IEEE Transactions on Acoustics, Speech and Signal Processing Vol. ASSP 27, n° 3, juin 1979 p. 247-254. "Predictive Coding of speech signals and subjective error criteria".

ATAL B, SCHROEDER M, (1979 b) IEEE conférence p. 453-455 Juillet 1979. "Optimizing predictive coders for minimum audible noise".

M. CARTIER, P. GRAILLOT, A. PISSARD (1977). Qualité de la parole codée : Evaluation objective, évaluation subjective Vol. 4 Recherches Acoustique p. 55-57.

H. LEWITT (1971) Transformed up-down methods in Psychoacoustics J.A.S.A Vol. 49, n° 2 (Part 2) 467-477.

W.A MUNSON, J.E. KARLIN (1962) Isopreference method for evaluating Speech Transmission circuits. J.A.S.A Vol 34, p. 762-774.

WETHERILL, H. LEWITT (1965) "Sequential Estimation of Points on a Psychometric Function BRIT. J. MATH. STATIST. Psychol. 18, 1-10.

ZWICKER E. (1954) Die Verdeckung Von Schmalbandgeräschen durch sinus töne Acustica 4, 415-420.

ZWICKER E. and SCHARF B (1965) A model of loudness summation Psych. Rev. 72, 3-26.

ZWICKER E, FELDJKELLER R, (1967) Das Ohr als Nacrichten Empfänger STUTTGART S HIRZELVERLAG.

ZWICKER E, (1956) Die Elementaren Grundlagen zur Bestimung der Informations kapazität des Gehörs Acustica 6, 325-381.

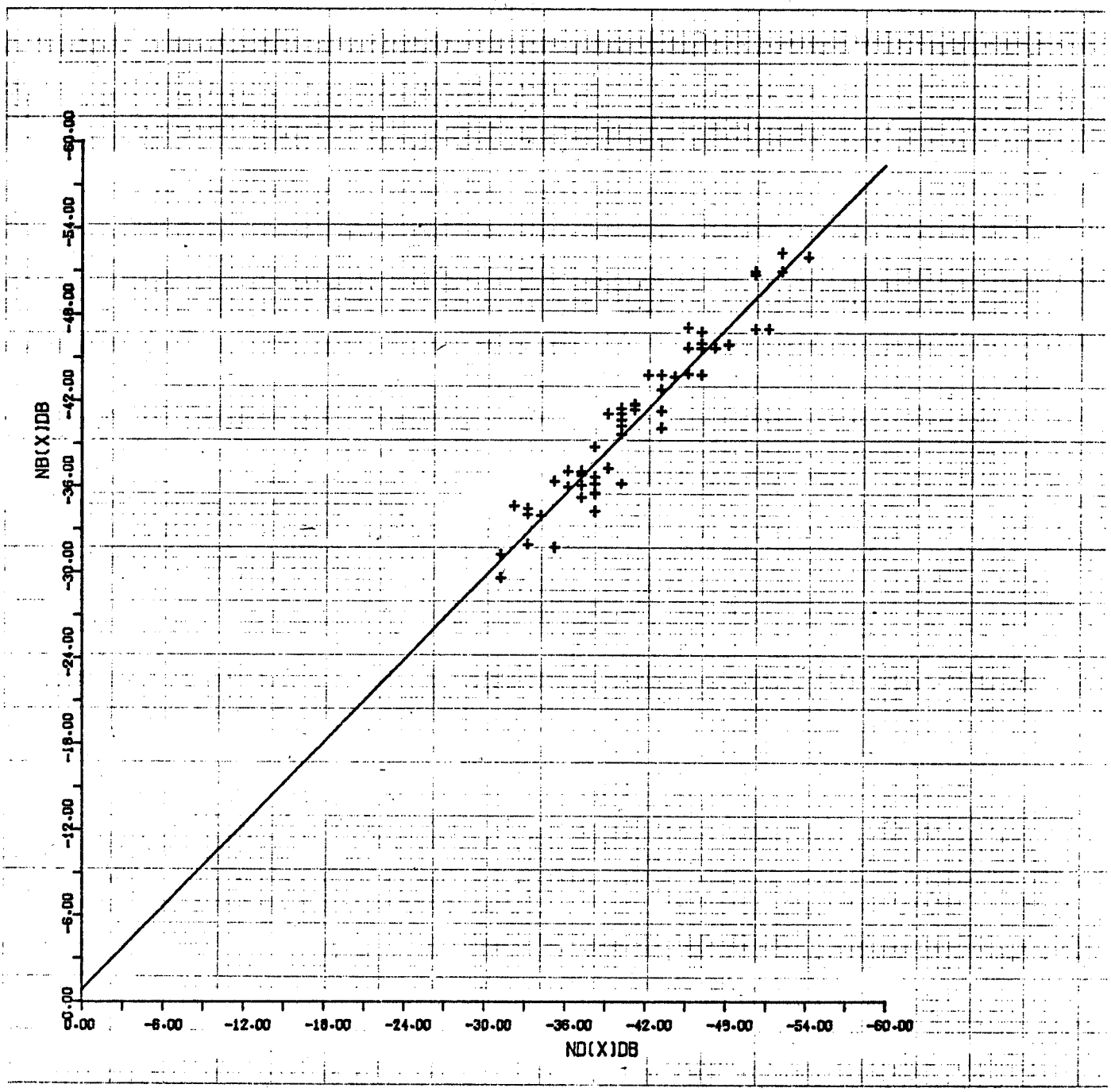


Figure 1 Relation entre NB(X) et ND(X)

Locuteur	$\bar{x}$ (dB)	$S_{\bar{x}}$ (dB)
CS	44	1,2
MA	42	1,2
RB	44	2,2
JYM	42	1,7
JL	44	1,4
JD	39	1,6
FE(N)	45	1,6
FE(F)	42	1,5
DP	43	1,4
MF(N)	40	1,4
MF(F)	43	2,0
DL	40	1,2

Tableau n°1 voyelle"a"

Locuteur	$\bar{x}$ (dB)	$S_{\bar{x}}$ (dB)
CS	42	1,7
MA	51	1,3
RB	41	0,9
JYM	45	1,3
JL	46	1,2
JD	37	1,9
FE(N)	50	2,1
FE(F)	40	1,3
DP	43	1,5
MF(N)	38	1,9
MF(F)	44	2,4
DL	41	1,4

Tableau n°2 voyelle "e"

Locuteur	$\bar{x}$ dB	$S_{\bar{x}}$ dB
CS	49	0,9
MA	44	1,2
RB	43	1,5
JYM	43	1,4
JL	40	1,3
JD	40	0,9
FE(N)	49	1,0
FE(F)	50	1,6
DP	44	1,8
MF(N)	40	0,7
MF(F)	38	0,6
DL	44	1,9

Tableau n°3 voyelle "i"

Locuteur	$\bar{x}$ dB	$S_{\bar{x}}$ dB
CS	54	1,7
MA	54	1,5
RB	50	1,0
JYM	51	1,0
JL	56	1,3
JD	49	0,9
FE(N)	58	0,4
FE(F)	50	1,2
DP	55	1,3
MF(N)	51	1,4
MF(F)	49	1,3
DL	49	1,2

Tableau n°4 voyelle "o"

Locuteur	$\bar{x}$ dB	$S_{\bar{x}}$ dB
CS	54	1,5
MA	46	1,9
RB	45	1,4
JYM	45	1,6
JL	45	1,3
JD	40	1,1
FE(N)	49	0,9
FE(F)	47	1,2
DP	40	1,4
MF(N)	41	1,4
MF(F)	40	1,8
DL	40	1,7

Tableau n°5

Voyelle "u"

$\bar{x}$: moyenne de seuils.

$S_{\bar{x}}$: ec.typ. de $\bar{x}$

Voyelle	$\bar{x}$ H dB	$\bar{x}$ F dB	$S_{\bar{x}}$ H dB	$S_{\bar{x}}$ F dB	t	Ho $\bar{x}H \neq \bar{x}F$
a (a)	42,7	42,1	0,6	0,7	0,7	pas significatif 55%
e (e)	41,8	44,7	0,8	0,8	2,5	significatif 98%
i (i)	40,9	47,1	0,5	0,7	7,5	significatif 99%
o (o)	51,1	53,5	0,5	0,6	2,7	significatif 99%
u (y)	42,7	45,9	0,6	0,9	3,0	significatif 99%

Tableau n°6 Comparaison Homme - Femme

$\bar{x}$ H : moyenne desseuils hommes. $S\bar{x}H$ écart type de $\bar{x}H$

$\bar{x}$ F : moyenne des seuils femmes. $S\bar{x}$ écart type de $\bar{x}F$

t : Signification de la différence entre $\bar{x}_H$ et $\bar{x}_F$ (degré de liberté 59)

Voyelles	$\bar{x}$ dB	$S\bar{x}_{dB}$	$\bar{x}1, \bar{x}2$	t	% de signification de $x1 \neq x2$
a	42,4	0,46	a/e	1,12	Significatif à 75 %
e	43,3	0,59	e/i	0,90	Significatif à 68 %
i	44,0	0,71	i/a	2,25	Significatif à 95 %
u	44,3	0,55	i/u	0,39	Pas significatif 30 %
o	52,3	0,43	u/o	11,44	Significatif à 99 %

Tableau n°7 Comparaison entre voyelles.

$\bar{x}$: moyenne des seuils sur tous les locuteurs $S_{\bar{x}}$ écart type de $\bar{x}$

t : signification des différences entre moyenne $\bar{x}1$, $\bar{x}2$ (degré de liberté 119)

Voyelles	$\bar{x}7_{dB}$	$\bar{x}8_{dB}$	t	%	$\bar{x}10_{dB}$	$\bar{x}11_{dB}$	t	%
a	45,3	41,7	1,63	88 %	39,8	42,9	1,31	78 %
e	50,4	40,3	4,04	99 %	38,3	43,9	1,84	90 %
i	49,3	50,2	0,5	30 %	40,1	37,9	2,25	94 %
o	57,7	50,4	5,8	99 %	50,6	49,5	0,5	30 %
u	48,6	46,9	1,1	70 %	41,3	39,9	0,62	32 %

Tableau n°8 Comparaison voix normale, voie forcée.

$\bar{x}7$, $\bar{x}8$: Seuil de détection pour la voix normale et forcée de la locutrice(F.E)

$\bar{x}10$, $\bar{x}11$: Seuil de détection pour la voix,normale et forcée du locuteur (M.F)

t : Signification des différences $\bar{x}7 \neq \bar{x}8$ ou $\bar{x}10 \neq \bar{x}11$.

% : % de signification de l'hypothèse $\bar{x}7 \neq \bar{x}8$ ou de l'hypothèse $\bar{x}10 \neq \bar{x}11$.

Locuteur	NB(X)	ND(X)
CS	-36,2	-36
MA	-35,4	-39
RB	-38,7	-44
JYM	-35,8	-40
JL	-36,9	-39
JD	-31,6	-33
FE(N)	-38,4	-38
FE(F)	-35,7	-37
DP	-36,7	-37
MF(N)	-31,4	-35
MF(F)	-35,2	-38
DL	-36,4	-37

Tableau n°9 voyelle "a"

corrélation 75 %.

Locuteur	NB(X)	NB(X)
CS	-34,3	-??
MA	-47,6	-50
RB	-35,1	-38
JYM	-40,9	-43
JL	-40,9	-43
JD	-29,3	-31
FE(N)	-45,3	-46
FE(F)	-35,6	-36
DP	-39,3	-40
MF(N)	-30,9	-31
MF(F)	-39,7	-43
DL	-36,7	-36

Tableau n°10 voyelle "é"

corrélation 87 %.

Locuteurs	NB(x)	ND(x)
CS	-46,4	-46
MA	-41,0	-41
RB	-40,3	-40
JYM	-39,9	-40
JL	-36,3	-38
JD	-33,9	-38
FE(N)	-45,5	-48
FE(F)	-46,7	-45
DP	-41,1	-40
MF(N)	-35,8	-38
MF(F)	-33,7	-33
DL	-40,7	-39

Tableau n°11 voyelle "i"

corrélation 82 %

Locuteurs	NB(x)	ND(x)
CS	-50,6	-52
MA	-50,4	-50
RB	-43,5	-45
JYM	-46,6	-51
JL	-51,6	-54
JD	-43,4	-43
FE(N)	-52,7	-52
FE(F)	-45,3	-45
DP	-50,6	-50
MF(N)	-43,4	-46
MF(F)	-45,3	-47
DL	-43,4	-42

Tableau n°12 voyelle "o"

corrélation 80 %

Locuteur	CS	MA	RB	JYM	JL	JD	FE(N)	FE(F)	DP	MF(N)	MF(F)	DL
NB(x)	-51,9	-42,4	-41,4	-40,7	-41,3	-33,6	-45,6	-42,3	-36	-34,9	-34,1	36,5
ND(x)	-52	-43	-41	-40	-41	-34	-46	-44	-35	-37	-33	-37

Tableau n°13 Voyelle "u"

Corrélation 90 %

Xlèmes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

COMPARAISON DE PROCEDURES D'EVALUATION DE LA QUALITE DE LA PAROLE CODEE.

Patrick GRAILLOT CNET - LANNION

RESUME

Les mesures objectives (principalement rapport signal sur bruit) ne sont pas suffisantes pour exprimer la qualité d'un système de codage de la parole ; il est nécessaire de faire intervenir le jugement humain.

Cette note décrit différentes procédures d'essais d'écoute : isopréférence, jugement par catégories, préférence directe par paires, netteté aux logatomes. Chaque procédure permet de définir un indice de qualité.

On donne ensuite les résultats obtenus pour diverses configurations d'une douzaine de codeurs situés dans la gamme 2,4-32 kbit/s

COMPARAISON BETWEEN DIFFERENT PROCESSES USED FOR THE EVALUATION OF THE QUALITY OF CODED SPEECH.

SUMMARY

Objective measurements (mainly signal-to-noise ratio) are not sufficient to evaluate the quality of a certain class of speech coding systems. It is necessary to be evaluated by subjective tests.

This article describes different evaluation experiments : isopreference test, category judgment test, relative preference test and an articulation test with non sense syllabes.

Each process provides a quality score.

Then, results are given for various coders with different configurations and informations rate between 2,4 and 32 kbit/s.

INTRODUCTION

La qualité d'un système de codage de la parole peut être caractérisée soit par des mesures objectives (principalement rapport signal sur bruit), soit par des méthodes subjectives (essais d'écoute). Ces dernières sont nécessaires pour deux raisons :

- la qualité d'un système destiné à transmettre de la parole ne peut pas être évaluée de façon sérieuse si l'on ne fait pas intervenir le jugement humain.

- si le rapport signal à bruit est bien corrélé avec la qualité pour un codeur de bonne qualité (par exemple M.I.C), il n'en est pas de même pour les codeurs de débit inférieur à 32 kbit/s.

La qualité subjective d'un signal de parole (RISSET, 1971 ; CARTIER, 1979) s'exprime suivant plusieurs critères : intelligibilité, agrément et fidélité à la voix du locuteur. A chacun de ces critères on fait correspondre un indice ; le problème est de déterminer à quel point chacun de ces indices concourt à la qualité globale.

Notre but a été de déterminer, au moyen de différentes procédures, des indices d'intelligibilité et d'agrément pour des systèmes de codage de la parole (originaires ou non du C.N.E.T) dont le débit s'étale entre 2,4 et 32 kbit/s.

I - LES PROCEDURES UTILISEES POUR LES ESSAIS DE QUALITE

A - Intelligilité

Pour estimer l'intelligibilité nous avons utilisé (PERSON, 1973) les essais aux logatomes (mots sans signification, de structure consonne-voyelle-consonne) effectués par une équipe permanente d'opérateurs de téléphonométrie.
Le résultat de chaque essai provient de l'écoute de 8 listes de 50 logatomes par 6 opérateurs et s'exprime en pourcentage de logatomes reconnus.

B - Agrément

1 - Isopréférence

L'essai d'isopréférence (I.E.E.E., 1969) consiste à faire écouter à l'auditeur une paire comprenant d'une part la référence passant dans le codeur à tester et d'autre part la référence à laquelle on ajoute du bruit (en général multiplicatif).

La valeur du rapport signal sur bruit pour laquelle l'agrément d'écoute est comparable entre les deux éléments de la paire, permet de caractériser le codeur.

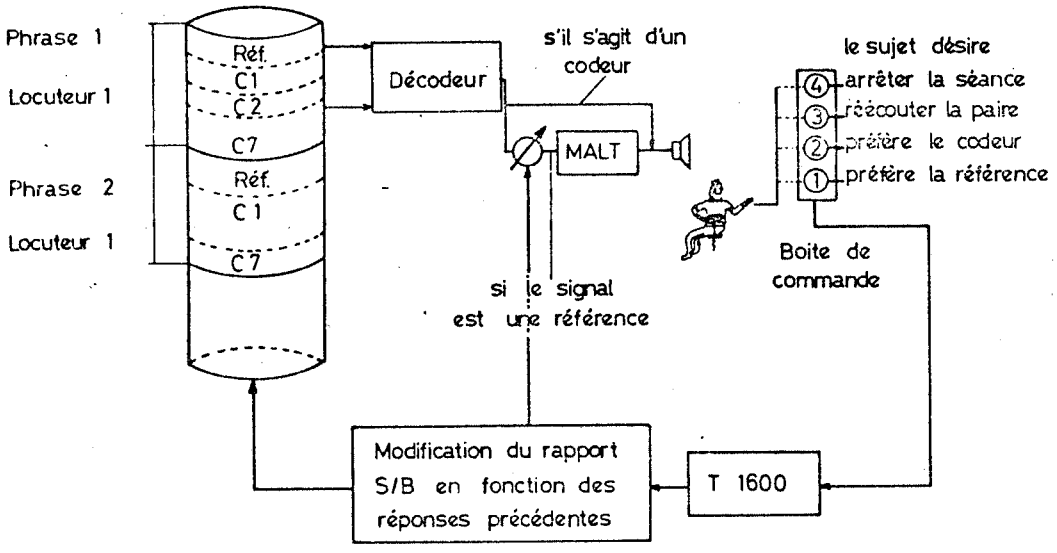
Les signaux d'essais sont constitués de plusieurs phrases de courte durée (2 à 2,5 s) prononcées par plusieurs locuteurs, hommes et femmes.

L'appareil, MNRU dans la littérature anglaise (Schroeder, 1968), qui ajoute du bruit au signal est dénommé MALT (Mesure d'Agrément des Liaisons Téléphoniques) dans sa version CNET.

Nous avons implanté sur calculateur T 1600, en stockant les signaux numérisés sur un disque magnétique, l'essai d'isopréférence suivant deux présentations possibles :

a) Par ajustement : le sujet modifie lui-même le potentiomètre du MALT afin de trouver l'équilibre d'agrément.

b) Par une "transformed up-down method" : cette procédure a nécessité la définition d'un nouvel appareil MALT. Dans cette version, il ajoute au signal la somme de trois bruits différents : additif, multiplicatif et extérieur. Il peut fonctionner en mode manuel (commandes par roues codeuses) ou recevoir ses commandes du calculateur par un mot de 16 bits qui lui transmet les validations, le type de bruit à modifier et le rapport signal sur bruit.



Pour une phrase de test, nous appellerons A la référence bruitée et B le signal provenant du codeur. Le sujet écoutera la paire présentée de manière aléatoire sous la forme ABAB ou BABA et indiquera, en appuyant sur un bouton, s'il a trouvé l'agrément d'écoute supérieur pour le premier ou le second élément de la paire (il n'existe pas de réponse "équivalente"). Un algorithme modifie alors le rapport signal sur bruit en fonction des réponses précédentes et permet

d'obtenir, pour chaque phrase du test, une valeur moyenne d'isopréférence et un intervalle d'incertitude autour de cette moyenne (zone à l'intérieur de laquelle le sujet a préféré tantôt la référence bruitée, tantôt le signal passant à travers les codeurs).

Pour rendre cette procédure plus souple, nous avons en outre deux types de fichiers (stockés sur disque magnétique) :

- un fichier descriptif du test, qui est lu en tête du programme. Il contient le type du bruit à faire varier, la plage de variation et l'ordre des phrases à présenter.

- un fichier par sujet et qui conserve ses réponses. Ceci permet au sujet d'arrêter la séance de test lorsqu'il le désire.

2 - Jugement par catégories

A étant la référence et B le signal passant à travers le codeur, le sujet écoute, pour chaque phrase-test, la séquence ABAB. Il note pour chaque paire la dégradation apportée par le codeur par rapport à la référence suivant une échelle de 5 à 1 correspondante aux adjectifs suivants : imperceptible, perceptible mais non gênant, perceptible et légèrement gênant, gênant et très gênant. Chaque codeur est caractérisé par la note moyenne obtenue.

Préalablement à cette procédure, le sujet est soumis à une séance d'entraînement.

3 - Préférence directe par paires.

Cette procédure est implantée sur calculateur. On présente au sujet des paires constituées de deux codages différents de la même référence et on lui demande quel élément de la paire lui paraît avoir le meilleur agrément d'écoute. Une réponse "équivalent" est autorisée.

On voit que cette procédure s'oppose aux deux précédentes pour lesquelles les comparaisons sont indirectes.

Pour chaque paire de codeurs Q et Q', on obtient un pourcentage de préférence de l'un par rapport à l'autre :

n réponses $\begin{cases} m \text{ "préfère Q"} \\ p \text{ "préfère Q'"} \\ q \text{ "équivalent"} \end{cases}$ $\begin{cases} \text{pourcentage de préférence de Q par rapport à Q'} = \dfrac{2m + q}{2n} \\ \text{pourcentage de préférence de Q' par rapport à Q} = \dfrac{2p + q}{2n} \end{cases}$

II - DEFINITION DES INDICES DE QUALITE

Nous avons vu que notre but était d'extraire de chaque procédure d'essais subjectifs un indice de qualité. Il est donc nécessaire que chaque indice s'exprime suivant la même unité. Nous avons choisi le décibel et nous nous sommes servis du fait que le MALT, référence dans la procédure d'isopréférence, peut être considéré comme un codeur particulier dans les procédures des logatomes et du jugement par catégorie.

- l'essai d'isopréférence caractérise un codeur par une grandeur, exprimée en décibels, que nous nommerons R_{iso}.

- nous pouvons inclure dans un jugement par catégorie, parmi les codeurs, différentes valeurs de MALT. On peut alors définir un indice de dégradation R deg (en décibels) en faisant une interpolation linéaire avec les deux valeurs de MALT dont les notes sont les plus proches de celle du codeur.

MALT x dB ⟶ note m

codeur Q ⟶ note q $\quad R\ deg\ (Q) = y + \left[(x-y)\ (q-n/m-n)\right]$

MALT y dB ⟶ note n

$n < q < m$

- nous avons effectué des essais aux logatomes avec différentes valeurs de MALT. Ceci nous permet, pour chaque codeur, de définir un indice de netteté R net en faisant une interpolation linéaire sur les pourcentages de logatomes reconnus.

III - LES DIFFERENTS CODEURS TESTES

A - Débit 10-24 Kbit/s

Nous avons testé quatre types de codage :

- un vocodeur (ZURCHER, CARTIER et FRICHOU, 1975) à excitation vocale à canaux (V.E.V.C), sous forme de maquette, aux débits de 9,6 ; 12,5 et 14,5 kbit/s. Le codage (en MIC) de la bande basse représentait les 4/5 du débit total. La bande haute (800 -3400 Hz) était analysée par un banc de filtres.

- un vocodeur (LE GUYADER, 1978) à excitation vocale à prédiction linéaire (V.E.V.P.L) : bande basse en MIC, bande haute (800 - 3 400 Hz) codée par prédiction linaire (8 coefficients, facteur de gain). Plusieurs simulations ont été testées, elles utilisaient un signal d'excitation propre (V.E.V.P.L.P) et pour l'une l'excitation d'un vocodeur à canaux (V.E.V.P.L.V).

- une maquette de codeur △adaptatif aux débit de 16, 20, 24 et 32 kbit/s. Il est à noter que ce codeur était conçu pour fonctionner à 32 kbit/s et donc non optimisé pour les débits inférieurs.

- un autre vocodeur, au débit de 16 kbit/s, sur lequel nous resterons discrets (V.E.X).

B - Codeurs à 32 Kbit/s

Notre but était d'évaluer des codeurs étudiés dans notre groupe :

- un codeur (ZURCHER, PISSARD, 1978) différentiel adaptatif par analyse spectrale (MICDAS) à 32 kbit/s. L'adaptation du prédicteur se fait en fonction d'une analyse spectrale, obtenue par un banc de filtres, du signal de référence. Un prototype de ce codeur est réalisé avec un quantificateur linéaire.

- des codeurs différentiels classiques dont l'adaptation du gain du quantificateur se fait à partir d'une estimation de la variance à l'entrée de ce quantificateur (GOODMAN, WILKINSON, 1975). Le prédicteur est soit fixe (codeur F), soit adaptatif (codeur G = méthode du gradient, codeur T = méthode du filtre en treillis). Dans le codeur T, les coefficients de corrélation partielle sont calculés séquentiellement à partir du signal reconstitué (algorithme d'Itakura et Saïto), la synthèse et la prédiction se faisant à partir d'un filtre à treillis à deux multiplieurs.

C - Vocodeur à canaux (V.C.) : Référence "immortelle" du CNET :

14 canaux, codés sur 4 bits, dans la bande téléphonique ; débit de 4800 bits/s (FERRIEU, PERSON, 1968).

IV - RESULTATS DES ESSAIS SUBJECTIFS

A - Procédure d'isopréférence par ajustement

Les premiers essais subjectifs "informatisés" de qualité ont utilisé la procédure d'isopréférence par ajustement. Le but (CARTIER, GRAILLOT, PISSARD, 1977) était de comparer les codeurs à débit moyen (autour de 16 kbit/s.). On ne se trouvait donc pas dans la zone optimale pour l'utilisation de l'essai d'isopréférence. En effet, celui-ci présente deux inconvénients :

- on compare le signal codé à la référence à laquelle on ajoute du bruit multiplicatif. Or, si le bruit d'un codeur comme le M.I.C. est très proche d'un bruit multiplicatif , au fur et à mesure que le débit diminue, le bruit du codeur devient de plus en plus différent d'un bruit multiplicatif. Donc, pour l'auditeur, la notion "agrément d'écoute" va alors répondre à des critères personnels. Ce qui explique que l'écart-type des réponses augmente au fur et à mesure que le débit diminue.

- inversement, pour les codeurs de très bonne qualité, on se heurte au problème de la différence du seuil de détection de la dégradation, apportée par le MALT, suivant les phrases et les locuteurs (COMBESCURE, BOYER, 1979). La conséquence est qu'au fur et à mesure que le débit augmente, le corpus utilisé (phrases et locuteurs) doit augmenter.

B - Procédure d'isopréférence avec modification automatique du rapport S/B

1 - Codeurs à débit moyen

Nous avons testé 5 codeurs : V.E.V.C 12,5 kbit/s ; △ à 16kbit/s ; V.E.V.P.L.P 16 kbit/s V.E.V.P.L.V 16 kbit/s et V.E.X 16 kbit/s. Les résultats sont représentés sur la Fig. 1.

Le corpus utilisé était constitué de 4 phrases (voir jugement par catégories) prononcées par 4 locuteurs (2 hommes et 2 femmes).

		Moyenne, voix d'hommes	Moyenne, voix de femmes	Moyenne générale
V.E.V.C	12,5 kbit/s	16,2 dB	14,4 dB	15,3 dB
△	16 kbit/s	19,8 dB	19 dB	19,4 dB
V.E.V.P.L.P	16 kbit/s	19,2 dB	22,2 dB	20,7 dB
V.E.V.P.L.V	16 kbit/s	18,7 dB	21,3 dB	20 dB
V.E.X	16 kbit/s	20,6 dB	19,6 dB	20,1 dB

Fig. 1 : Résultats de l'essais d'isopréférence.

Il est intéressant de noter que l'essai d'isopréférence détecte une différence significative d'agrément entre les voix de femmes et d'hommes (LE GUYADER, 1978 ; pages 180-181).

Sur le même corpus et avec les mêmes auditeurs, on a réalisé un jugement direct par paires (fig.2).

POURCENTAGE DE PREFERENCE DE :

PAR RAPPORT A	V.E.V.C.	△	V.E.V.P.L.V.	V.E.V.P.L.P.	V.E.X
V.E.V.C		67,9 %	82,1 %	92,9 %	96,4 %
△	32,1 %		75 %	75 %	89,3 %
V.E.V.P.L.V.	17,9 %	25 %		64,3 %	78,6 %
V.E.V.P.L.P.	7,1 %	23 %	35,7 %		32,1 %
V.E.X	3,6 %	10,7 %	21,4 %	67,9 %	

Fig.2 : Résultat du jugement direct par paires (par exemple : pourcentage de préférence du codeur sur le V.E.V.C = 67,9 %.

Notons les points suivants :

- Cette méthode est surtout intéressante pour les codeurs proches. Ainsi, on n'avait une différence de Riso entre V.E.V.P.L.P et V.E.V.P.L.V que de 0,7 dB alors que le pourcentage de préférence de V.E.V.P.L.P par rapport à V.E.V.P.L.V est de 64,3 %. En détaillant les résultats on voit que cette différence est causée par la supériorité de l'excitation propre pour les voix de femmes (déjà notée à l'isopréférence).

- La différence entre comparaison directe et indirecte. Ainsi, % de préférence de V.E.V.P.L.P. à V.E.X = 67,9 %.

Mais % préf. (V.E.V.P.L.P/Q) < % préf. (V.E.X/Q) pour les trois autres codeurs Q.

Il est simple de se définir une relation d'ordre entre les codeurs à partir de la matrice de préférence. On choisira, lors de l'établissement de la formule mathématique, de privilégier les comparaisons directes ou indirectes.

2 - Codeur à 32 kbit/s.

Nous donnons les résultats de cet essai d'isopréférence, dont le corpus était constitué de 2 phrases prononcées par trois locuteurs (2 femmes F1,F2 et un homme H). Ce tableau (Fig.3) montre l'intérêt de l'essai d'isopréférence, procédure longue à réaliser et difficile pour les sujets, qui est de fournir des informations "fines" par les différences obtenues entre les phrases et les locuteurs et de permettre ainsi de préciser les caractéristiques des codeurs (LE GUYADER, PISSARD ; 1978 et 1979).

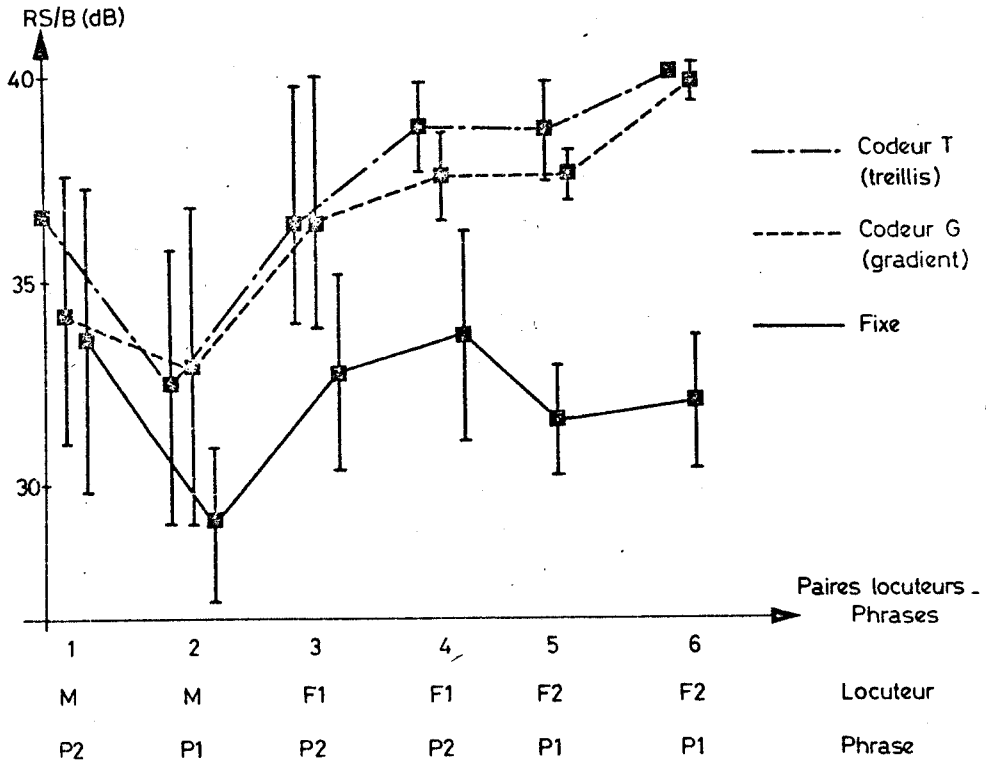


Fig.3 : Résultat de l'essai d'isopréférence.

C - Jugement par catégories.

Nous avons réalisé deux essais subjectifs de jugement par catégories. Pour les deux essais les conditions d'écoute, le corpus et les sujets étaient les mêmes. Le corpus était constitué de 4 phrases prononcées chacune par un locuteur différent (2 hommes, 2 femmes). A chaque essai on testait 12 codeurs. En outre on présente la référence comme un codeur quelconque, ce qui permet de juger l'attention des sujets (la dégradation de la référence par rapport à elle-même devant être jugée imperceptible). Chaque couple référence-codeur est testé deux fois. Donc chaque sujet écoute :

$$104 \text{ paires} = \underbrace{(12+1)}_{\text{codeur}} \times \underbrace{(4)}_{\text{références}} \times \underbrace{(2)}_{\text{présentations}}$$

Le nombre d'auditeurs était de douze, la moitié ayant l'habitude de la parole codée et l'autre moitié étant "naïve".

On a testé, sur les deux essais, dix-huit codeurs représentant une large gamme de débits et de techniques (x 4 signifie en cascade 4 fois) :

1er essai : MICDAS 32 kbit/s ; MALT 30 dB ; MALT 25 dB ; Δ32 kbit/s ; MICDAS x 8 ; MALT 20 dB ; Δ32 x 4 ; V.E.V.C 12,5 kbit/s ; MALT 15 dB ; V.C. 2,4 Kbit/s ; MALT 10 dB ; MALT 5 dB.

2ème essai : MICDAS ; MALT 30 dB ; MALT 25 dB ; V.E.V.P.L.P ; V.E.V ; Δ16 kbit/s ; V.E.V.P.L.V ; Δ32 x 4 ; MALT 15 dB ; Δ32 x 8 ; MICDAS x 4 ; MALT 5 dB.

	Moyenne	Ecart-type
Référence	4.975	0.02
MICDAS	4.83	0.12
MALT 30	4.68	0.29
MALT 25	4.20	0.56
△ 32	4.18	0.49
VEVPLP	3.85	0.81
VEX	3.63	0.41
△ 16	3.04	0.54
VEVPLV	3.02	0.97
MALT 20	2.91	0.35
△32 x 4	2.47	0.46
VEVC	2.46	0.56
MALT 15	2.30	0.52
△32 x 8	2.17	0.56
MICDAS x 4	2.00	0.40
V.C.	1.40	0.30
MALT 10	1.40	0.32
MALT 5	1.06	0.05
MICDAS x 8	1.05	0.04

Fig. 3 : Résultats du jugement par catégories

V - COMPARAISON DES DIFFERENTS INDICES.

Des essais de netteté aux logatomes ont été effectués sur certains codeurs et sur un grand nombre de configurations de MALT. Les essais aux logatomes présentent le défaut de n'être significatifs que pour les codeurs de qualité moyenne ; au-delà on atteint rapidement une asymptote. Donnons les valeurs des trois indices pour quatre codeurs :

	V.E.V.C. 12,5 Kbit/s	16 Kbit/s	VEVPLP	32 Kbit/s
R net (logatomes)	10 dB	10 dB	14 dB	référence
R iso (isopréférence)	15,3 dB	19,2 dB	20,7 dB	24 dB
R deg (jug. par cat.)	16,2 dB	19,5 dB	23,8 dB	25 dB

La différence entre R net et R deg ou Riso est particulièrement significative et ne peut être expliquée par la variance des résultats.

Signalons trois points :

- La question posée aux sujets lors d'un jugement par catégorie est "trouvez-vous la dégradation..." ; la comparaison de ces indices montre qu'une dégradation est jugée en perte d'agrément d'écoute et non en perte d'intelligibilité.

- Aucune des trois procédures ne tient compte de la fidélité au timbre du locuteur. Le MALT, qui dégrade moins le timbre que les codeurs à débit moyen, devrait être "avantagé" dans une comparaison indirecte comme le jugement par catégories (Référence - MALT et Référence-Codeur) par rapport à une comparaison directe comme l'isopréférence (MALT - Codeur). Or R deg est supérieur à R iso !!!

- Caractériser un codeur à débit moyen par ces trois indices revient à ne le juger que par rapport à un bruit multiplicatif. Mais il faut convenir qu'il est difficile de prendre un autre type de bruit comme référence.

Signalons que pour les codeurs au débit de 32 Kbit/s nous trouvons des résultats proches de ceux décrits dans un article récent du BSTJ (Daumer, Cavanaugh, 1978) où un certain nombre de codeurs sont introduits, seuls ou en cascade, dans un grand nombre de communications téléphoniques dans diverses conditions.

VI - CONCLUSION

Ces différents essais subjectifs ont permis de définir les conditions d'utilisation et les avantages des différentes procédures. Le principal défaut des expériences décrites ici vient du fait que les corpus utilisés (phrases et locuteurs) sont trop restreints. Des contraintes pratiques en sont la cause : durée du test pour les auditeurs (nécessité d'avoir des auditeurs rémunérés) et configuration actuelle du calculateur utilisé (capacité du disque magnétique).

Le jugement par catégories présente deux avantages :

- Facilité d'utilisation.

- Cette procédure permet de noter un codeur quelle que soit sa place sur l'échelle de qualité. Alors que pour les essais aux logatomes et d'isopréférence si le codeur est de bonne qualité, l'indice d'isopréférence est très fiable (bruit du codeur proche d'un bruit multiplicatif) mais on n'a pas de résultats exploitables aux logatomes. Dans ce cas, il faudrait alors utiliser des procédures d'intelligibilité plus "difficiles" (NAKATANI, 1973). Par contre, si le codeur est de qualité moyenne, l'indice d'intelligibilité aux logatomes est fiable mais l'indice d'isopréférence ne l'est plus.

Discutons enfin de la variance des résultats, c'est-à-dire de la reproductibilité de tels essais. Signalons d'abord que les variances présentées sont inter-auditeurs ; ce qui donne une idée très "pessimiste" de la dispersion (en particulier pour le jugement par catégories pour lequel il faudrait intégrer une variance intra-auditeur). Nous n'avons présenté ici qu'une partie des essais réalisés lors de ces deux dernières années au C.N.E.T. La conclusion est que des essais différés dans le temps , effectués avec la même procédure, les mêmes conditions d'écoute, les mêmes auditeurs (sinon, on se recentrera au moyen de codeurs déjà testés) et sur le même corpus , donnent des résultats très proches. La reproductibilité est plus difficile si l'on modifie l'un de ses paramètres (pour l'influence du choix du corpus ; voir COMBESCURE, BOYER, 1979). L'importance réciproque de ces paramètres est difficile à évaluer exactement si l'on ne recourt pas à des essais systématiques.

VII - BIBLIOGRAPHIE.

CARTIER (M.), Janv. 1979, Le codage de la parole. L'écho des recherches.

CARTIER(M.), GRAILLOT (P.), PISSARD (A.), 1977, Qualité de la parole codée : évaluation objective et évaluation subjective. Recherches acoustiques. Vol. IV

COMBESCURE (P.), BOYER (M.), 1979, Les dégradation du signal de parole par l'addition d'un bruit de type MALT et la méthode d'isopréférence pour l'évaluation de la qualité de la parole codée. NT/DAS/ETA/63,

DAUMER (W.R.), CAVAUNAGH (J.R.), 1978, A subjective comparison of selected digital codecs for speech. BSTJ, Vol. 57, n°9, pp. 3119-3165.

FERRIEU (G.), PERSON (J.M.), 1968, Etude d'un vocodeur. Etude CRL 2013/ETA.

I.E.E.E., 1969, Recommended practice for speech quality measurements". I.E.E.E. Trans. on Audio Electroacoustic AU 17 n°3.

LE GUYADER (A.), 1978, Etude d'un vocodeur à excitation vocale et à base de prédiction linéaire. Thèse Université Rennes.

LE GUYADER (A.), PISSARD (A.), 1978, Codage différentiel de la parole à 32 kbit/s. Recherches Acoustiques. Vol. V.

LE GUYADER (A.), PISSARD(A.), 1979, Codage différentiel adaptatif de la parole pour le réseau téléphonique, 7ème colloque sur le traitement du signal et ses applications, Nice .

NAKATANI(L.H.), 1973, Functional evaluation of speech quality Symposium F.A.S.E. Liège.

PERSON (J.M.), 1973, Essais de netteté pratiqués par l'Administration Française des P.T.T. Symposium F.A.S.E. Liège.

RISSET (J.C)., 1971, Méthodes d'évaluation de l'intelligibilité et de la qualité de la parole. 2ème J.E.P. Aix.

SCHROEDER (M.R.), 1968, Reference signal for signal quality studies. J.A.S.A, Vol. 44, n°6, p. 1736.

ZURCHER (F.), CARTIER (M.), FRICHOU (M.), 1975, Vocodeur à bande de base. Recherches Acoustiques. Vol. II.

ZURCHER (F.), PISSARD (A.), Procédé et circuit de transmission de type MIC différentiel à prédiction adaptative utilisant un filtrage par sous-bandes et une analyse spectrale". Brevet France n°78-36-709.

THEME III : Variabilité inter et intra locuteurs (aux niveaux articulatoire et acoustique) :

a) observation et analyse,

b) adaptation des sytèmes de reconnaissance automatique aux locuteurs,

c) vérification et identification du locuteur.

THEME III :

a) Observation et analyse

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

ETUDE COMPARATIVE DES TRAJECTOIRES DU F2 DANS LA PAROLE DES DEFICIENTS AUDITIFS ET DANS CELLE DES ENTENDANTS

S. BARTH, R. CHULLIAT

INSTITUT NATIONAL DE JEUNES SOURDS DE CHAMBERY
B.P. 15 73 160 COGNIN

RESUME : Nous avons tenté, à partir de l'analyse des trajectoires du second formant (F2) extraites de phrases énoncées par des jeunes entendants et des enfants sourds, de cerner l'impact d'une surdité sur certains paramètres acoustiques (durée globale, durée des segments stables et des transitions, vitesse de transition, fréquences des F2 des voyelles...)

SUMMARY : From the analysis of the second formant trajectories in two sentences pronounced by normal hearing children and deaf ones, we have tried to show the effects of deafness on some acoustical parameters : overall duration, duration of stable segments and of transitions, transitions rates, second formant frequencies in vowels.

I INTRODUCTION

On trouve souvent, dans la littérature, des résultats de mesures faites sur de la parole de sourd. Ceux-ci concernent généralement des groupements consonne-voyelle et ne donnent guère de renseignements sur ce que peut être la parole continue de nos élèves. Il nous a donc semblé intéressant de faire quelques expérimentations dans ce domaine et plus particulièrement de chercher certains paramètres susceptibles d'objectiver les difficultés articulatoires de ceux-ci.

A cet effet, nous avons donc choisi d'analyser les trajectoires du deuxième formant (F2) extraites de phrases simples.

II EXPERIMENTATION

II 1 Choix des groupes de sujets

On admet, en général, que la démutisation de nos élèves est achevée vers huit ou dix ans et que le pédagogue intervient alors pour des exercices d'entretien de la parole. Nous avons donc choisi nos sujets comme suit :

- 10 entendants (ENT)
- 10 déficients auditifs sévères (DAS)
- 10 déficients auditifs profonds (DAP)

Le degré de surdité des déficients auditifs a été déterminé suivant les normes du Bureau International d'Audio-Phonologie (B.I.A.P) à partir de la moyenne des pertes sur les fréquences conversationnelles. A l'intérieur de chaque groupe de sourds, les audiogrammes présentaient peu de différences.

Les locuteurs sont du sexe masculin.

II 2 Choix du matériel phonétique

Nous avons choisi deux phrases facilement assimilables par nos élèves, de manière à éviter la lecture de celles ci au cours des enregistrements :

- "le vase est rouge"
- "ce petit bébé va tomber"

Elles ont, de plus, l'avantage de posséder des voyelles bien opposées (i, a, u) et des voyelles d'articulation plus délicate ($\varepsilon, ə, \tilde{ɔ}$).

II 3 Enregistrement

Celui-ci a été pratiqué dans la chambre sourde de notre laboratoire avec un magnétophone REVOX A77 équipé d'un micro SENNHEISER MD 441 et d'une bande SONY HL.

Chaque séance commençait par un rapide entretien, entre l'enfant et l'expérimentateur, suivi de l'énoncé des deux phrases apprises. Nous avons décidé de ne pas pratiquer parallèlement de correction afin de conserver un certain naturel à l'exercice.

II 4 Analyse des enregistrements

Les bandes obtenues ont été analysées par l'intermédiaire d'un programme développé en T.S.L. (TIME SERIES LANGUAGE) permettant de donner sur un sonagramme numérique(quantification temporelle 10 ms, fenêtre d'analyse 300 Hz, quantification fréquentielle 100 Hz) les traces des maxima spectraux.

Celles-ci ont parfois été corrigées à la main, afin d'obtenir des trajectoires de F2 possédant une certaine continuité.

D'autre part, ne voulant comparer que des choses comparables, nous avons été obligés de ne retenir qu'une partie du corpus, à savoir : ləva et ʃəpətibebe qui présentait, dans l'ensemble, un certain degré de correction (séquences intelligibles mais pas toujours bien timbrées). Ces portions de phrases ont été sélectionnées après avoir fait entendre les enregistrements à trois professeurs de notre établissement qui ont noté phonétiquement ce qui avait été émis en détectant les erreurs d'articulation.

III RESULTATS

III 1 Valeurs du F2 des voyelles sur leurs parties stables

Ces mesures ont été faites à des instants qui correspondaient aux maxima d'une fonction de stabilité spectrale (au sens de J.S. LIENARD) calculée sur une fenêtre de 40 ms.

Le tableau ci-après donne les moyennes en Hz des valeurs du F2 des voyelles retenues, ainsi que les $\bar{\sigma}$ et σ/m correspondants, pour chaque groupe de sujets définis en II 1.

	ENT	DAP	DAS
ə (lə)	m = 1840 σ = 120 σ/m = 6,52 %	1666 169 10,20 %	1690 104 6,18 %
a (va)	1540 220 14,29 %	1670 210 12,57 %	1722 147 8,56 %
ə (sə)	2040 135 6,65 %	1712 169 9,87 %	1655 206 12,45 %
ə (pə)	2155 170 7,92 %	1588 99 6,26 %	1662 254 15,32 %
i (ti)	3190 364 11,43 %	3150 237 7,5 %	3100 279 9 %
e (be)	2650 162 6,13 %	2160 310 14,37 %	2280 649 31,22 %
e (be)	2560 220 8,59 %	2050 320 15,6 %	2340 253 10,84 %

Le test de WILCOXON appliqué à nos échantillons montre que :

- il n'existe pas de différence significative entre les productions des sourds et celles des entendants pour les voyelles i et a.
- cette différence est significative pour les autres voyelles, sauf pour le deuxième e, pour lequel les échantillons ENT et DAS sont homogènes.
- mis à part le cas ci-dessus, on ne trouve pas de différence significative entre les deux groupes de sourds.

On remarquera que le F2 des voyelles des déficients auditifs est, en général, plus grave que celui des entendants.

Les variations de longueur du C.V. n'entrent probablement pas en jeu pour expliquer cette différence. Celle ci est sans doute induite par des modifications de la

fonction d'aire du C.V. dûe à des modalités de contrôle articulatoire différents.

Pour le groupe ENT, on constate une augmentation du F2 en fonction du contexte consonnantique (lə, sə, pə). Cette propriété n'apparait pas pour les groupes DAS et DAP.

III 2 Problèmes de durée

Les procédures de mesure des durées ont été les suivantes :

durée globale : intervalle de temps séparant les deux minima extrêmes (début et fin) de la fonction de stabilité spectrale.

durée d'une période stable : intervalle de temps séparant deux pointeurs qui déterminent les limites des parties en plateau de la fonction de stabilité spectrale lissée.

durée d'une transition : mesurée manuellement par inspection du sonagramme entre le début du formant et l'instant où un F2 stable est atteint.

a) durée globale (en ms) (moyenne, écart-type, σ/m) et débit (en syllabes/seconde).

	ENT	DAP	DAS
lava	= 338	1282	795
	= 37	309	227
	= 11 %	24 %	28,6 %
d =	5.92 syll s^{-1}	1.56 syll s^{-1}	2.52 syll s^{-1}
sapatibebe	= 836	3035	2171
	= 118	749	645
	= 14 %	24,7 %	29 %
d =	5.96 syll s^{-1}	1.64 syll s^{-1}	2.30 syll s^{-1}

On remarquera que les débits sont sensiblement équivalents d'une séquence à l'autre.

Le test de WILCOXON exhibe une différence significative entre les trois groupes (ENT, DAP, DAS).

En prenant comme référence les durées mesurées sur le groupe des entendants, on obtient les facteurs multi-

plicatifs moyens à appliquer pour obtenir les durées des autres groupes :

- 3.67 pour les DAP
- 2.48 pour les DAS

La durée globale d'une émission semble donc être un bon indice de classification de nos groupes en fonction de leur atteinte audiologique.

b) durées globales absolues et relatives des périodes stables (en ms et %)

	ENT	DAP	DAS
lava	m = 221 ms 65 %	1125 ms 90 %	685 ms 85 %
sapatibebe	m = 536 ms 64 %	2625 ms 86 %	1770 ms 82 %

S'il existe une différence significative entre les durées absolues des périodes stables (tenue des voyelles) des trois groupes, celle-ci n'existe qu'entre les entendants et les groupes de sourds pris globalement pour les durées relatives.

Ainsi, entre les deux groupes de sourds, il ne semble apparaître qu'une simple modification de l'échelle temporelle.

Notons que cette compression de l'échelle de temps observée entre les groupes DAS et DAP ne s'accompagne pas d'une dilatation de celle des fréquences.

En effet, au & III 1, nous avons indiqué qu'il n'existe pas de différences significatives entre les valeurs des F2 des voyelles émises par les différents groupes de sourds.

c) analyse des durées moyennes des périodes stables (en ms)

(voir tableau ci-après)

durée	ENT	DAP	DAS
l	45	186	66
ə	45	251	174
v	61	392	171
a	74	344	259
s	45	82	75
ə	34	242	210
silence avant p	70	336	229
ə	33	242	129
silence avant t	75	272	175
i	54	249	152
b (tenue)	29	392	261
e	55	256	161
b (tenue)	52	288	129
e	90	222	263

Chez les sourds, les voyelles, les consonnes tenues et les silences sont beaucoup plus longs que la normale. On remarque une différence significative entre les trois groupes, sauf pour /s/ ou DAP et DAS qui sont équivalents.

d) durées globales absolues et relatives des transitions (en ms)

Les périodes transitoires occupent les continuum temporels suivants :

	ENT	DAP	DAS
lava	117 35 %	134 10 %	110 15 %
sapatibebe	300 36 %	410 14 %	400 18 %

Pour /ləva/ il n'existe pas de différence significative entre les durées totales des périodes transitoires des trois groupes.

Par contre, en valeur relative, celle-ci est évidente entre les sourds (pris globalement) et les entendants.

Pour la deuxième séquence, la différence est nette en valeur absolue et relative.

e) <u>analyse des durées des transitions</u> (en ms)

	ENT	DAP	DAS
s ə	33	36	48
ə p	27	37	35
p ə	36	25	31
ə - silence	35	73	48
t i	40	52	57
i b	36	77	77
b e	37	39	37
e b	36	50	43
b e	33	36	43
l ə	21	34	26
ə v	56	49	49
v a	40	52	37

Mis à part quelques cas particuliers (ə → silence, ib), la différence n'est pas significative entre ces valeurs moyennes. On peut donc admettre que la durée intrinsèque des transitions ne joue pas un grand rôle dans la séparation de nos groupes.

Par contre, leurs durées rapportées à la durée totale permet de distinguer sourds et entendants sans amener de précisions sur le degré de surdité.

III 3 Fréquences de départs - Fréquences cibles - Variation transitoire du F2 (en Hz)

	ENT	DAP	DAS
lə	2040 - 1840 ΔF = 200	1900 - 1666 ΔF = 234	1988 - 1690 ΔF = 298
əv	1840 - 1190 ΔF = 650	1666 - 1200 ΔF = 466	1690 - 1270 ΔF = 420
və	1190 - 1540 ΔF = - 350	1200 - 1670 ΔF = - 420	1270 - 1722 ΔF = - 452
sə	2360 - 2040 ΔF = 320	2040 - 1712 ΔF = 328	2110 - 1655 ΔF = 455
ə - silence	2040 - 1440 ΔF = 600	1712 - 1322 ΔF = 390	1655 - 1280 ΔF = 375
pə	1540 - 2155 ΔF = - 615	1277 - 1588 ΔF = - 311	1437 - 1662 ΔF = - 225
ə - silence	2155 - 2571 ΔF = - 416	1588 - 2220 ΔF = - 632	1662 - 2175 ΔF = - 513
ti	2744 - 3190 ΔF = - 446	2390 - 3150 ΔF = - 760	2250 - 3100 ΔF = - 850
ib	3190 - 1870 ΔF = 1320	3150 - 2144 ΔF = 1006	3100 - 1710 ΔF = 1390
be	1860 - 2650 ΔF = - 790	1520 - 2160 ΔF = - 640	1680 - 2280 ΔF = - 600
eb	2650 - 1810 ΔF = 840	2160 - 1275 ΔF = 885	2280 - 1420 ΔF = 860
be	1940 - 2560 ΔF = - 620	1460 - 2050 ΔF = - 590	1590 - 2340 ΔF = - 750

On constate que les fréquences de départ et les fréquences cibles du F2 dans les transitions sont en général plus basses pour les sourds que pour les entendants.

Les plages fréquentielles utilisées sont comparables, sauf pour les plosives sourdes où apparaissent généralement des problèmes de tension musculaire.

IV CONCLUSION

L'étude de la trajectoire du F2 dans la parole continue semble permettre de recueillir des résultats susceptibles :

- de confirmer les mesures audiologiques,

- de faciliter la prévision des limites des méthodes de rééducation.

Sur le plan audiologique, la durée globale d'une séquence et la durée moyenne des segments stables nous paraissent donner accès au degré de surdité. Les autres paramètres mesurés, excepté la durée des transitions pour laquelle nous n'avons pas observé de différences significatives, ne permettent qu'une classification plus grossière sourd - entendant.

On peut s'interroger sur l'origine de l'allongement des durées des périodes stables en fonction du degré de perte auditive. Il est possible d'incriminer les méthodes de démutisation. Dans notre cas, cette hypothèse n'est pas suffisante pour expliquer la différence qui existe entre les groupes DAP et DAS dont les sujets ont été démutisés dans le même établissement suivant une méthode mixte. Il est fort probable que les modalités du feed-back audition-phonation diffèrent d'un groupe à l'autre.

Notons que nos résultats semblent montrer que les mécanismes de contrôle de l'articulation ne sont pas tous dépendants du degré de perte auditive :

- la durée des transitions ne varie pas beaucoup d'un groupe à l'autre,

- la fréquence du F2 des voyelles n'est pas significativement différente entre les groupes DAP et DAS.

On peut donc supposer que le contrôle de la parole n'utilise pas le feed-back auditif d'une manière continue. Cette hypothèse est en bonne concordance avec les phénomènes d'anticipation mettant en jeu des modes de contrôle programmés par l'apprentissage.

BIBLIOGRAPHIE

R.B. MONSEN : second formant transitions of selected consonant-vowel combinations in the speech of the deaf and normal hearing children. Journal of speech and Hearing Research Vol 19 (2) June 76.

R.B. MONSEN : durational effects of vowel production in the speech of deaf children - JSHR Vol 17 (3) Sept. 74

G.D. ALLEN : vowel duration measurement : a reliability study - J.A.S.A. vol 63 (4) April 78

P. KEATING, S.E. BLUMSTEIN : effects of transition length on the perception of stop consonants - J.A.S.A. vol 64 (1) July 78

T. GAY : effect of speaking rate on vowel formant movements J.A.S.A. vol 63 (1) Jan 78

S. BARTH, R. CHULLIAT : étude anamorphotique des voyelles françaises - IXe JEP - LANNION - 1978

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

S. BARTH, R. BEN FADHEL, G. MAJO

INSTITUT NATIONAL DE JEUNES SOURDS DE CHAMBERY

B.P. 15
73 160 COGNIN

RESUME : Nous présentons les résultats de mesures du V.O.T., faites à partir de parole émise par des enfants sourds et des entendants, en cherchant des corrélations entre celles-ci et des facteurs caractérisant les déficients auditifs (degrés de surdité, méthode de démutisation).

SUMMARY : The aim of this paper is to present values of the voice onset time we have measured with deaf and normal hearing speakers, we have tried to show some correlations between these results and the degree of deafness. our work also compare two speech training methods.

I INTRODUCTION

De nombreux travaux dans le domaine de la communication parlée ont montré l'importance du **V.O.T.** pour la perception catégorielle de la parole (ABRAMSON, LISKER 1967, CORBIT, EIMAS, COOPER 1973, BECKMANS 1977).

Un des problèmes principaux de l'apprentissage de la parole par nos élèves est celui de l'acquisition de l'opposition voisé - non voisé. En étudiant les mesures publiées par PRESTON et YENI - KOMSHIAN (1967), PORT et PRESTON (1969), EGUSHI et HIRSH (1969), PORT et PRESTON (1972) et ZLATIN (1972), on constate qu'il existe une génèse du **V.O.T.** parallèle à la maturation neuro-musculaire de l'enfant. Le contrôle moteur de la parole s'affine lentement chez l'entendant pour atteindre une précision optimale au stade pré-pubertaire.

Pour le sourd de naissance ou le devenu sourd dans les toutes premières années de la vie, cet apprentissage dépend fortement de l'appareillage, c'est-à-dire de la qualité du couplage prothèse auditive - oreille pathologique.

Aucune mesure de V.O.T. n'existant pour les sourds de langue française, il nous a semblé intéressant d'étudier celui-ci sous l'aspect d'indice de la coordination articulo-laryngée et d'essayer d'en tirer des données sur la pathologie de la parole liée à une déficience auditive.

II EXPERIMENTATION

Nous avons cherché à déterminer :

a) l'existence d'une corrélation entre les valeurs du **V.O.T.** et le degré de perte auditive.

b) les actes orthophoniques susceptibles d'améliorer les valeurs de **V.O.T.** pathologiques et, dans une certaine mesure, l'intelligibilité de la parole de nos élèves.

c) les différences susceptibles d'exister dans les réalisations de D.E.V. d'enfants sourds ayant été démutisés suivant des méthodes différentes.

II 1 Choix du corpus phonétique

Compte-tenu des difficultés que l'on rencontre avec la parole des déficients auditifs, nous avons demandé à nos sujets d'émettre les consonnes p, t, k, d, g, v, z, ʒ associées aux voyelles a, i et u.

Une bande enregistrée avec un adulte entendant du sexe masculin a été préparée afin de permettre des tests de perception à partir du même corpus.

II 2 Détermination des groupes d'enfants

KENT suggérant que la stabilité du V.O.T. est acquise vers huit ans et l'apprentissage des mécanismes de base de la parole étant normalement achevé à cet âge par nos élèves, nous avons sélectionné quarante déficients auditifs (4 groupes de 10) nés entre 1967 et 1969).

groupe	déficients auditifs sévères	déficients auditifs profonds	démutisation à dominante globale	démutisation à dominante analytique
1 : DAS 2	X		X	
2 : DAS 1	X			X
3 : DAP 2	-	X	X	
4 : DAP 1		X		X

Les déficients auditifs sévères ont une perte moyenne, sur les fréquences conversationnelles, comprise entre 70 et 90 dB. Les pertes des déficients auditifs profonds sont supérieurs à 90 dB. A l'intérieur de chacun de ces groupes audiologiques, les audiogrammes étaient très peu différents.

La dominante analytique ou globale indique qu'un choix a été fait pour la méthode de rééducation, les méthodes analytiques (plus exactement de construction) insistant plus sur l'élément phonétique au détriment de la structuration globale.

A ces groupes, nous avons ajouté 10 enfants entendants du même âge.

II 3 Enregistrement

Celui-ci a été pratiqué dans la chambre sourde de l'Institut National de Jeunes-Sourds de Chambéry,) l'aide d'un magnétophone REVOX A 77 équipé d'une bande SONY HL, d'un micro SENNHEISER MD 441 et d'une conque ELLIPSON 35 W pour le contrôle.

Chaque sujet a été pris individuellement, l'expérimentateur lui demandant de lire à haute voix les groupes convonne-voyelle (C.V.) qu'il lui indiquait, à raison d'une émission toutes les 5 secondes.

Pour les sujets sourds, chaque enregistrement était précédé par une lecture préalable de la liste des C.V. accompagnée de corrections appropriées et d'explications graphiques ou verbales destinées à faire comprendre le protocole. Seules les émissions jugées correctes devant être conservées (pour pouvoir comparer des choses comparables), nous avons recueilli, grâce à ces phases préparatoires, elles-mêmes enregistrées, une masse importante de documents sonores utilisables pour la phase b) de notre expérimentation.

II 4 Analyse des enregistrements

Les signaux issus du magnétophone ont été traités grâce à un système d'analyse spectrale développé en T.S.L. (Time Series Language) autour d'un PDP 11/20 équipé d'un cabinet TIME-DATA XO et d'une console de deux disques RK 05. Les documents restitués étaient :

- un sonagramme numérique (quantification fréquentielle 100 Hz, quantification temporelle 10 ms, fenêtre d'analyse 300 Hz).

- des spectres instantanés (0,6 KHZ) et des tranches de signal(oscillogrammes) choisis par l'expérimentateur.

La détermination du V.O.T. de chacune des émissions a été faite manuellement à partir de ceux-ci en prenant comme origine des temps l'achèvement de la consonne(*) suivant l'optique de KENT et non pas le début de celle-ci comme le suggèrent COHEN et MASSARO. En effet, les réalisations des consonnes par les déficients auditifs, et notamment des fricatives, durent deux à trois fois plus longtemps que chez l'entendant (S. BARTH, R. CHULLIAT, suivie de la trajectoire du F2, dans le même volume), ce qui entâche les résultats d'une manière non négligeable.

La précision obtenue était de l'ordre de ± 2,5 ms pour les 1440 valeurs extraites (25 ms visualisés sur 5 carreaux de l'écran).

(*) L'achèvement de la consonne a été estimé sur des oscillogrammes comme étant l'instant où l'amplitude du signal correspondant à la consonne n'est plus discernable de celle du bruit de fond ; un contrôle systématique a été fait par inspection des sonagrammes numériques et visualisation des spectres instantanés.

II 5 Test de perception

Nous avons présenté la bande test à 10 sujets déficients auditifs sévères, suivant deux modes d'écoute :

- utilisation d'un amplificateur INTERACOUSTICS DS 4 et d'un casque KHOS,

- utilisation des prothèses individuelles.

L'expérimentateur demandait aux sujets de remplir une grille en sautant une case en cas de doute. Les groupes C.V. étaient présentés à la cadence de 12 par minute.

Le dépouillement s'est fait en ne tenant pas compte des confusions phonémiques dans une même catégorie de consonnes (voisées ou non voisées), le nombre d'erreurs sur les traits continu et interrompu ayant été nul.

III RESULTATS

III 1 Valeurs moyennes du V.O.T.

La figure III,1 fournit les valeurs moyennes de V.O.T. de chaque voyelle pour les entendants et les groupes de sujets sourds définis en II 2. Nous n'avons pas tenu compte des variations du V.O.T. en fonction de la voyelle associée dans cette étude préliminaire.

Les courbes obtenues sont encadrées par des courbes à $\pm \sigma$. Chez l'entendant, on remarque que le V.O.T. :

- augmente lorsque l'on passe de p à k et de f à ʃ ,

- augmente en valeur absolue de b à g,

- diminue en valeur absolue de v à ʒ .

Le lecteur pourra se rapporter, à titre de comparaison, aux données trouvées par M. WAJSKOP pour les occlusives françaises (R.A. BRUXELLES 12/2 p. 70-98 fig. 3 et 5 b).

a) cas des plosives sourdes

La croissance du D.E.V. de p à k n'est pas observée systématiquement chez les déficients auditifs et lorsqu'elle se produit (groupe 1, 4), l'amplitude de la variation entre t et k est amoindrie.

b) cas des fricatives sourdes

La croissance du D.E.V. de f à ʃ n'existe que dans les groupes de déficients auditifs sévères. De plus, la variation des valeurs entre s et ʃ est moins importante que chez l'entendant.

c) cas des plosives et des fricatives sonores

Pour les consonnes sonores, les graphes correspondants aux groupes de déficients auditifs ne sont pas comparables à celui obtenu pour les entendants.

Pour les déficients auditifs profonds, il existe très peu de différences entre les valeurs mesurées pour b, d, et g. L'allure des courbes, pour ces consonnes, est très semblable dans le cas des déficients auditifs sévères.

d) conclusion

L'observation des tracés de la figure III fait apparaître des modifications importantes des valeurs du V.O.T. dans la parole du sourd. Les problèmes se situent au niveau des fricatives sourdes et des consonnes voisées. Ces résultats sont en bonne concordance avec les difficultés que rencontre tout rééducateur de sourds dans le domaine de l'apprentissage de la parole.

Cette objectivation de données empiriquement connues nous permet d'insister sur le fait que la qualité de la parole de nos élèves dépend essentiellement de l'acuité de la coordination articulo-laryngée dont le V.O.T. est un indice.

La figure III 1 d) donne une idée de la dispersion des valeurs de V.O.T. pour chacun des groupes de sujets. Elle représente les valeurs moyennes de σ/m pour les consonnes voisées et les consonnes sourdes. On remarquera que les méthodes de démutisation à dominante globale (groupe 1 et 3) tendent à diminuer l'impact du degré de surdité sur la dispersion des valeurs de de D.E.V.

III 2 Test de perception des traits voisé et non-voisé

Le tableau ci-après donne les moyennes, les écarts-types et les dispersions des taux de reconnaissance exacte et des erreurs.

	perception exacte du trait voisé	perception exacte du trait non non voisé	confusion voisé-non voisé	confusion non voisé-voisé
amplificateur HIFI	m = 69 % σ = 12 % σ/m = 17 %	75 % 10 % 13 %	31 % 6 % 19 %	24 % 4 % 17 %
prothèses individuelles	44 % 13 % 30 %	59 % 15 % 25 %	56 % 13 % 23 %	41 % 14 % 34 %

On constate une différence significative entre l'amplificateur de table et les prothèses individuelles en faveur du premier, avec augmentation de la dispersion des résultats lors de l'utilisation des appareils de correction auditive.

Ces données justifient l'emploi d'amplificateurs de qualité pour les séances de démutisation et d'orthophonie au cours desquelles on doit apprendre à l'enfant à affiner son contrôle moteur de la parole en utilisant ses restes auditifs au mieux.

Les tableaux suivants fournissent les moyennes de reconnaissance exacte et d'erreurs obtenues pour les fricatives et les occlusives suivant les deux modes d'écoute.

ampli HIFI	perception exacte du trait voisé	perception exacte du trait non voisé	confusion voisé-non voisé	confusion non voisé voisé
fricatives	67 %	70 %	33 %	30 %
occlusives	70 %	81 %	30 %	19 %

prothèses individuelles				
fricatives	44 %	62 %	55 %	37 %
occlusives	51 %	70 %	48 %	30 %

Ce sont les consonnes voisées qui sont les moins bien reconnues avec un taux d'erreur de l'ordre de 30 % pour l'ampmificateur de table et de 50 % pour les prothèses individuelles.

Notons que l'amplificateur à large bande améliore la perception des occlusives sourdes (19 % d'erreurs au lieu de 30 % pour les fricatives sourdes).

Cette constatation est à mettre en parallèle avec les résultats présentés en III 1 où les valeurs de V.O.T. pour les consonnes voisées des sujets sourds montraient certaines anomalies par rapport à celles des entendants.

III 3 Procédés d'orthophonie améliorant le D.E.V. des sujets sourds

Comme nous l'avons indiqué en II.3, nous disposions d'une masse importante de documents concernant les séances de rééducation. Le traitement complet de ces données n'est pas encore terminé mais les premiers résultats indiquent que la perception par toucher comparatif :

- des vibrations laryngiennes pour les consonnes voisées,
- de l'absence de vibrations laryngiennes pour les consonnes non-voisées,
- du souffle et de la tension pour les fricatives,

tend à améliorer les valeurs pathologiques de V.O.T. et par là même, la qualité de la parole de nos élèves.

Nous devons noter qu'une fois la valeur correcte de V.O.T. obtenue, celle-ci se maintient au cours de la séance et n'est pas perturbée par des corrections particulières visant à améliorer le timbre des voyelles.

IV CONCLUSION

Le V.O.T. qui, comme nous l'avons vu, semble nécessiter un contrôle auditif pour sa réalisation normale, paraît être un outil intéressant pour juger objectivement de la qualité de la parole de nos élèves. Sa mesure permet de mieux comprendre les problèmes de coordination articulo-laryngée et d'envisager la recherche de séries raisonnées d'actes orthophoniques offrant des possibilités de mener au mieux notre travail de rééducateur.

Figure III. 1

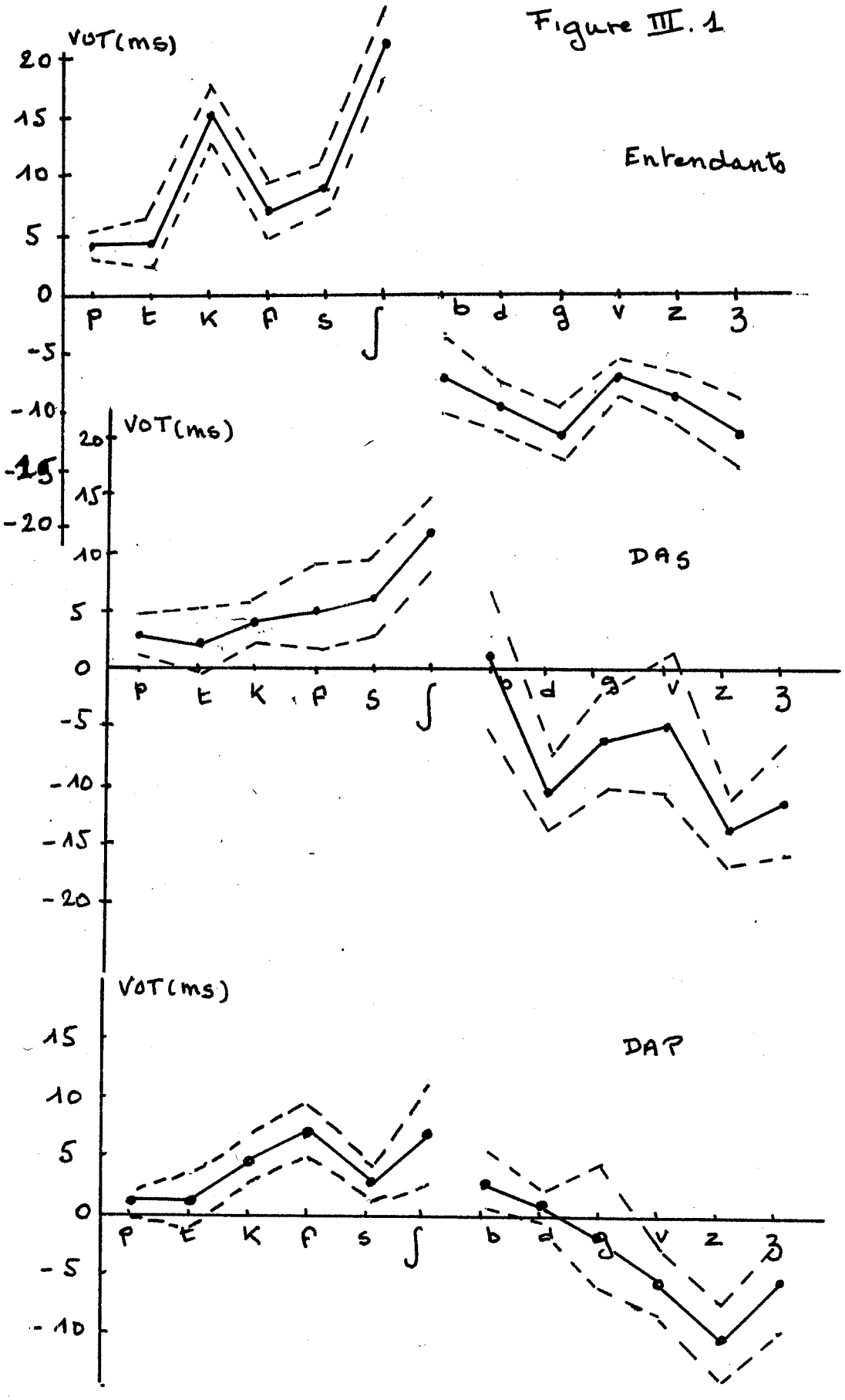
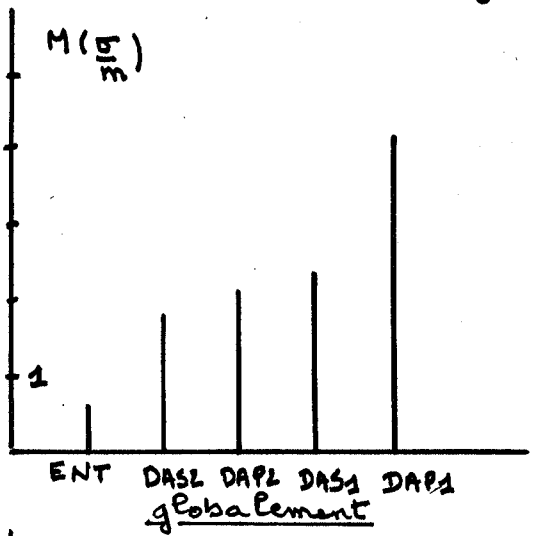
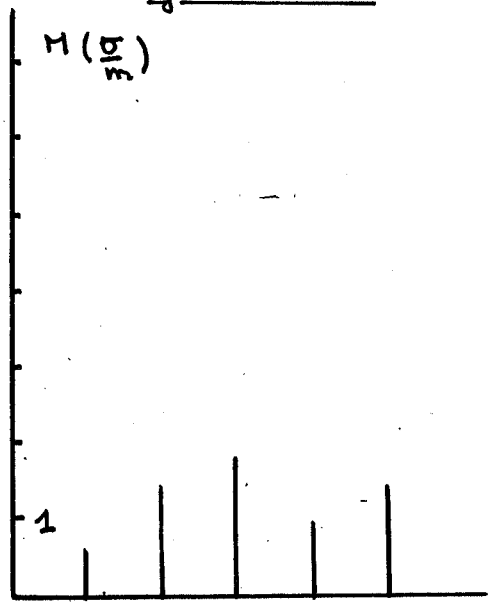

VOT(ms)
20
15
10
5
0
-5
-10
-15
-20
p t k f s ʃ b d g v z ʒ
Entendants
VOT(ms)
DAS
p t k f s ʃ b d g v z ʒ
VOT(ms)
DAP
p t k f s ʃ b d g v z ʒ

Fig III. 1. d

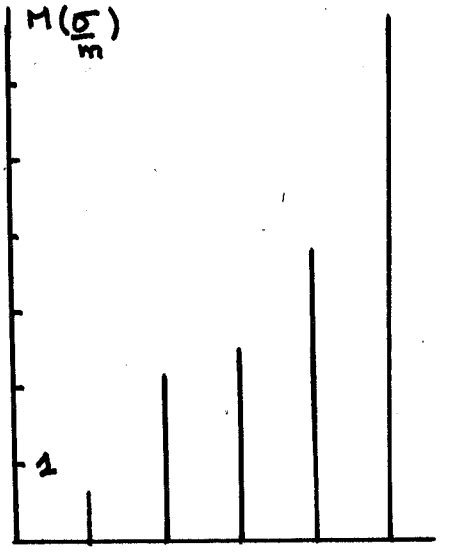


globalement

valeur moyenne de $\frac{\sigma}{m}$ dans chaque groupe



consonnes sourdes



consonnes voisees

La dispersion des résultats est plus grande pour les méthodes à dominante constructive (pour les consonnes voisées particulièrement)

Le degré de surdité influe moins sur celle-ci dans le cas de méthodes à dominante globale (DAS2, DAP2)

BIBLIOGRAPHIE

A.S. ABRAMSON, L. LISKER : voice onset time in stop consonnants. Acoustic analysis and synthesis - 5e Congrès International d'Acoustique - LIEGE 1965.

R. BECKMANS : structuration perceptive de deux indices acoustiques et perception catégorielle - 9e J.E.P. LANNION 1978.

L.J. BOE : anatomie et physiologie de la phonation - Travaux de l'Institut de Phonétique de Grenoble - 1977.

COHEN, MASSARO : the contribution of fundamental frequency and voice onset time to /ZI/ - /SI/ distinction JASA vol 60 (3) sept. 1976.

S. EGUSHI, I.J. HIRSH : development of speech sounds in children - Acta otolaryng. suppl 257 - 1969.

P.D. EIMAS, J.D. CORBIT : some properties of linguistic feature detectors - Perception and Psychophysics - vol 13 - 1973.

R.D. KENT : anatomical and neuro muscular maturation of speech mechaniom : evidence from acoustic studies. Journal of speech and Hearing Research vol 19 n° 34 sep 1976.

A.M. LIBERMAN : some results of research on speech perception. JASA vol 29, 1957.

L. LISKER : is it VOT or first formant transition detector. JASA vol 57, 1975.

D.B. PISONI : indentification and discrimination of the relative onset time of two component tones - Implication for voicing perception in stops. JASA vol 61 - 1977.

D.K. PORT, M.S. PRESTON : early apical stop production. A voice onset time analysis. Haskins Laboratoires Report on speech research SR - 29/30. New Haven, conn : Haskins Laboratoires 1972.

M.S. PRESTON, G. YENI-KOMSHIAN : studies on the development of stop consonants in children - Haskins Laboratories 1967.

J. WOOD : discriminability response bias and phoneme categories in discrimination of V.O.T. JASA vol 60 (6) - 1976.

ZLATIN : voicing contrast : perceptual and productive V.O.T. characteristics of adults. JASA vol 56 (3) 1974.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

CENTRE DE GRAVITE FREQUENTIEL ET MOMENT D'ORDRE DEUX DU SPECTRE DE VOYELLE DU FRANCAIS.

DEMARS C.

Département de Recherches Linguistiques
UNIVERSITE DE PARIS VII

RESUME EN FRANÇAIS

Cette communication concerne les tout premiers résultats d'une étude du centre de gravité fréquentiel et du moment d'ordre 2 du spectre des voyelles du Français.

Envisagé dès 1951 par JAKOBSON, le calcul du centre de gravité a été récemment repris par différents auteurs : dans cette étude-ci, les calculs et du centre de gravité M1 et du moment d'ordre 2, M2, sont effectués sur tout le spectre disponible engendré par 255 filtres d'égale largeur de bande, régulièrement distribués de 0 à 5000 Hz, chaque composante étant prise en valeur et non en logarithme (échelles linéaires).

A chaque voyelle correspond un couple (M1,M2). Les résultats concernent seulement 5 voyelles produites par 18 locuteurs différents d'un corpus qui comportait 16 logatomes de type CV CV. Ils montrent :
1) que la dispersion autour de la moyenne des paramètres M1 et M2 est relativement élevée.
2) que dans le plan (M1,M2) trois catégories principales peuvent être distinguées.

Ce travail préliminaire, destiné à préciser une méthodologie sera développé ultérieurement dans une étude approfondie.

CENTER OF GRAVITY AND SECOND MOMENT OF THE SPECTRUM OF VOWELS.

DEMARS C.

Département de Recherches Linguistiques
Université de Paris VII.

This report describes the very first results of a study on the center of gravity and second moment of vowel spectrum in French.

Listed, among others, by JAKOBSON and al (1951) as a measure for the feature "acuteness", the computation of the center of gravity has been recently developped by different researchers. For CHISTOVICH (1979) the computation is limited to the first two formants, for STALHAMMAR (1978) it only concerns the frequencies above the first formant. It is also partial for PERENNOU (1978), it is extended to the whole spectrum by MERCIER (1978) and LIENARD (1979). The computation of the second moment has been suggested byJAKOBSON (1951) as a possible measure for the feature of compactness.

In this report, the computation of the center of gravity and the second moment about the mean, was made on the whole spectrum available, generated by 255 filters of equal bandwidth regularly distributed from 0 to 5000 cps. After multiplying the signal by a Gaussian window,a spectrum was performed every 12.8 millisecond for each word of the corpus (the rate of overlapping for one spectrum to another being 75%). After visualizing the signal and making sure we were in a stable part of the vowel, we kept a set of two numbers, M1 for the frequency of the center of gravity and M2 for the second moment.

Theresults concern only 18 productions of 5 vowels /i/, /u/, /o/, /ɔ/, /a/ taken from (/puʔi/,/Rala/,/tose/, /bivu/, /gɔʒɛ /, /kɔʃɛ/) a corpus which included 16 logatomes of the type CV CV pronounced by 21 speakers (3 chidren, 18 adults - 10 men and 8 women).

It appears from these first results that the dispersion about the mean of the parameters M1 and M2 is rather high (It is the very problem of variability between speakers). The representation in the plan (M1,M2) and STUDENT's test show that on the average, three main classes /o/, /a/,/i/ can be found.

This preliminary study was a test experiment on a limited corpus to define a method for further studies already planned.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

CENTRE DE GRAVITE FREQUENTIEL ET MOMENT D'ORDRE DEUX DU SPECTRE DE VOYELLE DU FRANÇAIS.

DEMARS C. Département de Recherches Linguistiques
Université de Paris VII

I. INTRODUCTION

Cette communication présente les premiers résultats d'une étude du centre de gravité fréquentiel et du moment d'ordre 2 du spectre de voyelles du Français. Seules 5 voyelles : /u/, /o/, /ɔ/, /a/, /i/, du corpus sont analysées dans ce travail-ci.

Cette étude a été faite, en vue d'une comparaison avec une étude analogue, actuellement en cours, faite sur le même corpus, du centre de gravité et du moment d'ordre 2 de la distribution d'énergie le long de l'axe des fréquences instantanées des voyelles. La méthode de la fréquence instantanee est basée sur la notion de fréquence considérée comme la vitesse angulaire d'un point dans le plan, une des coordonnées étant le signal lui-même, l'autre le signal en quadrature (VILLE,1948). A un signal, on fait correspondre une distribution T (ϖ,t) où ϖ est la fréquence instantanée et t le temps. Par intégration de T (ϖ,t) sur le temps, on obtient une fonction de la variable ϖ, notée $\lambda^2(\varpi)$ qui représente une distribution de l'énergie du signal sur l'axe des fréquences instantanées. On démontre que les centre de gravité du spectre et de la distribution $\lambda^2(\varpi)$ sont confondus (BERTHOMIER, 1976).

Envisagé par JAKOBSON et al. (1951) comme un possible indice d'acuité, le calcul du centre de gravité du spectre a été repris par différents auteurs. Pour LIENARD (1978) le calcul est effectué sur un spectre qui va de 200 à 7000 Hz, le banc de 32 filtres de bande passante de l'ordre de 300 Hz, étant réglé de façon quasi logarithmique, l'amplitude de chaque composante étant prise en logarithme, sauf pour les très faible valeurs; les spectres, calculés toutes les 10ms, étant moyennés sur 1,25s dans la partie stable; cet indice ne permet qu'une séparation en trois classes phonétiques des 18 sons étudiés et fournit un classement sur l'échelle grave-aigu satisfaisant si l'on considère les moyennes des 10 locuteurs.

Pour MERCIER (1978), le calcul du centre de gravité est limité à la partie du spectre comprise entre 300 et 4200 Hz correspondant aux 14 filtres d'inégales largeurs de bande de son vocodeur à canaux. Il est effectué tous les 13ms. Il est utilisé pour la détection des fricatives non voisées.

Chez CAELEN (1978), les calculs sont limités respectivement à

la somme des énergies des 3 premiers canaux et des 3 derniers canaux (et au rapport GA de ces deux nombres), d'un spectre calculé toutes les 8ms par un analyseur (modèle d'oreille) inventé par l'auteur. La valeur de GA est utilisée pour l'étiquetage en catégories acoustiques.

Dans une étude de psychoacoustique, CHISTOVICH (1978) fait des expériences de perception de voyelles synthétiques pour préciser l'effet de centre de gravité dans la perception de voyelles suggéré par différents auteurs : c'est à dire pour préciser de quelle manière une voyelle de deux formants F1 et F2 peut être rendue perceptuellement semblable à un stimulus à un seul formant F^* de fréquence (F1+F2)/2 par ajustement du rapport A2/A1 des amplitudes des formants. L'auteur met en évidence l'existence d'une distance critique entre les formants F1 et F2 au dessus de laquelle cet effet ne se produit pas et discute de l'existence possible d'une relation entre l'extraction des pics du spectre et les effets de centre de gravité.

Dans le cadre d'expériences de psychoacoutisque STALHAMMAR (1978) calcule le centre de gravité G_2^1 de la partie du spectre supérieure à celle du premier formant F1, et la différence G_2^1 - F1. Il montre que cette différence est corrélée de façon positive avec les paramètres de position de la langue (hauteur et antériorité). Il la compare à celle obtenue par application des formules de BLADON et FANT (1978) et montre que du point de vue de la reconnaissance automatique de la parole elle est plus intéressante que la différence F2-F1, sans toutefois couvrir tous les cas possibles. Il cherche alors, à l'aide d'une synthèse à 2 où à 4 formants, si ce n'est pas un autre facteur, qu'il appelle facteur de forme (du spectre), qui est significatif dans la perception des voyelles.

En ce qui concerne le moment d'ordre 2, il a été envisagé par JAKOBSON et al (1951) comme un possible indice de "compacité".

Dans ce travail les calculs du centre de gravité M1 et du moment centré d'ordre 2, M2, ont été effectués sur tout le spectre engendré par 255 filtres d'égale largeur de bande, répartis linéairement de 0 à 5000 Hz, l'amplitude de chaque composante du spectre de puissance étant prise en valeur et non en logarithme (échelle linéaire). Le moment d'ordre 2 est effectivement calculé et les voyelles sont représentées dans le plan (M1,M2).

II. MATERIEL ET METHODES.

a) Le corpus.

On a construit un corpus de 16 logatomes, de 2 voyelles et 2 consonnes, de type CV CV, sans signification pour éliminer au maximum l'effet du sens sur la prononciation (HOUSE 1953), transcrits en français pour faciliter la prononciation par des locuteurs non phonéticiens et notamment l'obtention de /ɔ/ ("cochait"= kɔʃɛ , "gohjait"= gɔʒɛ). Ces mots ont été prononcés dans une salle ordinaire silencieuse, près du micro par 21 locuteurs (3 enfants, 18 adultes - 10 hommes et 8 femmes dont une d'origine étrangère agés de 25 à 79 ans). Pour chaque mot, les locuteurs déclenchaient eux-mêmes par l'intermédiaire d'un microordinateur l'enregistrement qui était d'une durée de 0,4096 seconde. Le signal était filtré entre 20 et 5000 Hz, puis échantilloné à la fréquence de 10 kHz et digitalisé par un convertisseur analogique digital et quantifié sur 8 bits par un codage en amplitude linéaire (CARTIER 1979). Le signal était en même temps visualisé sur un oscilloscope. L'enregistrement était répété si l'émission était mal cadrée dans la fenêtre temporelle, trop faible ou trop forte (saturation) la dynamique étant limitée à 48 db. S'il était satisfaisant les données correspondantes étaient

stockées sur disquettes.

On ne garantit pas que des nuances fines par exemple entre /o/ et /ɔ/ aient été pleinement réalisées par tous les locuteurs.

b) Traitement des données.

Pour chaque mot on a effectué sur les 1920 premières millisecondes une analyse de FOURIER, en calculant toutes les 12,8 millisecondes le spectre sur 512 points, le taux de recouvrement étant ainsi d'un spectre sur l'autre de 75% (cf Fig 1). Avant le calcul du spectre le signal était multiplié par une fenêtre gaussienne (HARRIS 1978). Compte tenu de l'influence de la fenêtre la résolution en fréquence de l'analyse de FOURIER était donc de l'ordre de 35 Hz.

Pour chaque spectre on a calculé

1) Son moment d'ordre zéro, MØ, c'est à dire l'énergie totale, somme des carrés des parties réelles et imaginaires du spectre, étendue à tout le domaine du spectre.

$$M\,\emptyset = \sum_k E(k) = \sum_k \left(Re^2(k) + Im^2(k)\right)$$

2) Son moment d'ordre un, M1, ou centre de gravité fréquentiel barycentre des fréquences affectées de leur énergie

$$M1 = \frac{\sum_k k \times E(k)}{M\emptyset}$$

3) Son moment d'ordre deux, M2, centré au centre de gravité précédent

$$M2 = \frac{\sum_k (k - M1)^2 \times E(k)}{M\emptyset}$$

en utilisant une échelle de fréquence linéaire, chaque fréquence étant prise par son énergie prise en valeur absolue et non en décibel.

On remarquera que M1 et M2 sont normalisés.

Pour chaque début de mot, dont on a vérifié qu'il comportait bien la première voyelle, on avait donc un ensemble de 45 valeurs (3 pour chacun des 15 spectres). Les voyelles étaient repérées par le maximum de l'énergie MØ, la stabilité relative de M1, le minimum de M2.

Après visualisation du signal enregistré, sur une table traçante on a retenu pour chaque voyelle un spectre et donc un triplet MØ, M1, M2 en employant les critères suivants : minimum de M2 et maximum (ou proximité du maximum) de MØ, éloignement des consonnes voisines encadrant la voyelle, absence de parasites sur la partie correspondante du signal (tops synchro ou autres). On a vérifié que l'on était dans la partie la plus stable de la voyelle. Pour chaque voyelle on calcule la valeur moyenne et l'écart-type des paramètres M1 et M2 pour l'ensemble des locuteurs et on a compraré ces moyennes par le test de STUDENT.

III. RESULTATS

Les résultats concernent 18 productions de 5 voyelles : le /u/ de /puʔi/, le /o/ de /tose/ et /doze/, le /ɔ/ de /kɔʃɛ / et /gɔʒɛ /, le premier /a/ de /mana/, et /Rala/, le /i/ de /bivu/. Des statistiques sont faites toutes catégories confondues.

L'ensemble des résultats est présenté sur la figure 2 pour le

centre de gravité M1, et sur la figure 3 pour le moment d'ordre 2 : M2. On notera que pour les deux paramètres, la dispersion en valeur absolue est élevée.

La figure 4 donne dans le plan (M1,M2) la position des valeurs moyennes des deux paramètres M1 et M2 pour les 8 voyelles étudiées ainsi qu'une visualisation des ecarts-types correspondants.

Pour les deux paramètres, on a comparé les moyennes à l'aide du test de STUDENT qui donne la probabilité en quelque sorte pour les différentes classes dêtre "semblable" et cela pour les deux paramètres M1 et M2 séparément.

On remarque que le test (en accord avec la figure 4), au seuil de 1% pour les deux paramètres regroupe les /a/ de /mana/ et /Rala/, les /ɔ/ de /gɔʒɛ / et /kɔʃɛ /, les /ɔ/ de /tɔse/ et les /ɔ/ de /gɔʒɛ / et /kɔʃɛ /, ainsi que /u/ de /puʔi/ et le /o/ de /doze/.

Au seuil de 1%, un des paramètres assure, en moyenne la séparation en classes distinctes, du /u/ de /puʔi/ du /o/ de /tose/, du/o/ de /doze/ et respectivement du /o/ de / tose/ du /ɔ/ de /gɔʒɛ/ et /kɔʃɛ/, du /i/ de /bivu/ avec le /o/ de /doze/ et /tose/ du /ɔ/ de /gɔʒɛ / et /kɔʃɛ /. De même pour le /a/ de /Rala/ et le /ɔ/ de /kɔʃɛ/.

Discussion.

La petite taille du système informatique a posé beaucoup de problèmes. C'est la raison pour laquelle, au lieu d'effectuer des variations systématiques, on a pris un nombre de combinaisons limitées. C'est ainsi que l'on n'a pas de paires minimales. Pour cette même raison chaque logatome avait une durée limitée de 4096 ms correspondant à 4 K (kilooctets), ce qui ne nous garantit pas que les voyelles aient eu une durée suffisante.

D'autre part, s'agissant d'un travail préliminaire sur un corpus limité, destiné à préciser une méthodologie, c'est volontairement que l'on n'a pas effectué de statistiques plus ou moins savantes, sur les données recueillies.

On se propose d'étendre les résultats de ce travail préliminaire à l'ensemble des locuteurs et des voyelles du corpus (afin en particulier d'aborder les phénomènes de coarticulation)et de les préciser en effectuant des calculs statistiques complémentaires et détaillés suivant les différentes catégories de locuteurs. On se propose aussi de chercher à améliorer la méthode en en faisant varier les différents paramètres, aussi bien les paramètres liés à la digitalisation du signal (niveau de quantification, fréquence d'échantillonage) ou à la fenêtre (choix de la fenêtre, longueur) que ceux liés à l'analyse de FOURIER (résolution en fréquence) ou au traitement des données (pas du calcul des moments, échelles de mesure des fréquences et des énergies).

BIBLIOGRAPHIE

BERTHOMIER (C), 1976, Représentation dans un plan fréquence instantanée-temps, d'un signal. Thèse Doctorat d'Etat Université de Paris VI.

BLADON (R.A.W.), PANT (G), 1978, A two formant model and the cardinal vowels, STL-QPSR 1/1978 pp 1-8.

CARTIER (M), 1979, Le codage de la parole, l'Echo des Recherches Janvier 1979 pp 4-11.

CHISTOVICH (L.A), LUBLINSKAYA (V.V), 1979, The "Center of Gravity" effect in vowel spectra and critical distance between the formants : Psychoacoustical study of the perception of vowels-like stimuli Hearing Research I (1979) pp 185-195.

HARRIS (P.J), 1978, On the use of windows for harmonic analysis with the discrete Fourier transform, IEEE (ASSP). Vol 66, N.Q.I., January 78, pp 51-83.

HOUSE (A.S.), FAIRBANKS (G), 1953, The influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels, JASA, January 1953, Vol 25, number 1, pp 105-113.

JAKOBSON (R), FANT (G.M), HALLE (M), 1961, Preliminaries to Speech Analysis, MIT Press, 4th printing, 1961.

LIENARD (J.S), 1979, Speech Characterization from a Rough Spectral Analysis, XXXe IEEE International Conference on Acoustics, Speech and signal processing. ICASSP, Washington 1979, pp 595-599.

MERCIER (G), 1978, Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL 9è Journées d'Etudes sur la Parole GALF LANNION 1978, pp321-341.

CAELEN (J), PERENNOU (G), 1978, Indices et traits acoustiques dans un système de reconnaissance de la parole continue 9è Journées d'Etudes sur la Parole GALF, LANNION 1978, pp 95-103.

STALHAMNAR (J.W.J.), 1978, Form Factors for Power Spectra of Vowel Nuclei, K.T.H., 1978, pp 23-34.

VILLE (J), 1978, Théorie et applications de la notion de signal analytique, Cables et Transmissions, Janvier 1948 I pp 61-74 (Hors commerce).

Fig.1 Représentation de la forme d'onde d'un même mot (gohjait /gɔʒɛ/) pour trois locuteurs différents (VII0, VIII, VII2). On a indiqué quelques unes des positions successives de la fenêtre (1er, 2è, 3è, 15è).
Seuls les 2400 premiers points de chaque enregistrement ont été représentés.
Pour VII2 on a indiqué approximativement la position des différents éléments phonétiques (/g/, /ɔ/, /ʒ/).
Sur les 3 graphiques on distingue à gauche un top parasite.

Fig.2 Résultats du calcul du centre de gravité M1 pour les différents locuteurs et les différentes productions des voyelles étudiées.
La barre verticale indique la valeur moyenne.

Fig.3 Résultats du calcul du moment d'ordre M2, centré, pour les différents locuteurs et les différentes productions des voyelles étudiées.
La barre verticale indique la valeur moyenne.

Fig.4 Représentation dans le plan M1,M2 des valeurs moyennes des paramètres M1 et M2 pour les différentes productions des voyelles étudiées. La longueur des segments est égale à deux fois l'écart-type.

Fig.1 Représentation de la forme d'onde d'un même mot (gohjait /gɔʒɛ/) pour trois locuteurs différents (VII0, VIII, VII2). On a indiqué quelques unes des positions successives de la fenêtre (1er, 2è, 3è, 15è).
Seuls les 2400 premiers points de chaque enregistrement ont été représentés.
Pour V112 on a indiqué approximativement la position des différents éléments phonétiques (/g/, /ɔ/, /ʒ/).
Sur les 3 graphiques on distingue à gauche un top parasite.

Fig.2 Résultats du calcul du centre de gravité M1 pour les différents locuteurs et les différentes productions des voyelles étudiées.
La barre verticale indique la valeur moyenne.

Fig.3 Résultats du calcul du moment d'ordre M2, centré, pour les différents locuteurs et les différentes productions des voyelles étudiées.
La barre verticale indique la valeur moyenne.

Fig.4 Représentation dans le plan M1,M2 des valeurs moyennes des paramètres M1 et M2 pour les différentes productions des voyelles étudiées. La longueur des segments est égale à deux fois l'écart-type.

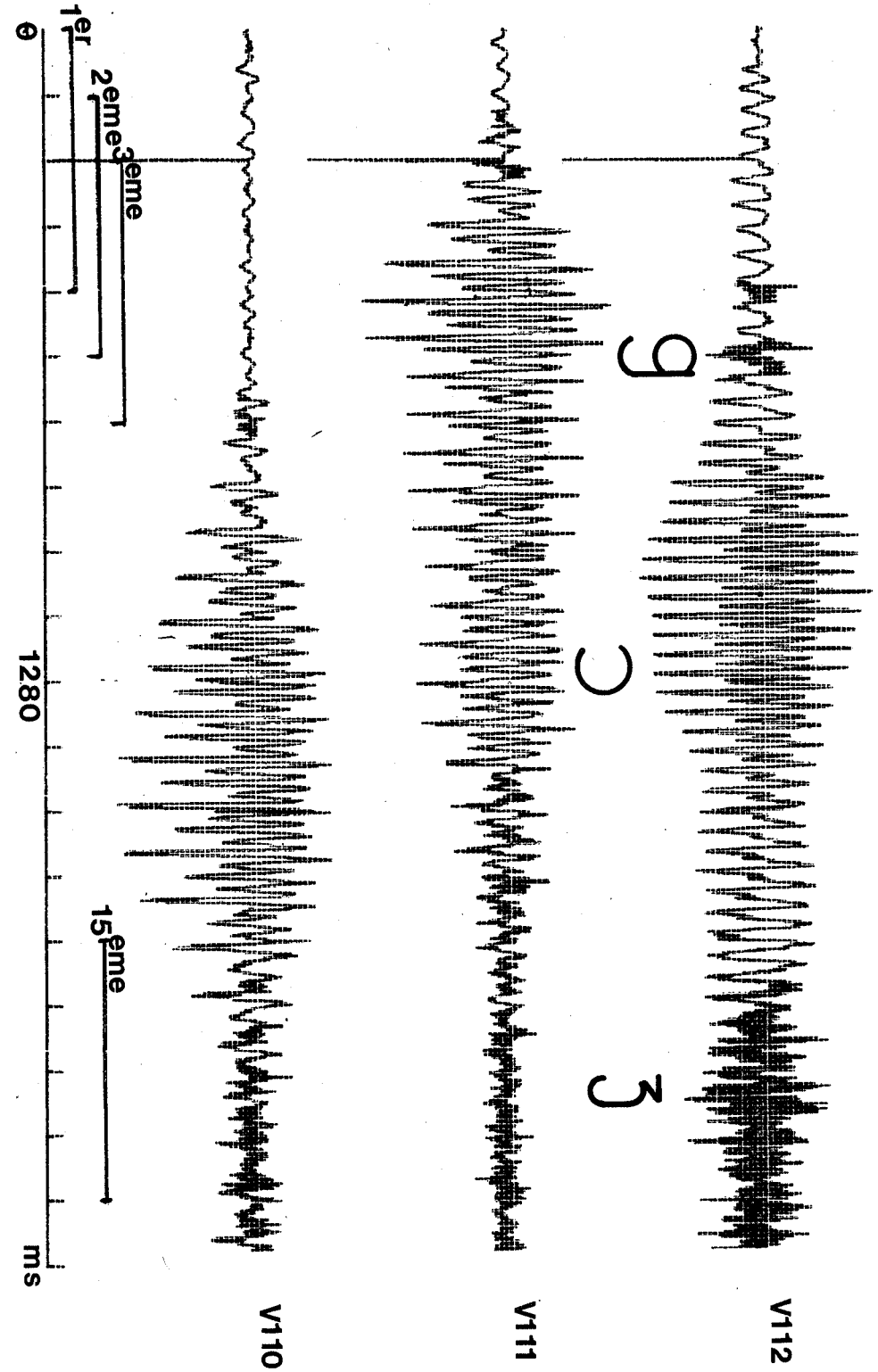


Fig. 1

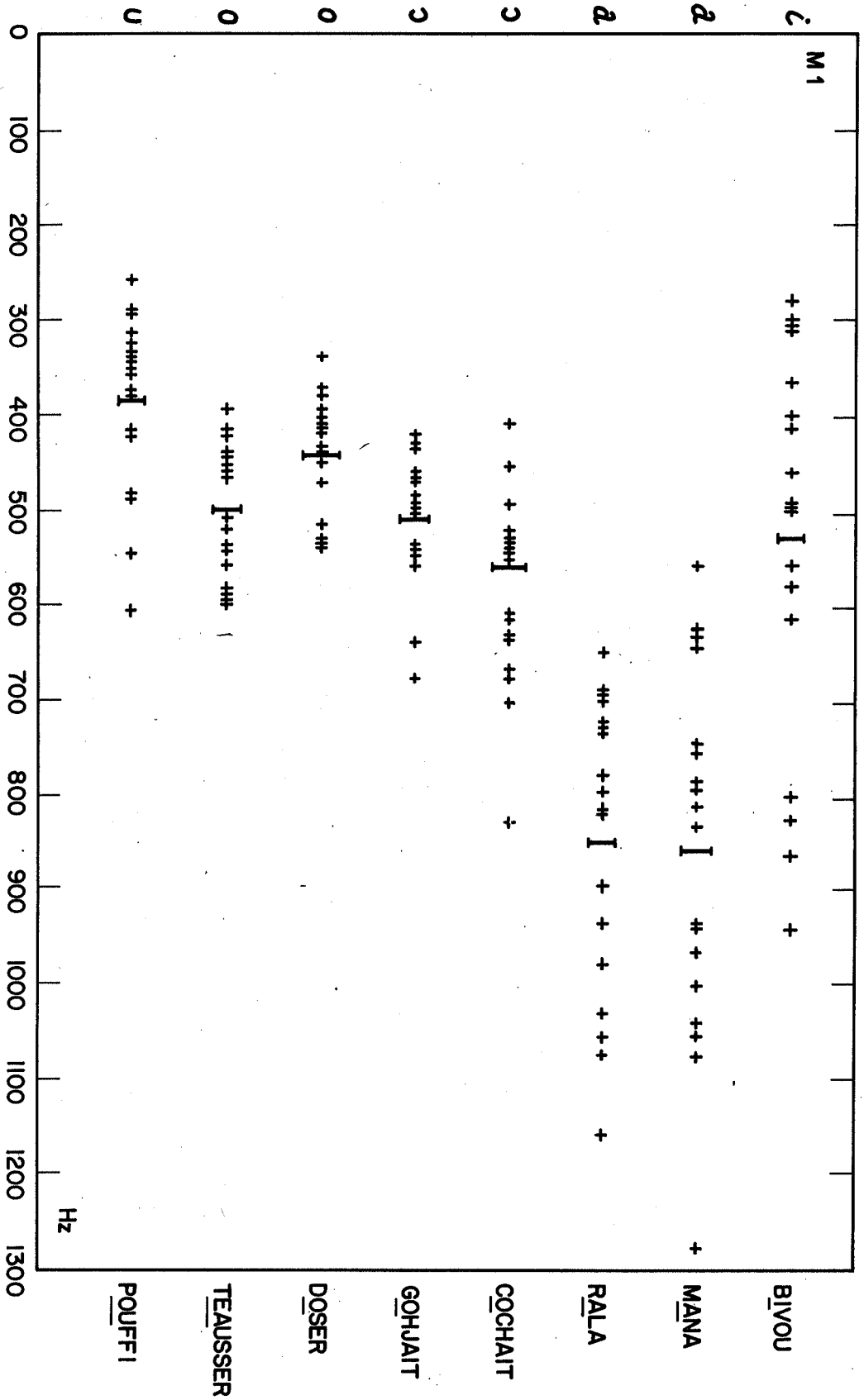

M1
BIVOU
MANA
RALA
COCHAIT
GOHJAIT
DOSER
TEAUSSER
POUFFI
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Hz

Fig. 2

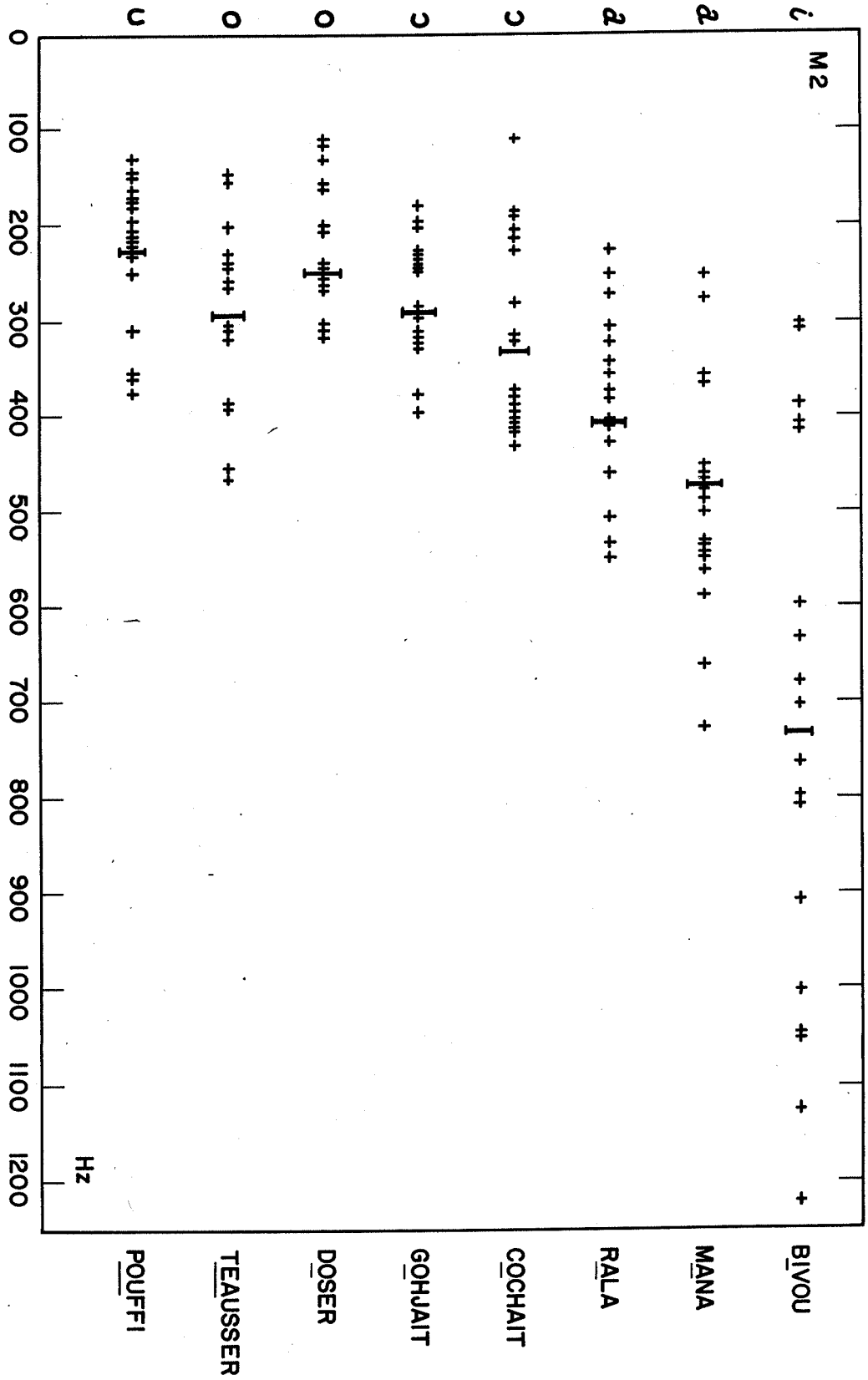

M2
i
a
a
c
c
o
o
u
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Hz
BIVOU
MANA
RALA
COCHAIT
GOHJAIT
DOSER
TEAUSSER
POUFFI

Fig. 3

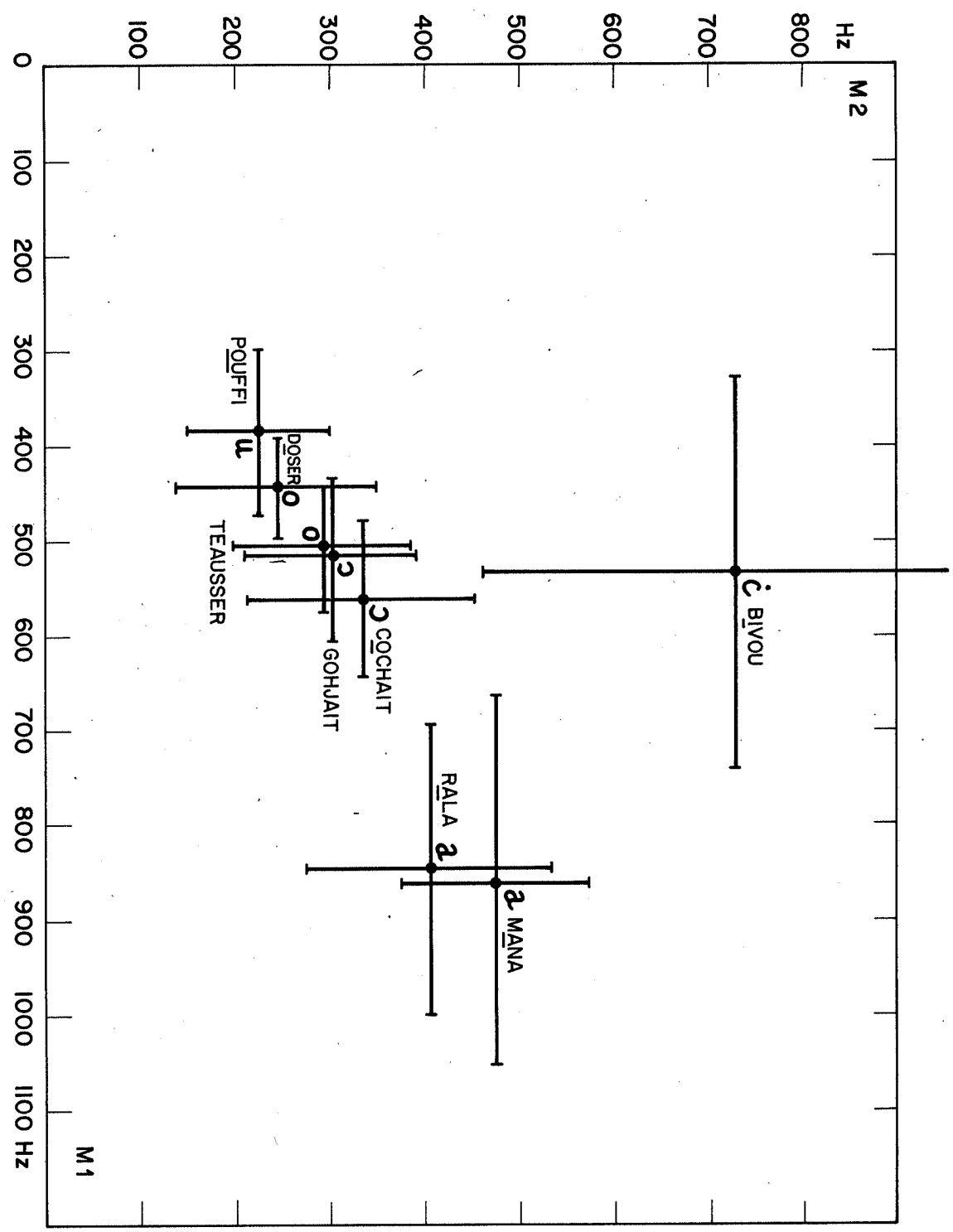

Hz
M2
100
200
300
400
500
600
700
800
0
100
200
300
400
500
600
700
800
900
1000
1100 Hz
M1
POUFFI
DOSER
TEAUSSER
GOHJAIT
COCHAIT
BIVOU
RALA
MANA

Fig. 4

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

VARIABILITE ACOUSTIQUE ET INTEGRATION PERCEPTIVE DES CIBLES PROSODIQUES

Albert DI CRISTO, Institut de Phonétique d'AIX EN PROVENCE

RESUME

Dans des études antérieures (DI CRISTO, I977, I978) nous avons avancé l'hypothèse que l'intégration perceptive et fonctionnelle des groupes intonatifs se fonderait essentiellement sur la prise en compte de certains points-clés (ou cibles prosodiques), qui correspondent respectivement : a) à l'attaque, b) à la prétonique et c) au point de hauteur du contour du groupe intonatif. Nous avons affirmé, d'autre part (DI CRISTO, I977), que les cibles prosodiques, qui reflètent les instructions linguistiques des centres supérieurs, sont programmées en termes de syllabes. Nous avons observé, en effet, que les variations paramétriques fonctionnelles qui actualisent les instructions linguistiques sont déclenchées dès le début de l'émission de la consonne dans les syllabes CV. Toutefois, la réalisation des cibles prosodiques s'accompagne inévitablement de variations microprosodiques (DI CRISTO, I978) qui manifestent diverses contraintes et qui constituent une source de variabilité importante.

Nous nous proposons, dans ce travail, d'analyser cette variabilité et d'évaluer ses effets sur l'intégration perceptive des cibles prosodiques. Nous avons choisi dans ce but d'étudier les réalisations du contour continuatif (intonème progrédient) dans les syllabes CV et CVC, en faisant varier systématiquement la nature de V (haute ~ basse) et celle des consonnes pré et post-vocaliques (occlusives ~ constrictive, voisée ~ non voisée).

L'analyse des configurations de Fo et d'intensité effectuée dans la première partie de cette étude, montre que l'association DN (glissando de Fo direct / et glissement d'intensité négatif \) est significativement plus fréquente dans le contexte (C_1 non voisée + V). Nous formons l'hypothèse que ce glissement négatif d'intensité permet d'atténuer la perceptibilité de la fin du glissando et qu'il contribue ainsi à réduire, sur le plan perceptif, les dépassements éventuels du point de hauteur du contour occasionnés par les consonnes initiales non voisées.

Dans la seconde partie de notre recherche, les résultats font apparaître que les variations intrinsèques et co-intrinsèques de Fo et de durée n'exercent pas une influence notable sur la réalisation du point de hauteur du contour. Ce dernier se place régulièrement dans le niveau infra-aigu et varie uniquement en fonction de la nature de V (hauteur intrinsèque). Cette relative invariabilité

est due à la mise en oeuvre d'un processus de réajustement qui ajuste la pente du contour en fonction de l'entourage consonantique.

Dans les recherches intonologiques, l'interprétation des indices et des traits ne peut être fondée sur l'observation des données objectives. Elle nécessite une stylisation préalable, basée sur l'effacement de la microprosodie et sur le transcodage perceptif des données acoustiques. Il importe également de prendre en compte, pour que la démarche demeure valable, les stratégies d'ajustement décrites dans cette étude.

SUMMARY

In previous research (DI CRISTO, 1977, 1978) we suggested that perceptual and functional integration of intonation groups is founded essentially on the recognition of certain key points (or prosodic targets) corresponding to a) the onset, b) the pretonic and c) the pitch-point of the intonation contour. We also claimed (DI CRISTO, 1977) that prosodic targets, reflecting higher-order linguistic instructions are programmed in terms of syllables. We have observed that functional linguistic variations actualising linguistic instructions are set off from the beginning of the consonant in CV syllables. The production of prosodic targets, however, is inevitably accompanied by microprosodic variations (DI CRISTO, 1978) showing a number of constraints and constituting an important source of variability.

In this study we propose to analyse this variability and evaluate its effects on the perceptual integration of prosodic targets. With this aim, we have chosen to examine realisation of non-final intonation patterns (major continuation) on CV and CVC syllables, systematically varying the nature of V (high ~ low) and of the initial and final consonants (stop ~ fricative, voiced ~ unvoiced).

The analysis of Fo and intensity patterns in the first part of this study shows that a combination DN (direct Fo glissando / and negative intensity glide ↘) is significatively more frequent in the context (C1 unvoiced + V). We propose the hypothesis that the negative intensity glide decreases the perceptibility of the end of the glissando thus helping to reduce,perceptually , the effect of overshooting the pitch-point of the contour caused by the initial unvoiced consonants.

In the second part, the results show that intrinsic and cointrinsic variations of Fo and duration do not have any appreciable effect on the realisation of the pitch-point of the contour. The latter is situated regularly in the mid-high tone level and varies only in function of the nature of V (intrinsic pitch). This relative invariability is due to the action of a process of readjustment adjusting the slope of the contour in function of surrounding consonants.

In intonation research, the interpretation of cues and features cannot be based on the observation of objective data. A preliminary stylising of the data is necessary, based on the elimination of microprosodic effects and the perceptual re-coding of acoustic data. It is also important, for the procedure to be valid, to take into account the strategies of adjustment described in this study.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

VARIABILITE ACOUSTIQUE ET INTEGRATION PERCEPTIVE DES CIBLES PROSODIQUES

Albert DI CRISTO, Institut de Phonétique d'AIX EN PROVENCE

I. - INTRODUCTION

Dans des recherches antérieures sur l'intonation du français (DI CRISTO, I976, I977, I978) nous avons présenté deux hypothèses. La première est que la reconnaissance des groupes intonatifs se fonderait principalement sur la prise en compte de certains points-clés (ou cibles prosodiques) qui correspondent à l'attaque, à la prétonique et au contour final. L'agencement syntagmatique de ces points-clés, qui sert de base à l'intégration perceptive et fonctionnelle des structures intonologiques, est un indicateur puissant de la hiérarchie des constituants syntaxiques de l'énoncé (DI CRISTO I975) En accord avec cette hypothèse, on pourrait supposer que lors de l'intégration auditive, il est procédé à une simple interpolation entre les cibles prosodiques qui reflètent les instructions linguistiques émanant des centres supérieurs (DI CRISTO, I976; HIRST, I980). D'après notre seconde hypothèse, les cibles prosodiques seraient programmées en termes de syllabes. Il est apparu, en effet, dans nos travaux (DI CRISTO, I977) que les instructions linguistiques qui engendrent les variations paramétriques des cibles prosodiques sont déclenchées dès le début de l'émission de la consonne, dans une syllabe CV par exemple (1). Nous savons, d'autre part, que parmi les variations paramétriques qui accompagnent la réalisation des cibles prosodiques, certaines reflètent diverses contraintes inhérentes à l'émission de la parole (DI CRISTO, I978).

Le but de cette recherche est d'analyser les variations qui dépendent de ces contraintes et d'évaluer les incidences de cette variabilité sur la perception des cibles prosodiques.

2. - PROTOCOLE EXPERIMENTAL.

Dans cette recherche préliminaire, nous avons volontairement limité nos investigations à l'étude des réalisations de l'intonème progrédient (2) dans les syllabes CV et CVC. Deux corpus ont

---

(1) Voir les résultats des travaux de HIRST (HIRST, I980) sur l'anglais qui donnent lieu à des conclusions similaires.

(2) cf. DI CRISTO et CHAFCOULOFF (I980) : l'intonème progrédient en français : caractéristiques intrinsèques et extrinsèques, à paraître dans les Hommages à Georges FAURE, Studia Phonetica DIDIER. Nous appelons intonème progrédient l'unité mélodique à valeur continuative qui est réalisée au terme des groupes intonatifs non terminaux

été conçus dans ce but : le premier comprend des mots bissyllabiques dont la syllabe finale CV est formée de l'une des consonnes suivantes [p, t, k, b, d, g, f, s, ʃ, v, z, ʒ] et de [a] ou [i] . Le second est constitué de bissyllabes CVC comprenant des voyelles identiques à celles du premier corpus, tandis que les consonnes C1 et C2 sont représentées par les occlusives [t, d] et par les constrictives [s, z]. Des combinaisons différentes permettent d'obtenir des contextes consonantiques symétriques [t-t] , [d-d] etc... ou dissymétriques [t-s] ,[d-t], [z-s] , etc... Les mots appartenant aux corpus 1 et 2 sont intégrés dans un énoncé dont la structure demeure invariable : "il a prononcé le mot - quand il a parlé". Les énoncés du premier corpus ont été enregistrés 5 fois et ceux du second corpus 3 fois, par 3 sujets. Les plans d'expérience des corpus I et 2 sont donc respectivement les suivants :

(I) $L_3 * C_3 * O_2 * X_2 * T_2 * R_5$

(avec L : les locuteurs; C : les consonnes prises individuellement; O : la distinction occlusive constrictive; X : le voisement de C1; T : le type de voyelle (haute basse) et R : les répétitions).

(2) $L_3 * O_2 * X_2 * Q_2 * Z_2 * T_2 * R_3$

(avec L : les locuteurs; O : le mode occlusif ou constrictif de C1; X : le voisement de C1; Q : le mode occlusif ou constrictif de C_2.; Z : le voisement de C_2; T : le type vocalique (voyelle haute/voyelle basse) et R : les répétitions).

L'analyse des paramètres acoustiques a été réalisée à l'aide du programme MELINT mis au point par R. ESPESSER en I978.

3. - RESULTATS.

A) Description des configurations de Fo et d'intensité .

La première étape de ce travail, a pour objet l'étude de la variabilité des configurations acoustiques de Fo et d'intensité par lesquelles se réalise l'intonème. Les configurations de Fo du contour se répartissent en deux classes. La première comprend les configurations directes (D), caractérisées par une pente positive régulière. La seconde est constituée des configurations convexes (C) : pente plus rapide dans la partie initiale qu'à la fin du contour (figure 1). Le tableau I montre que la forme de la configuration est fortement influencée par la nature de la consonne C 1.On observe qu'une consonne initiale voisée entraîne la réalisation d'une configuration convexe, alors qu'une consonne non voisée engendre une montée directe. Cette bipartition s'explique aisément si l'on considère les effets microprosodiques décrits par plusieurs chercheurs (LEHISTE and PETERSON, I96I; LEA, I973; GANDOUR, I974; JEEL, I975; HOMBERT and LADEFOGED, I976; DI CRISTO, I977; DI CRISTO et CHAFCOULOFF, I977). Nous savons, à la suite de ces travaux, qu'une consonne initiale non voisée provoque une augmentation importante de la Fo initiale de la voyelle subséquente, tandis qu'une consonne voisée produit l'effet inverse.

Le relèvement de la Fo initiale accentue le caractère direct de la configuration ascendante. Il peut même engendrer parfois la réalisation d'un modelé concave. Lorsque la consonne initiale est voisée, la voyelle subséquente est affectée d'une Fo initiale relativement basse, ainsi que d'une configuration fortement ascendante dans sa première partie, ce qui contribue, toutes choses égales par ailleurs, à accroître la convexité du contour.

En ce qui concerne les configurations d'intensité du contour, nous avons observé qu'elles peuvent également être réparties selon deux classes. Celles de la première se différencient par leur stabilité (absence de glissement), c'est pourquoi nous les qualifions de "tracés en plateau" (P). Celles de la seconde se caractérisent par la présence d'un glissement négatif (N) dont l'ampleur est variable (de 4 à 7 dB, en moyenne, pour les différents locuteurs).

Compte tenu du caractère indissociable des paramètres Fo et intensité dans l'intégration perceptive des cibles prosodiques, il nous a paru intéressant d'étudier conjointement leur variabilité. Cette démarche nous a conduit à regrouper les réalisations dans les quatre catégories suivantes :

- D/N : association d'un glissando direct de Fo et d'une chute d'intensité
- D/P : association d'un glissando direct de Fo et d'un plateau d'intensité
- C/N : association d'une configuration convexe de Fo et d'une chute d'intensité
- C/P : association d'une configuration convexe de Fo et d'un plateau d'intensité.

Les pourcentages respectifs de ces combinaisons, en fonction de la nature de C1 et de celle de la voyelle, sont présentés dans le tableau II. Les résultats montrent que la combinaison DN est plus fréquente lorsque C1 est non voisée, alors que CP et CN sont plus nombreuses lorsque C1 est voisée (effets significatifs à .OI). L'analyse des réalisations de chaque locuteur corrobore cette tendante (fig. 2) dont nous tenterons d'expliquer les fondements dans la dernière partie de notre étude.

B) Estimation des effets intrinsèques et co-intrinsèques dans la réalisation des différents paramètres.

Nous nous proposons d'étudier plus précisément dans la seconde étape de ce travail, les effets du contexte consonantique et de la nature de la voyelle sur la réalisation des paramètres Fo et durée. Les valeurs retenues pour cette analyse sont les suivantes (fig. 3) :

- Fo initiale du contour : Fo INI.
- Fo au point de hauteur du contour : Fo 2/3 (ROSSI, I97I, I978).
- Fo au terme du contour : Fo FIN.
- durée de la voyelle tonique : d V
- Fo de la voyelle prétonique : Fo VP.

a) Fo initiale du contour (Fo INI).

L'examen du tableau III fait apparaître que la variation moyenne de Fo INI en fonction du voisement de CI est de 22 % pour le corpus I et de I8 % pour le corpus 2 (corpus I : $F_{(I-4)} = 422, \alpha < .OOI$., corpus 2 = $F_{(I-2)} = 2873, \alpha < .OOI$). On observe également que la Fo INI du contour est plus élevée pour [i] que pour [a] (effet significatif à .OI dans le corpus I et à . 05 dans le corpus 2). D'autre part, il convient de préciser que l'interaction des facteurs : voisement de C_1 et nature de V est significative pour les 2 corpus (à .OOI et .05, respectivement). On peut donc en déduire que l'effet du voisement est plus accusé pour la voyelle haute que pour la voyelle basse.

Le caractère occlusif ou constrictif de C_1 affecte aussi les valeurs de Fo INI. Nous remarquerons effectivement dans les tableaux III et IV que les écarts de Fo INI dus au voisement de C1 sont sensiblement plus importants avec les constrictives qu'avec les occlusives.

Toutefois, les effets relatifs à la nature de V et au mode occlusif ou constrictif de C1 demeurent mineurs comparativement à ceux du voisement qui donnent lieu à des écarts moyens de Fo INI voisins de 20 %.

b) Fo 2/3 et Fo FIN

L'intonème progrédient est réalisé par un glissando positif de Fo dont la configuration et l'étendue varient en fonction de divers facteurs. Il nous a paru intéressant de chercher à définir les marges de dispersion des valeurs atteintes au terme du glissando (Fo FIN) et au point de hauteur (1), qui se situe approximativement aux deux tiers de la montée (ROSSI, I97I). Nous avons calculé, dans ce but, la moyenne (m), l'écart-type (σ), le coefficient de variation (σ/m) et les limites de confiance (1.c) des valeurs de ces points. Afin d'isoler les effets intrinsèques dus à la voyelle, nous avons calculé séparément les valeurs de [i] et celles de [a] . Le tableau IV, dans lequel est consigné les résultats, montre que les valeurs de σ, de σ/m et de 1.c. restent faibles. La valeur moyenne du rapport σ/m est inférieure à I/2 ton (soit moins de 6 %). Ce résultat paraît indiquer que, pour une voyelle donnée, les marges de variation du point de hauteur et de la fin du contour demeurent étroites et inférieures au seuil différentiel (2) de fréquence fondamentale défini par ROSSI et CHAFCOULOFF (I972 a). En revanche, les latitudes des variations qui dépendent de la nature de la voyelle (fréquence intrinsèque) sont plus importantes (tableau V). Nous retiendrons, par exemple, que les écarts de Fo au point de hauteur sont supérieurs au seuil différentiel de fréquence, mais qu'ils restent inférieurs à la dynamique tonale du niveau infra-aigu , qui est de 3 demi-tons (3). Les résultats des analyses de variance (4) révèlent que seul le facteur T (nature de la voyelle) est significatif dans tous les cas (5). Il existe, d'autre part, une interaction significative entre les facteurs T et L (locuteurs) qui indique que l'ampleur des variations intrinsèques dépendant de la nature de V n'est pas la même pour tous les sujets.

Il ressort nettement de ces analyses que le contexte consonantique n'affecte pas de façon significative les valeurs du point de hauteur du contour. Ce résultat peut paraître surprenant à première vue, car il a été montré, dans de nombreuses recherches, que le contexte consonantique exerce une influence importante sur la Fo des voyelles adjacentes (HOUSE and FAIRBANKS, I953; MOHR, I97I; LÖFQVIST, I975; DI CRISTO et CHAFCOULOFF, I977; NISHINUMA, I978). L'absence d'effet constaté dans notre étude nous conduit à penser qu'il s'agit d'un phénomène propre à la réalisation de l'intonème.

---

(1) Nous appelons "point de hauteur" la valeur finale perçue du glissando de fréquence.

(2) La valeur du seuil différentiel de fréquence fondamentale $\Delta \frac{Fo}{Fo}$ retenue ici est de 6 %.

(3) Voir à ce sujet : ROSSI et CHAFCOULOFF (I972 b) : Les niveaux intonatifs, p. I74.

(4) Cf. infra les plans d'expériences : Protocole expérimental.

(5) Corpus 1 : Fo 2/3 = F (I,4) = 228, $\alpha <$ .OOI., Fo FIN = F (I,4) = I40, $\alpha <$ .OOI. Corpus 2 : Fo 2/3 : F (I,2) = 278, $\alpha <$ .OI., Fo FIN : F (I,2) = I83, $\alpha <$ .OI.

c) Durée de la voyelle tonique (dV).

A l'inverse de Fo, la durée de la voyelle du contour est soumise à de multiples influences. Si l'on considère, par exemple, les résultats de l'analyse de variance des données du second corpus, nous constatons que les facteurs O (caractère occlusif ou constrictif de C1) et Q (caractère occlusif ou constrictif de C2) sont significatifs au seuil $\alpha < .01$ et que les facteurs : X (voisement de C1) Z (voisement de C2) et T (nature de la voyelle) le sont au seuil $\alpha < .001$. Nous ne pouvons, faute de place, nous attarder à commenter ici toutes les données d'une analyse très complexe, dont les détails seront présentés dans une étude ultérieure. (figure 4).

d) Etude des corrélations entre les différentes variables.

Nous avons calculé, dans la dernière étape de notre travail, les matrices d'inter-corrélation entre les variables : Fo INI., Fo 2/3 (1), Fo FIN, dV et Fo VP (tableau VI). Nous notons un nombre élevé de corrélations négatives entre FO INI et dV (variables 2 et 5). Ce résultat s'explique facilement, car l'accroîssement de Fo INI est dû à la présence d'une consonne C1 non voisée qui abrège la voyelle.

La corrélation positive entre Fo 2/3 et Fo FIN (variables 3 et 4) qui s'avère significative dans tous les cas, n'appelle aucun commentaire.

Plus intéressante à considérer est la corrélation positive entre Fo 2/3 et Fo VP (3 et I) dont la valeur significative dans de nombreux cas dénote l'existence d'un lien étroit entre ces deux variables.

Si nous prenons un seuil de probabilité égal ou supérieur à .01, nous ne relevons qu'un seul cas de corrélation significative entre Fo FIN et dV (4 et 5) et aucune corrélation entre Fo 2/3 et dV (3 et 5). Ces faits confirment ainsi notre conclusion du paragraphe (b), à savoir que les valeurs finales du contour (Fo 2/3 et Fo FIN) demeurent insensibles aux variations de la durée de V qui sont conditionnées par le contexte consonantqiue (fig. 5).

4. - DISCUSSION

Nous avons observé, au cours de l'interprétation des résultats de l'analyse expérimentale, que la réalisation de l'intonème progrédient est sujette à une forte variabilité. Cette dernière se manifeste notamment par la diversité des configurations du contour et par l'importance des marges de dispersion des valeurs des paramètres prosodiques (Fo, durée, intensité).

La question posée est alors de savoir comment cette variabilité peut être compatible avec les contraintes de l'intégration auditive des cibles prosodiques qui reflètent les instructions linguistiques et la compétence intonologique du locuteur-auditeur.

Dans la première partie de notre travail, nous avons montré que l'association (DN) du glissando de Fo direct (D) et du glissement négatif d'intensité (N) s'observe généralement dans le contexte (C1 [- voisée] + V). En revanche, les associations (CP) et (CN) demeurent les plus fréquentes dans le contexte (C1 [+ voisée] + V). Nous savons qu'un glissement négatif d'intensité, même non perceptible (ZWICKER, I962), modifie la perception d'un glissando de fréquence croissant ou décroissant (ROSSI, I978). Ce phénomène constitue la base d'une explication plausible des

(1) Fo 2/3 = Fo Ph c'est à dire Fo au point de hauteur du glissando.

interactions de Fo et d'intensité dont nous venons de rendre compte. En ce qui concerne les réalisations (CN) et (CP) (1), comme le contour de Fo se termine généralement par un palier d'une durée égale ou supérieure à 40 msec., on peut supposer que la modulation d'intensité n'affecte pas la perception du glissando dont le terme est intégré grâce à l'effet d'ancrage qu'exerce le palier final. En revanche, le glissement négatif d'intensité atténue la perceptibilité de la partie finale du contour dans (DN). Nous avons pu établir (DI CRISTO, 1978) qu'en français la consonne non voisée provoque, comme c'est le cas dans de nombreuses langues, une augmentation sensible de la Fo globale de la voyelle subséquente. Si l'on admet, par ailleurs, que les niveaux intonatifs (ROSSI et CHAFCOULOFF, 1972 b) assument un rôle décisif dans l'identification des cibles prosodiques, on conviendra que les marges de variation de ces niveaux doivent être relativement étroites. Compte tenu de ces remarques, nous pouvons émettre l'hypothèse que le glissement négatif d'intensité qui s'observe régulièrement dans le contexte (C1 [- voisée] + V) serait un effet compensatoire destiné à réduire, sur le plan perceptif, les dépassements éventuels (overshoot) du point de hauteur qui sont occasionnés par les consonnes initiales non voisées. (2).

Le second fait saillant qui émerge de notre analyse est que les variations intrinsèques et co-intrinsèques de la durée du contour n'exercent pas une influence notable sur la réalisation du point de hauteur. Ce dernier se place régulièrement dans le niveau infra-aigu (fig. 6) et varie uniquement en fonction de la nature de la voyelle (fréquence intrinsèque). Nous avons fait remarquer plus haut que l'ampleur de cette variation demeure toutefois inférieure à la dynamique tonale de l'infra-aigu qui est de 3 demi-tons.

L'absence d'effet du contexte consonantique sur le point de hauteur que nous relevons dans notre étude, paraît être en contradiction avec les résultats obtenus par d'autres chercheurs, notamment avec ceux de ERIKSON and ALSTERMARK (1972) ou de LÖFQVIST (1975). Ces derniers constatent en effet dans leurs travaux que la réalisation des contours du suédois est fortement influencée par les variations du contexte consonantique et de la durée vocalique. C'est ainsi, par exemple, que la présence d'une consonne non voisée abrège la voyelle précédente et provoque une troncation du contour de Fo de cette voyelle. Si la règle de troncation s'appliquait au français, on s'attendrait à ce que les valeurs de Fo FIN soient systématiquement plus élevées dans les contextes tVz et sVz que dans dVt et zVt . Or, cela est loin d'être le cas ainsi que le révèlent les données du tableau VII.

L'invalidité de la règle de troncation en ce qui concerne les exemples de notre corpus nous fait pencher en faveur d'une autre hypothèse : celle du réajustement de la pente du contour. On peut penser, en effet, que si les valeurs de Fo Ph (3) et de Fo FIN ne dépendent pas du contexte, c'est peut être parce qu'un effet de réajustement est mis en oeuvre pour modifier la pente du glissando en fonction de l'entourage consonantique. Dans cette éventualité, la pente du glissando devrait être relativement lente quand C 1 est [-voisée]. Comme cette consonne provoque

---

(1) CP : glissando positif de Fo convexe et plateau d'intensité,
CN : glissando positif de Fo convexe et glissement négatif d'intensité

(2) Il convient toutefois de considérer cette hypothèse avec prudence, car il est possible que le phénomène observé soit dû simplement à un facteur physiologique dont nous ignorerions la nature exacte.

(3) Fo Ph : Fo du point de hauteur, correspondant approximativement aux 2/3 de la durée du contour.

une hausse importante de la Fo INI de la voyelle subséquente, une pente trop rapide entraînerait inévitablement le dépassement du point de hauteur requis pour la réalisation de la cible prosodique. Par contre, la pente devrait être plus rapide quand CI est [+ voisée] (principalement quand C2 est [-voisée]), car s'il en était autrement, la valeur du point de hauteur correspondant au niveau intonatif de la cible ne serait pas atteinte. Le calcul de la pente du contour dans les contextes (C1 [-voisée] + V) et (C1 [+ voisée] + V) confirme l'hypothèse du réajustement. Il se trouve effectivement, comme le montre le tableau VIII, que cette pente est environ deux fois plus rapide pour les voyelles qui sont précédées d'une consonne voisée. Nous constatons, en outre, (tableau IX) que les pentes les plus lentes se réalisent quand C2 est une consonne dite allongeante, comme [v] ou [z]. Si la pente était rapide dans ces contextes, il devient évident que les valeurs de Fo Ph et de Fo INI seraient franchement supérieures à celles des autres réalisations. Cette dernière observation corrobore ainsi le bien-fondé de l'hypothèse du réajustement.

Malgrè l'importance des variations paramétriques qui affectent la réalisation des cibles prosodiques, le point de hauteur demeure d'une grande stabilité et se réalise toujours dans le niveau infra-aigu. La constance de ce phénomène, qui s'exprime indépendamment des locuteurs, constitue un invariant perceptif grâce auquel l'intonème progrédient est identifié. La stabilité du point de hauteur est obtenue par la mise en oeuvre, de la part du locuteur-auditeur, de divers processus d'ajustement, qui pourraient être décrits à l'aide de règles prosodicotactiques (DI CRISTO, I978) et qui relèvent essentiellement de la compétence du sujet parlant. Parmi ces processus d'ajustement, nous avons noté, par exemple, l'interaction de Fo et de I (intensité) ainsi que la neutralisation des effets du contexte consonantique et de la durée sur Fo. On peut également penser que l'effacement perceptif des variations intrinsèques de Fo, qui pourraient représenter une autre source de variabilité du point de hauteur, s'opère dans des conditions similaires à celles qui ont été mises en évidence par ROSSI (I97I b) pour l'intensité et par PETERSEN (I974) pour la durée.

Dans les recherches intonologiques, l'interprétation des indices et des traits ne peut être fondée directement sur l'observation des données objectives qui traduisent simultanément les instructions linguistiques et diverses formes de contraintes. Cette interprétation nécessite une stylisation préliminaire (DI CRISTO et al. I979 , NISHINUMA et ROSSI I979), qui consiste essentiellement à effacer les variations microprosodiques et à procéder au transcodage perceptif des données acoustiques. Toutefois, il importe de prendre en compte, pour que cette démarche reste valable, les stratégies d'ajustement que nous avons décrites dans cette étude.

REFERENCES BIBLIOGRAPHIQUES

- DI CRISTO A.(I975) : Recherches sur la structuration de la phrase française, Actes VIèmes Journées d'Etudes sur la Parole (Toulouse), I : 96-II6.

- DI CRISTO A.(I976) : Application d'un modèle d'analyse prosodique à l'étude du vocatif, Trav. Inst. Pho. Aix, 3 : 2I3-358.

- DI CRISTO A.(I977) : Les Faits Microprosodiques. Thèse de Doctorat de 3ème Cycle (Université de Provence) : 359p.

- DI CRISTO A.(I978) : De la Microprosodie à l'Intonosyntaxe, Thèse de Doctorat d'Etat, (Université de Provence) : I274 p.

- DI CRISTO A.et CHAFCOULOFF M.(I977) : Les faits microprosodiques du français : voyelles, consonnes et coarticulation, Actes VIIIème Journées d'Etudes sur la Parole (Aix en Provence), I : I47-58.

- DI CRISTO A., ESPESSER R., et NISHINUMA Y. (I979) : Présentation d'une méthode de stylisation prosodique, Comm. IXème Congrès Internat. Sciences Phonétiques (Copenhague).

- DI CRISTO A. et CHAFCOULOFF M. (I980) : L'intonème progrédient en français : caractéristiques intrinsèques et extrinsèques, Hommage à G. FAURE, Studia Phonetica (sous presse).

- ERIKSON Y. and ALSTERMARK M. (I972) : Fundamental frequency correlates of the grave word accent in Swedish. The effect of vowel duration, Q.P.S.R. Speech Transmission Lab. Stockholm, 2-3 : 53-60.

- GANDOUR J. (I974) : Consonant types and tone in siamese, J. of Phonetics, 2 : 337-50.

- HIRST D. J. (I980) :L'estimation des paramètres des trajectoires de fréquence fondamentale. Comm. présentée aux XIèmes Journées d'Etudes sur la Parole (Strasbourg).

- HOMBERT J. M. and LADEFOGED P. (I976) : The effect of aspiration on the fundamental frequency of the following vowel, U.C.L.A. Working Papers in Phonetics, 36.

- HOUSE A. J. and FAIRBANKS G. (I953) : The influence of consonant environment upon the secondary acoustical characteristics of vowels, J.A.S.A., 25 : I05-I3.

- JEEL V. (I975) : An investigation of the fundamental frequency of vowels after various Danish consonants, in particular stop consonants, A.R.I.P.U.C., 9 : I9I-2II.

- LEA W. A. (I973) : Segmental and suprasegmental influences of fundamental frequency contours, in : Consonant Types and Tone : I7-70.

- LEHISTE I. and PETERSON G. E. (I96I) : Some basic considerations in the analysis of intonation, J.A.S.A., 33 (4) : 4I9-25.

- LÖFQVIST A. (I975) : Intrinsic and extrinsic Fo variations in Swedish tonal accents, Phonetica, 3I : 228-47.

- MOHR B. (I97I) : Intrinsic variations in the speech signal, Phonetica, 23 : 65-93.

- NISHINUMA Y. (I979) : Un modèle d'analyse automatique de la prosodie Editions du C.N.R.S.

- NISHINUMA Y. et ROSSI M. (I979) : Essai d'automatisation de l'analyse prosodique du français, Comm. IXème Cong. Int. Sciences Phonétiques (Copenhague).

- PETERSEN N. R. (I974) : The influence of tongue height on the perception of vowel duration in Danish, A.R.I.P.U.C., 8 : I-IO.

- ROSSI M. (I97I a) : Le seuil du glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonetica, 23 : I-33.

- ROSSI M. (I97I b) : L'intensité spécifique des voyelles, Phonetica, 24 : I29-6I.

- ROSSI M. (I978) : Interaction of intensity glides and frequency glissandos : In Honor to D.B. FRY (sous presse).

- ROSSI M. et CHAFCOULOFF M. (I972 a) : Recherches sur le seuil différentiel de fréquence fondamentale dans la parole, Trav. Inst. Pho. Aix, I : I77-84.

- ROSSI M. et CHAFCOULOFF M; (I972 b) : Les niveaux intonatifs, Trav. Inst. Pho. Aix. I : I67-76.

- ZWICKER E. (I962) : Direct comparison between the sensations produced by frequency modulation and amplitude modulation, J.A.S.A., 34 (8) : I425-30.

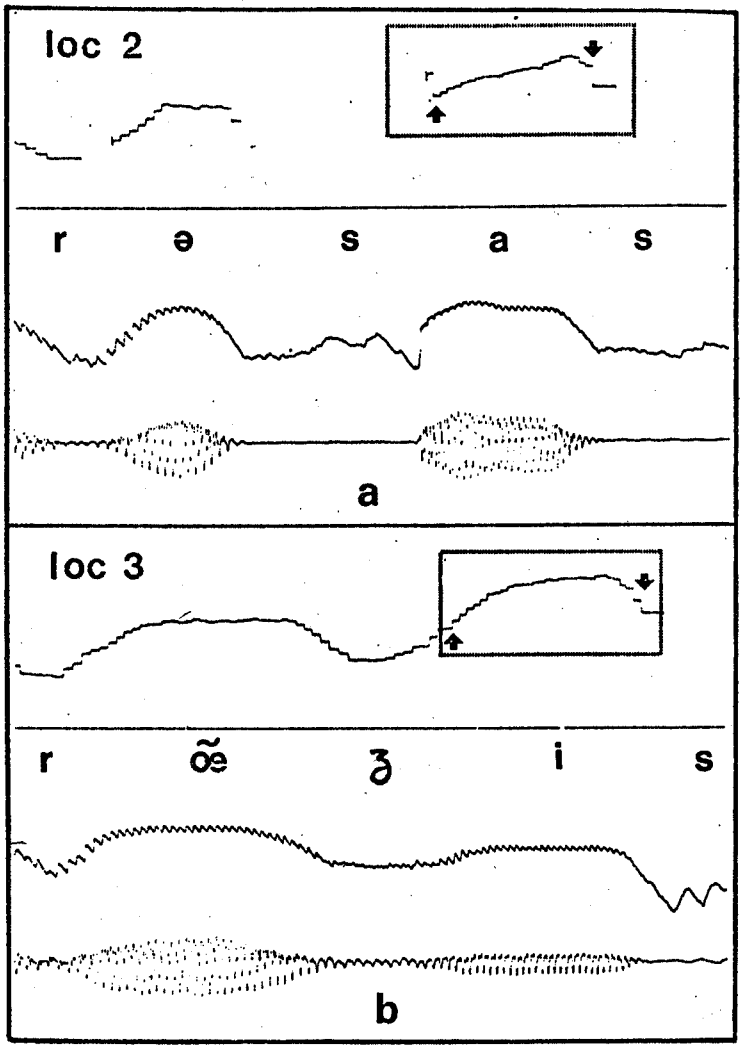


Figure 1. Exemples de configurations de Fo correspondant à diverses réalisations de l'intonème progrédient. a) configuration directe. b) configuration convexe.

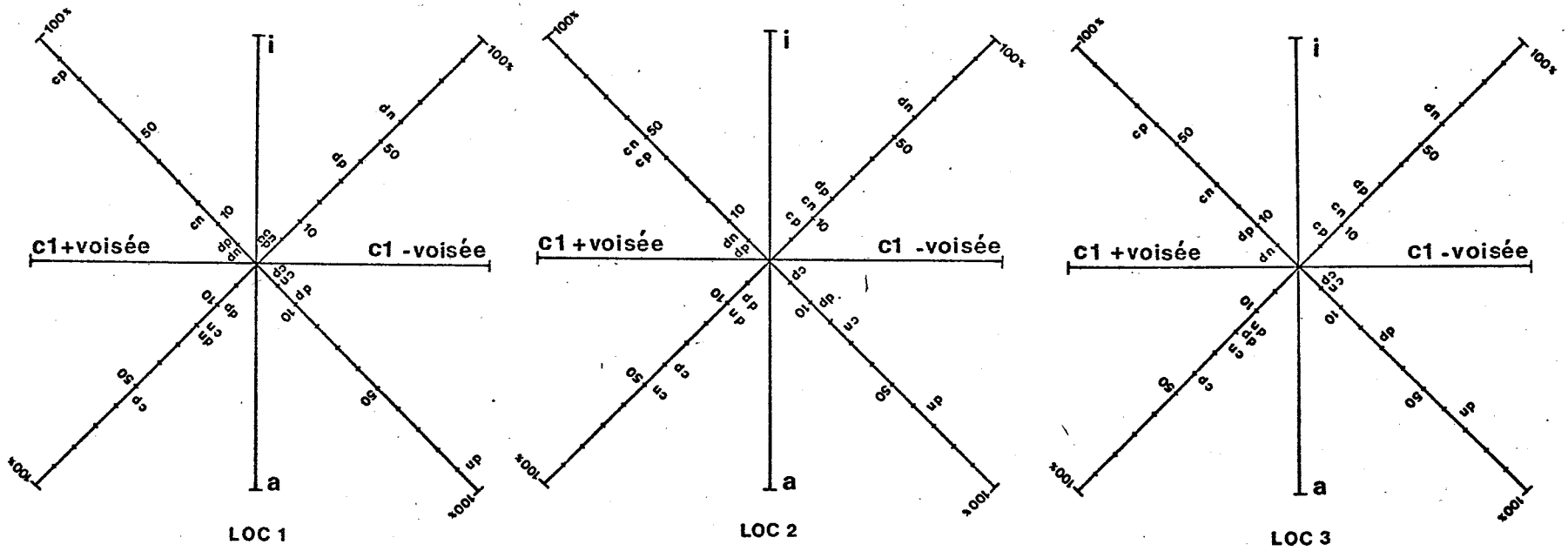


Figure 2. Pourcentages des réalisations DN, DP, CN et CP en fonction de la nature de V et du voisement de C1.

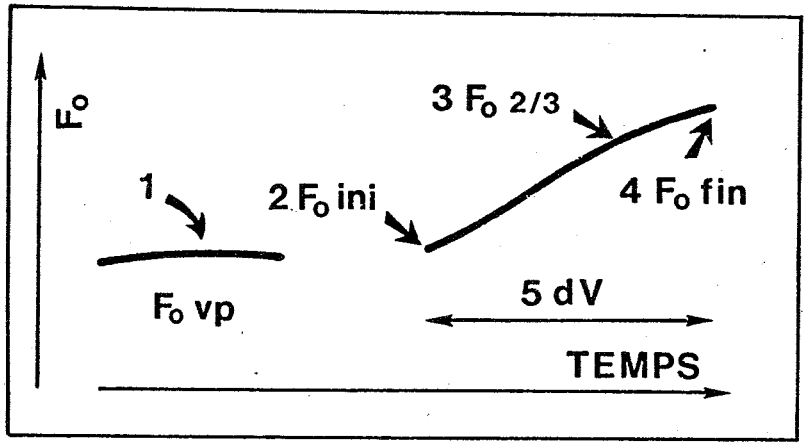


Figure 3. Illustration des paramètres retenus dans l'étude acoustique:

1.- Fo VP: Fo de la voyelle précédent le contour (prétonique).
2.- Fo INI: Fo à l'initiale du contour.
3.- Fo 2/3: Fo au point de hauteur (PH) du contour.
4.- Fo FIN: Fo au terme du contour.
5.- dV: durée de la voyelle du contour.

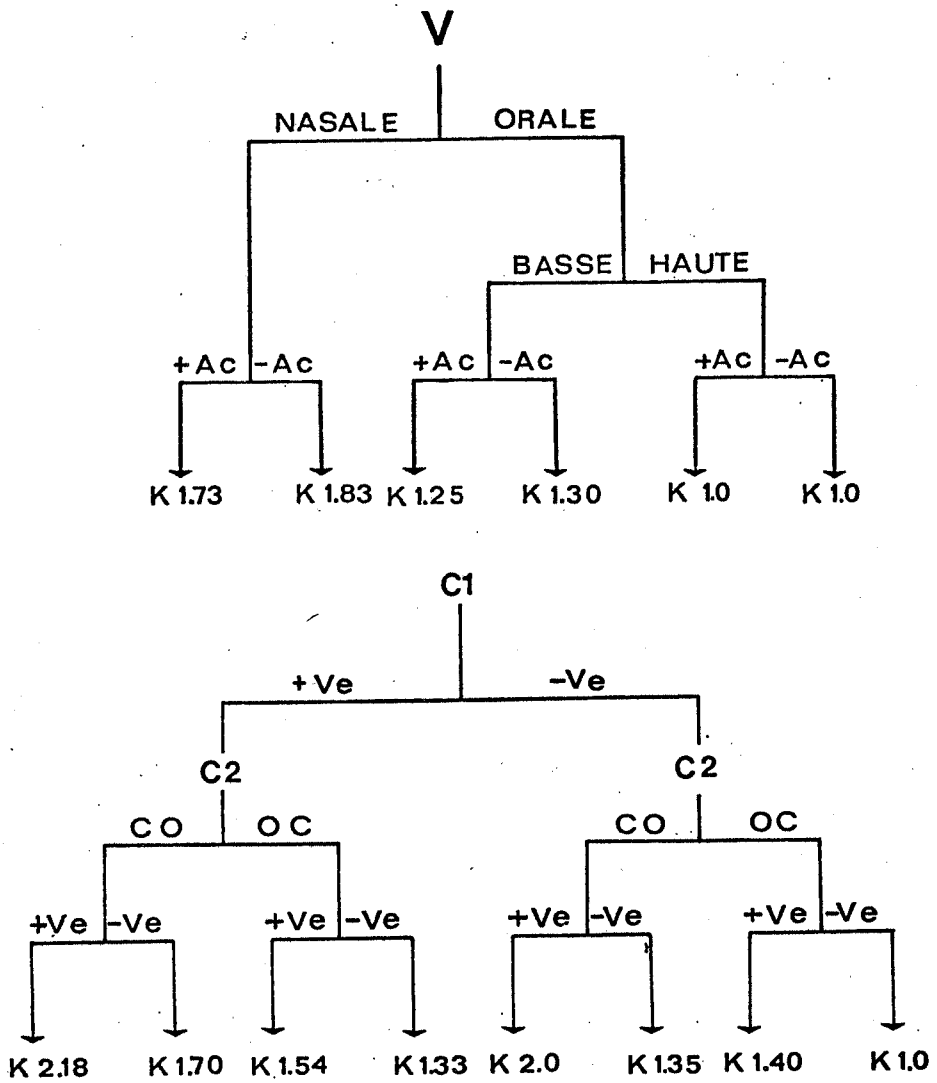


Figure 4. Variabilité de la durée de la voyelle du contour en fonction de son mode d'articulation et des consonnes adjacentes. Les coefficients qui indiquent l'allongement par rapport à une valeur de référense doivent se lire de droite à gauche (d'après DI CRISTO, 1978).

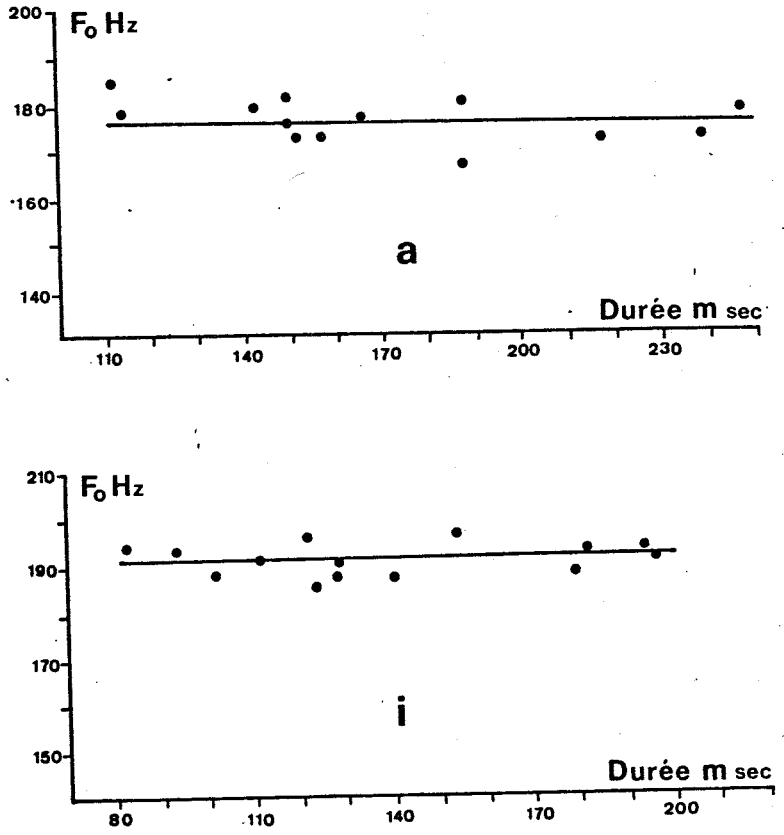


Figure 5.- Relations entre les valeurs de Fo (Fin du glissando) et les durées du contour (la figure illustre l'absence de corrélation entre les deux paramètres).

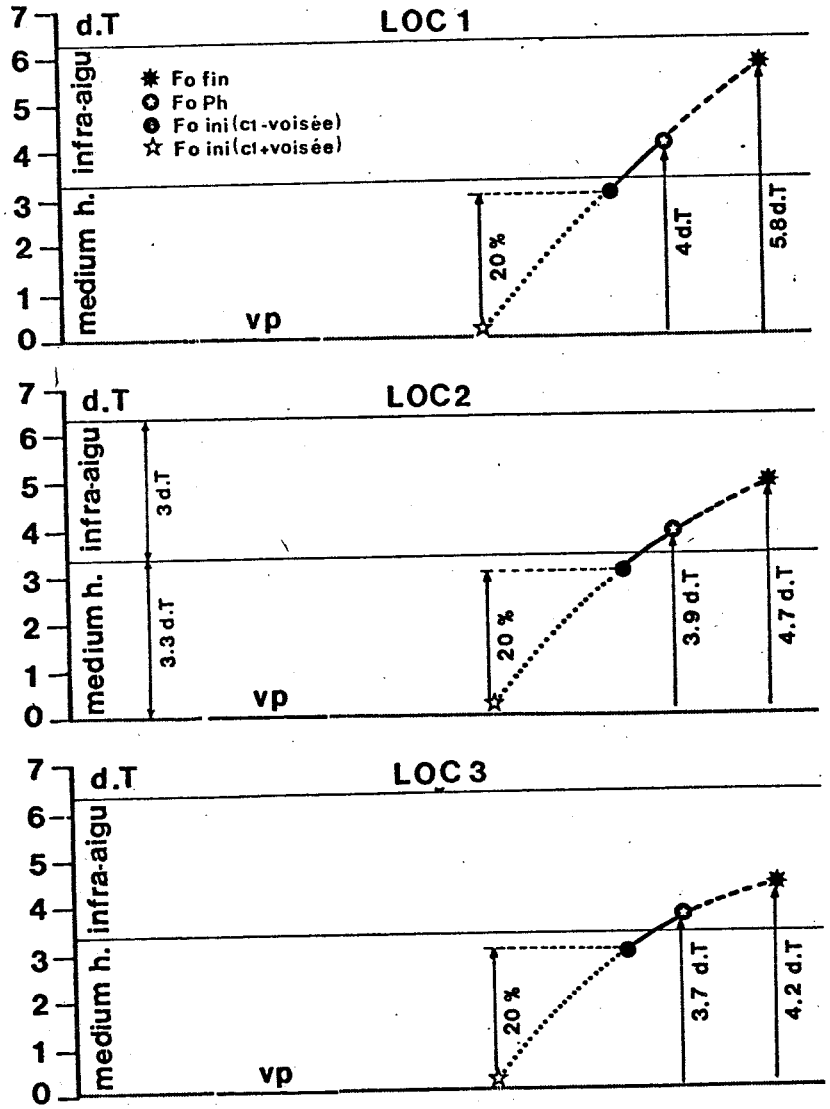


Figure 6.- Réalisations du contour de l'intonème progrédient en fonction des niveaux intonatifs (après effacement des effets intrinsèques).Les intervalles sont exprimés en demi-tons (d.T).

CONFIGURATION	C 	D 
C_1 + Voisée	86	12
C_1 - Voisée	14	86

Tableau I. Pourcentages de réalisation des configurations de Fo: convexes (C) et directes (D) en fonction de la nature de C1.

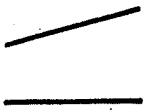
CONFIGURATION Fo / I	D N 	D P 	C N 	C P 
C_1 + Voisée	9	6.5	31	54
C_1 - voisée	67	21	9	3

Tableau II. Pourcentages de réalisation des configurations de Fo et d'intensité (I) en fonction de la nature de C1 dans les syllabes CV des contours. D: configuration de Fo directe; C: configuration de Fo convexe; N: configuration d'intensité négative; P: configuration d'intensité en plateau.

	CORPUS 1	CORPUS 2	CORPUS 1 + 2
OCCLUSIVES	21	16	18.5
CONSTRICTIVES	24	21	22.5
a	18	17	17.5
i	26	20	23
TOUS CONTEXTES	22	18	20

Tableau III. Rapports moyens (en %) entre la Fo INI des voyelles précédées d'une consonne non voisée et celle des voyelles précédées d'une consonne voisée (ensemble des contextes et des locuteurs).

CORPUS 1 — F_0 2/3

	i				a			
	m	σ	σ/m	l.c.	m	σ	σ/m	l.c.
LOC1	215	11	5	4	191	5	5	3.4
LOC2	177	8	4.5	3	174	8	4.5	3
LOC3	166	9	5.4	3	149	10	6.7	3

CORPUS 1 — F_0 FIN

	i				a			
	m	σ	σ/m	l.c.	m	σ	σ/m	l.c.
LOC 1	219	11	5	4	199	10	5	3.4
LOC 2	181	8	4.4	3	180	8	4.4	3
LOC 3	170	9	5	3	155	9	6	3

CORPUS 2 — F_0 2/3

	i				a			
	m	σ	σ/m	l. c.	m	σ	σ/m	l. c.
LOC 1	219	10	4.5	4.3	186	11	6	4.6
LOC 2	183	6	3.3	2.5	171	6	3.5	2.5
LOC3	147	7	5	3.3	137	7	5	3

CORPUS 2 — F_0 FIN

	i				a			
	m	σ	σ/m	l. c.	m	σ	σ/m	l. c.
LOC1	230	10	4.5	4.3	202	12	6	5
LOC 2	186	6	3.2	2.5	178	6	3.4	2.5
LOC 3	155	8	5	3.5	148	9	6	3.8

Tableau IV. Moyenne (m), écart-type (σ), coefficient de variation (σ /m) et limites de confiance (l.c.) des points Fo 2/3 (PH) et Fo FIN du contour. Toutes les valeurs sont exprimées en Hz, à l'exception de (σ /m) qui est exprimée en pourcentages. Les limites de confiance sont calculées en prenant dans la table de "t" la valeur seuil: .01.

	CORPUS 1	CORPUS 2	CORPUS 1 + 2
Fo INI	7	6	6.5
Fo P. hauteur	9.5	11	10
Fo FIN	7.7	8	8

Tableau V. Ecarts intrinsèques moyens de Fo (en %) aux points: Fo INI, Fo PH et Fo FIN du contour (ensemble des locuteurs).

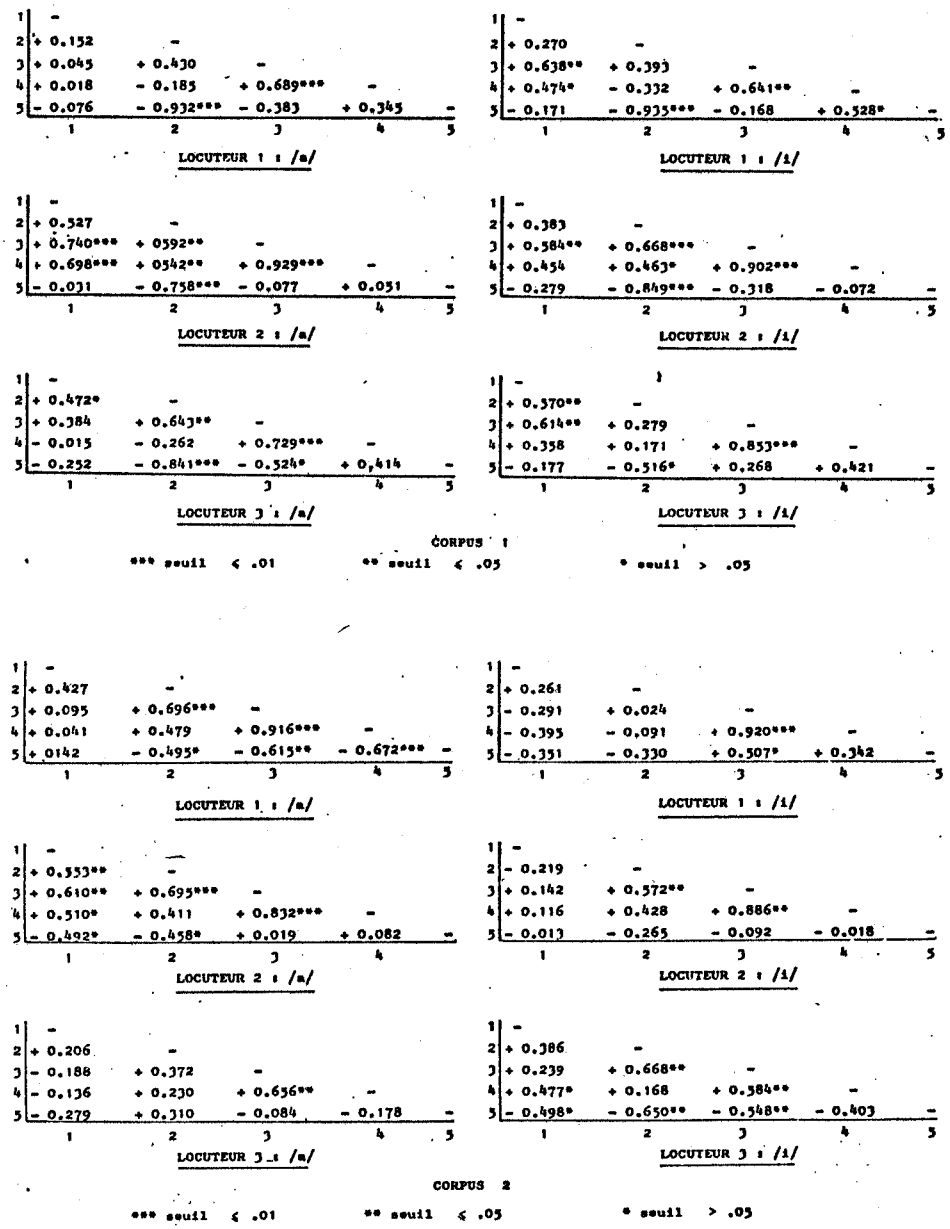


Tableau VI. Matrices d'inter-corrélation entre les variables: 1: Fo VP, 2: Fo INI, 3: Fo 2/3, 4: Fo FIN et 5: dV (cf. la figure 3). Chaque valeur exprime un coefficient de corrélation (r) entre deux variables. Les seuils de signification sont donnés par les astérisques.

LOCUTEURS / SYLLABES	LOCUTEUR 1	LOCUTEUR 2	LOCUTEUR 3	LOCUTEURS 1 à 3
d V t	218	182	159	186
z V t	214	175	155	181
t V z	210	182	159	184
s V z	212	180	148	180

Tableau VII. Valeurs moyennes (en Hz) de Fo FIN dans les contextes: /d V t/, /z V t/, / t V z/ et /s V z/.

		LOCUTEUR 1	LOCUTEUR 2	LOCUTEUR 3	LOCUTEURS 1 à 3
C 1 - VOISEE	P.H.	1.8	2.7	1.8	2.1
	FoFIN	1.8	2.2	1.8	1.9
C 1 + VOISEE	P.H.	6.2	4.9	3.5	4.9
	FoFIN	4.8	3.7	3.4	4.0

Tableau VIII. Valeurs moyennes de la pente du glissando du contour (en Hz/Csec.) en fonction du voisement de C1.

		LOCUTEUR 1	LOCUTEUR 2	LOCUTEUR 3	LOCUTEURS 1 à 3
CORPUS 1	$\binom{t}{s}$ V t	2.2	2.7	1.4	2.1
	$\binom{t}{s}$ V z	1.4	1.6	1.0	1.8
CORPUS 2	$\binom{d}{z}$ V t	5.2	4.1	3.5	4.3
	$\binom{d}{z}$ V z	3.5	2.5	1.9	2.6

Tableau IX. Valeurs moyennes de la pente du glissando du contour (en Hz/Csec.) en fonction du voisement de C1 et de C2.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

VARIATIONS PROSODIQUES INTER ET INTRA LOCUTEURS

MARTIN, Philippe

INSTITUT DE PHONETIQUE
UNIVERSITE DE PROVENCE
EXPERIMENTAL PHONETICS LABORATORY
UNIVERSITY OF TORONTO

RESUME

On a fait lire à 7 locuteurs francophones, d'origines socio-géographiques diverses, 11 phrases correspondant aux 11 différentes hiérarchies prosodiques projectives selon lesquelles peuvent s'organiser 4 mots prosodiques (correspondant à 4 unités syntaxiques accentuées). Ces phrases ont été lues 3 fois à 2 reprises, en demandant aux locuteurs de simuler à chaque fois des émotions correspondant sensiblement à la joie, la colère et la tristesse. Les 462 enregistrements obtenus ont été soumis à l'analyse instrumentale de manière à pouvoir en étudier les réalisations prosodiques, et plus particulièrement les variations mélodiques. Ces réalisations ont été confrontées à chaque fois aux prédictions d'un modèle théorique utilisant les traits ± Extrême, ± Montant, ± Ample, ± Bas pour décrire les caractéristiques des contours mélodiques placés sur les syllabes accentuées des différentes phrases. Les résultats obtenus montrent une variation d'écart par rapport au modèle allant de 6,5 % à 15,7 % selon le locuteur, de 5,3 %) 17,1 % selon le type de hiérarchie syntaxique, et de 7,4 % à 12,2 % selon le type d'émotion.

SUMMARY

Seven native speakers of French from various socio- geographical backgrounds were asked to read 11 different sentences corresponding to 11 different syntactic hierarchies. When reading these sentences, the speakers were asked to simulate 3 emotions related to joy (neutral), anger and sadness. An acoustical analysis of this material was then performed and Fo contours pertaining to each stressed syllabe were considered.

These data were compared to various predictions obtained from a theorical model, which use the features ± Extreme, + Rising, ± Ample and ± Low to describe the contours. The results show a discrepancy from the model ranging between 6,5 % and 15,7 % according to the particular speaker, from 5,3 % to 17,1 % according to the type of syntactic hierarchy, and ranging from 7,4 % to 12,2 % according to the type of emotion.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

VARIATIONS PROSODIQUES INTER ET INTRA LOCUTEURS

MARTIN, Philippe

INSTITUT DE PHONETIQUE
UNIVERSITE DE PROVENCE
EXPERIMENTAL PHONETICS LABORATORY
UNIVERSITY OF TORONTO

INTRODUCTION

Le rôle de la prosodie dans l'indication de la structure syntaxique de la phrase faisant encore l'objet de nombreux débats parmi les phonéticiens et linguistes, l'étude des variantes de réalisations des paramètres prosodiques constitue sans doute un domaine privilégié dans la recherche d'invariants portés par le signal de parole. Comme dans le cas d'autres paramètres phonétiques, les variations des paramètres prosodiques peuvent être observés soit par analyse des dispersions autour de réalisations "moyennes" au sens statistique, soit par l'examen des écarts par rapport à un modèle théorique.

La première méthode, utilisée par exemple par BOE, CONTINI et RAKOTOFIRINGA (1975) paraissant trop dépendante du choix préalable des paramètres étudiés, on a préféré se référer ici aux prédictions d'un dispositif théorique censé rendre compte d'une fonction particulière de la prosodie.

Le modèle théorique utilisé pose l'existence, dans la phrase, d'une structure prosodique a priori indépendante de la structure syntaxique. Les rapports entre ces 2 structures s'expriment:

- par une règle d'accentuation, qui fait correspondre chaque syllabe accentuée (en dehors de l'accent d'insistance) à un mot prosodique;
- par la propriété de projectivité de la hiérarchie prosodique, qui rend impossible la correspondance avec une hiérarchie syntaxique qui serait non projective (type "Marie, le matin, déjeune").

Dans certaines conditions particulières de production du discours, comme la lecture, on peut admettre de plus que les deux hiérarchies prosodique et syntaxique sont congruentes.

En français, les marques qui indiquent la structure prosodique sont manifestées de façon privilégiée par des contours intonatifs apparaîssant sur les syllabes accentuées. Le fonctionnement de ces marques peut être décrit en utilisant, par exemple, les traits:

± Bas (le terme bas se référant au point extrême du contour)

± Montant (sens de variation mélodique du contour)

± Ample (amplitude de variation mélodique du contour)

ce qui permet d'obtenir les relations de phonologiques qui suivent:

Règle du contour final extrême: $\boxed{-\text{Ext} \rightarrow +\text{Ext}}$

le contour final, porté par la dernière syllabe accentuée de la phrase atteint le point mélodique (ie. la fréquence fondamentale) le plus bas (structures prosodiques déclaratives) ou le plus haut (structures prosodiques interrogatives)

Règle d'inversion de pente: $\boxed{-\alpha\ \text{Montant} \rightarrow +\alpha\ \text{Montant}}$

un contour donné est de pente mélodique opposée à celui dont il dépend à droite ($\alpha \pm +$ ou $-$)

Règle d'abaissement d'amplitude ou de niveau:

$$\boxed{\begin{array}{l} +\text{Amp} \leftarrow -\text{Amp} \\ -\text{Bas} \leftarrow +\text{Bas} \end{array}}$$

un contour donné se différentie d'un contour de même pente mais en position différente dans la hiérarchie prosodique et situé à sa gauche par une amplitude ou un niveau plus bas.

Ces règles permettent d'établir la séquence de contours prosodiques correspondant à une structure donnée, chaque contour étant décrit par un faisceau de traits ± Extrême, ± Montant, ± Ample, ± Bas, dont la valeur est attribuée selon la relation entretenue avec ses voisins de droite et de gauche situés plus haut ou au même niveau dans la hiérarchie prosodique.

Ainsi, la hiérarchie ((A) (B)) ((C) (D)) d'une structure prosodique déclarative est indiquée par la séquence de contours

$$\left(\left(\begin{matrix}A\\ \begin{bmatrix}-\text{Ext}\\ -\text{Mt}\end{bmatrix}\end{matrix}\right)\left(\begin{matrix}B\\ \begin{bmatrix}-\text{Ext}\\ +\text{Mt}\\ +\text{A}\end{bmatrix}\end{matrix}\right)\right)\left(\left(\begin{matrix}C\\ \begin{bmatrix}-\text{Ext}\\ \\ -\text{A}\end{bmatrix}\end{matrix}\right)\left(\begin{matrix}D\\ \begin{bmatrix}+\text{Ext}\\ -\text{Mt}\end{bmatrix}\end{matrix}\right)\right)$$

L'application des différentes règles est détaillée comme suit:

1) A B C D

$$\begin{bmatrix}-\text{Ext}\end{bmatrix} \qquad \begin{bmatrix}+\text{Ext}\\ -\text{Mt}\end{bmatrix}$$

Contour final D extrême et descendant
Contour de premier niveau B non extrême

2) A B C D

$$\begin{bmatrix}-\text{Ext}\\ -\text{Mt}\end{bmatrix} \quad \begin{bmatrix}-\text{Ext}\\ +\text{Mt}\end{bmatrix} \qquad \begin{bmatrix}+\text{Ext}\\ -\text{Mt}\end{bmatrix}$$

Contour A de deuxième niveau, de pente opposée au contour B dont il dépend, et non extrême puisque non final

3) A B C D

$$\begin{bmatrix} -\text{Ext} \\ -\text{Mt} \end{bmatrix} \quad \begin{bmatrix} -\text{Ext} \\ +\text{Mt} \\ +\text{A} \end{bmatrix} \quad \begin{bmatrix} -\text{Ext} \\ \\ -\text{A} \end{bmatrix} \quad \begin{bmatrix} +\text{Ext} \\ -\text{Mt} \end{bmatrix}$$

Contour C de deuxième niveau, non extrême et d'amplitude ou de niveau inférieur au contour de niveau supérieur et situé à sa gauche, B

Des contours situés à un même niveau dans la hiérarchie ne peuvent s'opposer par les traits utilisés dans les règles. Ainsi dans une hiérarchie (A) (B) (C) (D), les contours A, B et C doivent être de même pente et présenter la même valeur du trait d'amplitude.

DONNEES

On a fait lire à 7 locuteurs francophones, d'origine socio-géographiques diverses, 11 phrases correspondant aux 11 différentes hiérarchies prosodiques projectives composées de 4 mots prosodiques. Ces phrases ont été lues 3 fois à 2 reprises, en demandant aux locuteurs de simuler à chaque fois les émotions correspondant à la joie, à la colère et à la tristesse. Ces catégories ont été choisies de manière à représenter les émotions "de base" de type +1 , +2 et - établies par Davitz (1969). Les 7 x 11 x 6 = 462 enregistrements obtenus ont été soumis à l'analyse instrumentale (analyseur de mélodie) de manière à pouvoir en étudier les réalisations prosodiques.

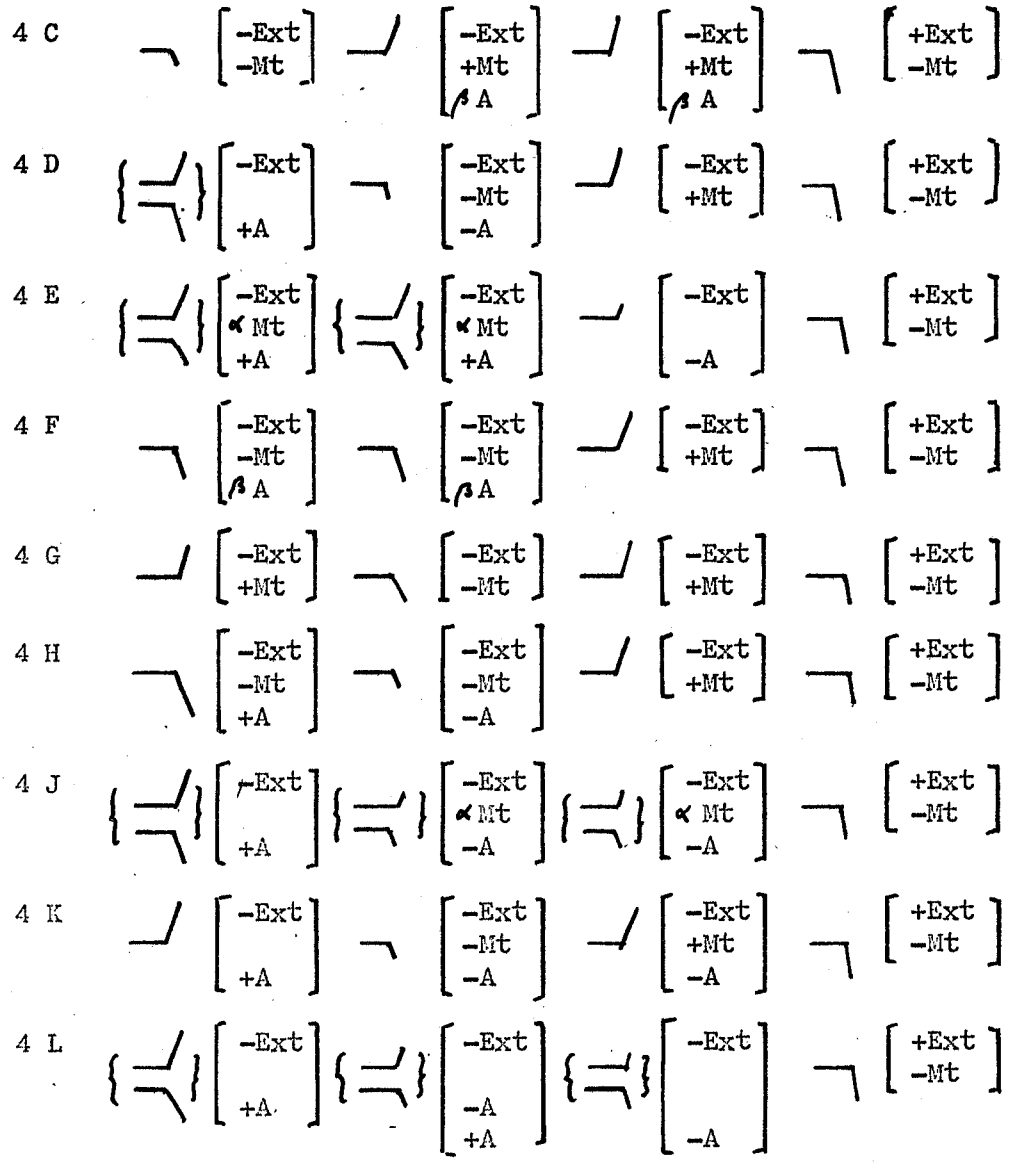
Liste des phrases
Le parenthésage indique la hiérarchie prosodique projective à 4 unités congruentes à la hiérarchie syntaxique.

4 A (Metro) (boulot) (Pernod) (dodo)
4 B ((L'oncle) (de Maurice)) ((mange) (de la confiture))
4 C ((Le frère) (de Marie)) (le matin) (déjeune)
4 D (Maurice) ((le père) (de Julie)) (s'en va)
4 E (Aristide) (chaque_soir) ((avale) (son café))
4 F ((Si Brigitte) (comme je le crains) (revient)) (tu m'appelles)
4 G (((La cousine) (du marin)) (d'en face)) (est partie)
4 H ((Si Alexandre)) ((prend) (ses pilules))) (ça va)
4 J (le matin) ((Patrick) (mal réveillé) (se fâche))
4 K (le lundi) (((le bistrot) (d'en face)) (est fermé))
4 L (Finalement) ((Juliette) ((a pris) (sa valise))

Séquences prosodiques
Séquences de contours correspondant à ces 11 hiérarchies
(α , β = + ou -)

4 A $\left\{ \begin{matrix} \_/ \\ \neg \end{matrix} \right\}$ $\begin{bmatrix} -\text{Ext} \\ \text{Mt} \\ \text{A} \end{bmatrix}$ $\left\{ \begin{matrix} \_/ \\ \neg \end{matrix} \right\}$ $\begin{bmatrix} -\text{Ext} \\ \text{Mt} \\ \text{A} \end{bmatrix}$ $\left\{ \begin{matrix} \_/ \\ \neg \end{matrix} \right\}$ $\begin{bmatrix} -\text{Ext} \\ \text{Mt} \\ \text{A} \end{bmatrix}$ $\neg$ $\begin{bmatrix} +\text{Ext} \\ -\text{Mt} \end{bmatrix}$

4 B $\neg$ $\begin{bmatrix} -\text{Ext} \\ -\text{Mt} \end{bmatrix}$ $\_/$ $\begin{bmatrix} -\text{Ext} \\ +\text{Mt} \\ +\text{A} \end{bmatrix}$ $\_/$ $\begin{bmatrix} -\text{Ext} \\ \\ -\text{A} \end{bmatrix}$ $\neg$ $\begin{bmatrix} +\text{Ext} \\ -\text{Mt} \end{bmatrix}$



(dans cette séquence, il y a 2 niveaux de contrastes d'amplitude)

Les contours observés ont été ensuite comparés aux prédictions théoriques données plus haut. Les résultats apparaîssent dans le tableau p. , donnent le nombre de contours réalisés différamment des contours théoriques, rapportés au total des contours observés. Les rapports ont été calculés pour chaque type de phrase et selon les réalisations de chaque locuteur dans les trois catégories d'émotions choisies (T: tristesse; C: colère; N: neutre). En fin de ligne et en bas de colonne on a rapporté les totaux relatifs à chaque catégorie, ainsi que les pourcentages correspondants.

RESULTATS

Le tableau donne:

1) les variations totales (nombre de réalisations différentes des contours théoriques, divisé par le nombre total de contours observés;)

2) Variations inter-locuteurs (nombre de réalisations "non-théoriques" / nombre total de contours observés pour chaque locuteur, toutes catégories d'émotions confondues);

3) Variations intra-locuteurs (nombre de réalisations non-théoriques / nombre total selon le type de hiérarchie syntaxique);

4) Variations inter-locuteurs selon le type d'émotion simulée (pourcentage de réalisations non théoriques selon la catégorie d'émotion).

Les résultats obtenus montrent que le taux de pourcentage d'écart par rapport aux prédictions théoriques est relativement peu variable:

Variation du taux inter-locuteur: de 6,5% à 15,7%
(émotions confondues)

Variations du taux intra-locuteur: de 5,3% à 17,1%
(selon le type de hiérarchie syntaxique)

Variations du taux intra-locuteur: de 7,4% à 12,2%
(selon le type d'émotion)

Les invariants produits par le modèle théorique semblent donc recevoir une bonne confirmation expérimentale. On peut également noter que les niveaux d'écart semblent liés au type de phrase, au locuteur, et à la catégorie d'émotion. Ainsi, pour la majorité des locuteurs, le taux d'écart de la hiérarchie 4L (Finalement, Juliette a pris sa valise) est plus élevé que pour 4D (Maurice, le père de Julie, s'en va). Ceci peut s'expliquer par le rapport différent entretenu par les hiérarchies prosodique et syntaxique dans ces deux phrases: la non congruence nécessaire due à la non projectivité de la hiérarchie syntaxique empêche la neutralisation des contours comme dans le cas 4D.

De même, certains locuteurs présentent une tendance à un taux d'écart plus ou moins grand, les cas MN (4,6% en moyenne) et PL (15,7%) étant des exemples extrêmes. Les variations d'écart global (toutes phrases et toutes catégories confondues) selon les catégories d'émotions simulées suggèrent qu'un degré d'activité émotive croissant (allant de la tristesse à la colère) correspond à un degré de contrôle dans les réalisations prosodiques décroissant et se manifeste ici par un taux d'écart par rapport au modèle qui va croissant.

On peut donc penser que la congruence entre prosodie et syntaxe, bien réalisée à la lecture d'un texte, sera plus ou moins vérifiée selon les conditions de production du discours (lecture, parole spontanée, état émotif, contexte, etc.) rencontrées par le locuteur.

Locuteur	Phrase	A	B	C	D	E	F	G	H	J	K	L	TOTAUX		
DM	**T**	**-**	**1/7**	**2/8**	**0/8**	**0/8**	**0/7**	**0/6**	**0/6**	**0/5**	**0/8**	**0/7**	**3/70**		
	C	-	1/7	2/6	1/7	0/8	1/7	0/8	0/7	0/8	0/8	0/7	5/74	14/213	6,5%
	N	-	2/8	0/6	0/8	0/7	3/7	0/7	1/5	0/6	0/7	3/8	6/69		
MC	T	0/8	0/7	0/7	0/8	0/8	0/7	0/8	0/6	0/7	0/8	3/5	3/77		
	C	1/8	1/7	4/6	1/7	3/5	1/8	1/7	0/6	1/8	0/8	1/8	14/78	28/225	12,4%
	N	4/4	0/4	1/5	1/8	0/6	1/8	0/7	2/4	0/8	0/8	2/8			
MN	T	0/7	2/6	1/8	0/8	0/5	0/8	0/6	0/6	2/8	0/7	0/6	5/75		
	C	0/3	0/6	0/7	0/4	-	0/8	0/6	0/6	-	0/7	0/6	0/53	9/192	4,6%
	N	0/4	1/8	0/5	0/4	0/5	0/6	0/7	0/6	0/7	1/4	2/8	4/64		
AL	T	-	0/8	0/4	0/8	0/8	0/8	0/8	2/8	2/6	0/7	0/6	4/71		
	C	0/8	2/8	0/8	1/7	2/8	1/8	0/6	0/8	0/5	2/8	2/6	10/80	21/224	9,3%
	N	2/7	0/4	1/7	0/8	1/8	0/8	0/6	1/7	1/7	0/6	1/7	7/73		
RB	T	1/6	2/8	1/7	0/8	0/8	2/7	0/8	2/7	0/8	1/8	0/8	9/83		
	C	1/3	2/8	1/8	1/7	2/5	0/7	0/7	0/6	1/8	0/8	1/8	9/75	23/232	9,9%
	N	0/6	1/4	1/8	0/7	0/8	0/5	0/6	1/6	0/8	0/8	1/8	4/74		
PL	T	1/4	0/4	0/4	0/4	0/4	1/4	-	-	-	-	-	1/24		
	C	2/4	0/4	1/2	0/2	0/3	1/4	1/3	1/4	1/4	-	1/4	8/35	14/89	15,7%
	N	1/4	1/4	1/3	0/2	-	1/3	0/4	1/4	0/4	-	-	5/30		
AB	T	0/4	2/8	1/8	0/4	0/7	2/7	0/8	2/8	1/8	1/8	0/8	9/76		
	C	-	1/4	1/4	-	0/4	0/4	0/4	1/4	0/4	0/4	0/3	3/35	18/177	9,9%
	N	0/7	3/8	1/8	0/4	0/4	0/4	0/4	0/7	0/6	1/8	1/8	6/68		
MR	T	-	2/8	0/8	0/7	0/6	2/8	0/8	0/4	2/8	-	1/6	7/63		
	C	2/8	1/7	0/7	2/6	1/6	0/8	6/8	0/7	-	0/5	0/8	12/70	29/205	14%
	N	0/8	1/5	1/8	0/8	2/4	2/8	1/5	0/7	1/8	1/5	1/6	10/72		
TOTAL		15/108	26/152	20/152	7/146	10/137	18/157	9/144	14/139	10/141	7/130	20/151	T:42/537	7,4 %	
													C:61/500	12,2 %	
		14,5%	17,1%	13,1%	4,8%	7,3%	11,4%	6,2%	10%	7 %	5,3%	13,2%	N:53/520	10,2 %	

TABLEAU DES RESULTATS

Total général: 156/1557 - 10 %

Le tableau donne le nombre de désaccords entre les contours expérimentaux et les contours théoriques, rapportés au nombre total de contours observés dans chaque catégorie, les contours étant décrits par les traits ± Extrême, ± Montant, ± Ample, ± Bas.

Les résultats sont ventilés selon le type de hiérarchie syntaxique (colonnes A à L), les locuteurs (lignes DM à MR), et le type d'émotion (lignes T, tristesse, C, colère, N, neutre).

This table gives the number of discrepancies between the theoretical and the experimental melodic contours, as described by the features ± Extreme, ± Rising, ± Ample, and ± Low.

The columns correspond to different syntactic hierarchies and the rows correspond to the different speakers. The emotions are given by the symbols T: sadness, C: anger and N: neutral.

REMERCIEMENTS

Je remercie vivement Mr. Daniel LEPETIT, assistant de recherche à l'Université de Toronto, pour son aide précieuse dans l'analyse et le dépouillement des données.

REFERENCES

BOE, L-J., CONTINI, M. et RAKOTOFIRINGA, H. (1975) "Etude statistique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques français Phonetica 32 pp. 1-23.

DAVITZ, J.R. (1969) The Language of Emotions, Academic Press, New-York.

MARTIN, Ph., (1978) "Questions de phonosyntaxe et de phonosémantique en français" Linguisticae Investigationes, II pp. 93-126.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

CORRELATIONS ENTRE VARIABILITE ARTICULATOIRE ET VARIABILITE ACOUSTIQUE CHEZ DEUX LOCUTEURS

ZERLING Jean-Pierre — Laboratoire de Phonétique de Lille

RESUME

L'analyse porte sur les images d'un film radiologique et sur des photographies latérales et frontales de l'orifice labial. Différents paramètres articulatoires sont comparés chez deux locuteurs: le diamètre sagittal du conduit en certains points caractéristiques, la forme de l'orifice labial, la position verticale du larynx et la longueur du conduit vocal.

On insiste particulièrement sur les différences entre les deux sujets et sur les différences inhérentes à chaque sujet. L'analyse de l'onde sonore permet de déterminer dans quelle mesure les variations articulatoires ont des conséquences au plan acoustique.

Les premiers résultats d'une expérience en cours sont présentés: elle a pour but d'évaluer l'influence des mouvements crâniens sur la réalisation acoustique des voyelles.

SUMMARY

This study is based on the analysis of an X-ray film showing lateral views of the vocal-tract, and on a series of frontal and lateral pictures of the lips.

Differences between the two subjects are discussed along with differences inherent to each subject. Sound wave analysis allows to check whether articulatory variations have acoustical consequences or if they are compensed by motor equivalence.

Several articulatory variables are compared for two speakers: sagittal vocal-tract width as measured in some important locations, shape of mouth opening, vertical position of the larynx and vocal-tract length.

Some special aspects of stops [b] and [g] are described: e.g. the influence of [b] upon articulatory and acoustical realisations of vowels. Or the effects of vocalic context upon tongue shape and closure location during [g], which brings to an articulatory explanation of formants transition variability for this consonant.

Preliminary results of a new experiment are given. It is an attempt to estimate the influence of head movements (skull angle measurements) upon the acoustical characteristics of vowels.

CORRELATIONS ENTRE VARIABILITE ARTICULATOIRE ET VARIABILITE ACOUSTIQUE CHEZ DEUX LOCUTEURS

ZERLING Jean-Pierre Laboratoire de Phonétique de Lille

I - INTRODUCTION

Cette étude rassemble des observations émises dans notre Thèse de 3° Cycle (ZERLING, 1979.b) à propos de plusieurs paramètres articulatoires: les formes latérale et frontale du résonateur labial, l'angle maxillaire, la forme de la langue, la position du larynx, et le diamètre et la longueur du conduit vocal. Un autre paramètre fait l'objet d'une étude additionnelle: l'angle crânien.

Nous distinguons deux sortes de différences inter-locuteurs: celles liées à la morphologie du sujet et celles qui dépendent de l'articulation. Pour les variations intra-locuteur, la part est faite entre celles qui sont compensées et celles qui ont des conséquences au plan acoustique.

II - PROCEDURE EXPERIMENTALE

L'étude comporte une analyse graphique des mouvements du conduit vocal vu latéralement des lèvres jusqu'au larynx. Les profils articulatoires ont été obtenus à partir de deux films radiologiques accompagnés d'une bande-son synchrone. Ces films représentent l'articulation de 15 logatomes du type /əCVC/ pour les voyelles [i,ɛ,a,ɔ,u] et les occlusives [b,d,g] , pour deux sujets masculins (FL et JPZ). La cadence de prise de vue est de 60 images par seconde.

Ces données sont complétées par une série de photographies réalisées pour les mêmes sujets pendant l'émission des mêmes logatomes, avec en plus les voyelles [y] et [œ] . Chaque cliché représente une vue frontale et une vue latérale des lèvres pendant la réalisation d'une voyelle tenue en contexte. Comme pour les vues cinéradiographiques, des repères étalonnés fournissent une échelle de mesures précise. L'analyse spectrale de la bande sonore montre que l'allongement nécessaire à la prise de vue n'a pas modifié de manière sensible l'articulation des voyelles.

Enfin, une étude complémentaire a été entreprise ultérieurement pour l'un des sujets. Les logatomes ont été prononcés pour trois positions différentes de la tête: normale (N), haute (H) et basse (B). L'angle crânien par rapport à la normale est mesuré à partir de la tangente au maxillaire supérieur, matérialisée par une baguette serrée entre les molaires supérieures et inférieures et sortant de la bouche vers l'avant. Les déviations observées pour les positions extrèmes sont respectivement de 40° pour H et 35° pour B, par rapport à la normale. Ces valeurs ne sont données ici qu'à titre indicatif. La bande sonore a permis de réaliser une analyse spectrale et mélodique des voyelles.

III-DIFFERENCES MORPHOLOGIQUES ET VARIATIONS ARTICULATOIRES INTER-LOCUTEURS

Lorsque l'on compare les conduits vocaux de plusieurs locuteurs, il est nécessaire de distinguer l'aspect morphologique de l'aspect articulatoire. Chaque sujet est doté d'un appareil phonatoire aux caractéristiques propres: taille, volume et forme. Des sujets différents parviennent à émettre des réalisations aux composantes acoustiques relativement proches. Les particularités anatomiques et physiologiques de chacun sont donc compensées par l'articulation, c'est-à-dire par la position des organes pendant la

phonation. L'adaptation articulatoire exigée par les différences morphologiques entre plusieurs locuteurs est un phénomène analogue à celui de la compensation articulatoire observée chez un même sujet. En plus de leur origine morphologique, les variations articulatoires inter-locuteurs peuvent également dépendre des habitudes ou des comportements phonatoires propres à chaque individu .

III-1 - Diamètre du conduit vocal

La composition spectrale de l'onde acoustique dépend des rapports $\frac{A_i}{A_{i+1}}$ liant l'aire de chaque section du conduit à la suivante (ZERLING,1974). Si deux conduits vocaux diffèrent en volume dans un rapport k, la suite des rapports $\frac{k.A_i}{k.A_{i+1}}$ est identique à la précédente. Le spectre de l'onde émise est donc théoriquement inchangé. Si l'on admet que le diamètre sagittal en un point du conduit est directement proportionnel à l'aire en ce point, on constate que des différences systématiques du diamètre pour deux sujets distincts peuvent n'avoir à elles seules aucune influence au plan acoustique. Les données pour nos deux sujets confirment ces suppositions. La taille du conduit est directement liée à la morphologie du locuteur: le diamètre est systématiquement plus grand chez le sujet FL que chez JPZ. La différence est maximale dans les régions des lèvres et du pharynx.

III-2 - Forme et aire de l'orifice labial

La forme et l'aire de l'orifice labial peuvent être décrites à l'aide d'un petit nombre de paramètres: l'écart horizontal entre les commissures (A), l'espace vertical entre les lèvres (B), l'aire frontale (S) et la protrusion (P) (FROMKIN, 1964; DESCOUT et al,1978; ABRY et al,1979;ZERLING,1980)

On observe une corrélation entre les paramètres A, B et S chez les deux sujets. Néanmoins, la forme frontale de l'orifice labial est systématiquement différente: il est toujours plus plat chez le sujet JPZ. L'aire S est toujours supérieure chez FL: elle atteint parfois une valeur double pour une même voyelle (fig. 1). Pour les raisons évoquées au paragraphe précédent ces différences de diamètre et d'aire de l'orifice peuvent ne jouer qu'un faible rôle si elles sont accompagnées de différences analogues au niveau des autres sections du conduit vocal.

La contraction du muscle orbicularis oris a pour effet de réduire l'aire de l'orifice labial et d'augmenter la protrusion. Chez le sujet JPZ, l'aire atteint des valeurs très faibles avec les voyelles arrondies, et avec [u] en particulier. L'avancement des lèvres est alors très marqué.Chez ce sujet, le paramètre P suffit à diviser les voyelles en deux catégories: arrondies et non-arrondies. En revanche, chez l'autre sujet, les variations de P sont progressives et moins catégorielles. La protrusion ne semble donc pas être un élément fondamental en phonation. Elle n'est peut-être que la conséquence de l'activité musculaire qui vise à réduire l'aire de l'orifice.

III-3 - Position verticale du larynx

Les mouvements du larynx sont complexes (ROSSI et AUTESSERRE, 1979). Ils résultent de la combinaison de plusieurs forces indépendantes: couplage direct (attraction linguale), couplage inverse (attraction hyoïdale) et contrôle moteur (ZERLING, 1979.b, p.161). Il est permis de penser que ces mouvements ont pour but d'ajuster la forme du conduit vocal afin d'affiner la production acoustique. En particulier, les déplacements du larynx permettent de modifier la longueur du conduit.

Les positions relatives du larynx selon les voyelles diffèrent chez les deux sujets: elles sont distinctes chez JPZ et groupées chez FL. L'observation des variations de Fo permet d'expliquer ces différences. Le sujet FL a prononcé les logatomes en maintenant une fréquence fondamentale quasiment constante: Fo = 125 ± 5 Hz. Ce contrôle de la fréquence laryngée a sans doute été obtenu en imposant une contrainte musculaire qui a indirec-

tement limité les déplacements du larynx. Ce genre de contrainte existe chaque fois que Fo est fortement contrôlée: voix chantée ou tenue (BOTHOREL, 1979). Ce phénomène permet de comprendre pourquoi les observations concernant la position du larynx sont parfois si différentes selon les auteurs.

III-4 - Longueur du conduit vocal.

Il semble que la longueur "efficace" du conduit vocal corresponde à la distance entre le larynx et la commissure des lèvres (LONCHAMP, 1978). Les variations de la longueur sont dues à celles de trois paramètres: la position horizontale des commissures, l'emplacement vertical du larynx et les lieu et degré de rétrécissement du conduit vocal (ZERLING, 1979.a). Le troisième paramètre modifie la longueur en changeant le rayon du conduit et en modifiant la circonférence dans sa partie courbe. Son rôle est très important: malgré les très faibles mouvements des extrémités du conduit observés pour le sujet FL et les déplacements de grande amplitude pour le sujet JPZ, les variations de longueur, de voyelle à voyelle, sont identiques chez les deux sujets (fig. 2). Le troisième paramètre permet donc de compenser dans une large mesure les faibles mouvements des extrémités.

Par ailleurs, on note que le conduit vocal du sujet FL est systématiquement plus long que celui de JPZ (10 mm environ). On sait, pour l'avoir observé(PETERSON et BARNEY, 1952), ou calculé en synthèse (ZERLING, 1974), que les variations de longueur du conduit affectent les valeurs formantiques des voyelles. Il n'est donc pas surprenant de constater que ces valeurs sont supérieures pour le sujet JPZ. La différence est d'autant plus grande que les valeurs sont élevées.

III-5 - Déplacement du lieu occlusif de [g]

Il est souvent admis que le lieu occlusif de [g] est lié au degré d'antériorité de la voyelle voisine. L'observation des profils articulatoires de nos deux sujets et de ceux publiés par P. SIMON(1967) et ROCHETTE (1973) révèle que celà est faux pour la voyelle [a] dite antérieure. Selon nous, le lieu occlusif de [g] dépend du lieu articulatoire de la voyelle, pris au sens de point de rétrécissement maximal du conduit vocal. En présence des voyelles alvéolaires [i,e,y,ø] le contact occlusif peut avancer; en présence des pharyngales [a,ɔ] ou des vélaires [u,o] , il conserve une position post-palatale. La figure 3 révèle que ce phénomène varie d'un sujet à l'autre: il est beaucoup plus net chez FL que chez JPZ. L'occlusion est prépalatalisée chez le premier en contexte [i] , alors qu'elle demeure presque toujours post-palatale chez le second. Sur le plan acoustique, ces différences ont peu d'importance, mais elles s'ajoutent à d'autres qui sont évoquées plus loin (§ IV.2.c).

IV-VARIATIONS INTRA-LOCUTEUR:INFLUENCE DU CONTEXTE,COMPENSATION ARTICULATOIRE

S'il est fréquent d'observer des différences morphologiques, articulatoires et acoustiques entre deux sujets, on relève également chez un même locuteur des variations entre différentes réalisations d'un même son. En général, elles sont fonction du contexte. Nos observations nous amènent à distinguer les modifications articulatoires qui ont des conséquences acoustiques de celles qui sont compensées afin d'obtenir une réalisation acoustique inchangée.

IV-1 - Variations articulatoires compensées

IV-1-a - Protrusion labiale et aire aux lèvres

Nous avons dit plus haut que si l'aire labiale est un paramètre fondamental de l'émission des sons, en revanche, la protrusion labiale n'est peut-être que le reflet de la contraction de l'orbicularis oris visant à diminuer l'aire. La comparaison des profils articulatoires des voyelles [a] et [ɔ] prononcées par chaque sujet dans trois contextes différents révèle

que lorsque le profil labial est quasiment identique, la position linguale est nettement différente: le rétrécissement pharyngal est placé plus haut pour [ɔ] que pour [a]. Au contraire, lorsque la différence aux lèvres est importante, le contour lingual est alors très voisin pour les deux voyelles. Il s'agit là d'un phénomène d'équivalence motrice.

Lors de l'étude de l'aire aux lèvres, plusieurs photographies ont été réalisées à des moments différents pour les mêmes voyelles. Bien que les composantes spectrales de chacune d'elles soient très voisines, l'aire labiale pour une même voyelle et pour un même sujet peut varier du simple au double: par exemple, avec la voyelle [i] (fig. 1). La latitude des variations est supérieure pour les voyelles non-arrondies.

On montre en synthèse qu'une grande différence d'aire labiale peut être compensée par une faible variation de la longueur du conduit (LONCHAMP et ZERLING, 1980). Les données montrent que les voyelles non-arrondies sont labialisées en contexte [b]. L'aire aux lèvres est alors plus faible que dans d'autres contextes (fig. 4). Or, on observe que la longueur du conduit (larynx-commissures) pour [i] est légèrement plus faible en contexte [b], malgré la labialisation, donc malgré une protrusion plus importante des lèvres (fig. 2 et 4). Pour cette voyelle, la labialisation de l'aire semble donc compensée par une réduction de la longueur du conduit vocal.

IV-1-b - Angle maxillaire

La mâchoire inférieure joue un rôle important dans les phénomènes de compensation articulatoire (LINDBLOM, SUNDBERG, 1971; HUGGHES, ABBS, 1976; LINDBLOM et al, 1977; KIM et SOHN, 1979). Un des exemples les plus significatifs tiré de notre étude se manifeste chez les deux sujets dans les groupes [gag] . Pendant la tenue de $[g]_1$, la mâchoire est en position haute afin de maintenir l'occlusion. La langue n'est pas libre de s'abaisser; elle ne peut qu'anticiper la forme apicale non-rétractée caractéristique de [a] (cf § IV.2.c). Le passage à la voyelle est obtenu par un abaissement du maxillaire nettement supérieur à celui observé habituellement pour [a] (fig.5a). La très grande ouverture de l'angle permet donc, à elle seule, de compenser la position haute de la langue et suffit à la tirer vers le pharynx.

IV-1-c - Hauteur du larynx.

Il ressort de notre observation et de l'analyse d'une vingtaine d'articles récents (cf ZERLING, 1979.b, p.161) que l'ordre de placement du larynx de haut en bas, pour une voix parlée naturelle et pour un sujet donné, est [a, i, u] . Une contrainte quelconque imposée au larynx (action musculaire ou contrôle de la fréquence laryngée), peut entraîner des modifications de sa position. Le résultat acoustique souhaité est donc obtenu par une équivalence motrice à d'autres niveaux: forme des lèvres, position de la langue et longueur du conduit.

IV-2 - Variations articulatoires non compensées

Parfois, les différences articulatoires liées à la nature du contexte ne sont pas compensées et se manifestent sur le plan acoustique. Ce phénomène se produit notamment dans les cas d'assimilation. Quatre exemples permettent de l'illustrer de manière significative.

IV-2-a - Influence de l'occlusion labiale sur la position de la langue

Dans les groupes [bVb], il apparait que la langue est toujours plus basse qu'en présence des occlusive linguales, puisqu'elle n'est pas sollicitée pour une occlusion (fig. 5.b). Il est donc permis de penser que les positions vocaliques atteintes en présence de [b] sont proches de ce qu'on a coutume d'appeler "positions cibles" (target vowels).

IV-2-b - Influence de [b] sur les voyelles non-arrondies

Dans tous les cas, et chez les deux sujets, il se produit une labialisation des voyelles non-arrondies [i, ɛ, a] dans les groupes [bVb] : la lèvre inférieure est plus haute et plus avancée.

FANT (1960) a constaté que la protrusion labiale, la diminution de l'aire aux lèvres et l'augmentation de la cavité alvéo-palatale sont trois causes de la diminution des valeurs de F2. Dans nos données, les valeurs légèrement plus faibles du second formant en présence de [b] reflètent sans aucun doute l'influence articulatoire de la consonne (labialisation et position linguale plus basse)(fig. 5).

IV-2-c - Influence de l'activité musculaire linguale vocalique sur [g]

L'étude des profils articulatoires révèle que les voyelles se divisent en deux classes caractérisées chacune par une forme apicale particulière: pour les axes I-A et Y-A (voyelles [i,e,ɛ,a] et [y,ø,œ]), l'apex est en position plate, non-rétractée; pour l'axe U-A (voy. [u,o,ɔ]), il est rétracté, découvrant le plancher buccal.(ZERLING, 1979.b,p.76-93). Cette caractéristique est liée aux activités musculaires intrinséque et extrinsèque de la langue.

Pendant l'occlusion de [g], la caractéristique apicale de la voyelle subséquente est anticipée (fig. 7). Cette assimilation entraine une différenciation de la forme et du volume de la cavité alvéo-palatale selon le contexte vocalique. Dans certains cas, elle contribue également au déplacement du contact occlusif (cf §III.5).

Nous insistons sur le fait que cette catégorisation occlusive dépend de la caractéristique apicale de la voyelle, et non de son lieu d'articulation. En outre, cette observation permet enfin de comprendre les particularités des transitions formantiques de [g] souvent décrites mais jamais encore expliquées (DELATTRE et al, 1955).

IV-2-d - Conséquences acoustiques des variations de l'angle crânien(premiers résultats)

Dans la plupart des études cinéradiographiques, la position crânienne est maintenue constante pendant les prises de vue et de son. Pour les analyses purement acoustiques, cette précaution est plus rarement prise. L'angle crânien constitue un paramètre articulatoire non négligeable puisqu'il est susceptible de modifier sensiblement la forme globale du conduit vocal. L'émission acoustique peut donc également subir des variations. Si certains auteurs ont pris soin de neutraliser cette variable, aucun à notre connaissance n' a essayé de la contrôler et d'évaluer ses conséquences acoustiques.

Nous avons entrepris une étude préliminaire pour un seul locuteur. Le sujet a prononcé l'ensemble des logatomes décrits précédemment de trois manières différentes: position crânienne naturelle (N), aucune contrainte n'étant imposée; position tête haute (H), avec une élévation maximale du menton; et position tête basse (B), le menton étant quasiment en contact avec la poitrine. Les angles mesurés entre les positions H et B et la position N sont respectivement de 40 et 35 degrés. Comme lors des expériences précédentes, il a été demandé au sujet de conserver une fréquence fondamentale constante.

Les hypothèses à vérifier étaient les suivantes: les variations de l'angle crânien ont-elles pour conséquences

- des variations inconscientes de Fo (par exemple, une élévation avec la position H et un abaissement avec la position B);
- des modifications du spectre acoustique des voyelles (par exemple, une élévation des formants dans la position H);
- des différences acoustiques perceptibles, dues à un timbre différent ?

Les premiers résultats de l'expérience permettent de tirer les conclusions suivantes qui ont, rappelons le, un caractère préliminaire.

1° - Contrairement à notre attente, le sujet a maintenu un Fo constant indépendamment de l'angle crânien. Aucune influence subjective ("haut" ou "bas") n'est venue modifier la représentation mentale de la fréquence laryngée. Notons néanmoins que le sujet est un phonéticien exercé.

2° - L'analyse spectrale révèle des modifications systématiques du timbre de toutes les voyelles, indépendemment du contexte consonantique:
- position H (tête levée), renforcement sensible et élévation du quatrième formant dont la fréquence augmente de 100 à 300 Hz.
- position B (tête basse), abaissement de la fréquence du quatrième formant de 100 à 200 Hz, accompagné parfois d'une diminution de l'amplitude.
- pour la voyelle [u] , les positions H et B semblent avoir pour conséquence un affaiblissement du second et surtout du troisième formant.

3° - Les différences de timbre sont faiblement perçues. Néanmoins, des auditeurs à l'oreille exercée attribuent à la voix H un timbre plus métallique et plus aigu, et à la voix B un timbre plus sombre et des résonances plus graves.Ces différences perceptives concordent avec les modifications spectrales observées à propos de F4.

Cette première approche du problème révèle donc que les variations de l'angle crânien sont responsables de légères modifications de l'amplitude et de la fréquence du quatrième formant (situé entre 3000 et 4000 Hz). Ces modifications n'ont que de faibles conséquences sur la perception des voyelles.

Aucune variation sensible des trois premiers formants n'a été observée. Une méthode d'analyse plus précise que la spectrographie serait nécessaire pour vérifier ce dernier point. Deux hypothèses peuvent être émises pour le moment: - les variations de l'angle crânien n'ont pas de conséquences sur les formants bas;

- les variations de l'angle crânien ont des conséquences qui sont compensées par une adaptation articulatoire des organes phonatoires.

Cette seconde solution nous parait a priori la plus plausible et elle montre qu'il est indispensable de mieux connaitre le rôle de ce paramètre lorsqu'on désire comparer des données articulatoires. Sur le plan purement acoustique, il semble pouvoir être négligé.

V - CONCLUSIONS

Cette étude rassemble plusieurs exemples de variations articulatoires inter et intra-locuteurs observées lors de la réalisation d'une thèse de 3° Cycle (ZERLING, 1979.b).

Les différences entre deux locuteurs sont généralement liées à leurs "habitudes" articulatoires (variabilité du lieu occlusif de [g]) ou à leurs différences morphologiques (taille et volume du conduit vocal). La plupart des variations articulatoires sont compensées soit lors de l'articulation, soit naturellement (variations du diamètre et de l'aire transversale du conduit ou de l'aire aux lèvres). En revanche, d'autres ont des conséquences au plan acoustique: les différences de longueur du conduit entraînent des variations spectrales systématiques.

Chez un même sujet, les variations articulatoires sont souvent liées au contexte (assimilations). Nombre d'entre elles sont dues à des phénomènes d'équivalence motrice et n'ont donc aucune influence acoustique perceptible: par exemple, les positions relatives de la mâchoire inférieure et de la langue sont très variées selon le contexte; ou encore, l'aire aux lèvres peut varier considérablement pour une même voyelle, les variations pouvant être compensées par de faibles modifications de la longueur du conduit.

Certaines articulations ne sont pas compensées et sont perceptibles au plan acoustique. L'occlusive [b] permet un abaissement de la langue pendant les voyelles voisines, et labialise les voyelles non-arrondies en augmentant l'arrondissement des lèvres, leur protrusion et en diminuant l'aire labiale. Ces modifications sont responsables d'un abaissement

systématique du second formant.

Pendant l'occlusion de [g] , l'apex anticipe une position plate ou rétractée en fonction du contexte vocalique subséquent. La différenciation de la forme et du volume de la cavité alvéo-palatale qui en découle est directement responsable de la catégorisation des transitions formantiques de [g] selon le contexte vocalique.

Les résultats préliminaires d'une étude des conséquences acoustiques des variations de l'angle crânien montrent que l'élévation de la tête pendant la phonation entraîne une augmentation et un renforcement du quatrième formant. Son abaissement provoque les effets inverses. Aucune influence importante de l'angle crânien n'a été observée sur les trois premiers formants ou sur la fréquence fondamentale Fo.

Ces diverses observations montrent, s'il en était besoin, combien les phénomènes de coarticulation sont complexes, et avec quelle prudence il faut avancer dans les domaines de la description et de la synthèse articulatoires. Il convient, à chaque pas, de déterminer dans quelle mesure les variations observées sont significatives sur le plan articulatoire et d'évaluer l'importance de leurs conséquences acoustiques.

o o o o o o o

REFERENCES BIBLIOGRAPHIQUES

ABRY C., BOË L.J., GENTIL M., DESCOUT R. et GRAILLOT P.(1979) "La géométrie des lèvres en français. Protrusion vocalique et protrusion consonantique", 10° J.E.P., Grenoble, p.99.

BOTHOREL A. (1979) "Déplacement de l'os hyoïde et Fo", Séminaire: Larynx et parole, GALF, Institut de Phonétique de Grenoble.

DESCOUT R., BOË L.J. et ABRY C. (1978) "Labialité vocalique et labialité consonantique en français, 1-ers résultats", 9° JEP,Nancy, p.177-189.

FANT G. (1960) Acoustic theory of speech production, Mouton, The Hague, 2°. ed., 1970.

FROMKIN V.A. (1964) "Lip positions in American English vowels", Language and Speech, 7, p.215-225.

DELATTRE P., LIBERMAN A.M. et COOPER F.C. (1955) "Acoustical loci and transitional cues for consonants", JASA 27,4, 769-773, in Readings in acoustic phonetics, LEHISTE ed., 1967.

HUGGHES O. et ABBS J.H. (1976) "Labial mandibular coarticulation in the production of speech: implications for the operation of motor equivalence", Phonetica, 33, p.199-221.

KIM C. et SOHN H. (1979) "A complementary relationship between lip and jaw movements during articulation", 9° ICPS, Copenhague, p.195.

LINDBLOM B., LUBKER J. et MC ALLISTER R. (1977) "Compensatory articulation and the modeling of normal speech production behavior", Articulatory modeling and phonetics symposium, Grenoble, p.147-160.

LINDBLOM B.E.F. et SUNDBERG J.E.F.(1971)" Acoustical consequences of lip tongue, jaw and larynx movements", JASA 50, 4, p.1166-1179.

LONCHAMP F. (1978) Recherche sur les indices perceptifs des voyelles orales et nasales, Thèse de 3° Cycle, Université de Nancy II.

LONCHAMP F. et ZERLING J.P. (1980) "Estimation de la longueur du conduit vocal et de l'aire labiale à partir des fréquences formantiques", Séminaire international: labialité, GALF, Lannion.

PETERSON G.E. et BARNEY H.L. (1952) "Control methods used in the study of vowels", JASA,24,2, p.175-184., in LEHISTE, Readings...

ROCHETTE C. (1973) Les groupes de consonnes en français,Klincksieck, Paris

ROSSI M. et AUTESSERRE D. (1979) "Mouvements de l'os hyoïde et du larynx et fréquence intrinsèque des voyelles", Séminaire Larynx et Parole, GALF, Institut de Phonétique de Grenoble.

SIMON P. (1967) Les consonnes du français.Mouvements et positions articulatoires à la lumière de la cinéradiographie, Klincksieck, Paris.

ZERLING J.P. (1974) "Etude d'un programme d'ordinateur permettant de déterminer les trois premiers formants à partir de la fonction d'aire", Travaux de l'Institut de Phonétique de Nancy, 1, p.257-291.

ZERLING J.P. (1979a)"Description de cinq voyelles orales du français en contexte, et nouvelle classification articulatoire", Travaux de l'Institut de Phonétique de Nancy,2, p. 55-87.

ZERLING J.P. (1979.b) Articulation et coarticulation dans des groupes occlusive-voyelle en français, Thèse de 3° Cycle, Université de NancyII.

ZERLING J.P. (1980) "Coarticulation labiale et aire aux lèvres dans des groupes occlusive-voyelle en français", Séminaire international: labialité, GALF, Lannion.

o o o o o o

FIGURES

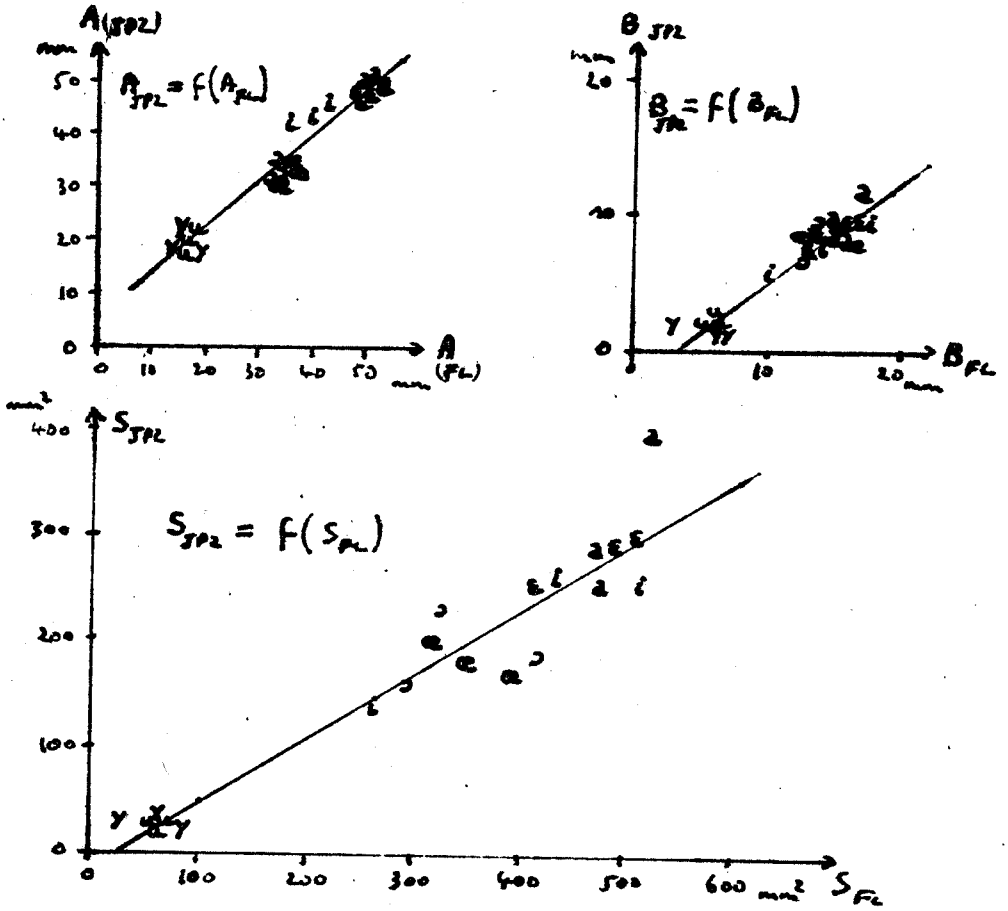


Fig. 1 – Corrélations entre les paramètres labiaux chez les deux sujets:
a – écartement des commissures (A)
b – espace inter-labial (B)
c – aire de l'orifice labial (S)

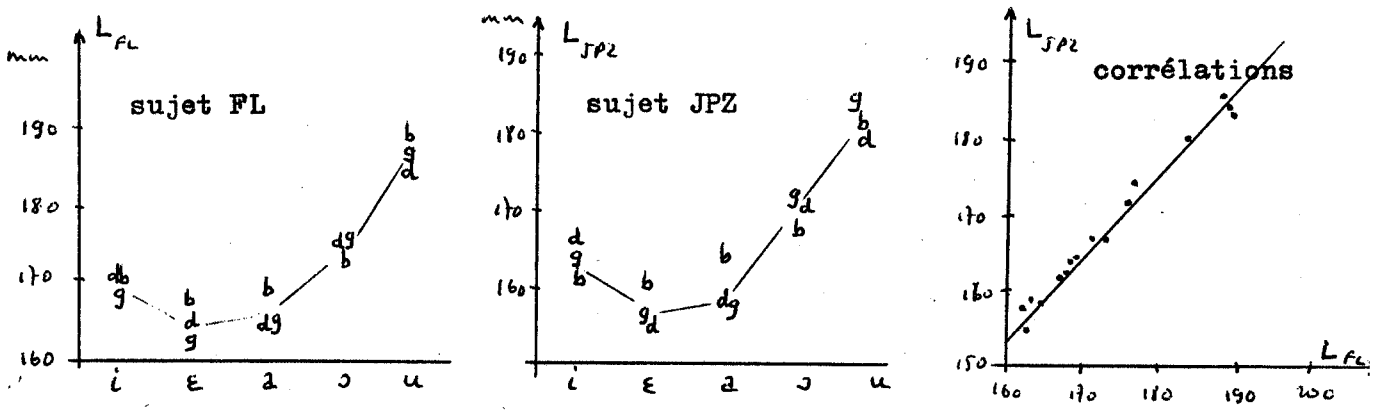
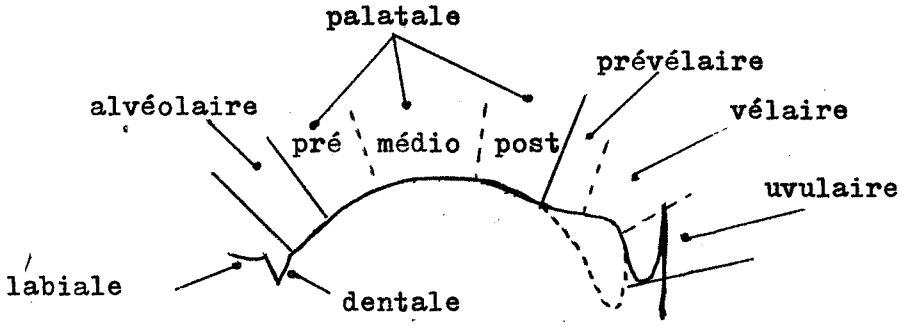


Fig. 2 - Variations de la longueur du conduit vocal (larynx-commissures) en fonction du contexte vocalique et consonantique, et corrélations des logueurs pour les deux sujets.



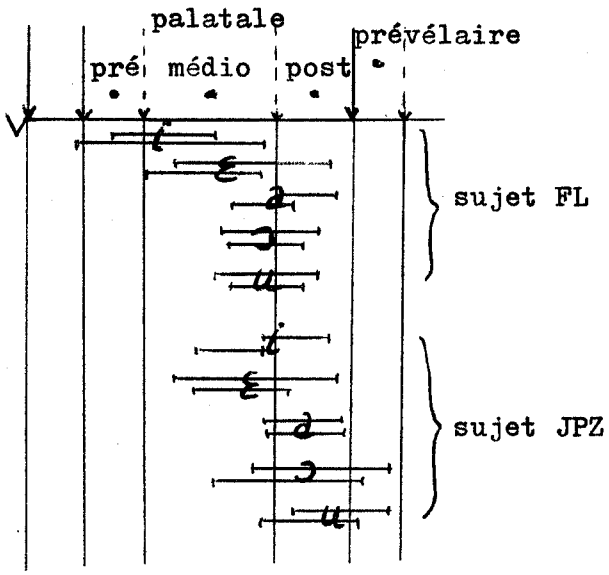


Fig. 3.b - Emplacement des zones de contact occlusif de
en fonction du contexte vocalique (chez les deux sujets).

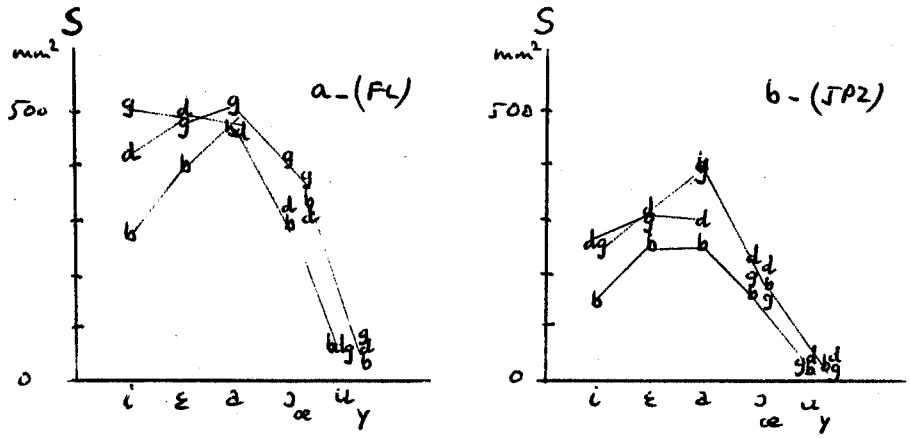


Fig. 4 - Variations de l'aire S de l'orifice labial pour chaque voyelle en fonction du contexte consonantique.
a - sujet FL b - sujet JPZ

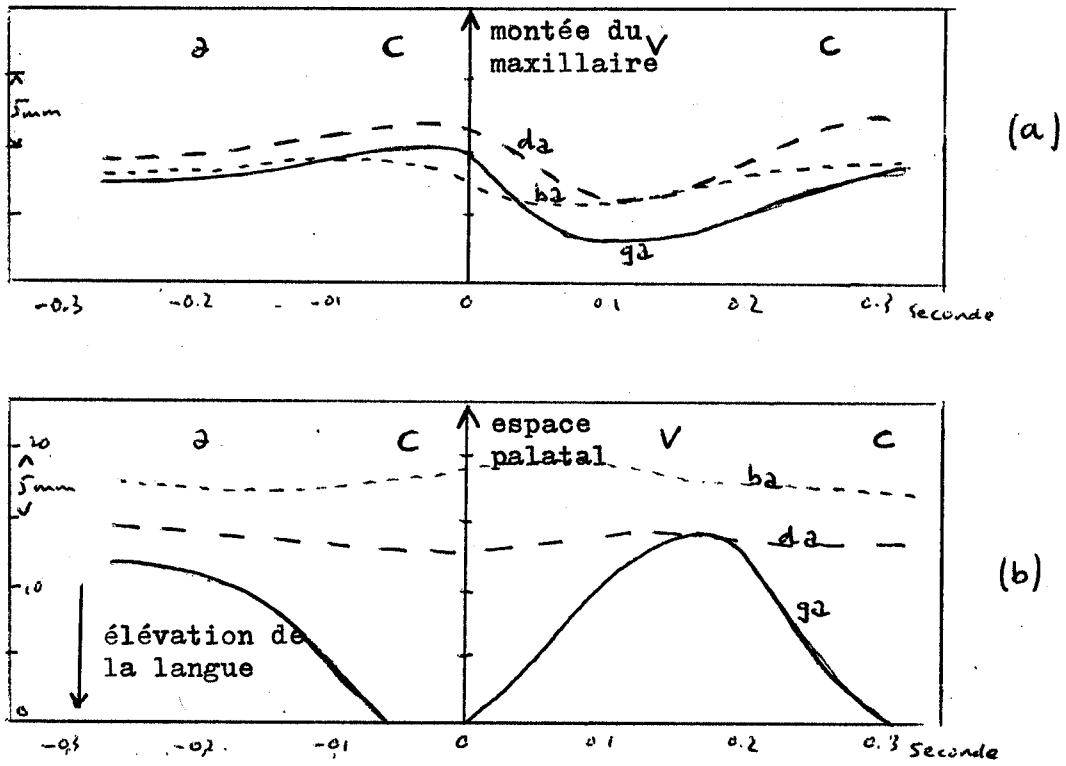


Fig. 5 - Variations de l'angle maxillaire (a) et de l'espace dorso-palatal (b) en fonction du temps dans des logatomes /əCaC/, sujet JPZ.

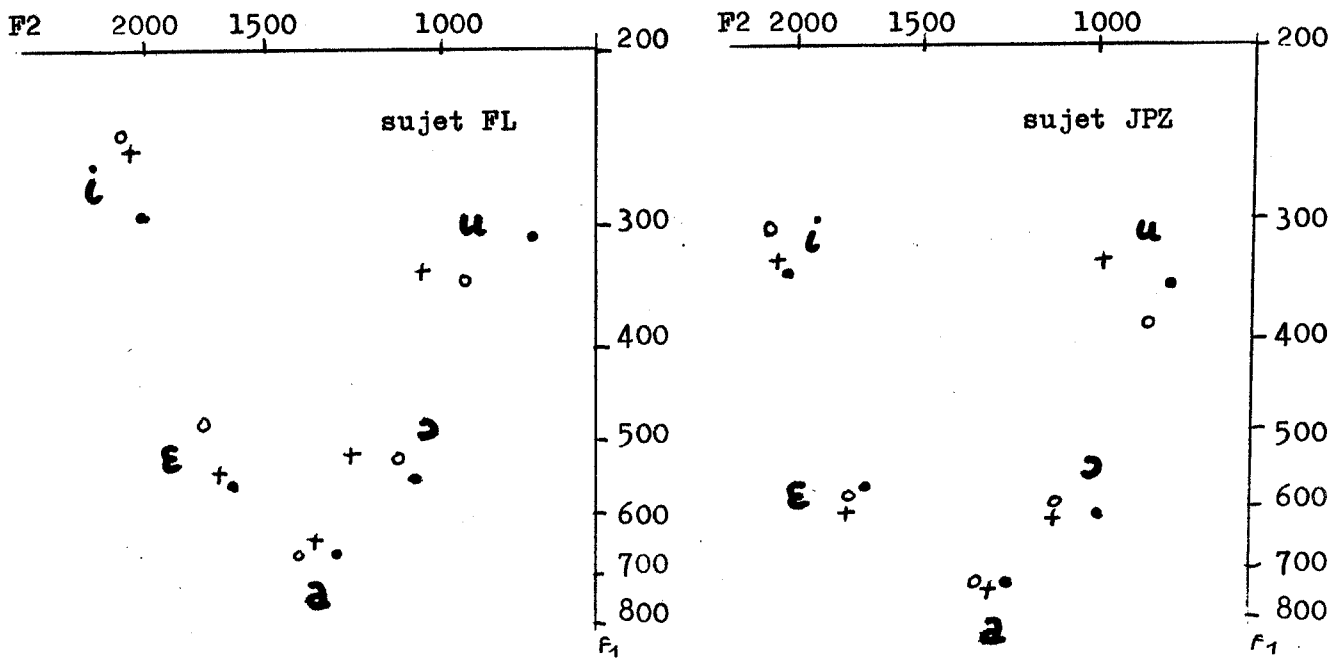


Fig. - 6 - Valeurs des deux premiers formants en milieu de voyelle chez les deux locuteurs

• [b] + [d] o [g]

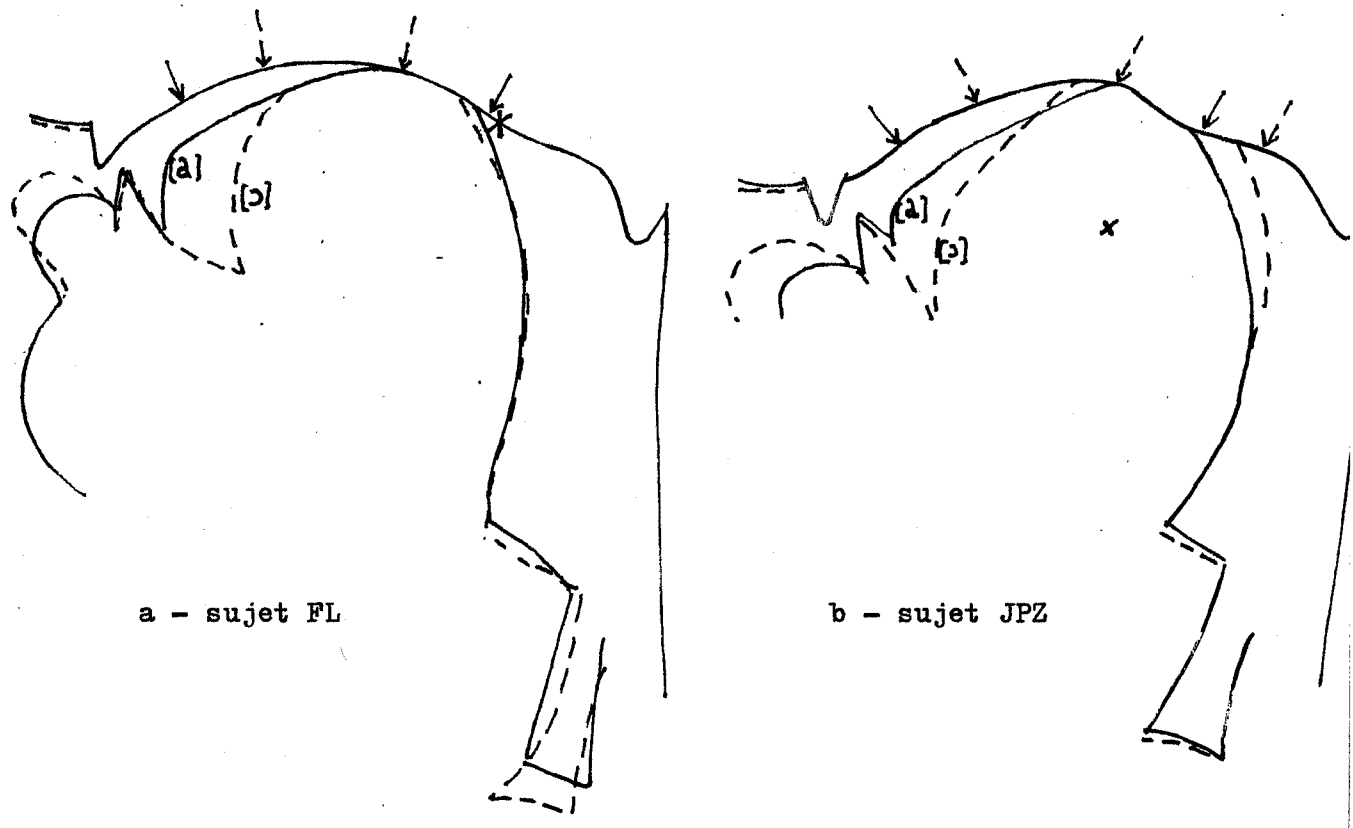


Fig. 7 - Comparaison des profils occlusifs de [g] dans les contextes vocaliques [a] et [ɔ]

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

SYSTEME PHONETIQUE, IDIOLECTE ET DIFFERENCES INDIVIDUELLES.

ABRY Christian — Institut de Phonétique
BOË Louis-Jean — Grenoble

RESUME

Les diversités inter- et intra-individuelles ont été évacuées des descriptions phonétiques. Cette démarche apparaît comme la conséquence implicite, du moins dans un premier temps, de la dichotomie saussurienne langue/parole. Celle-ci se devait de rejeter, de l'objet de la linguistique, toute possibilité qui n'appartenait pas au système.

L'intégration récente de la variabilité intrasystématique par les études de socio-phonétique permet d'envisager un traitement non réductionniste des différences idiolectales.

A notre avis ces différences ont été jusqu'à présent, trop vite ramenées à des différences de système idiolectal alors qu'il est possible de montrer, comme nous essaierons de le faire en proposant diverses procédures, que ce sont des variantes inter-individuelles intrasystématiques.

SUMMARY

Inter- and intra-speaker variability has been dismissed from phonetic descriptions. This attitude seems to be a consequence of the Saussurian dichotomy langue/parole : everything and especially the variability that did not pertain to the system had to be rejected from linguistics.

Recent integration of intra-systematic variability in socio-phonetic studies offers the opportunity of non-reductionist treatment of idiolect differences.

We think for our own that such differences have been too soon attributed to differences of system between idiolects. We shall try to show that it is possible to propose a set of procedures in order to treat this individual but intra-systematic difference.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

SYSTEME PHONETIQUE, IDIOLECTE ET DIFFERENCES INDIVIDUELLES.

ABRY Christian — Institut de Phonétique
BOË Louis-Jean — Grenoble

VARIABILITE ET SYSTEME

Alors que de très nombreuses études ont, non seulement décrit[1] (BOË & CORSI, à paraître), mais exploité la variabilité interlocuteur (procédures d'identification et de vérification du locuteur), les diversités inter- et intra-individuelles ont été évacuées des descriptions phonétiques. Cette démarche apparaît comme la conséquence implicite, du moins dans un premier temps, de la dichotomie saussurienne langue/parole. Celle-ci se devait de rejeter, de l'objet de la linguistique, toute variabilité qui n'appartenait pas au système.

Tout au plus a-t-on reconnu très vite, pour la fonction expressive (au sens de BÜHLER, c'est-à-dire symptomatique du locuteur, de sa localisation sociale, etc.), l'existence d'un code passible d'une branche des études phonologiques : la phonologie expressive (LAZICZIUS, repris par TROUBETZKOY, 1970, p. 17 sqq.). Celle-ci éliminait, cela va de soi, les caractéristiques individuelles[2] du locuteur liées à sa morphologie,

---

[1]. Dans ce domaine, il faut signaler le travail de précurseur de Louise KAISER (1939, 40, 41, 42, 44).

[2]. En réalité, on retrouve chez TROUBETZKOY, comme chez MARTINET, que code implique choix, ce qui élimine d'emblée toutes les caractéristiques dites intrinsèques, qui sont automatiquement conditionnées par la nature des segments, leur environnement, que celles-ci soient générales ou individuelles. Or, l'on sait bien qu'il existe une possibilité de codage perceptif de ces caractéristiques, comme l'ont montré les nombreuses expériences sur les indices acoustiques (cf. pour le voisement SLIS & COHEN, 1969, a.b.).

son sexe, son âge, son état émotionnel, etc., c'est-à-dire, ses indices phonobiologiques et psychologiques.

En se limitant à la fonction représentative, section de la phonologie qui a donné lieu aux plus riches développements, l'hypothèse de base a été et reste que la variabilité se traite hors système et que, si elle est intégrée au système, elle le modifie (changements d'inventaire phonématique, des règles phonotactiques ou génératives). Cette alternative a été pendant longtemps un obstacle épistémologique de taille au traitement structural de la variation dialectologique avant que WEINREICH (1954) n'instaure le concept de diasystème[3].

Mais intersystème ou supersystème, il n'en reste pas moins que les systèmes dialectaux ou idiolectaux, ainsi réunis, n'intègrent pas la variabilité intrasystématique. Celle-ci est alors pensée au niveau de la réalisation de chacun des systèmes dans laquelle doivent être maintenues des constantes relationnelles. Ainsi la plupart des procédures de normalisation des espaces vocaliques tendent-elles à se débarasser de la variance interlocuteurs, qu'elles fassent appel à des transformations par homothétie ou à d'autres plus sophistiquées (GERSTMAN, 1968; HARSHMAN, 1970; LOBANOV, 1971; SANKOFF & al., 1974; NORDSTRÖM & LINDBLOM, 1975; NEAREY, 1977; ROUSSEAU & SANKOFF, 1978).

Or nous pensons qu'il existe entre les différentes réalisations d'un même système par différents locuteurs, une variabilité irréductible. A notre avis, ces différences ont été, jusqu'à présent, trop vite situées au niveau du système idiolectal alors qu'il est possible de montrer,

---

[3]. Il n'existe pas de véritable pratique terminologique en ce qui concerne la typologie des relations entre les systèmes linguistiques. Sans envisager les systèmes réduits, les systèmes véhiculaires ou les proto-systèmes (systèmes génétiquement reconstruits) - systèmes qui ont en quelque sorte pour fonction de passer plus ou moins "par-dessus" les systèmes - on pourrait réserver le terme de diasystème aux situations linguistiques dans lesquelles il existe une pratique des correspondances systématiques. Il y aurait alors intersystème lorsque ces correspondances sont établies directement, sans que l'un au moins des deux systèmes en présence ait à être interprété ou modifié (prétraité linguistiquement en quelque sorte) pour que la communication s'établisse; supersystème dans le cas contraire.

comme nous essaierons de le faire en proposant diverses procédures, que ce sont des variantes inter-individuelles, intra-systématiques.

DU SYSTEME A SON INDIVIDUATION

L'étude des systèmes phoniques présuppose, en approche traditionnelle, au moins deux conditions qui n'ont jamais été véritablement éprouvées :

1° Pour faire la preuve existentielle d'un système, il est nécessaire de pouvoir quantifier l'intercompréhension.

2° Il est en outre indispensable de montrer que les supports symboliques, que met en oeuvre la pratique du système, présentent des caractéristiques communes (valeurs des paramètres) ou s'établissent dans un ensemble de rapports communs (une structure).

Alors que l'évidence de (1°) ne semble par faire défaut, il existe pour (2°) une accumulation de données contraire (STEVENS, 1971); parmi celles-ci citons l'indispensable adaptation au locuteur que doivent mettre en oeuvre, et à laquelle se heurtent, tous les systèmes de reconnaissance automatique de la parole.

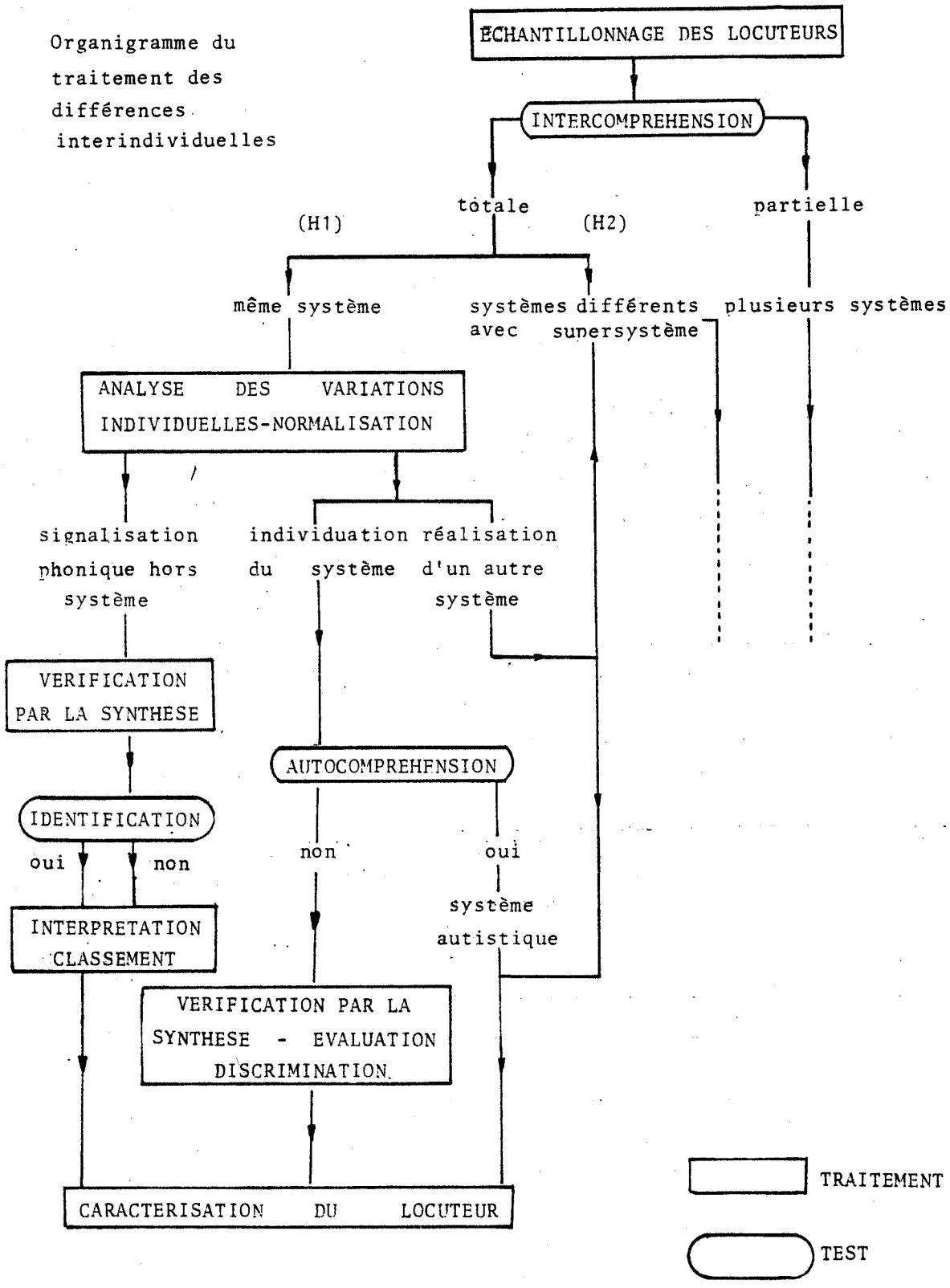
On s'est donc trouvé devant deux évidences apparemment contradictoires :

1° Il existe une indéniable pratique de l'intercompréhension.

2° Les signaux acoustiques utilisés dans la parole présentent des variations individuelles difficilement réductibles. Ces différences sont essentiellement de deux ordres. Les unes se situent aux niveaux anatomique et physiologique, elles sont relativement bien connues; les autres tiennent à l'utilisation proprement individuelle que le locuteur fait de son appareil vocal. C'est en somme, en ce qui concerne la parole, le correspondant des habitudes posturales et gestuelles que constituent les différentes techniques du corps. C'est sans aucun doute le concept de base articulatoire individuelle qui rendrait le mieux compte actuellement, au niveau phonétique, de cet aspect de l'hexis corporelle (BOURDIEU, 1972) qu'est l'habitus vocal.

Il s'agit maintenant de proposer une méthodologie permettant de mettre en évidence dans le signal de la parole les parts revenant res-

Organigramme du traitement des différences interindividuelles



pectivement à :
1° La fonction de signalisation phonique individuelle;
2° La réalisation individuée du système linguistique.

En fait cette dichotomie fonctionnelle peut permettre de séparer dans le matériau phonique, ce qui sert préférentiellement à la caractérisation individuelle et qui ne tient pas à la réalisation individuée du système, c'est-à-dire la signalisation phonique individuelle hors système linguistique. Le système individué quant à lui, outre sa fonction proprement linguistique, peut, bien entendu caractériser le locuteur[4].

PROPOSITIONS

La démarche que nous proposons est explicitée dans l'organigramme ci-contre.

- Première étape indispensable et supposée résolue, celle de l'échantillonnage des locuteurs, sur des bases linguistiques évidentes et avec le maximum de variables indépendantes extra-linguistiques (anatomiques, sociales, locales, etc.).

- Ensuite, l'élaboration de tests d'intercompréhension globale et sélective entre les locuteurs permettra l'établissement et la quantification du fonctionnement du code linguistique[5]. On aboutit ici à l'évaluation des distances entre les locuteurs (taux d'intercompréhension) qui peuvent être interprétées en termes de plus ou moins grande similarité des systèmes individuels. L'hypothèse (H 1), sur laquelle nous allons démarrer à ce stade, est que les taux de forte similarité nous permettent d'induire une forte identité des systèmes. Il nous apparaît impossible, à cette étape, d'explorer (H 2), une autre hypothèse de départ, qui est

---

4. "Outre les procédés purement expressifs, il en est d'autres qui remplissent en plus une fonction représentative spéciale. Souvent la prononciation d'un groupe de sujets parlants se distingue de la prononciation habituelle par le fait qu'elle néglige une opposition phonique distinguant des significations (donc de valeur représentative) ou qu'à l'inverse elle présente une telle opposition là où la prononciation des autres groupes de sujets parlants l'ignore". TROUBETZKOY (1970), p. 23.

5. On pourra prévoir certainement les effets d'adaptation en particulier l'ajustement au cadre vocalique interne (cf. LADEFOGED & BROADBENT, 1957).

celle d'une forte intercompréhension due, non pas à l'identité des systèmes, mais à l'existence d'un diasystème ou plus généralement d'un quelconque code de transduction (supersystème). Il faut admettre que cette possibilité donne lieu à des réalisations substantielles suffisamment différentes pour être rapidement repérées. Seule la suite de l'analyse permettra donc d'apporter des arguments en faveur de H1 ou H2.

- L'analyse des variations intra- et interlocuteurs menée sur n paramètres sera soumise à différentes procédures d'élimination des variations extra-systématiques. L'idéal serait de pouvoir posséder une théorie suffisamment élaborée, des causes de variabilité pour pouvoir éliminer sélectivement et exhaustivement, à chaque stade, les paramètres qui en dépendent (morphologie, habitus vocal, etc.). En réalité les procédures de normalisation disponibles réduisent aveuglément la variance, quelles qu'en soient les causes. Leur utilisation requiert quelques précautions qui ne tiennent pas seulement à l'introduction, par certaines d'entre elles, d'artefacts de procédure dans les données (FERRARI-DISNER, 1979) mais à leur trop grand pouvoir réductionnel (LABOV, 1979). Si nous nous limitons aux méthodes d'analyse des données qui fournissent les vecteurs propres de chaque locuteur (analyse en composantes principales) dans l'espace des n paramètres observés, il nous sera possible d'opérer - et dans le cas où pour tous les locuteurs c'est la même hiérarchie des paramètres qui rend compte de la variance - dans un même espace, différentes transformations pour passer des configurations d'un locuteur à celle d'un autre. Nous aboutirons à des relations sur les ensembles (union, intersection, complémentarité) définissant la configuration des réalisations systématiques propre à chaque locuteur.

Ainsi dans les réalisations des oppositions systématiques d'arrondissement examinées en français sur 5 locuteurs (ABRY & al., 1979 et BOË & al., 1980), il est clair que les différentes transformations sur l'espace que nous avons obtenu par analyse en composantes principales (espace qui se ramène au facteur de forme K_2 de l'orifice labial fonction de son aire S) n'empêchent pas le comportement d'un locuteur (D.L.) d'être radicalement différent de celui des autres. Ce dernier présente une dynamique propre plus importante (qui peut être aisément normalisée) mais ce n'est pas là l'irréductibilité; l'essentiel est qu'il est le seul à séparer, par le jeu des lèvres, outre ses voyelles arrondies et non arrondies, ses con-

sonnes sibilantes,et celles-ci par le seul facteur de forme. La question qui pourrait se poser est la suivante : n'aurions-nous pas affaire, dans le cas de ce locuteur, à un système de réalisation individué différent de celui des autres locuteurs (c'est-à-dire de celui du français commun) utilisant comme discriminant phonologique le trait d'arrondissement pour les consonnes comme pour les voyelles ? La réponse dans ce cas serait aisée à trouver puisqu'un test de lecture labiale montrerait sans doute que ce locuteur a le même taux de confusion en autocompréhension sur les oppositions de type sy/ʃy que les autres locuteurs francophones, pour lesquelles ces réalisations sont habituellement des sosies labiaux.

- A la lumière de cet exemple il est possible de mettre en place des procédures de décision viables (test d'autocompréhension) qui permettent de décider entre réalisation individuée d'un système inter-individuel et un système à la limite autistique, qui présente tout au plus une potentialité fonctionnelle que le sujet pourra utiliser à son propre compte, ou dans les rares cas où il tombera sur un interlocuteur présentant les mêmes particularités de code.

- Ayant ainsi dégagé la variation individuelle intra-systématique, les stratégies expérimentales que nous allons proposer auront pour but de mettre en évidence les fonctionnements de cette variation. Pour commencer, il semblera de bonne méthode de tester la tolérance perceptive des différentes normalisations apportées à l'espace propre de chaque locuteur par rapport à la transmission de leur système commun. Par exemple, si une augmentation ou une diminution de la dynamique d'un paramètre d'un locuteur aboutit à une dégradation importante de l'intelligibilité, il sera loisible de suspecter la normalisation.

- Il semble indispensable de mettre aussi en oeuvre des tests de résistivité de chaque système individué sous différentes conditions de dégradation (masquage, filtrage, etc.) pour évaluer leurs efficacités pragmatiques respectives.

- L'application des procédures habituelles qui permettent de tester la pertinence des caractéristiques dégagées à l'analyse sont plus délicates en ce qui concerne les traits des systèmes individués. Les tests d'identification ont déjà été éliminés (rappelons-le, le sujet n'utilisant pas

les différences dans un test d'autocompréhension); restent les tests de discrimination.

A QUAND UNE TYPOLOGIE IDIOLECTALE ...

La fonction d'intelligibilité ayant été ainsi soigneusement explorée, il sera donc possible d'utiliser les indices dégagés pour la caractérisation phonique en tenant compte des différents plans fonctionnels, ce qui doit permettre une véritable typologie idiolectale à laquelle ne sauraient prétendre les tentatives qui ont été faites jusqu'ici. les unes ont sacrifié à l'établissement du système linguistique toute méthodologie de l'individuation, les autres ont couru directement à l'identification ou à la vérification du locuteur par des procédures totalement aveugles à l'organisation de la communication linguistique individuée.

BIBLIOGRAPHIE

ABRY (C.), BOË (L.J.), GENTIL (M.) & DESCOUT (R.), GRAILLOT (P.), 1979, La géométrie des lèvres en français. Protrusion vocalique et protrusion consonantique. - 10e JEP, GCP du GALF, pp. 99-110.

BOË (L.J.), ABRY (C.) & CORSI (P.), 1980, Les problèmes de normalisation interlocuteurs : une méthode d'ajustement aux limites. - 11e JEP, GCP du GALF.

BOË (L.J.) & CORSI (P.), (à paraître), Reconnaissance du locuteur (identification et vérification). Présentation bibliographique.-

BOURDIEU (P.), 1972, Esquisse d'une théorie pratique. - Droz, Paris-Genève.

BÜHLER (K.), 1934, Axiomatik der Sprachwissenschaft. - Kant-Studien 38, 40 et Sprachtheorie 28, Iéna.

FERRARI-DISNER (S.), 1979, Cross-linguistic normalization. - 9th Int. Congr. Phonetic Sci. 1, p. 262.

GERSTAMN (L.-H.), 1968, Classification of self-normalized vowels. - IEEE Trans. Audio Electroacoust. AU-16, pp. 78-80.

HARSHMAN (R.), 1970, PARAFAC : Models and conditions for an "explanatory" multimodal factor analysis. - Working Papers in Phonetic 16, Phonetics Labs. UCLA

KAISER (L.), 1939, Biological and statistical research concerning the speech of 216 Dutch students. - I, Archives Néerlandaises de Phonétique Expérimentale 15, pp. 1-76.

KAISER (L.), 1940, Biological and statistical research concerning the speech of 216 Dutch students. - II, Archives Néerlandaises de Phonétique Expérimentale 16, pp. 77-136.

KAISER (L.), 1941, Biological and statistical research concerning the speech of 216 Dutch students. - III, Archives Néerlandaises de Phonétique Expérimentale 17, pp. 143-211.

KAISER (L.), 1942, Biological and statistical research concerning the speech of 216 Dutch students. - IV, Archives Néerlandaises de Phonétique Expérimentale 18, pp. 1-58.

KAISER (L.), 1944, Biological and statistical research concerning the speech of 216 Dutch students. - V, Archives Néerlandaises de Phonétique Expérimentale 19, pp. 37-78.

LADEFOGED (P.) & BROADBENT (D.-E.), 1957, Information conveyed by vowels. - J.A.S.A. 29, pp. 98-104.

LABOV (W.), 1979, A Sociolinguistic approach to the problem of normalisation.- 9th Int. Congr. Phonetic Sci. 1, Paper 441.

LAZICZIUS (J.-V.), 1935, A New category in phonology. - 2nd Int. Congr. Phonetic Sci., pp. 57-60.

LOBANOV (B.-M.), 1971, Classification of Russian vowels spoken by different speakers. - J.A.S.A. 49, pp. 606-608.

NEAREY (T.), 1977, Phonetic features systems for vowels. - Unpublished Diss. Univ. of Connecticut.

NORDSTRÖM (P.-E.) & LINDBLOM (B.), 1975, A Normalization procedure for vowel formant data. - 8th Int. Congr. Phonetic Sci. Paper 212.

ROUSSEAU (P.) & SANKOFF (D.), 1978, A Solution to the problem of grouping speakers. - In : Linguistic variation models and methods. New-York-San Francisco-London, SANKOFF Ed., Academic Press, pp. 97-117.

SANKOFF (D.), SHORROCK (R.W.) & Mc KAY (W.), 1974, Normalization of formant space through the least squares affine transformation. - Unpublished Program and Documentation.

SLIS (I.H.) & COHEN (A.), 1969 a., On the complex regulating the voiced-voiceless distinction. - I, Language & Speech 12, pp. 80-102.

SLIS (I.H.) & COHEN (A.), 1969 b., On the complex regulating the voiced-voiceless distinction. - II, Language & Speech 12, pp. 137-155.

STEVENS (K.N.), 1971, Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. - 7th Int. Congr. Phonetic Sci., pp. 206-231.

TROUBETZKOY (N.S.), 1970, Principes de phonologie. - Trad. fr. de Grundzüge der Phonologie, 1939. TCLP VII, 272 p., Prague, 1 ed. 1949, Klincksieck, Paris.

WEINREICH (U.), 1954, Is a structural dialectology possible ? Word 10, pp.388-

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

LES PROBLEMES DE NORMALISATION INTERLOCUTEURS. METHODE D'AJUSTEMENT AUX LIMITES.

BOË Louis-Jean — Institut de Phonétique
ABRY Christian — Grenoble
CORSI Patrick — Laboratoire d'Informatique et de Mathématiques Appliquées - Grenoble

RESUME

Les difficultés que rencontre l'adaptation au locuteur en reconnaissance automatique de la parole et les succès obtenus par les différentes procédures de vérification de ce même locuteur rendent véritablement indispensable la prise en compte théorique des multiples données de l'analyse sur les différences inter-individuelles.

Le traitement de cette différence n'a jamais été véritablement intégré dans l'analyse des systèmes phonétiques bien que des descriptions très minutieuses aient été fournies dès la fin du XIXe siècle, dans le champ des études dialectologiques.

Les seules procédures qui se soient développées - et qui n'ignorent pas purement et simplement cette différence - ont pour but de la réduire : c'est le cas des différentes normalisations proposées jusqu'ici.

Dans l'analyse des différences inter-individuelles, nous proposons :

1. De normaliser les paramètres descriptifs (déterminés dans une étape préliminaire de traitement des données) traités en logarithme, sur la base de leur dynamique maximale. Nous tenons compte ainsi de l'exploration maximale de l'espace des réalisations considérées (ici l'espace articulatoire des lèvres) et de la précision relative des gestes dans la production de la parole.
2. De tester les constantes structurales irréductibles des espaces individuels pour déceler s'il s'agit ou non d'un véritable système individuel ou d'une individuation du système (ABRY & BOË, 1980)*.

*. Système phonétique, idiolecte et différences individuelles (dans ces mêmes J.E.P.).

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

LES PROBLEMES DE NORMALISATION INTERLOCUTEURS. METHODE D'AJUSTEMENT AUX LIMITES.

BOË Louis-Jean — Institut de Phonétique
ABRY Christian — Grenoble
CORSI Patrick — Laboratoire d'Informatique et de Mathématiques Appliquées - Grenoble

SUMMARY

Difficulties in the field of automatic recognition of speech and success in talker recognition show at least a lack of theoretical treatment of individual differences.

Phonetic analysis of systems has not yet integrated those data, in spite of some careful descriptions available as soon as the end of the 19th century from dialect studies.

Until now, the only treatment that don't simply cancel those differences has been a somewhat reducing one : namely normalization procedures.

To analyze interindividual differences we propose :

1. To normalize descriptive parameters (first chosen by a preliminary data treatment) by expressing them in logarithmic function referenced to their maximum amplitude. We try so to account for a maximal exploration of the space under consideration (i.e. lip space for French) and of the relative precision of articulatory gestures in speech production.
2. To test irreductible structural constants of individual spaces to detect if we are concerned with a true idiolectal system or simply with an individuation of a common system.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

LES PROBLEMES DE NORMALISATION INTERLOCUTEURS. METHODE D'AJUSTEMENT AUX LIMITES.

BOË Louis-Jean — Institut de Phonétique
ABRY Christian — Grenoble
CORSI Patrick — Laboratoire d'Informatique et de Mathématiques Appliquées - Grenoble

INTRODUCTION

La diversité des réalisations interlocuteurs a pu être quantifiée au fur et à mesure que se sont développées les techniques d'analyse de la parole et qu'elles ont été utilisées pour de vastes échantillons. Les travaux de Louise KAISER (1939, 40, 41, 42, 44) constituent le véritable départ, malheureusement peu suivi, de ce que devraient être les études systématiques des variations idiolectales. Ces descriptions, qui pourraient passer pour pur pointillisme microlinguistique, forment en réalité une banque de données inestimable pour qui se trouve confronté aux problèmes de l'adaptation au locuteur, en reconnaissance automatique.

Les solutions qui peuvent être apportées sont de deux types :
- traiter le plus opérationnellement possible les différences interlocuteurs sans a priori théorique sur leurs causes biologiques ou linguistiques;
- construire une théorie de la variabilité interindividuelle.

Les problèmes que posent cette variabilité ont été longtemps évacués des descriptions linguistiques (ABRY & BOË, 1979) faute d'outil théorique permettant leur traitement. Les linguistes, qui ne semblent pas rencontrer le même type d'écueil que les praticiens de la reconnaissance, sont pourtant confrontés périodiquement au dilemme suivant :

- "Vous n'avez pris qu'un seul locuteur, vous ne décrivez par conséquent qu'un idiolecte".
- "Vous avez pris n locuteurs, quels sont vos critères d'homogénéité bio- et sociolinguistiques ?".

A en juger par les réponses données, ce réductionnisme de la variabilité ne jouit pas d'une grande quiétude épistémologique.

Par contre le pragmatisme des exploiteurs de la différence, qui travaillent dans le domaine de l'identification ou de la vérification du locuteur, débouche sur des résultats tangibles (CORSI, 1979).

Jusqu'à présent les problèmes théoriques ont été relancés sur deux plans :

- au niveau de la production, par l'explicitation des causes de diversité articulatoires et acoustiques (STEVENS, 1971);
- par les études de dialectologie ou de sociolinguistique : ROUSSELOT (1891), ..., WEINREICH (1954), ..., LABOV (1979).

Les procédures de normalisation, jusqu'ici proposées, ont pour but essentiel de réduire, plus ou moins pragmatiquement, la différence.

LES PROCEDURES DE NORMALISATION

L'essentiel des travaux de normalisation interlocuteurs portent sur les formants vocaliques[1] : ils sont basés sur des présupposés anatomiques (longueur et proportions du conduit vocal) ou se justifient par leur efficacité classificatoire dans le plan F1/F2 ou F1/F2-F1. Citons les travaux de : FANT, 1966; SUZUKI & al., 1967; GERSTMAN, 1968; KLEIN & al., 1970; HARSHMAN, 1970; LOBANOV, 1971; SCHWARTZ, 1971; LIENARD & al., 1973; MAISSIS, 1973; SANKOFF & al., 1974; FANT, 1975; NÖRSTROM & LINDBLOM, 1975; BERNSTEIN, 1977; BROAD & WAKITA, 1977; NEAREY, 1977; WAKITA, 1977; FERRARI-DISNER, 1979; LABOV, 1979.

---

1. Il serait injuste d'oublier les travaux de comparaison interlocuteurs sur une des caractéristiques individuelles les mieux explicables, la fréquence laryngienne. Voir, par exemple, JASSEM, 1971, Journal of the International Phonetics Association 1, pp. 59-68, et BOË & al., 1975, Phonetica 32, pp. 1-23.
En ce qui concerne la géométrie des lèvres et ses relations avec le plan acoustique le seul travail de normalisation que nous connaissons est celui de LINKER, 1979, qui porte sur une comparaison des voyelles de quatre langues (cantonais, finnois, français et suédois).

Certaines normalisations sont linéaires d'autres plus complexes et c'est là un débat ouvert par FANT dès 1959. En effet, lorsque l'on considère les différences anatomiques du conduit vocal et leurs conséquences acoustiques, on a de bonnes raisons de croire avec lui qu'une simple homothétie ne permet pas de passer du conduit vocal d'un homme adulte à celui d'une femme ou d'un enfant.

Dans ce qui suit, nous passerons rapidement en revue celles des procédures pour lesquelles nous disposons d'évaluations comparatives.

. La procédure (I) est une simple homothétie, la fréquence du formant i est normalisée F_i^N par rapport au maximum des fréquences formantiques de l'ensemble vocalique considéré :

$$F_i^N = F_i / F_i \text{ max.} \qquad \text{(I)}$$

. GERSTMAN utilise les valeurs minimum $F_i$min et maximum $F_i$max des formants relevés pour l'ensemble des voyelles d'un même locuteur. Les valeurs normalisées sont donc référencées par rapport aux extrêmes :

$$F_i^N = (F_i - F_i \text{ min}) / (F_i \text{ max} - F_i \text{ min}) \qquad \text{(II)}.$$

. LOBANOV normalise par rapport à la valeur moyenne F_i moy et l'écart type σ_i calculé pour l'ensemble des voyelles :

$$F_i^N = (F_i - F_i \text{ moy}) / \sigma_i \qquad \text{(III)}$$

Les formants de chaque locuteur sont donc évalués par rapport à une valeur centrale pour laquelle il y a peu d'opposition phonétique, ce qui amène à une meilleure discrimination entre les voyelles périphériques.

. NEAREY part aussi de l'hypothèse qu'il existe une relation simple :

$$F_{ijk} = K_{ij1} F_{ijk}$$

permettant de passer des valeurs formantiques i (i = 1, 2) de la voyelle j du locuteur k à celles du locuteur 1. Il propose donc, par locuteur, une normalisation du type :

$$F_{ijk}^N = \text{Log } (F_{ijk}) + C_{ik} \qquad \text{(IV)}$$

où C est une constante dépendant du locuteur k. NEAREY référence par rapport à la valeur moyenne de F_i pour chaque locuteur.

. La méthode PARAFAC développée par HARSHMAN est beaucoup plus générale, elle permet de comparer les paramètres de différents locuteurs et de plusieurs langues et de minimiser les différences. Elle a été exploitée par Sandra FERRARI-DISNER pour les formants des systèmes vocaliques de l'allemand, du hollandais, du norvégien, du suédois, du danois, de l'anglais-américain et de l'anglais-californien; Wendy LINKER l'a utilisée pour des paramètres géométriques des lèvres en cantonais, en finnois, en français et en suédois.
Explicitons-la dans le cas d'une application aux valeurs formantiques. Les différences d_{ijk} du formant i par rapport à sa moyenne pour la voyelle j du locuteur k s'écrit :

$$d_{ijk} = f_{i1}\ v_{j1}\ s_{k1} + f_{i2}\ v_{j2}\ s_{k2} + \varepsilon$$

f_i, v_j, s_k sont des pondérations relatives à toutes les valeurs du formant i, de la voyelle j, du locuteur k et ε est l'erreur d'ajustement. On obtient ainsi les valeurs normalisées :

$$F^N_{ijk} = f_{i1}\ v_{j1}\ s_1 moy + f_{i2}\ v_{j2}\ s_2 moy + f_{ik} moy \quad (V)$$

$s_1 moy$ et $s_2 moy$ sont les moyennes des termes s_{k1} et s_{k2} précédents, f_{ik} est calculée pour le formant i et pour toutes les voyelles du locuteur k. Pour se placer au plus près des conditions perceptives FERRARI-DISNER utilise l'échelle des mels.

LES CRITERES D'EVALUATION

Les critères d'évaluation des procédures de normalisation n'ont pas été systématiquement établis.

. LOBANOV détermine un coefficient moyen η, calculé pour l'ensemble des voyelles normalisées et tous locuteurs réunis qui tient compte de la compacité de chaque aire de dispersion et de la distance entre les n paires voisines (figure 1) :

$$\eta = \frac{1}{n} \sum_{k,1}^{n} \frac{R_{k1\ min}}{max\ (d_k max,\ d_1 max)}$$

Il s'agit d'un critère purement fonctionnel qui tire sa légitimité de l'évaluation qu'il donne de la discrimination.

Calculé pour les voyelles du russe réalisées par cinq locuteurs, LOBANOV obtient le classement suivant pour les procédures que nous avons présentées.

	I	II	III
η	2,1	2,5	4,1

η = 1 correspondant aux voyelles non normalisées.

Sur la base de η sa méthode semble donc la plus efficace.

. L'approche de FERRARI-DISNER nous semble fort intéressante : elle fait intervenir un critère de réduction de la variance mais aussi une minutieuse évaluation linguistique de la normalisation.

La première partie de l'évaluation est opérée à partir d'ellipses de dispersion dont les diamètres, orientés selon les composantes principales, ont pour valeur deux fois l'écart type mesuré sur l'ensemble des points d'une voyelle donnée; ces ellipses contiennent environ 95% de chaque population vocalique (DAVIS, 1976). Plus une normalisation sera puissante, plus la somme des aires des ellipses (calculée pour toutes les voyelles) sera faible. Pour prendre en compte l'effet que la procédure opère sur la discriminabilité, les différents ensembles d'ellipses sont retracés en gardant constante la somme de leurs distances. En prenant comme référence les données brutes (efficacité égale à 1) et en reprenant les résultats de FERRARI-DISNER, nous obtenons les efficacités suivantes :

	II	V	III	IV
Efficacité	1,08	1,34	1,73	2,85

Sur la base de cette évaluation, le pouvoir de réduction de la procédure proposée par NEAREY apparaît nettement, devançant celle de LOBANOV, PARAFAC n'arrivant qu'en troisième position.

Dans la deuxième partie, FERRARI-DISNER étudie les effets de la normalisation à la lumière des connaissances de la phonologie et de la phonétique contrastive. Il s'agit d'une interprétation minutieuse dont la démarche est difficilement généralisable (puisqu'elle dépend essentiellement des systèmes phonétiques étudiés) et qui ne se prête pas à une évaluation quantitative. Mais c'est évidemment la plus proche d'une démarche linguistique. Après une longue analyse, elle estime que PARAFAC,

bien que moins puissante, se prête mieux en l'occurrence à des études phonétiques.

REMARQUES

Il faut bien noter que toutes ces normalisations sont opérées à partir d'une hypothèse implicite relativement forte : les différences inter-locuteurs peuvent être réduites par une procédure plus ou moins complexe puisque ce sont les réalisations d'un même système. La possibilité d'une individuation du système (ABRY & BOË, 1979), c'est-à-dire d'une adaptation individuelle du système par le locuteur, n'est jamais soupçonnée. Lorsque les résultats de la normalisation sont peu évidents, leurs auteurs n'hésitent pas à en rejeter la responsabilité sur les données en suspectant leur homogénéité. Enfin, aucune méthodologie n'est proposée pour classer les locuteurs à partir des résidus de cette normalisation.

NORMALISATION AUX LIMITES - APPLICATION A LA GEOMETRIE DES LEVRES

Nous supposons que, dans un premier temps, on dispose pour un ensemble de réalisations données d'un nombre limité de paramètres dont on a mis en évidence le pouvoir descriptif. Cette première étape peut passer par une analyse factorielle suivie d'une analyse discriminante. Nous avons utilisé des données extraites d'une étude sur la labialité en français. Ont été mesurés l'aire aux lèvres S et le facteur de forme K2, rapport entre l'écartement et l'aperture intérolabiale, pour les voyelles et les consonnes dans le cas de syllabes CV avec V = [i, e, y, ø] et C = [s, z, ʃ, ʒ] . Nous avons tracé pour 5 locuteurs les ellipses de dispersion dans le plan log S / log K2 pour les 4 classes de son (figure 2). En effet un traitement statistique préliminaire[2] (analyse des correspondances) nous a permis de dégager, parmi nos 12 paramètres, S et K2 comme de bons interprétants des deux premiers facteurs.

---

[2] . Pour plus de détails voir : DESCOUT & al., 1978, 9e JEP, pp. 179-189; ABRY & al., 1979, 10e JEP, pp. 99-110; ABRY & al., 1979, 9th Int. Congr. Phon. Sci. I, p. 177; GRAILLOT & al., 1980, Séminaire Int. sur la Labialité, GALF, Lannion.

Dans l'espace ainsi déterminé (et nous nous limiterons à un espace à 2 dimensions), on recherche le plus petit convexe contenant les ellipses de dispersion des classes d'apprentissage. Pour ce faire, on utilise des droites parallèles aux axes. On contracte, pour chaque locuteur, ce convexe de telle façon que sa projection sur chaque axe ait une amplitude égale à une référence déterminée et on le translate de manière à obtenir des valeurs identiques pour tous les locuteurs.

Par ailleurs, nous avons choisi d'opérer sur des fonctions logarithmiques des paramètres p_i (ici S et K2). Cette normalisation s'écrit donc pour le locuteur k et le paramètre p_i :

$$\log p^N_{ik} = a_{ik} \log p_{ik} + b_{ik}$$

avec $a_{ik} = \log (p_{ir}\ \max / p_{ir}\ \min) / \log (p_{ik}\ \max / p_{ik}\ \min)$

et $b_{ik} = \log p_{ir}\ \min - a_{ik} \log p_{ik}\ \min$

les valeurs p_{ir} correspondent à celles du locuteur de référence r choisi parce qu'il exploite la dynamique maximale; les valeurs p_i max et p_i min étant relevées par locuteur et pour l'ensemble des réalisations.

. En prenant des valeurs extrêmes et non des moyennes, on obtient l'espace maximum exploité pour une structure donnée, c'est ce que, avec CATFORD[3], nous appellerons l'exploitation linguistique des potentialités anthropophoniques. En adaptant un seuil de confiance en fonction de la nature des données et de la précision de leur obtention, il est possible de s'affranchir des valeurs par trop marginales.

Les ellipses de dispersion utilisées présupposent, par classe, une distribution normale des données, ce qui n'est pas une hypothèse trop contraignante. Leur équation est de la forme :

$$\frac{1}{1-R^2}\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 - 2R\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right) + \left(\frac{y-\bar{y}}{\sigma_y}\right)^2\right] = \chi^2_\alpha(2)$$

où R est le coefficient de corrélation et χ^2_α (2) désigne le point de pourcentage 100 α de la distribution du χ^2 à deux degrés de liberté.

---

[3]. CATFORD, 1968, In : Manual of Phonetics, pp. 309-332. North Holland Publishing Co. Amsterdam.

Un des axes de ces ellipses est la droite de régression orthogonale. Noter que les points de tangence des ellipses avec le convexe ne dépendent pas des échelles.

. Le choix des valeurs logarithmiques peut se justifier en fonction de la nature des paramètres traités. Pour la fréquence laryngienne ou pour les formants, elles peuvent être logiquement utilisées pour des raisons perceptives. Dans le cas de la normalisation que nous avons choisi d'illustrer, il s'agit de paramètres directement déductibles de mesures géométriques du conduit vocal. Très généralement nous pensons que dans le cas d'un geste articulatoire, les dispersions ne doivent pas être considérées en valeurs absolues mais en précision relative par rapport à l'amplitude du geste lui-même[4]. L'utilisation des résultats en échelle logarithmique nous semble donc judicieuse pour ce type de données. Plus précisément, dans notre cas, la modélisation de l'anatomie des lèvres proposé par LINDBLOM & SUNDBERG étant :

$$y = \pm \frac{B}{2}\left[1 - \left(\frac{2}{A}|x|\right)^{p}\right]$$

avec $p = K1 / (1-K1)$ et $K1 = S / AB$.
A et B étant respectivement l'écartement et l'aperture et S l'aire aux lèvres. Nous constatons que notre procédure permet de normaliser deux locuteurs ayant des paramètres A, B et p différents.

. D'autre part, l'influence de l'option logarithmique sur les résultats d'une évaluation de notre normalisation (critères de LOBANOV et de FERRARI-DISNER) ne peut que la rapprocher des performances de la procédure de NEAREY.

En ce qui concerne les critères plus proprement linguistiques, les commodités d'une telle normalisation pour la typologie des réalisations d'un système par les locuteurs (son individuation) peut se faire jour, dans un premier temps, en nous permettant de dégager, par des rela-

---

[4]. Cf. les travaux de FITTS, 1954, The Information capacity of the human motor system in controlling the amplitude of movement. - Journal of Experimental Psychology 47, pp. 381-391.

[5]. LINDBLOM & SUNDBERG, 1971, JASA 50, pp. 1166-1179.

tions ensemblistes sur les ellipses, les propriétés essentielles de la struture de ces réalisations, les premières à être notoirement pertinentes pour relier cette structure à celle du système.

Dans cette optique, un des premiers critères sera bien entendu la disjonction des ensembles correspondant aux oppositions phonémiques. On constate (figure 3) que tous nos locuteurs séparent, par le jeu des lèvres, les voyelles [+ rond] i, e, des voyelles [- rond] y, ø, tous contextes confondus. Ce critère peut être bien entendu affiné par un second qui est celui de la distance entre ellipses (distance "aux bords") permettant d'évaluer l'efficacité du principe de distinctivité[6]. Ainsi les mieux "distancés" pour cette opposition d'arrondissement sont-ils, dans l'ordre, les locuteurs FE puis JC, DL, DA et enfin PC.

Pour les consonnes, le premier critère ne fonctionne que pour un locuteur DL dont nous allons examiner plus en détail le comportement. En effet, si l'individuation d'un système peut se manifester, comme nous l'avons vu, par la bonne "tenue à distance" des oppositions phonémiques, le comportement des locuteurs dans les zones où le système est réputé redondant, peut être particulièrement révélateur. Rappelons que, pour les consonnes ʃ, ʒ et s, z, leur opposition n'est pas censée reposer sur le comportement des lèvres mais sur celui de la langue. DL en surdifférenciant labialement cette opposition nous met devant l'alternative suivante : ce locuteur utilise-t-il ou non son particularisme dans la communication linguistique ? On sait que l'information visuelle labiale et plus généralement faciale est intégrée à l'information acoustique[7]. Il s'agirait donc de démontrer par un test d'identification que ce sujet n'est pas capable de différencier en labiolecture [sy] de [ʃy] bien qu'il ait les moyens de le faire.

C'est par des procédures de cet ordre[8] que l'on pourrait éviter à l'avenir les discussions à échappatoires sur l'homogénéité systématique des locuteurs alors qu'il peut s'agir de véritables individuations.

---

6. LINDBLOM, 1978, Bulletin de l'Institut de Phonétique de Grenoble 7, pp. 1-23.

7. SUMMERFIELD, 1980, Phonetica 36, pp. 314-331.

8. ABRY & BOË, 1979, et communication dans ces mêmes JEP.

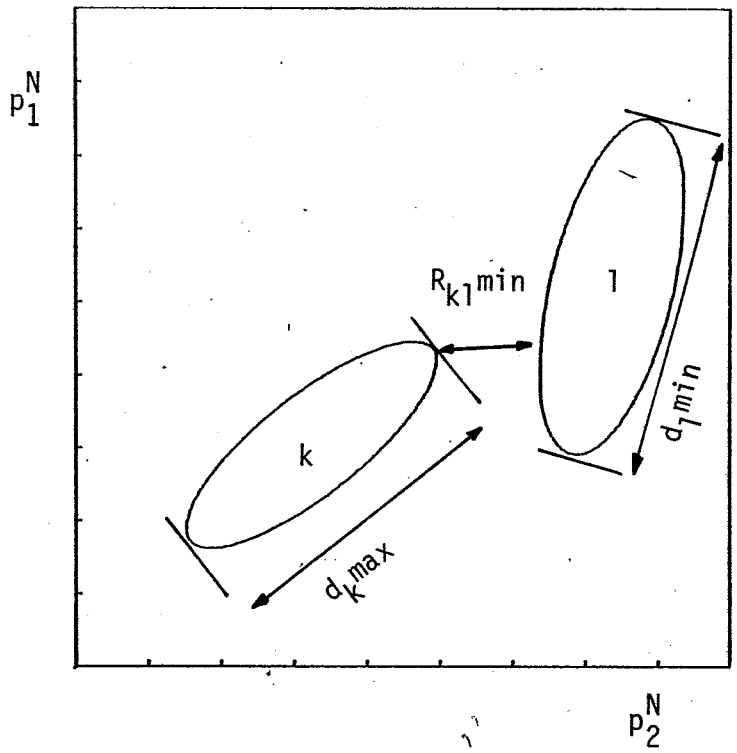


Figure 1 - Paramètres entrant en jeu dans la détermination du coefficient η proposé par LOBANOV pour évaluer les procédures de normalisation.

p_i^N : paramètre normalisé

k, 1 : paires de réalisations caractérisées par leurs ellipses de dispersion E_k et E_1

R_{k1} min : distance minimale entre E_k et E_1

$d_{k,1}$ max : grands axes E_k et E_1

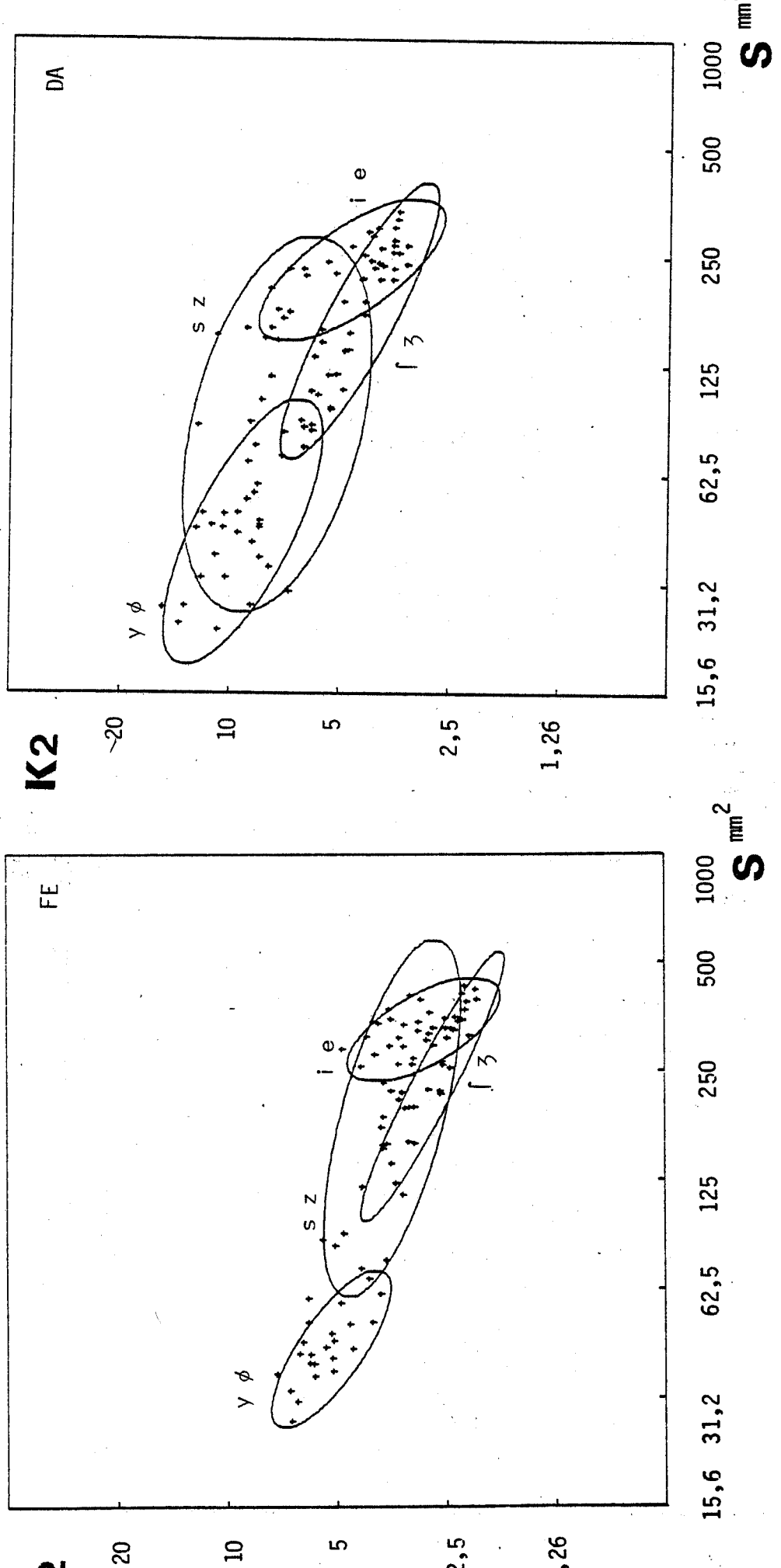


Figure 2 - Répartition dans le plan S/K2 des réalisations [i e], [y ø], [s z], [ʃ ʒ]
Locuteurs DA et FE. Résultats non normalisés.

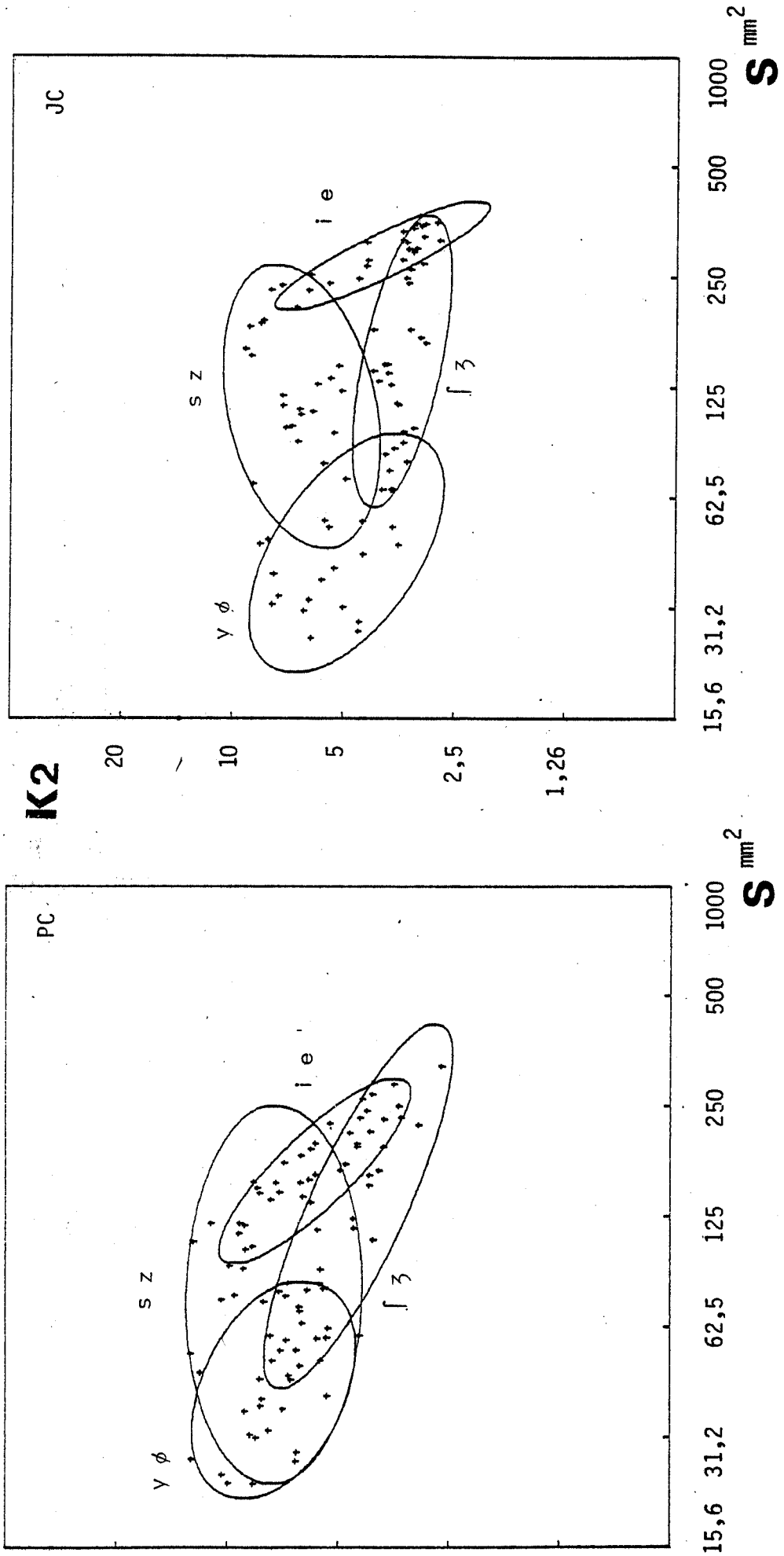


Figure 2 - Répartition dans le plan S/K2 des réalisations [i e], [y ø], [s z], [ʃ ʒ]
Locuteurs PC et JC. Résultats non normalisés.

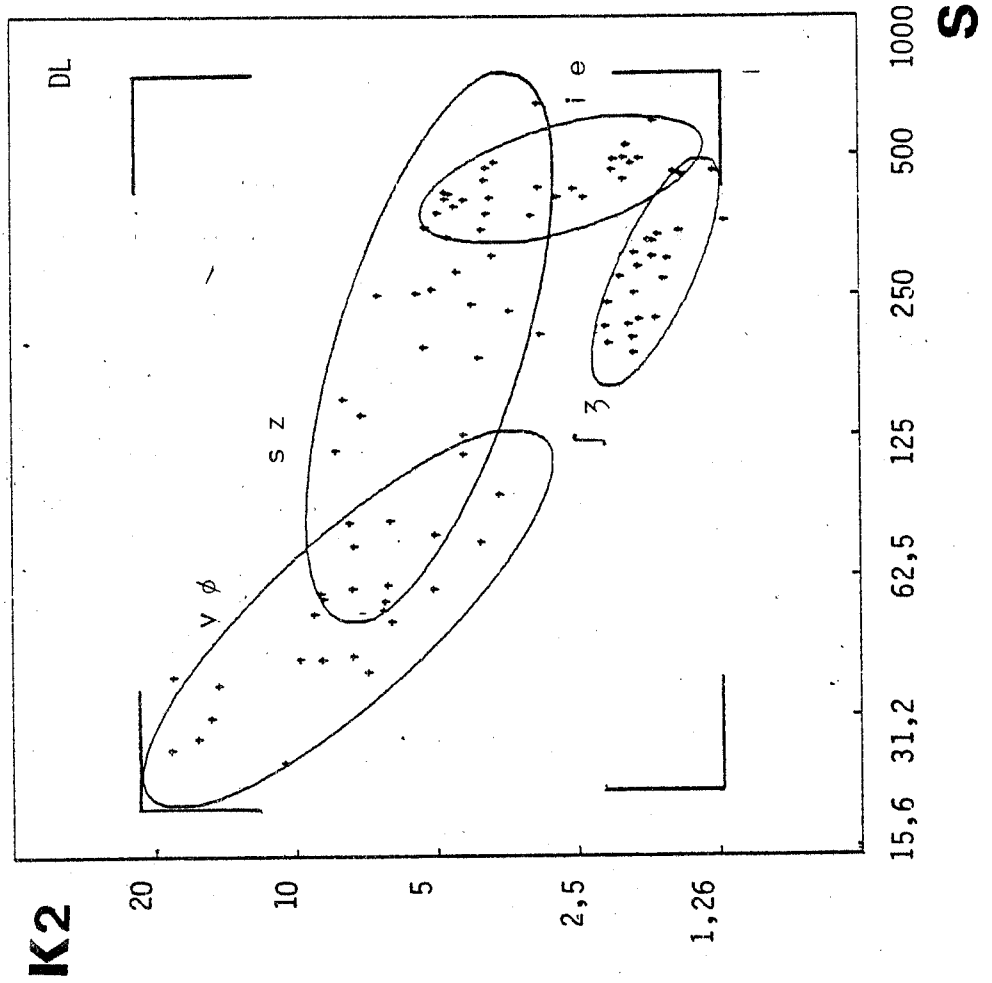


Figure 2 - Répartition dans le plan S/K2 des réalisations [i e], [y ø], [s z], [ʃ ʒ]
Locuteur DL. Ce locuteur va nous servir de référence.

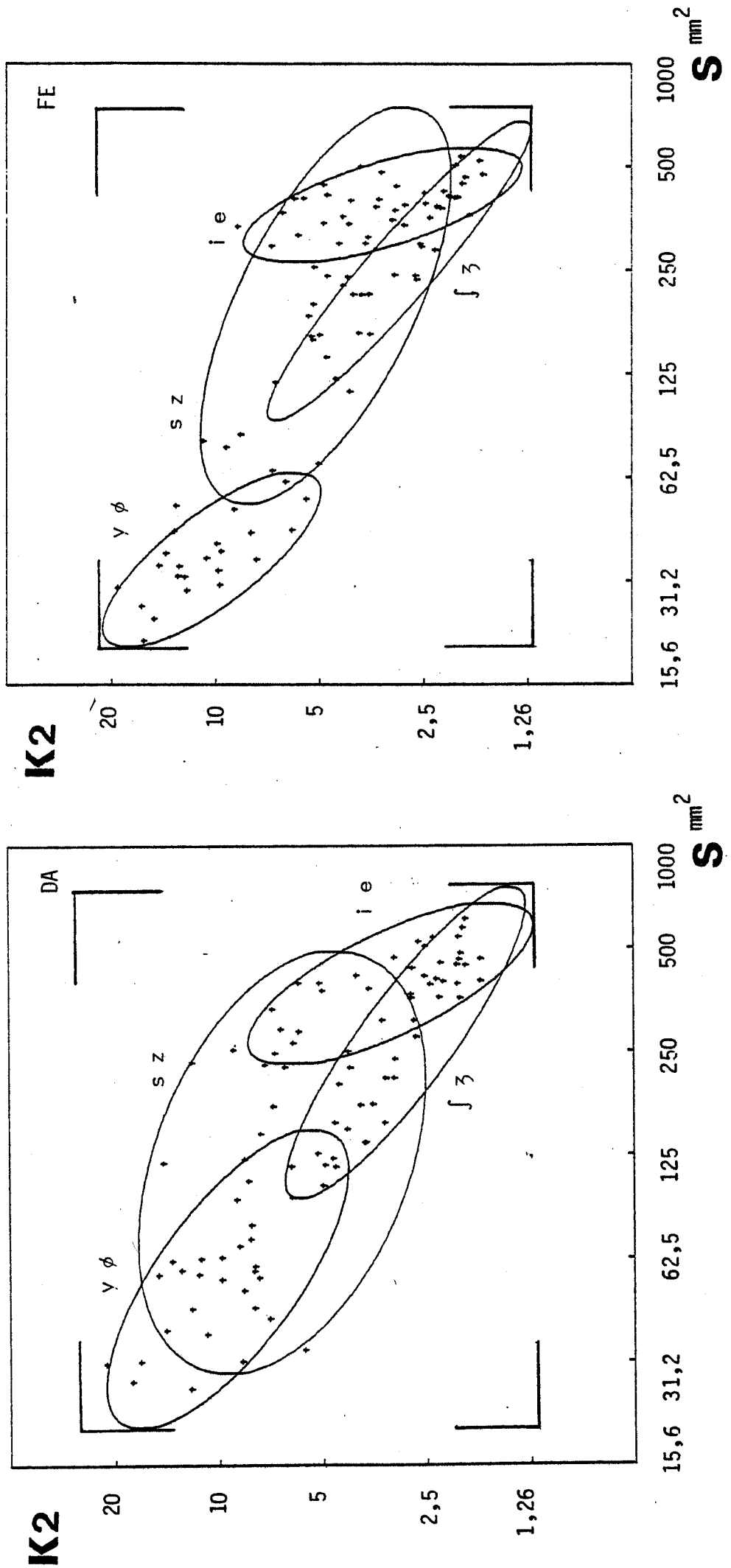


Figure 3 - Répartition dans le plan S/K2 des réalisations [i e], [y ø], [s z], [ʃ ʒ]
Locuteurs DA et FE normalisés sur la référence DL.

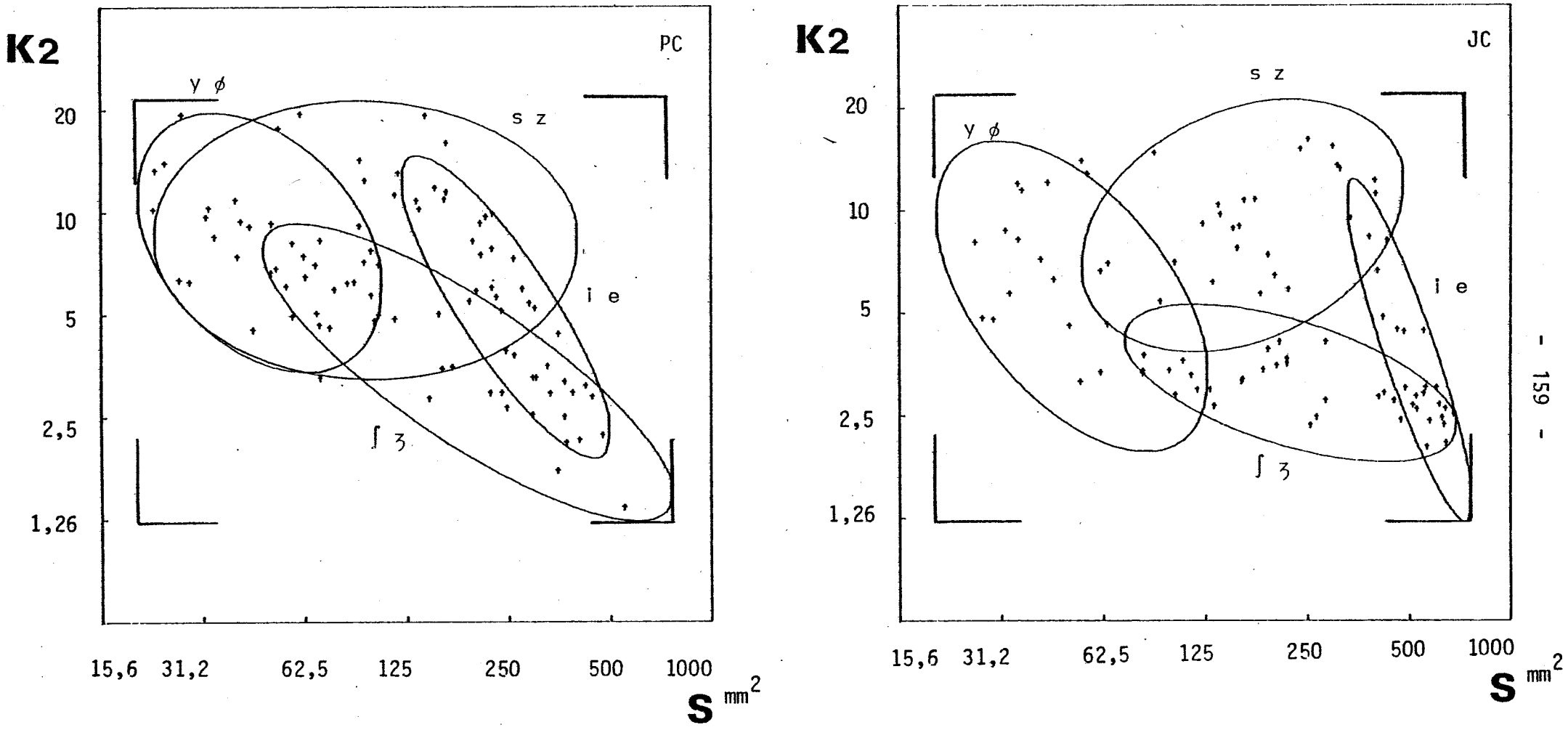


Figure 3 - Répartition dans le plan S/K2 des réalisations [i e], [y ø], [s z], [ʃ ʒ]
Locuteurs PC et JC normalisés sur la référence DL.

Les programmes utilisés pour cette procédure de normalisation et pour la visualisation graphique ont été écrits et implantés sur LSI 11/2 à l'Institut de Phonétique de Grenoble.

BIBLIOGRAPHIE

ABRY (C.) & BOË (L.J.), 1979, Pour une idiolectologie : aspects phonétiques de l'identité. - Colloque Int. Production et Affirmation de l'Identité, Toulouse.

BARTH (S.), CHULLIAT (R.), 1978, Modifications inter-locuteurs de l'échelle formantique. - 9e JEP, GCP du GALF, pp. 149-156.

BERDAN (R.), 1978, Multidimensional analysis of vowel variation. - In : Linguistic variation, pp. 149-160 - New-York, San Francisco, London - Ed. by D. SANKOFF.

BERNSTEIN (J.), 1977, Vocoid psychoacoustics, articulation and vowel phonology. - Phonetics Lab., The Univ. of Michigan.

BROAD (D.J.) & WAKITA (H.), 1977, Piecewise-planar representation of vowel formant frequencies. - JASA 62, pp. 1467-1473.

CARRE (R.), LANCIA (R.), WAJSKOP (M.), 1968, Sur la production des voyelles par les locuteurs hommes et femmes. - 6th Int. Congr. Acoustics, Paper B-1-10.

CORSI (P.), 1979, Reconnaissance automatique du locuteur : présentation générale, méthodologie et expérimentation, perspectives d'application. - Thèse Docteur-Ingénieur. Institut National Polytechnique de Grenoble.

DAVIS (S.B.), 1976, Computer evaluation of laryngeal pathology based on inverse filtering of speech. - SCRL Monograph 13, California, Santa Barbara.

FANT (G.), 1959, Acoustic analysis and synthesis of speech with applications to Swedish. - Ericson Technics 15.

FANT (G.), 1966, A Note on vocal tract size and non-uniform F-pattern scalings. - RIT STL QPSR 4, pp. 22-30.

FANT (G.), 1975, Non-uniform vowel normalization. - RIT STL QPSR 2-3, pp. 1-19.

FERRARI-DISNER (S.), 1979, Cross-linguistic normalization. - 9th Int. Congr. Phonetic Sci., I, p. 262.

FERRARI-DISNER (S.), 1979, Evaluation of normalizations. - Gotland Workshop (Sweden) : Vowels : Production and perception.

GERSTMAN (L.H.), 1967, Classification of self-normalized vowels. - Conf. on Speech Comm. and Process, Paper B6 - Bedford Mass.

GERSTMAN (L.H.), 1968, Classification of self-normalized vowels. - IEEE Trans-Audio Electroacoust., AU-16, pp. 78-80.

HARSHMAN (R.), 1970, PARAFAC : Models and conditions for an "explanatory" multi-modal factor analysis. - WPP 16, Phonetics Lab. UCLA.

HINDLE (D.), 1978, Approaches to vowel normalization in the study of natural speech. - In : Linguistic Variation, pp. 161-171, New-York - San Francisco - London, Ed. by. D. SANKOFF.

KAISER (L.), 1939, Biological and statistical research concerning the speech of 216 Dutch students. - I, Archives Néerlandaises de Phonétique Expérimentale 15, pp. 1-76.

KAISER (L.), 1940, Biological and statistical research concerning the speech of 216 Dutch students. - II, Archives Néerlandaises de Phonétique Expérimentale 16, pp. 77-136.

KAISER (L.), 1941, Biological and statistical research concerning the speech of 216 Dutch students. - III, Archives Néerlandaises de Phonétique Expérimentale 17, pp. 143-211.

KAISER (L.), 1942, Biological and statistical research concerning the speech of 216 Dutch students. - IV, Archives Néerlandaises de Phonétique Expérimentale 18, pp. 1-58.

KAISER (L.), 1944, Biological and statistical research concerning the speech of 216 Dutch students. - V, Archives Néerlandaises de Phonétique Expérimentale 19, pp. 37-78.

KLEIN (W.), PLOMP (R.), & POLS (L.C.W.), 1970, Vowel spectra, vowel spaces and vowel identification. - JASA 48, pp. 999-1009.

LABOV (W.), 1979, A Sociolinguistic approach to the problem of normalization. - 9th Int. Congr. Phonetic Sci. I, p. 441.

LIENARD (J.S.), SAPALY (J.), MLOUKA (M.), 1973, Normalisation fréquentielle de la parole. - 4e JEP GCP du GALF, pp. 173-183.

LINKER (W.), 1979, A cross-linguistic study of lip positions in vowels. - 9th Int. Congr. of Phon. Sci. I, p. 201.

LOBANOV (B.M.), 1971, Classification of Russian vowels spoken by different speakers. - JASA 49, pp. 606-608 (L).

MAISSIS (A.H.), 1973, Normalisation des paramètres phonémiques en reconnaissance automatique de la parole. - 4e JEP GPC du GALF, pp. 165-210.

MOL (H.), 1970, Fundamental of phonetics. II : Acoustical models generating the formants of the vowels phonemes. - The Hague - Paris, Mouton.

NEAREY (T.), 1977, Phonetic feature systems for vowels. - Unpublished. Doct. Diss. University of Connecticut.

NORDSTRÖM (P.-E.) & LINDBLOM (B.), 1975, A Normalization procedure for vowel formant data. - 8th Int. Congr. Phonetic Sci., Paper 292.

ROUSSEAU (P.) & SANKOFF (D.), 1978, A Solution to the problem of grouping speakers. - In : Linguistic variation models and methods, Montreal, SANKOFF Ed., Academic Press.

ROUSSELOT (P.), 1891-1892, Les modifications phonétiques du langage étudiées dans le patois d'une famille de Cellefrouin (Charente), 2° Partie : Modifications historiques de l'ancien fonds du patois. - Revue des Patois gallo-romans 14-15, pp. 65-208; 19-20, pp. 209-380.

SANKOFF (D.), SHORROCK (R.W.) & Mc KAY (W.), 1974, Normalization of formant space through the least squares affine transformation. - Unpublished.Program and Documentation.

SCHWARTZ (R.M.), 1971, Automatic normalization for recognition of vowels of all speakers. - M.S. Thesis, Cambridge. M.I.T.

STEVENS (K.N.), 1971, Sources inter- and intra-speaker variability in the acoustic properties of speech sounds. - 7th Int. Congr. Phonetic Sci., pp. 206-231.

SUZUKI (H.), KASUYA (H.) & KIDO (K.), 1967, The Acoustic parameters for vowel recognition without distinction of speakers. - Conf. Speech Communication and Processing, Bedford. Paper B5.

WAKITA (K.), 1977, Normalization of vowels by vocal-tract length and its application to vowel identification. - IEEE Trans. on Acoustics, Speech and Signal Process, ASSP-25, 2, pp. 183-192.

WEINREICH (U.), 1954, Is a structural dialectology possible ? - Word 10, pp. 388-400.

THEME III :

b) adaptation des systèmes de reconnaissance automatique aux locuteurs

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

Titre : UTILISATION DE LA PREDICTION LINEAIRE EN RECONNAISSANCE ET ADAPTATION AU LOCUTEUR

Nom : GRENIER Prénom : Yves

Ecole Nationale Supérieure des Télécommunications
Département Systèmes et Communications
46, rue Barrault, 75634, PARIS
Cedex 13 - FRANCE

Résumé :

On présente ici une méthode permettant de réaliser l'adaptation à un nouveau locuteur d'un système de reconnaissance de phonèmes, dont l'apprentissage avait été réalisé par un autre locuteur. Une méthode d'identification du locuteur est également présentée. Ces deux méthodes utilisent le même prétraitement sous la forme d'une modélisation par prédiction linéaire des segments successifs d'une phrase.

Summary :

We describe a method for the adaptation of a phonemic recognition system, to a new speaker, the training phase having been performed from the voice of another speaker. We also describe a method for speaker identification. These two methods have a common basis which is a modelisation by linear prediction of successive segments of a sentence.

1 - La variabilité due au locuteur

Le signal de parole véhicule de nombreuses informations qui se répartissent en trois catégories :

- des informations sur le sens du message,
- des informations sur l'identité du locuteur,
- des informations sur l'état du locuteur.

On s'intéressera ici à la deuxième catégorie, à savoir ce qui dans le signal vocal est lié à la personne du locuteur. Une tâche importante à accomplir lors de tout traitement d'un signal de parole est la séparation de ces diverses informations, soit que l'on désire ne conserver que l'information relative au locuteur, dans un but de reconnaissance (identification ou vérification) de l'identité de ce locuteur, soit que l'on désire au contraire éliminer cette information considérée comme un brouillage du sens du texte, lorsque c'est ce sens que l'on désire reconstruire ; on se trouve dans ce dernier cas devant une tâche de normalisation ou d'adaptation au locuteur d'un système de reconnaissance de la parole.

Dans ce qui suit seront traités les deux aspects de ces études concernant l'influence du locuteur telle qu'elle est visible dans le signal émis. Nous étudierons tout d'abord l'identification du locuteur par deux méthodes, dont la première a déjà été présentée (1), la seconde est inédite. Il est à noter que le cadre restreint dans lequel ont été réalisées les expériences décrites fait du résultat de ces expériences non pas le prototype d'un système d'identification complet, mais uniquement le modèle d'une représentation possible de l'information relative au locuteur. Nous étudierons ensuite l'adaptation au locuteur par une méthode déjà présentée (2) aussi nous attacherons nous à en décrire les améliorations. Ces études sont en fait étroitement liées, par la façon dont chacune aborde le signal de parole en réalisant une modélisation de ce signal par prédiction linéaire. Un segment de parole (phrase) est ainsi décomposé en zones sur lesquelles le signal peut être considéré comme stationnaire et donc modélisable aisément, le segment est ensuite représenté par le tableau formé par l'évolution des paramètres du modèle aux instants successifs de découpage et modélisation.

2 - Modélisation d'une phrase par prédiction linéaire

La prédiction linéaire étant une technique déjà largement utilisée, nous n'en rappellons les principes que dans le cas d'un processus vectoriel, le modèle autorégressif vectoriel s'avérant utile pour une représentation de la voix d'un locuteur.
Soit y_t un processus vectoriel à valeur dans $\mathbb{R}^m$ ($t \in \mathbb{Z}$). Si ce processus est autoregressif, d'ordre p, il vérifie :

$$\varepsilon_t = y_t + A_1 y_{t-1} + \ldots + A_p y_{t-p}$$

où ε_t est un bruit blanc vectoriel dans $\mathbb{R}^m$.

Les matrices $A_1 \ldots A_p$ sont les coefficients de la prédiction linéaire :

$$\tilde{y}_t = -A_1 y_{t-1} - \ldots - A_p y_{t-p}.$$

Ces matrices (m x m) sont déterminées par minimisation de la trace de la covariance du bruit ε_t (méthode des moindre carrés) ou du déterminant de cette matrice de covariance (méthode du maximum de vraissemblance). Comme dans le cas d'un processus scalaire on obtient dans les deux méthodes, la minimisation au moyen d'équations de Yule Walker :

$$(I\ A_1 \ldots A_p) \begin{pmatrix} R_o & & R_p \\ & \diagdown & \\ R_{-p} & & R_o \end{pmatrix} = (D_p\ 0 \ldots\ldots\ldots 0)$$

où $R_{i-j} = E(y_i\ y_j^T)$, (T désigne la transposée d'une matrice). D_p est la matrice de covariance de ε_t : $D_p = E(\varepsilon_t\ \varepsilon_t^T)$.

La résolution de ces équations nécessite de considérer en plus du processus autorégressif direct, un processus autorégressif rétrograde, pour lequel le sens du temps a été inversé. Les deux processus sont alors :

$$\begin{cases} \varepsilon_t^+ = y_t + A_1^+ y_{t-1} + \ldots + A_p^+ y_{t-p} \\ \varepsilon_t^- = y_t + A_1^- y_{t+1} + \ldots + A_p^- y_{t+p} \end{cases}$$

Les équations d'optimalité s'écrivent :

$$\begin{cases} [I\ A_1^+ \ldots A_p^+]\ \mathcal{R}_p = [D_p^+\ 0 \ldots 0] \\ [A_p^- \ldots A_1^-\ I]\ \mathcal{R}_p = [0 \ldots 0\ D_p^-] \end{cases}$$

avec $\mathcal{R}_p$ la matrice de Toeplitz par blocs de l'autocorrélation du processus vectoriel.

La résolution, comme dans le cas scalaire, se fait par récurrence sur l'ordre p des modèles (3,4,5). Pour passer du modèle d'ordre p au modèle d'ordre p+1, on tente d'abord de rajouter un zéro (matrice m x m nulle) aux modèles :

$$\begin{cases} [I\ A_{p,1}^+ \ldots A_{p,p}^+\ 0]\ \mathcal{R}_p = [D_p^+\ 0 \ldots 0\ \alpha_p] \\ [0\ A_{p,p}^- \ldots A_{p,1}^-\ I]\ \mathcal{R}_p = [\beta_p\ 0 \ldots 0\ D_p^-] \end{cases}$$

Puis on forme les combinaisons linéaires des deux équations avec les coefficients matriciels I et K^+_{p+1} d'une part, K^-_p et I d'autre part, d'où :

$$[I, A^+_{p,1} + K^+_{p+1} A^-_{p,p}, \ldots, K^+_{p+1}] \mathcal{R}_p = [D^+_p + K^+_{p+1} \beta_p, o, \ldots, o, \alpha_p + K^+_{p+1} D^-_p]$$

(et une relation analogue pour le système rétrograde).

On voit alors qu'en posant :

$$\begin{cases} K^+_{p+1} = -\alpha_p \cdot [D^-_p]^{-1} \\ K^-_{p+1} = -\beta_p [D^+_p]^{-1} \end{cases}$$

on annulle les termes $\alpha_p + K^+_{p+1} D^-_p$ et $\beta_p + K^-_{p+1} D^+_p$, obtenant par conséquent les modèles d'ordre p+1 optimaux.

$$\begin{cases} A^+_{p+1,p+1} = K^+_{p+1} \\ A^+_{p+1,i} = A^+_{p,i} + K^+_{p+1} A^-_{p,p+1-i} \qquad i = 1, \ldots, p \end{cases}$$

$$\begin{cases} A^-_{p+1,p+1} = K^-_{p+1} \\ A^-_{p+1,i} = A^-_{p,i} + K^-_{p+1} A^+_{p,p+1-i} \qquad i = 1, \ldots, p \end{cases}$$

L'algorithme est complété par la remarque :

$$\alpha_p = \beta^T_p \qquad \text{et} \qquad \begin{cases} D^+_{p+1} = D^+_p - \alpha_p [D^-_p]^{-1} \beta_p \\ D^-_{p+1} = D^-_p - \beta_p [D^+_p]^{-1} \alpha_p \end{cases}$$

Le modèle autorégressif vectoriel est donc estimé par une procédure très proche de l'algorithme de Levinson pour les modèles autorégressifs scalaires. La différence essentielle est le recours imposé aux prédicteurs direct et rétrograde, du au fait que la matrice $\mathcal{R}_p$ bien que symétrique est en tant que matrice Toeplitz par blocs, non symétrique. Dans le cas scalaire les deux prédicteurs sont identiques. Les résultats classiques sur la stabilité des modèles trouvent aussi leur équivalent dans le modèle vectoriel (6).

La représentation d'une phrase au moyen de la prédiction linéaire se fait de la façon suivante :

- découpage des échantillons (ceux-ci étant acquis à la cadence de 10 kHz) en N segments de 256 échantillons, se succédant avec recouvrement tous les 100 échantillons.

- calcul de la corrélation de chacun de ces N segments par l'estimateur ergodique, après application d'une fenêtre (de Hamming) sur le segment.

- calcul du modèle autorégressif scalaire sur chaque segment à un ordre p valant suivant les cas 10, 14 ou 16.

- calcul des coefficients cepstraux sur chaque segment. Ceux-ci se déduisent (7) des coefficients prédicteurs a_i par les relations :

$$\begin{cases} c_1 = a_1 & \\ j\, c_j = j\, a_j - \sum_{i=1}^{j-1} i\, c_i\, a_{j-i} & j = 2, \ldots, p \\ j\, c_j = -\sum_{i=1}^{p} (j-i)\, c_{j-i}\, a_i & j > p \end{cases}$$

Le résultat de ces traitements est un tableau rectangulaire de coefficients cepstraux (ou de prédiction, ou encore de corrélation partielle) sur lequel vont pouvoir être effectuées les analyses ultérieures.

4 - Utilisation en reconnaissance du locuteur

Une fois en possession du tableau rectangulaire des coefficients ainsi calculés, il est possible d'en extraire une représentation simplifiée de ce qui dans ce tableau est du à l'identité du locuteur. Interprétant à cet effet le tableau comme les composantes de N vecteurs de l'espace vectoriel R^p, il est alors possible de le considérer comme une réalisation d'un processus stochastique vectoriel. Ce dernier est représentable par ses différents moments, en particulier dans l'hypothèse où il serait stationnaire du second ordre : sa moyenne et sa matrice de covariance. Il peut paraître abusif quant à la terminologie, de considérer comme un processus aléatoire des mesures qui sont dérivées de la corrélation d'un premier processus, le signal de parole, la corrélation étant de nature déterministe, mais cela se justifie si on se rappelle que ces mesures sont en réalité des estimations de paramètres, et donc des fonctionnelles déterministes de variables aléatoires, par conséquent elles-même aléatoires.

Une telle représentation a été utilisée (1) dans un but de reconnaissance du locuteur. Dans une première expérience, la moyenne était conservée telle quelle, et concatenée au vecteur colinéaire à l'axe factoriel principal (vecteur propre associé à la plus grande valeur propre) de la matrice de covariance. Le vecteur ainsi obtenu était ensuite utilisé d'une façon que l'on peut qualifier de classique en reconnaissance des formes : comparaison aux références (préalablement acquises) de chaque locuteur potentiel, et reconnaissance de l'occurence proposée comme ayant été prononcée par le locuteur dont la référence est la plus proche au sens d'une métrique de Mahalanobis.

Une expérience réalisée sur une population de 11 locuteurs ayant prononcé chacun 10 fois la même phrase avait donné moins de

2 erreurs sur les 110 occurences, montrant qu'une telle paramétrisation pouvait raisonnablement fournir la base d'un système d'identification du locuteur.

Des expériences ultérieures ont montré que le taux d'erreur augmentait très peu quand les locuteurs prononçaient des phrases à chaque fois différentes. Un résultat analogue a été établi par A. Collins en réalisant la même expérience sur des locuteurs de langue anglaise prononçant 14 phrases différentes (8). D'autres expériences (1) réalisées à partir des matrices de covariance, utilisant une distance entre matrices, en terme de rapport de vraisemblances ne montrent pas de différences significatives avec les précédentes.

En conservant l'interprétation des mesures comme étant une réalisation d'un processus stochastique vectoriel, il est cependant possible d'aller plus loin. On peut faire sur ce processus une hypothèse plus forte que celle de stationnarité du second ordre. On supposera par exemple que le processus est autorégressif ; ceci impose au processus d'être centré, dans la pratique on se contentera de le centrer par soustraction de sa moyenne.

Soient donc $y_1\ y_2 \ldots y_N$ les colonnes du tableau des mesures (après centrage), on cherchera à les représenter par le modèle :

$$y_t + A_1\, y_{t-1} + \ldots + A_p\, y_{t-p} = \varepsilon_t,$$

les coefficients matriciels A_i seront chargés de représenter les caractéristiques vocales du locuteur. Le système de reconnaissance du locuteur qui pourrait utiliser cette représentation a été simulé lors d'une expérience de classification réalisée de la façon suivante :

- pour chaque locuteur a été calculé un modèle $[A_1, A_2, \ldots, A_p]$ sur un ensemble de 14 phrases toutes différentes.

- chacune des phrases a été filtrée par les modèles inverses des locuteurs et classée en fonction de la trace de la matrice "résidu" obtenue.

Sur les 70 essais réalisés aucune erreur n'a été constatée.

Le filtrage inverse rappellons-le consiste, dans l'emploi qui en est fait ici (et qui correspond à l'extrapolation au cas vectoriel du vocodeur à canaux adaptés (9)), à filtrer un processus y_t par plusieurs modèles $[A_1^{(k)} \ldots A_p^{(k)}]$, ce qui donne les résidus $\varepsilon_t^{(k)}$:

$$\varepsilon_t^{(k)} = y_t + A_1^{(k)}\, y_{t-1} + \ldots + A_p^{(k)}\, y_{t-p}$$

On forme ensuite les matrices E_k

$$E_k = E\,(\varepsilon_t^{(k)}\ (\varepsilon_t^{(k)})^T)$$

Le modèle correspondant à la trace minimale $tr(E_k)$ est celui du processus le plus proche de y_t parmi les processus $y_t^{(1)} \ldots y_t^{(n)}$, ayant servi au calcul des modèles (1) à (n). L'interprétation spectrale de cette proximité est moins simple que dans le cas scalaire. Dans le cas particulier traité ici, on peut considérer que le modèle calculé sur la voix d'un locuteur, modèlise les capacités articulatoires du locuteur, du moins en première approximation.

5 - Utilisation en adaptation au locuteur

La question ici posée est la suivante : peut-on utiliser l'information relative au locuteur contenue dans une phrase pour adapter à ce locuteur un étage de reconnaissance phonémique dont l'apprentissage a été réalisé sur un autre locuteur (2). Supposons disponible l'enregistrement de la même phrase prononcée par le locuteur standard (d'apprentissage), si il est possible de trouver une application optimale de la phrase standard sur la phrase nouvelle, la même application agissant sur les références standard devrait donner des références adaptées au nouveau locuteur.
Lors de la mise en oeuvre de ce dispositif (qui est de portée plus générale) au cas de la reconnaissance par prédiction linéaire, il est souhaitable de travailler sur les coefficients cepstraux ; une distance euclidienne sur ces coefficients ayant en effet une signification dans le domaine spectral, les transformations linéaires utilisées seront plus justifiées que sur d'autres coefficients.

L'application cherchée sera déterminée comme la composition d'une projection de l'espace standard dans un espace intermédiaire, avec une projection inverse de l'espace intermédiaire vers l'espace nouveau. Ces projections sont calculées de façon telle que dans l'espace intermédiaire, les deux processus (standard et nouveau) mesurés sur les phrases coïncident au mieux (analyse canonique des corrélations) (10). Soient y_t^o et y_t^1 les processus aléatoires (au sens décrit précedemment) obtenus pour le locuteur standard et le nouveau respectivement. On cherche deux matrices carrées (p x p) U et V telles que :

$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \Sigma_{00} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{11} \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix} = \begin{pmatrix} I & \Lambda \\ \Lambda^T & I \end{pmatrix}$$

où les matrices Σ représentent la covariance de y_t^o, y_t^1 et leur intercovariance.

$$\begin{pmatrix} \Sigma_{00} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{11} \end{pmatrix} = E\left[\begin{pmatrix} y_t^o \\ y_t^1 \end{pmatrix} \begin{pmatrix} y_t^o \\ y_t^1 \end{pmatrix}^T \right]$$

I est la matrice de covariance des projections de y_t^o et y_t^1, égale à la matrice identité.

Λ qui est la matrice d'intercovariance des processus projetés, est une matrice diagonale, dont les éléments diagonaux sont positifs, rangés en ordre décroissant. Plus la matrice Λ sera proche de la matrice unité, meilleure sera la superposition des projections. Soient c_o et c_1 les moyennes des mesures standards et nouvelles, chaque référence x_k^o standard deviendra x_k^1 définie par :

$$x_k^1 = V^{-1} U (x_k^o - c_o) + c_1$$

En tant que vérification de la validité de ce schéma a été réalisée l'expérience (biaisée) consistant à utiliser comme mesures y^o et y^1 les moyennes des phonèmes calculées sur de nombreux échantillons, c'est-à-dire les références correctes pour la reconnaissance phonémique. Le rendement de l'adaptation est évalué au moyen du rapport :

$$\rho = \frac{\text{score après adaptation - score sans adaptation}}{\text{score avec références apprises - score sans adaptation}}$$

Ce coefficient a été mesuré à la valeur 0,84 (moyenne pour 6 locuteurs).

Une difficulté subsiste encore, avant de pouvoir utiliser les mesures réalisées sur une phrase, à savoir la différence du rythme d'élocution qui oblige à recourir à une anamorphose temporelle pour faire coïncider à chaque instant t les deux phrases afin que les mesures y_t^o et y_t^1 concernent le même phonème. Cette anamorphose est calculée par une méthode de programmation dynamique sur l'énergie du signal.

6 - Conclusion

Plusieurs procédures ont été décrites ici permettant de prendre en compte ce qui après une analyse du signal par prédiction linéaire, représente dans les coefficients calculés l'influence de la voix du locuteur. La moyenne, la covariance des coefficients calculés est significative de l'identité du locuteur. Un modèle autorégressif vectoriel estimé sur ces coefficients en porte également la marque. Une procédure d'anamorphose puis de projection, appliquée sur la suite des modèles linéaires calculés permet d'adapter un étage de reconnaissance phonémique, d'un locuteur à un autre.

REFERENCES

GRENIER (Y.), 1977, Identification du locuteur par prédiction linéaire 1° Congrès AFCET-IRIA de Reconnaissance des Formes et Intelligence Artificielle, Chatenay-Malabry.

GRENIER (Y.), MAURIN (J.C.), 1979, Adaptation au locuteur par analyse canonique des corrélations. 2° Congrès AFCET-IRIA de Reconnaissance des Formes et Intelligence Artificielle, Toulouse.

WHITTLE (P.), 1963, On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix. -Biometrika, Vol. 50, pp. 129-134.

WIGGINS (R.A.), ROBINSON (E.A.), 1965, Recursive solution to the multichannel filtering problem. -J. Geophys. Res., Vol. 70, pp. 1885-1891, (Avril).

KAILATH (T.), 1974, A view of three decades of linear filtering theory, IEEE Trans. on IT, Vol. 20, n° 2, pp. 146-181, (Mars)

GAMBOTTO (J.P.), 1979, Méthodes d'estimation linéaires multidimensionnelles. Application à la reconnaissance et à la segmentation des textures. -Thèse de Docteur Ingénieur, / ENST.

GRAY (A.H.), MARKEL (J.D.), 1976, Distance measures for speech processing. -IEEE Trans. on ASSP, Vol. 24, n° 5, (October).

COLLINS (A.), 1979, Communication personnelle, Janvier.

GUEGUEN (C.), CARAYANNIS (G.), FARJAUDON (T.), LE CHEVALIER (F.), 1975, Un vocodeur à canaux adaptés. -L'Onde Electrique, Vol. 55, n° 7.

HOTTELING (H.), 1936, Relations between two sets of variables Biometrika, n° 28.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

ENTRAINEMENT LEXICAL SEMI-AUTOMATIQUE D'UN SYSTEME DE RECONNAISSANCE A BASE SYLLABIQUE

M. Lennig et P. Mermelstein
Recherches Bell-Northern, 3 Place du Comemrce
Ile des Soeurs, Québec, Canada H3E 1H6

Un des problèmes les plus difficiles dans les systèmes de reconnaissance de la parole à base syllabique est la variabilité dans la segmentation en syllabes. La variabilité intrinsèque du composant de segmentation syllabique fait que parfois, dans un mot, il manque des frontières syllabiques et, parfois, il y en a trop. Cette variabilité de segmentation n'est pas limitée à l'intérieur du mot mais peut aussi occasionner des interactions phonétiques à travers les frontières de mot.

La présente communication propose une solution des problèmes de variation de segmentation. La solution proposée consiste à incorporer l'information pertinente à cette variation, dans le lexique. Ces informations peuvent être extraites du même ensemble de données et selon le même procédé d'apprentissage que l'on emploie pour la génération des gabarits acoustiques de référence.

La méthode de spécification syntaxique que nous utilisons est celle de la grammaire sous forme de réseau de transition augmenté [1]. Nous utilisons des conditions et des actions définies sur les arcs pour ajouter la sensibilité de contexte permettant de tenir compte de la migration consonantique progressive et régressive à travers les frontières de mot. Nous tenons compte également des variations à l'intérieur du mot en incorporant diverses possibilités d'achemi-nement dans chaque sous-réseau lexical.

Nous avons entraîné le système sur un ensemble de 100 phrases prononcées par une locutrice montréalaise. La même locutrice a ensuite enregistré un nouvel ensemble de 100 phrases aléatoires. De ces 100 nouvelles phrases, le système en a correctement reconnu 76.

[1] W.A. Woods, 1970, Transition Network Grammar for Natural Language Analysis, Communications of the Association for Computing Machinery, Vol. 13, pp. 591-606.

ENGLISH SUMMARY OF "ENTRAINEMENT LEXICAL SEMI-AUTOMATIQUE D'UN SYSTEME DE RECONNAISSANCE A BASE SYLLABIQUE"

M. Lennig and P. Mermelstein
Bell-Northern Research, 3 Place du Commerce
Nuns' Island, Quebec, Canada H3E 1H6

It is desirable for a continuous speech recognition system to be able to adapt easily to a new speaker. The Harpy system is advanced in this regard: it can be trained to recognize a new speaker by having the speaker read just 20 sentences. However, the tuning performed by Harpy to accommodate a new speaker is limited to modifications of its acoustic templates. Harpy is not easily adaptable to new dialects which differ phonologically from the ones on which it was originally trained.

In this paper, we present a method of system training which, although less automatic than that of Harpy, is more general. Not only does it allow modification of the acoustic templates but also of the lexicon. The ability to modify the lexicon during system training means that the phonemic representation of words can be adapted to the individual speaker.

The recognition system we have implemented is syllable based. After acoustic preprocessing, the input speech is segmented into syllable-sized units before any recognition is attempted. Acoustic templates composed of multiple training tokens are matched against the unknown syllable using a dynamic programming algorithm.

One of the most serious problems in syllable-based speech recognition is variability in segmentation performance. Due to intra-speaker variability in speech production, the segmentation algorithm sometimes omits syllable boundary markers where they should normally occur or inserts them where they should not occur. Sometimes a syllable boundary marker is shifted from its normal position. Segmentation variability is not limited to the interior of a word but may also occur across word boundaries.

In addition to facilitating adaptation of the system to new dialects, the training method we propose offers a solution to the problem of intra-speaker segmentation variability. The solution consists of incorporating information pertinent to such variability into the lexicon during system training. In the proposed method, this information is gathered from the same data set and in the same training procedure used for the generation of acoustic reference templates.

The syntactic frameword employed is that of an Augmented Transition Network (Woods 1970). Conditions and actions on arcs are used to add the necessary context sensitivity to the grammar in order to account for forward and backward migration of consonants across word boundaries. Within-word variability is handled by including alternate paths through each lexical subnetwork.

The system was trained on 100 sentences spoken by a female speaker of Montreal French. A test set of 100 new randomly generated sentences were recorded by the same speaker. The system recognized 76 of these 100 test sentences.

Xlemes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

ENTRAINEMENT LEXICAL SEMI-AUTOMATIQUE D'UN SYSTEME DE RECONNAISSANCE A BASE SYLLABIQUE

LENNIG Matthew
MERMELSTEIN Paul

Recherches Bell-Northern, 3 Place du Commerce
Ile des Soeurs, Québec, Canada H3E 1H6

Il est souhaitable que les systèmes de reconnaissance de la parole s'adaptent facilement aux nouveaux locuteurs. Le système Harpy est bien avancé à cet égard: on peut l'adapter à un nouveau locuteur en lui faisant prononcer seulement vingt phrases (Klatt, 1977; Lowerre 1977). Cependant, l'adaptation que fait Harpy est limité au niveau des gabarits acoustiques. Harpy n'est pas capable de s'adapter facilement aux nouveaux dialectes qui diffèrent au niveau phonologique.

Dans cette communication nous présentons une méthode d'adaptation au locuteur qui, bien que moins automatique que celle de Harpy, est plus générale. Non seulement est-il possible de modifier les gabarits acoustiques par cette méthode, mais aussi de modifier le lexique en même temps. La possibilité de modifier le lexique pendant le procédé d'apprentissage implique qu'on peut adapter les représentations phonémiques des mots au locuteur.

Le système de reconnaissance de la parole continue que nous avons réalisé utilise la syllabe comme unité de segmentation et de reconnaissance. L'avantage d'un système à base syllabique est que ses gabarits acoustiques incorporent déjà une grande partie de la variation allophonique. Cependant, à cause de la variabilité intra-locuteur dans la production de la parole, la segmentation en syllabe n'est pas toujours faite de la même façon. La méthode d'apprentissage que nous proposons pour l'adaptation au locuteur offre aussi une solution au problème de la variabilité de segmentation.

DESCRIPTION GENERALE DU SYSTEME DE RECONNAISSANCE

Notre système de reconnaissance automatique de la parole continue consiste en quatre composants: un composant de prétraitement, qui extrait les paramètres acoustiques du signal de la parole utilisés pour la reconnaissance, un composant de syllabation, qui segmente la parole parametrisée en syllabes (Mermelstein, 1975), un reconnaisseur de phrase, qui dirige l'exploration de l'espace syntaxique afin de déterminer l'identité la plus probable de la phrase inconnue, et un comparateur syllabique, capable de calculer une mesure de distance (ou de dissimilarité) entre une syllabe inconnue et un gabarit de référence. Pour reconnaître une phrase inconnue, la phrase est d'abord prétraitée et segmentée en syllabes. Puis, le reconnaisseur de phrase dirige une exploration parallèle de tous les sentiers syntaxiques possibles en acceptant une syllabe à la fois de la phrase inconnue et en proposant plusieurs gabarits de référence pour lui être comparés.

En se basant sur les distances cumulatives de plusieurs sentiers parallèles d'analyse syntaxique, le reconnaisseur de phrase est capable d'éliminer certains sentiers d'analyse peu probables. Quand le reconnaisseur arrive à la fin de la phrase d'entrée, le sentier ayant la plus petite distance cumulative est choisie comme étant l'analyse la plus probable.

La méthode de spécification syntaxique que nous utilisons est celle de la grammaire sous forme de Réseau de Transition Augmenté (RTA) proposée par Woods (1970). Le RTA consiste en un système de réseaux de transition récursifs dont les arcs sont capables d'exécuter des actions et de tester des conditions arbitraires.

EXEMPLE: LES EXPRESSIONS DATE-HEURE

Le Graphique 1 représente le niveau le plus élevé d'une syntaxe sous forme de RTA qui accepte les expressions 'date-heure' en français. Ce réseau fait appel à deux sortes de sous-réseaux: les sous-réseaux lexicaux qui spécifient la structure des mots, et les sous-réseaux syntagmatiques, qui spécifient quelles suites de mots peuvent constituer des syntagmes. Par exemple, l'arc PUSH LE/ fait appel à un sous-réseau lexical qui spécifie les suites de syllabes possibles pour les différentes réalisations de surface du mot le. Quand l'arc PUSH LE/ est executé, la commande passe au réseau lexical LE/. Après que le réseau lexical accepte le mot le, la commande retourne au réseau du Graphique 1 dans l'état S/LE.

De l'état S/LE, le sentier d'analyse ou l'hypothèse peut prendre cinq arcs differents. Les arcs PUSH PREMIER/, PUSH VINGT/ et PUSH TRENTE/ font référence à des sous-réseaux lexicaux, tandis que les arcs PUSH TEEN/ et PUSH N29/ font référence à des sous-réseaux syntagmatiques. Le sous-réseau syntagmatique TEEN/ est représenté dans le Graphique 2. Il est capable d'accepter tous les nombres entre onze et dix-neuf en faisant appel à différents sous-réseaux lexicaux. Remarquez que l'arc qui porte l'étiquette JUMP permet qu'une hypothèse se déplace de l'état TEEN/DIX à l'état TEEN/END sans accepter de mot. (Cette trajectoire permet au sous-réseau TEEN/ d'accepter le mot dix.) L'arc POP fait retourner la commande au prochain niveau plus haut. Le Graphique 3 montre le sous-réseau syntagmatique N29/ qui accepte tous les nombres entre deux et neuf.

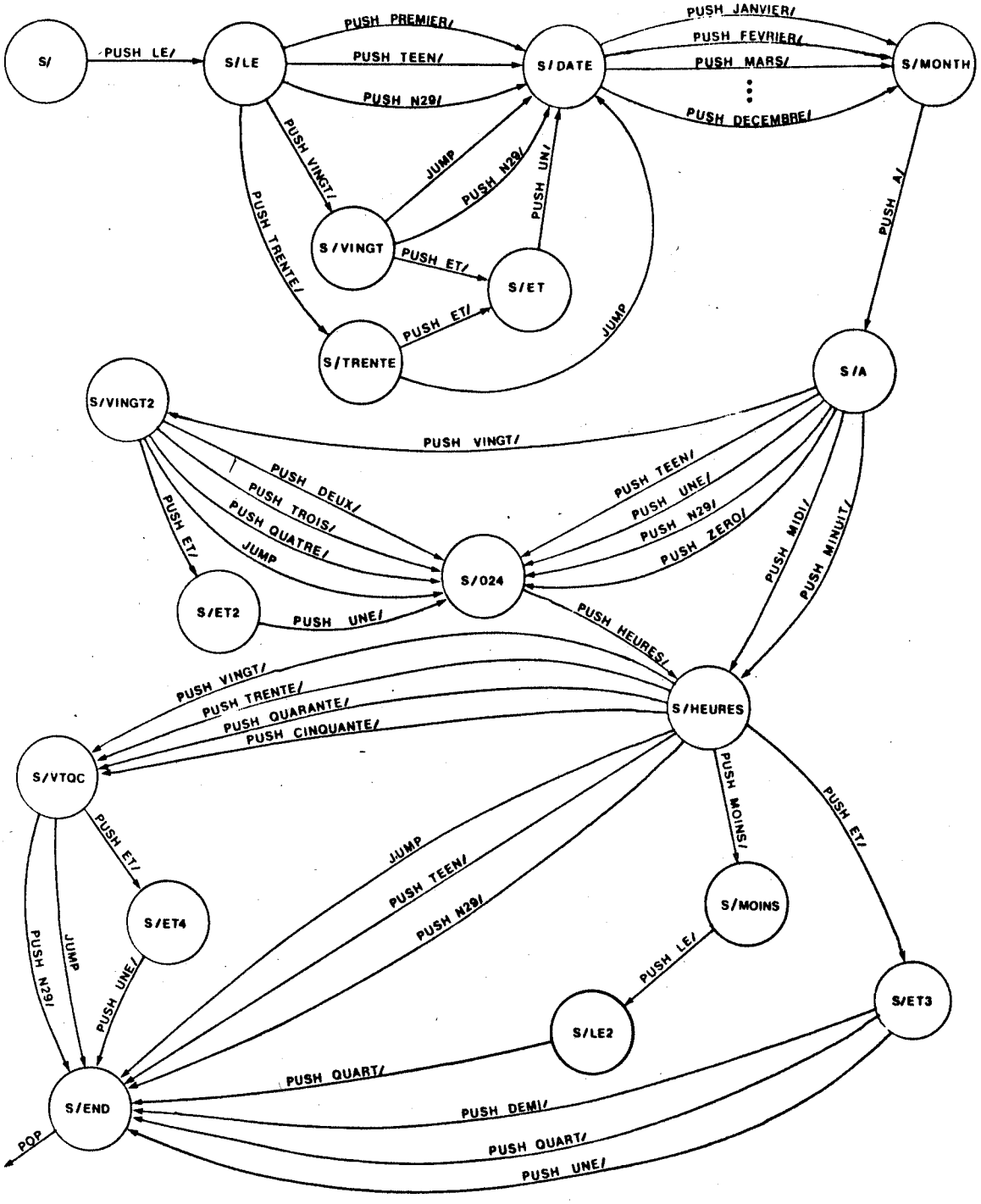
Pour voir comment fonctionnent les sous-réseaux lexicaux, prenons comme exemple celui qui accepte le mot premier, représenté dans le Graphique 4. On voit dans ce graphique que le mot premier peut avoir deux syllabations: il peut être segmenté en deux syllabes, [prə] suivi de [mje], ou bien il peut être segmenté en une seule syllabe: [prəmje]. Ainsi, la variabilité de syllabation peut être prise en main au niveau des sous-réseaux lexicaux.

Chaque fois qu'une hypothèse accepte une syllabe en traversant un arc ACCEPT, le reconnaisseur de phrases fait appel au comparateur syllabique pour calculer la distance acoustique entre la syllabe actuelle d'entrée et le gabarit correspondant à la transcription sur l'arc ACCEPT. Chaque hypothèse garde en mémoire la distance cumulative de toutes les syllabes qu'elle a acceptées.

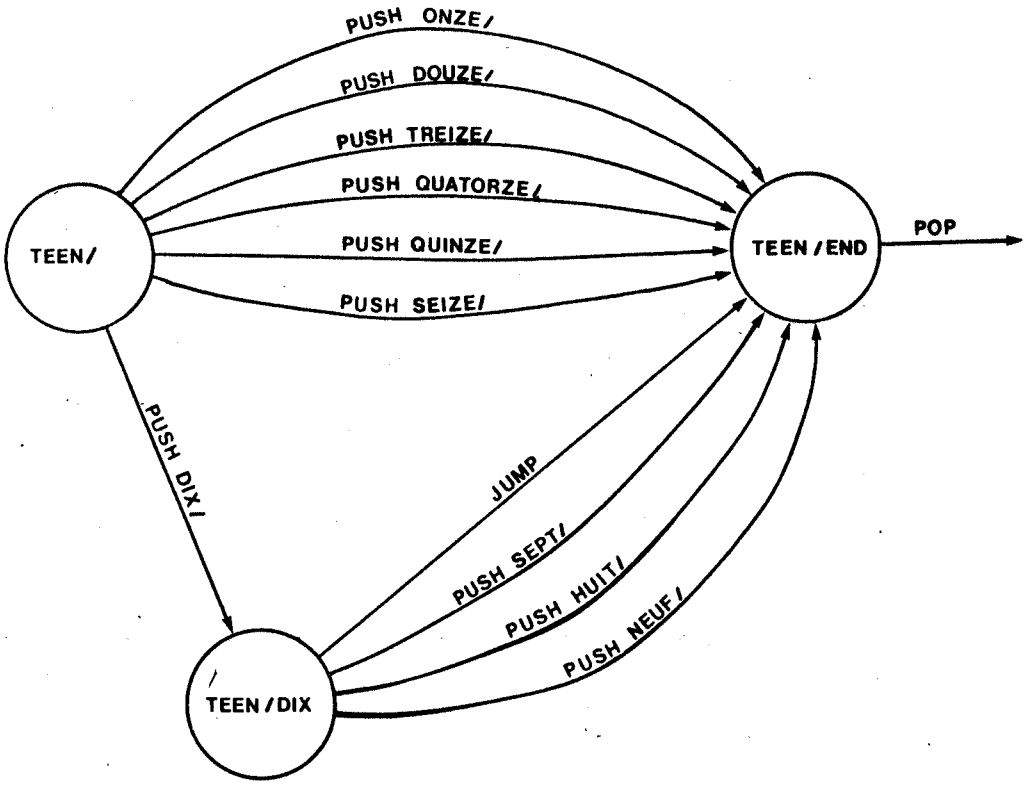
L'ADAPTATION DU LEXIQUE AU LOCUTEUR

L'apprentissage lexical consiste à créer un ensemble de réseaux lexicaux correspondant au vocabulaire du système et un ensemble de

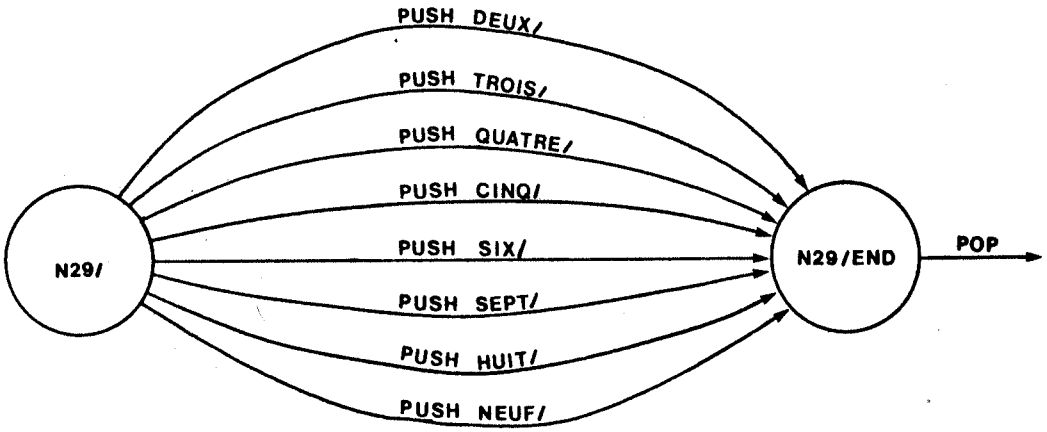
gabarits qui contient un gabarit pour chaque transcription phonétique figurant sur les arcs ACCEPT des sous-réseaux lexicaux. Dans la méthode d'apprentissage semi-automatique, nous générons ces deux ensembles d'informations à partir d'un ensemble de phrases d'apprentissage et à l'aide d'un transcripteur humain.



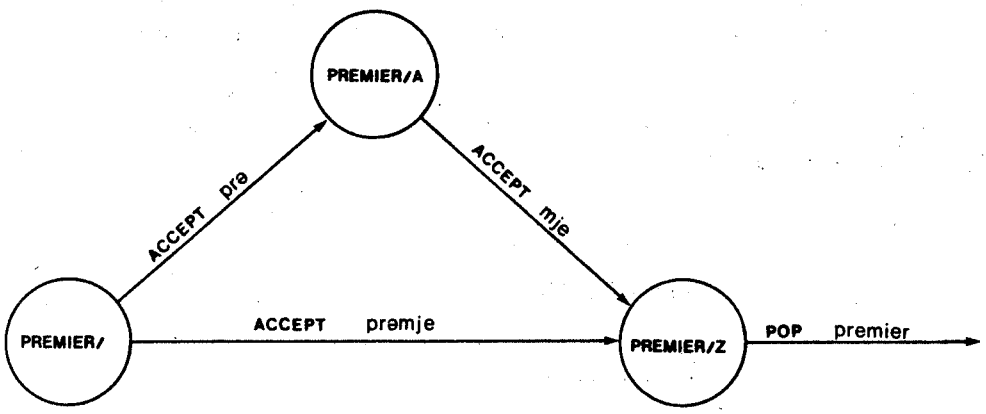
GRAPHIQUE 1. Réseau du niveau de la phrase, faisant partie de la spécification RTA des dates et heures françaises.



GRAPHIQUE 2. Sous-réseau syntagmatique TEEN/ qui accepte les nombres entre dix et dix-neuf.



GRAPHIQUE 3. Sous-réseau syntagmatique N29/ qui accepte les nombres entre deux et neuf.



GRAPHIQUE 4. Sous-réseau lexical qui accepte le mot premier, tenant compte des deux syllabations possibles.

Pour entraîner le système, l'ensemble des phrases d'apprentissage est tout d'abord segmenté en syllabes par le composant de syllabation. Un logiciel d'apprentissage reproduit à travers un haut-parleur le signal acoustique correspondant à chaque syllabe de chaque phrase, une syllabe à la fois. Après que l'ordinateur reproduit chaque syllabe, il pause et attend que le transcripteur tape une transcription phonétique de la syllabe au terminal.

Afin de pouvoir générer les sous-réseaux lexicaux directement, le transcripteur fait entrer non seulement la transcription phonétique de chaque syllabe des phrases d'apprentissage, mais aussi fait-il entrer une indication de la fin de chaque mot aussi bien que son orthographe standard. Nous prenons comme exemple la séquence d'apprentissage nécessaire pour créer le sous-réseau lexical pour le mot premier dans le Graphique 4. Supposons que les deux phrases suivantes figurent dans l'ensemble d'apprentissage:

(1) Le premier décembre à cinq heures dix.
(2) Le premier février à sept heures quatorze.

Supposons aussi que le mot premier dans (1) a été segmenté en une seule syllabe tandis que le mot premier dans (2) a été décomposé en deux syllabes. Ce qui suit est un exemple de la façon selon laquelle le transcripteur humain pourrait transcrire (1):

lə.LE
prəmje.PREMIER
de
sãbr.DECEMBRE
a.A
sẽk.CINQ
œ r.HEURES
dis.DIX

On voit dans cet exemple que le symbole point (.) est utilisé pour signifier la fin d'un mot et qu'il est suivi de l'orthographe normale du mot.

Supposons maintenant que (2) apparaît dans l'ensemble d'apprentissage, peut-être après plusieurs phrases intervenantes. La phrase (2) pourrait être transcrite comme suit:

lə.LE
prə
mje.PREMIER
fe
vri
je.FEVRIER
a.A
sɛt.SEPT
œ r.HEURES
ka
tɔrz.QUATORZE

Nous avons développé un compilateur lexical qui prend comme entrées des données de transcription comme celles-ci et produit comme sortie un lexique qui consiste en un réseau lexical correspondant à chaque mot unique figurant dans l'ensemble d'apprentissage. Le réseau lexical correspondant à chaque mot est capable d'accepter toutes les syllabations de ce mot qui ont apparu dans l'ensemble d'apprentissage.

Un logiciel associé crée un gabarit de référence correspondant à chaque transcription phonétique de syllabe qui peut apparaître sur un arc ACCEPT. Le gabarit consiste en une combinaison des paramètres acoustiques de toutes les syllabes de l'ensemble d'apprentissage qui porte cette transcription.

L'exemple du mot premier démontre une sorte de variation dans la décomposition syllabique: la division ou le manque de division variable à l'intérieur d'un mot. Plusieurs autres exemples de ce phénomène existent et peuvent être pris en compte par le même mécanisme. Maintenant voyons les sortes de variation dans la segmentation syllabique qui présentent le plus de difficultés: celles qui ont lieu à travers une frontière de mot.

MIGRATION PROGRESSIVE DES CONSONNES

Un phénomène fréquent dans la segmentation en syllabes et du français et de l'anglais est le transfert d'une consonne finale ou d'un groupe de consonnes finales au début du mot suivant. Ce phénomène, que nous désignons migration progressive, arrive surtout quand le mot suivant commence par une voyelle. Le Tableau 1 montre des exemples de la migration progressive.

Orthographe normale	Décomposition syllabique			
sept avril	sɛ	ta	vril	
sept avril	sɛt	ta	vril	
quatre heures	kat	trœr		
huit heures	ɥi	tœr		
seize heures	sɛz	zœr		
fifth October	fɪθ	θɔk	to	bə

TABLEAU 1. Cinq exemples français et un exemple anglais de la migration progressive.

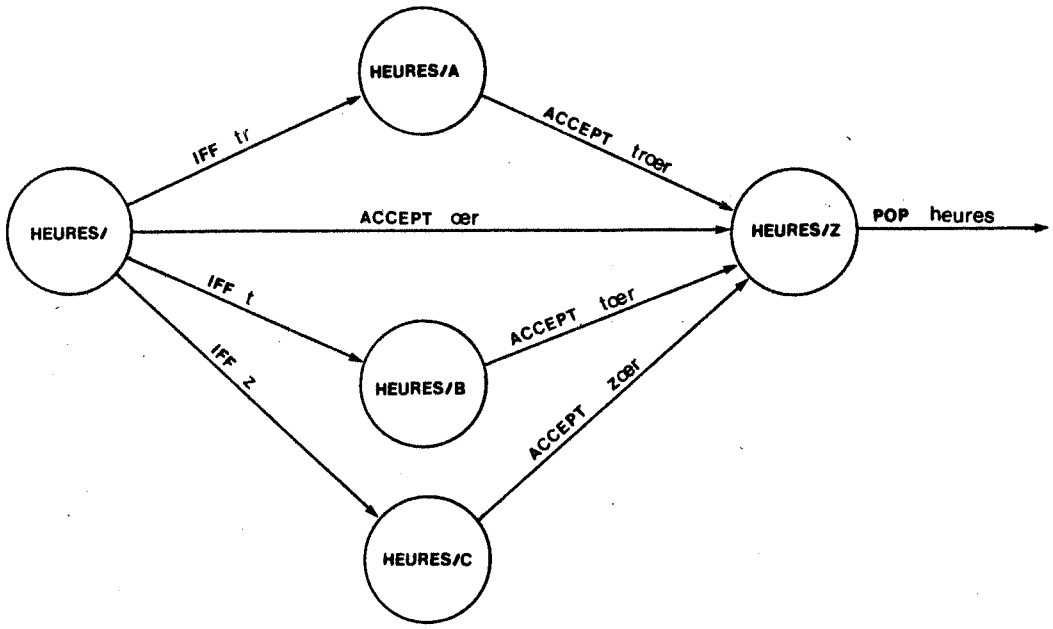
Afin de neutraliser ce genre de variabilité de segmentation, nous voudrions, par exemple, que le sous-réseau lexical pour le mot heures accepte la syllabe [trœr] mais seulement si le mot précédent accepté par l'hypothèse se termine en [tr]. Aussi, nous voudrions que le réseau accepte [tœr] comme une réalisation du mot heures, mais seulement au cas où le mot précédent se termine en [t]. De la même façon, le réseau doit accepter [zœr] si le mot précédent se termine en [z]. C'est-à-dire que pour tenir compte de la migration progressive nous sommes obligés d'ajouter de la sensibilité contextuelle à la grammaire. Les actions et les conditions sur les arcs peuvent servir à cette fin.

La façon la plus directe d'arriver à la sensibilité contextuelle nécessaire pour tenir compte de la migration progressive est de faire que chaque hypothèse garde en mémoire le groupe de consonnes finales du dernier mot qu'elle a accepté. Pour ce faire on définit une action qui emmagasine une valeur dans un registre de mémoire que nous appellerons FINAL. Chaque hypothèse aura un registre FINAL qui lui sera associé de la même façon qu'elle a associé avec elle un accumulateur pour garder sa distance cumulative. Nous définissons une action sur l'arc POP du niveau lexical qui met dans le registre FINAL une valeur qui correspond au groupe de consonnes finales du mot. Chaque fois que l'arc POP est executé pour faire retourner la commande au réseau de niveau plus haut, la valeur correspondante au groupe de consonnes finales du mot est emmagasinée dans le registre FINAL. Pour le mot quatre, par exemple, l'arc POP a comme argument le groupe final de consonnes [tr].

Pour compléter le mécanisme qui tient compte de la migration progressive, nous avons besoin de définir une condition qui teste le contenu du registre FINAL. Nous avons décidé de ce faire en définissant une nouvelle sorte d'arc que nous appelons IF FINAL (abrégé IFF) qui compare la valeur de son argument avec celle du registre FINAL de l'hypothèse qui le traverse. Si FINAL est égal à l'argument de l'arc IFF, l'hypothèse avance à sa destination, exactement comme s'il s'agissait d'un arc JUMP. Si, par contre, le contenu de FINAL n'est pas égal à l'argument de l'arc IFF, l'hypothèse est éliminée. Par exemple, le sous-réseau lexical pour le mot heures pourrait ressembler à celui du Graphique 5. Ce sous-réseau est capable d'accepter quatre formes alternatives du mot heures: [œr], [tœr], [zœr] ou [trœr]. La première de ces transcriptions est accessible à n'importe quelle hypothèse. Les autres exigent que le mot préalablement accepté se termine par un groupe précis de consonnes.

Afin que le compilateur lexical se serve de l'arc IFF et du nouvel argument de l'arc POP pour tenir compte de la migration progressive, nous étions obligés d'ajouter des capacités supplémentaires au procédé de transcription. Ainsi, nous utilisons le symbole virgule (,) pour séparer un groupe de consonnes qui a subi la migration progressive de la première syllabe du mot suivant. Par exemple, la séquence de transcription

.
.
.
kat.QUATRE
tr,œ r.HEURES
.
.
.



GRAPHIQUE 5. Sous-réseau lexical qui accepte quatre variantes du mot heures.

a l'effet de faire apparaître l'argument [tr] sur l'arc POP du mot quatre et de créer un sentier à travers le réseau heures qui accepte la syllabe [trœr] si le mot précédent s'est terminé en [tr] (comme dans le réseau du Graphique 5).

MIGRATION REGRESSIVE DES CONSONNES

Dans certains environnements phonologiques, les frontières syllabiques sont placées de façon qu'elles attachent le groupe initial de consonnes d'un mot à la fin du mot précédent. Cette situation arrive le plus souvent quand le mot précédent se termine par une voyelle. Le problème ressemble à celui décrit ci-haut, sauf que le groupe de consonnes émigre dans la direction inverse. Par exemple, le sept est parfois décomposé comme [ləs] [sɛt].

A cause de son analogie avec la migration progressive, nous avons choisi une notation de transcription similaire pour représenter la migration régressive. Le symbole trait d'union (-) sert à séparer le groupe de consonnes ayant subi la migration régressive du mot précédent. Ainsi, l'exemple le sept serait transcrit comme

.
.
.
lə-s.LE
sɛt.SEPT
.
.
.

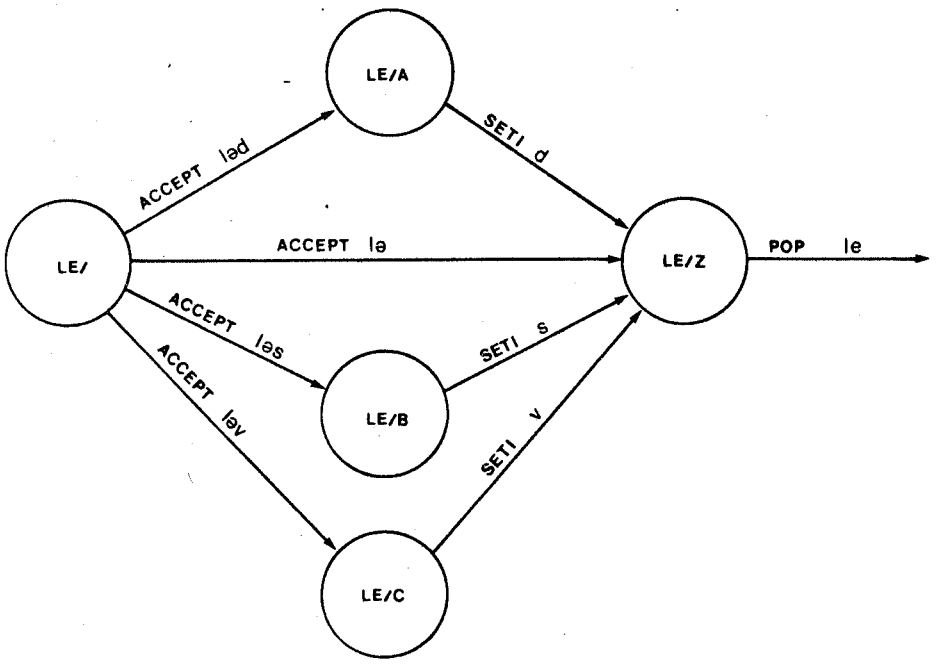
Le trait d'union indique que le [s] fait réellement partie du mot suivant mais qu'il a été attaché au mot le par le composant de

segmentation syllabique. Le compilateur interprète cette notation et se sert de cette information pour construire des sous-réseaux lexicaux pour le et deux qui reflètent la possibilité de cette segmentation. Dans cette partie, nous discutons la représentation de la sensibilité contextuelle nécessaire pour tenir compte de la migration régressive.

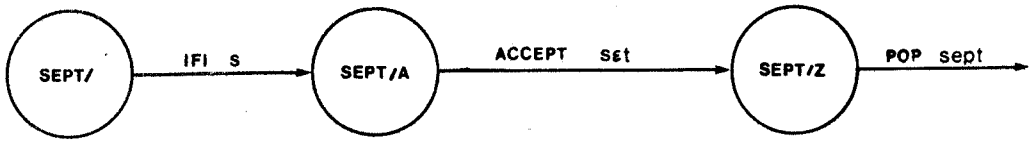
Nous voudrions accepter [ləs] comme une des formes de le, mais seulement si le mot suivant commence par [s]. Au premier abord, ceci paraît paradoxal: comment savoir à l'avance le sentier que suivra une hypothèse? La solution est de permettre que n'importe quelle hypothèse accepte [ləs] et puis de marquer cette hypothèse comme ayant besoin que son prochain mot accepté commence par [s]. Si l'hypothèse essaie d'accepter un mot suivant qui commence par une consonne autre que [s], l'hypothèse est éliminée immédiatement.

Pour marquer une hypothèse comme exigeant que le mot suivant commence par une consonne particulière, on emmagasine une valeur correspondante à la consonne initiale requise dans un registre de mémoire associé avec l'hypothèse. Le nom du registre servant à cette fin est INITIAL. On définit un arc spécial SET INITIAL (abrégé SETI) pour emmagasiner une valeur dans le registre INITIAL. Quand une hypothèse traverse un arc SETI, le reconnaisseur de phrase emmagasine l'argument de l'arc dans le registre SETI de l'hypothèse. Le Graphique 6 montre une configuration possible du réseau lexical du mot le qui utilise des arcs SETI pour forcer l'occurrence de certaines consonnes initiales dans le mot suivant. Si la forme non marquée [lə] est acceptée, le registre INITIAL est implicitement remis à zéro et n'importe quel mot peut suivre.

Afin de tester le contenu du registre INITIAL, nous définissons l'arc IF INITIAL (abrégé IFI). L'arc IFI est traité comme un arc JUMP si son argument est égal à la valeur du registre INITIAL de l'hypothèse qui essaie de la traverser. Sinon, l'arc bloque et l'hypothèse est éliminée. Le Graphique 7 illustre l'emploi de l'arc IFI dans le sous-réseau lexical pour le mot sept.



GRAPHIQUE 6. Sous-réseau lexical qui accepte différentes formes du mot le. Ce réseau illustre l'utilisation de l'arc SETI.



GRAPHIQUE 7. Sous-réseau lexical pour le mot sept illustrant l'emploi de l'arc IFI pour spécifier que la consonne initiale du mot est [s].

En employant des actions et des conditions sur les arcs, nous avons réussi à ajouter à nos descriptions syntaxiques la sensibilité contextuelle nécessaire pour tenir compte de la migration et progressive et régressive. Dans la prochaine partie nous verrons que, sans définir aucun nouveau mécanisme, nous pouvons tenir compte sur un niveau rudimentaire de la liaison française.

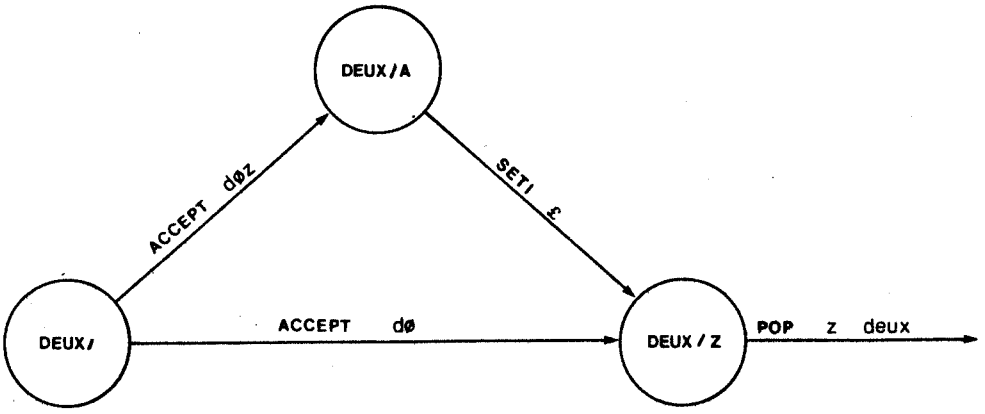
LA LIAISON FRANCAISE

Un exemple de la liaison se produit dans l'expression deux heures. Puisque [z] est la consonne finale sous-jacente du mot deux, il semble raisonnable d'essayer de la spécifier dans l'argument de la consonne finale de l'arc POP du sous-réseau lexical pour le mot deux. Le mécanisme de la spécification des consonnes finales sur les arcs POP existe déjà pour tenir compte de la migration progressive. Comme le mot deux se termine par une voyelle dans sa forme habituelle, l'argument de la consonne finale de l'arc POP reste libre pour emmagasiner la consonne finale sous-jacente. Un des sentiers traversant le sous-réseau pour heures consistera en un arc IFF z suivi d'un arc ACCEPT zœr. Ce sentier sera accessible à toute hypothèse qui a accepté comme mot précédent un mot se terminant en [z] ou bien ayant une consonne finale sous-jacente de [z].

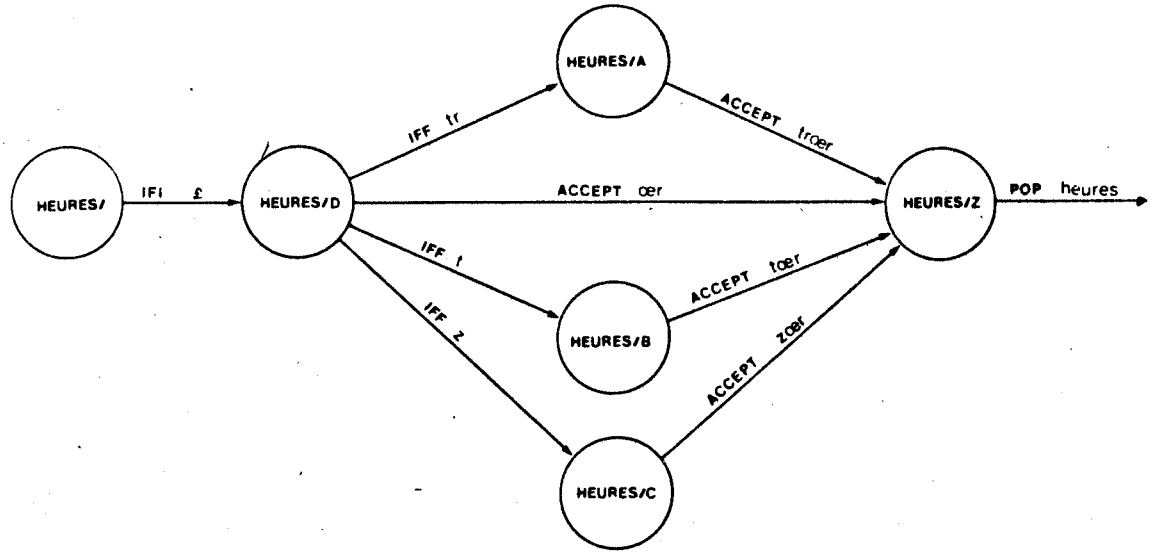
Pour tenir compte des cas où le [z] est prononcé à la fin du mot deux, il faut inclure un sentier qui accepte la syllabe [døz] dans le sous-réseau pour deux. Cependant, nous ne voulons pas accepter [døz] au cas où le mot suivant ne commence pas par un phonème qui déclenche la liaison. Par exemple, la séquence [døz] [me] pour deux mai est impossible.

Ce problème est analogue à celui de la migration régressive puisqu'on est obligé de prédire le type de segment par lequel commencera le mot suivant. Les mêmes méthodes peuvent s'appliquer. Dans le sous-réseau lexical pour deux on se sert d'un arc ACCEPT døz suivi d'un arc SETI £, où £ est un symbole spécial qui signifie 'segment déclenchant la liaison'. Le sous-réseau lexical pour heures (aussi bien que ceux pour tous les mots capables de déclencher la liaison) contient un arc IFI £ par lequel toute hypothèse entrant dans le réseau doit passer.

L'exemple deux heures est illustré dans le Graphique 8. Le Graphique 8A montre le sous-réseau lexical pour le mot deux, y compris l'arc SETI £ et la spécification de la consonne sous-jacente sur l'arc POP. Le Graphique 8B montre le sous-réseau lexical révisé pour heures avec l'addition de l'arc IFI £ pour indiquer que heures peut déclencher la liaison. Nous voyons qu'en utilisant des mécanismes déjà nécessaire



GRAPHIQUE 8A. Sous-réseau lexical pour le mot deux qui tient compte de sa forme de liaison.



GRAPHIQUE 8B. Sous-réseau lexical pour le mot heures, révisé pour tenir compte de la liaison.

pour tenir compte de la migration des consonnes nous pouvons également tenir compte d'une façon rudimentaire de la liaison.

Nous avons adopté une convention spéciale pour indiquer les occurrences de liaison dans la transcription d'apprentissage. On transcrit la consonne de liaison comme si elle n'appartenait à aucun des deux mots. Par exemple, si les mots deux heures étaient segmentés comme [dø z] [zœ r], alors la transcription d'apprentissage serait

.
.
.
dø-z.DEUX
z,œ r.HEURES
.
.
.

Le trait d'union indique que le z ne fait pas partie du mot deux en même temps que la virgule indique qu'il ne fait pas partie du mot heures. Cette contradiction de notation signale au compilateur lexical de construire les arcs nécessaires pour la liaison.

LA PERFORMANCE DU SYSTÈME

Nous avons entraîné le système d'abord sur un ensemble d'apprentissage de 100 phrases 'date-heure' prononcées par une locutrice du français montréalais. Quand les 100 phrases de l'ensemble d'entraînement ont été soumises au système, il en a correctement identifié 96.

Un nouvel ensemble de 100 phrases générées au hasard a été prononcé par la même locutrice et soumis au système. Cette fois le système a correctement identifié 76 des 100 nouvelles phrases.

Pour démontré la généralité du procédé d'apprentissage, nous l'avons utilisé pour adapter le système à reconnaître non pas un autre dialecte, mais une autre langue, cette fois choisissant un locuteur de l'anglais britannique. Pour que le système accepte les dates et les heures anglaises, il suffisait de spécifier manuellement les réseaux syntagmatiques: tous les nouveaux sous-réseaux lexicaux étaient créés lors de l'apprentissage semi-automatique. Le système a correctement identifié 50 sur 59 nouvelles phrases anglaises, soit 85%.

CONCLUSIONS

Nous avons réalisé un procédé semi-automatique d'adaptation au locuteur qui est puissant mais qui exige la participation d'un transcripteur humain. Cette méthode est suffisamment générale qu'elle peut ajouter du vocabulaire au lexique ou bien ajouter des représentations phonémiques au vocabulaire existant. Ainsi peut-elle servir de modifier le dialecte ou même le langage reconnu par le système. Cette méthode d'apprentissage incorpore dans le lexique une connaissance de la variabilité de segmentation.

Le problème le plus important de la méthode décrite ici est qu'elle exige une quantité suffisante de données d'apprentissage. Dans le système actuel, nous n'incorporons une variante dans un sous-réseau lexical que si elle apparaît dans l'ensemble d'apprentissage. Parce qu'on voudrait pouvoir entraîner un système avec un minimum de données d'apprentissage, il serait profitable de chercher des algorithmes qui pourront se généraliser à d'autres segmentations possibles à partir d'un ensemble limité d'apprentissage.

BIBLIOGRAPHIE

KLATT (D.H.), 1977, "Review of the ARPA speech understanding project". J. Acoust. Soc. Am. 62, pp. 1345-1364.

LOWERRE (B.T.), 1977, "Dynamic speaker adaptation in the Harpy speech recognition system". Conference Record of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, 9-11 May, pp. 788-790.

MERMELSTEIN (P.), 1975, "Acoustic segmentation of speech into syllabic units". J. Acoust. Soc. Am. 58, pp. 880-883.

WOODS (W.A.), 1970, "Transition network grammars for natural language analysis". Comm. Assoc. Computing Machinery 13, pp. 591-606.

THEME III :

c) Vérification et identification du locuteur

XIèmes JOURNÉES D'ÉTUDE SUR LA PAROLE
28, 29 et 30 mai 1980
STRASBOURG

VERIFICATION DE L'IDENTITE DE LOCUTEURS COOPERATIFS, A TRAVERS LE TELEPHONE, A L'AIDE D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE.

VIVES Roland CNET - LANNION

RESUME

Les systèmes de vérification de l'identité de locuteurs à partir de la voix qui sont actuellement les plus performants, sont ceux qui travaillent avec des phrases pré-établies prononcées par des locuteurs coopératifs. Nous nous sommes demandés s'il était possible dans ces conditions et en travaillant dans la bande téléphonique, d'obtenir des résultats similaires en utilisant des systèmes de reconnaissance de la parole. Avec de tels systèmes, on peut vérifier l'identité d'une personne en reconnaissant un mot de passe qu'elle a choisi et qu'elle prononce. 52 locuteurs ont été testés avec un système de reconnaissance de la parole fonctionnant en mode analytique. Les taux de mauvais rejet et de mauvaise vérification sont respectivement de 4,3 % et de 4 %.
Les premiers résultats obtenus avec un système de reconnaissance de mots isolés travaillant par comparaison directe au niveau acoustique sont du même ordre de grandeur.

VERIFICATION OF COOPERATIVE SPEAKERS, THROUGH THE TELEPHONE, WITH SPEECH RECOGNITION SYSTEMS.

SUMMARY

It seems to us interesting to evaluate the performance of a speech recognition system used as a speaker verification tool, assuming cooperative speakers uttering prescribed texts through a telephone. Each speaker first chooses a key sentence (composed of one to five words). The key sentence pronounced by the corresponding speaker is then recorded and two kinds of reference pattern are computed in order to compare two different types of speech recognition systems. The first pattern reference consists of a phonetic string calculated by the phonetic module of a general purpose speech recognition system KEAL. The second reference pattern is the output of a 14-channel vocoder.
In the verification process, the two kinds of reference patterns are used for comparison. In the first case the new utterance spoken by the claimed speaker is converted into a phonetic string by the phonetic module of KEAL. The resulting string is then compared with the stored reference pattern of the claimed speaker. A similarity score is calculated and the claimed identity is accepted if that score is sufficiently high.
52 speakers were tested (35 males and 17 females). On 138 tests, a customer reject rate of 4.3 percent is achieved with a 4 percent impostor accept rate. In the second experiment a dynamic programming algorithm is used to compare the input sentence to the reference pattern. Results were very similar.

I - INTRODUCTION

Est-ce une gageure de vouloir faire de la vérification de l'identité d'un correspondant à l'aide d'un système de reconnaissance automatique de la parole ?

Un tel système ne peut répondre en principe à la question "est-ce bien X qui a parlé ?". En général une machine ne travaille bien que pour la tâche pour laquelle elle a été conçue. Un système de reconnaissance de la parole est construit pour répondre à la question "qu'est-ce qu'il (elle) a dit ?". Certains systèmes travaillant relativement proche du niveau acoustique n'éliminent pas forcément les caractéristiques liées au locuteur (QUERRE et al., 1974) et peuvent laisser envisager une utilisation pour la vérification du locuteur, mais d'autres systèmes de reconnaissance, comme Keal (MERCIER et al., 1978) essaient de n'extraire du signal de parole que les éléments linguistiques communs à l'ensemble des personnes parlant la même langue : on comprend moins, pour ces derniers systèmes, comment ils pourraient servir en vérification du locuteur.

Nous allons montrer dans cet article l'intérêt que l'on peut trouver à utiliser ces deux classes de systèmes dans le domaine de la vérification de locuteurs.

En vérification du locuteur à partir de la voix, deux préoccupations essentielles ont guidé les chercheurs dans le choix des paramètres utilisés : premièrement extraire du signal de parole les caractéristiques propre au locuteur ce qui lui est intime, ce qui lui est difficile de modifier consciemment ; trouver, en second lieu des caractéristiques variant le moins possible dans le temps.

Cette recherche se heurte aux nombreuses causes de variation intra-locuteur : niveau et rythme d'élocution, état émotionnel, étant de santé, etc... D'un autre bord, les spécifications anatomiques ou la façon même de parler, l'accent, sont des caractéristiques qui, bien que plus difficiles à modifier consciemment, sont tout de même sujettes au vieillissement.

Les travaux dans ce domaine sont très nombreux dans le monde : ils ont utilisé l'intensité la fréquence du fondamental, le spectre à court terme, les coefficients de prédiction, la fréquence et la largeur des formants, la coarticulation nasale, la corrélation spectrale, la vitesse d'élocution et le rythme. (Voir en France : (GRENIER, 1977), (EL CHAFEI, 1979), (CORSI, 1979), (CORSI et BOE, 1979), (ABRY et BOE, 1979).

Les résultats sont encourageants :
ROSENBERG (1973) n'a t-il pas montré, dans une expérience effectuée sur quarante locuteurs, qu'une méthode de vérification automatique utilisant Fo et l'intensité, obtenait même de meilleurs résultats qu'un groupe d'auditeurs (98 % contre 96 %). Il faut cependant souligner que le corpus était constitué par une phrase pré-établie prononcée par des locuteurs coopératifs.

Un deuxième type de problème, lié au choix des paramètres, surgit immédiatement quand on envisage une application quelconque de la vérification du locuteur : il s'agit de l'adéquation des paramètres choisis aux conditions d'enregistrement et de transmission.

Pour le cas qui nous intéresse il faut se demander si les paramètres discriminatoires trouvés conservent leur efficacité à travers le réseau téléphonique.

Le troisième type de problème se place encore plus directement au niveau des applications : il s'agit de savoir si l'on envisage la vérification de locuteurs coopératifs ou la vérification d'identité de locuteurs qui chercheraient éventuellement à modifier leur voix pour ne pas être reconnu.
Il existe un bon nombre d'applications où le locuteur-utilisateur ne demande qu'à coopérer : qu'il s'agisse de transactions bancaires et commerciales par téléphone ou qu'il s'agisse de contrôle de l'accès à certaines informations confidentielles comme par exemple l'ouverture d'une case Phonex* (SOUBIGOU,1979). L'utilisateur désirant connaître le contenu de sa case Phonex, parlera de façon compréhensible quand le système lui demandera son mot de passe.

S'il y a à la limite adaptation volontaire au système, cet effort de ne pas bredouiller ne sera, en général, jamais perçu comme pénible.

La vérification d'identité d'un locuteur non coopératif est bien plus délicate par le fait même que le locuteur peut vouloir tromper la machine. Les champs d'application de tels systèmes appartiennent à la classe des outils supplémentaires pour l'appareil judiciaire et policier.

A cause de ces problèmes, les systèmes de vérification de locuteurs par la voix, les plus performants, sont actuellement ceux qui travaillent avec des phrases pré-établies prononcées par des locuteurs coopératifs.

Nous nous sommes demandés s'il était possible, dans ces conditions (phrases pré-établies, locuteurs coopératifs, bande téléphonique) d'obtenir des résultats équivalents en utilisant des systèmes de reconnaissance de la parole.

II - DESCRIPTION DE L'EXPERIENCE.

Faire de la vérification d'identité d'un locuteur avec un système de reconnaissance automatique dela parole nécessite un petit subterfuge.
Nous avons demandé à chaque locuteur de choisir une phrase clé, un "sésame ouvre-toi". Vérifier l'identité d'un locuteur revient donc avec cette hypothèse, à tester si la phrase prononcée par le locuteur est suffisamment ressemblante à celle qui est associée à son nom.

Deux séries de tests ont été effectuées en parallèle. Dans la première série 52 locuteurs (35 hommes et 17 femmes) ont été testés avec un système de reconnaissance de la parole travaillant en mode analytique (MERCIER et al., 1978) : segmentation du signal vocal en syllabes et phonèmes et analyse lexicale.

---

* Phonex est la poste restante du téléphone. Un abonné possédant une case Phonex peut consulter sa case de n'importe quel point du réseau téléphonique et entendre les messages qu'on aura pu lui laisser.

Le signal de parole était analysé par un vocodeur à 14 canaux travaillant dans la bande passant du téléphone (300 - 3 400 Hz).

Un détecteur de mélodie fournissait simultanément la fréquence fondamentale du signal. Nous nous sommes servis du programme d'analyse phonétique du système Keal pour passer du spectre acoustique donné par le vocodeur à sa représentation phonétique. Seul le phonème le plus probable est conservé pour étiqueter les segments phonétiques trouvés. On obtient ainsi pour le codage de référence d'un mot de passe, une forme phonétique qui peut être assez différente de la forme phonétique standard. La phrase clé : "Tes laitues naissent-elles ?" a par exemple été décodée en : /t y t e t l e l e s t a mə b/
alors que le codage phonétique standard de cette phrase est :
/t e l e t y nɛ s t ɛ l (ə) /.
La phase de test du corpus s'est déroulée entre 3 jours et 6 semaines après la phase d'enregistrement.
Chaque locuteur a été invité au moins deux fois à répéter son mot de passe à quelques jours d'intervalle.

Contrairement à la phase de confection des éléments de référence, on conserve un étiquetage multiple pour les segments phonétiques trouvés par l'analyseur phonétique. Le module lexical compare ensuite ce spectre phonétique au codage phonétique de référence de chaque mot de passe et donne une évaluation des cadrages réalisés sous forme d'un indice de ressemblance (VIVES, 1979).

Deux mesures permettent d'évaluer classiquement les performances d'un système de vérification de locuteurs : un taux de mauvais rejet représentant le nombre de cas pour lesquels la phrase prononcée par une personne n'a pas été jugée suffisamment ressemblante à la phrase qu'elle avait dite en référence et un taux de mauvaise vérification représentant le nombre de fois ou une personne peut se faire passer pour quelqu'un d'autre.

Sur 138 tests effectués on a obtenu un taux de mauvais rejet de 4,3 pour cent (6 mauvais rejets sur 138) et un taux de mauvaise vérification de 4 pour cent en faisant bien sûr l'hypothèse que l'imposteur n'a aucune connaissance préalable du mot de passe.

Chaque fois qu'un locuteur a prononcé son mot de passe, nous l'avons comparé à tous les mots de passe de référence. Il y a eu en moyenne un autre mot de passe qui était trouvé suffisamment ressemblant pour être accepté à la vérification. Nous avons évalué le taux de mauvaise vérification par le rapport théorique du nombre de cas défavorables (C^1_{51}) sur le nombre de cas possibles (C^2_{52}).

Actuellement, 24 parmi les 52 locuteurs précédents (17 hommes et 7 femmes) ont participé à la seconde série de tests qui mettait en oeuvre un système de reconnaissance de mots isolés travaillant par comparaison directe du spectre acoustique de la phrase clé telle qu'elle a été enregistrée au cours de la phase d'apprentissage, avec le spectre acoustique d'une prononciation ultérieure (QUERRE et al, 1974.). Les taux de mauvais rejet et de mauvaise vérification semblent être du même ordre de grandeur que ceux obtenus par la méthode analytique.

Discussion et conclusion.

Nous nous sommes posés la question de savoir si dans le contexte phrases préétablies, locuteurs coopératifs et bande téléphonique on pouvait obtenir, avec un système de reconnaissance de la parole, des résultats comparables à ceux annoncés par les équipes faisant de la vérification de locuteurs en extrayant du signal de parole des caractéristiques discriminatoires inter-locuteur. Dans la revue qu'il a faite sur les travaux en vérification du locuteur, ROSENBERG (1976) mentionne un système opérationnel utilisé à la compagnie Texas Instrument pour contrôler l'accès à un système informatique.

Les taux de mauvais rejet et de mauvaise vérification sont de 4 pour cent avec la prononciation d'un seul mot clé monosyllabique. Les performances s'améliorent si l'on demande à l'utilisateur de prononcer d'autres mots clés en cas de doute. Avec une stratégie de décision séquentielle nécessitant en moyenne la prononciation de 1,3 mots clés, les taux de mauvais rejet et de mauvaise vérification passent respectivement à 0,3 pour cent et 1 pour cent. Les résultats que nous avons obtenus sont bien comparables et peuvent aussi être améliorés par l'emploi d'une stratégie de décision séquentielle.

On peut nous reprocher de contourner le problème de la vérification d'identité d'un locuteur : dans notre expérience ce n'est effectivement pas par sa voix qu'une personne est identifiée, mais par ce qu'elle dit. Le choix d'une phrase clé ou d'un mot de passe est souvent compris comme quelque chose de très personnel, comme un signature. Il est sûr que si le fait de pouvoir partager un mot de passe avec les personnes que l'on désire peut présenter certains avantages, (délégation de pouvoir) on peut trouver dangereux d'être à la merci d'un imposteur qui aurait eu connaissance du mot de passe. Des tests effectués aux Laboratoires Bell sur la résistance au mime d'un système de vérification de locuteur par la voix à travers le téléphone montrent que le taux de mauvaise vérificiation peut passer de 1,5 pour cent à vingt sept pour cent dans le cas de bons mimes (ROSENBERG, 1976).

Dans l'expérience que nous avons faite le taux de mauvaise vérification passe de 4 pour cent à soixante pour cent dans le cas où un imposteur connaît le mot de passe d'une personne. Les résultats sont évidemment moins bons et ne feront que se dégrader au fur et à mesure que nos systèmes de reconnaissance de la parole s'amélioreront. Est-ce là un avantage sans appel de la vérification de locuteur à partir de la voix ? Il est quasi certain que l'on améliorera la résistance au mime des futurs systèmes mais en parallèle on pourra développer des systèmes de mime sophistiqués : à partir d'un corpus important de parole, prononcé par une personne, ne sera-t-il pas possible d'engendrer des phrases perceptivement identiques à celles que pourrait prononcer cette personne, en prenant des syllabes ou des mots entiers et en manipulant les paramètres prosodiques (durée, Fo), pour les replacer dans le nouveau contexte ? . Le monde dans lequel nous vivons ne manque pas d'offir à qui le voudra la possibilité de confectionner de larges corpus de parole d'autrui : enregistrement direct par magnétophone, enregistrement de conversations téléphoniques, de discours radio-télévisés etc...

Il reste un "avantage" que nous reconnaissons aux études de vérification de locuteurs par la voix : c'est de faire avancer les connaissances en identification et en vérification de locuteur non coopératifs.
Si certaines applications dans ces domaines peuvent être jugées "utiles" à notre société, il est facile d'imaginer ce que pourraient devenir de tels outils entre les mains de personnes, d'organismes ou d'états, peu enclins à respecter les libertés individuelles.

La dynamite n'a pas uniquement été utilisée pour creuser des tunnels ou extraire des matériaux de construction dans des carrières.
Les résultats d'une étude peuvent toujours être détournés à des fins dont les conséquences sont parfois loin d'être en rapport avec les avantages qui les ont suscités.

Au moment où la télématique s'installe dans notre société cet article propose une solution au problème de la vérification d'identité du locuteur. La lecture des mots de passe choisis par les personnes ayant effectué le test, montre que cette solution n'est pas dénuée de poésie (voir annexe). Nous avons aussi montré, que pour des performances équivalentes, la vérification d'un locuteur à partir de ce qu'il dit était plus conviviale qu'à partir de sa voix (ILLICH,1973).

Note : Je remercie, Christian GAGNOULET, pour les tests de vérification du locuteur, qu'il a effectué avec un système de reconnaissance de mots isolés, travaillant par comparaison directe des spectres au niveau acoustique.

III - BIBLIOGRAPHIE.

ABRY (C.), BOE (L.J), 1979, Pour une idiolectologie : Aspects Phonétiques de l'identité. Coll. Inter. Production et Affirmation de l'Identité - Toulouse.

CORSI (P.), 1979, Reconnaissance automatique du locuteur : présentation générale, méthodologies et expérimentation, perspectives d'application. Thèse docteur-ingénieur. Institut National Polytechnique de Grenoble.

CORSI (P.), BOE (L.J), 1979, Définition et sélection de caractéristiques temporelles en vue de la vérification automatique du locuteur. 2ème congrès AFCET-IRIA Reconnaissance des formes et intelligence artificielle. Toulouse Tome 2, pp. 324-333.

EL CHAFEI (C), 1979, Un système de reconnaissance automatique de locuteurs sur Mini-Ordinateur 2ème congrès AFCET-IRIA - Reconnaissance des formes et intelligence artificielle - Toulouse Tome 2, pp.305-312.

GRENIER (Y.), 1977, Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique. Thèse de docteur - ingénieur ENST - Paris.

ILLICH (I.), 1973, La Convivialité (édition du seuil).

MERCIER (G.), QUINTON (P.), VIVES (R.), 1978, KEAL : un système pour un dialogue avec une machine. Actes du congrès de l'AFCET, 13-15 novembre, GIF-SUR-YVETTE, pp. 304-314.

QUERRE (M.), MERCIER (G.), GRESSER (J.Y), 1974, Reconnaissance Automatique de la parole : Application de la programmation dynamique à l'identification de mots isolés ; résultats comparés sur 1 000 mots. Note technique CEI/CSI/44 CNET - Lannion.

ROSENBERG (A.E), 1973, Listener Performance in Speaker verification task. IEEE Trans. Audio Electroacoustic, Vol AU-21, pp.221-225.

ROSENBERG (A.E.), 1976, Automatic Speaker Verification : A Review- Proceeding of the IEEE Vol. 64, n°4, avril pp. 475-487.

SOUBIGOU (A.), 1979, Note technique sur le logiciel de Phonex NT/DAS/SST/13.

VIVES (R.), 1979, Utilisation de l'Information phonémique et syllabique pour la reconnaissance de mots prononcés isolément ou dans des phrases. 10èmes J.E.P. Grenoble 30 mai, 1er juin, pp. 375-384.

ANNEXE : Liste des phrases clé choisies par les personnes ayant participé aux tests.

Feutre bleu marine
A la quinte
Les histogrammes
Ecrans claviers
Je suis enrhumé
Il a fait beau samedi
Ici Charlot
Poul Palud
Au pied levé

Vite au secours
Son pull est gris
Juillet oh sainte chose
Ramadan
La lampe s'allume
Qui se plaint du sommeil
Le petit chat est mort
Il ne fait pas beau
Poète prends ton luth

Salut les copains
Félix le chat
Trébeurden
Quatre heure quatre
Les espaces infinis
129-D
Godillots godasses
Allez-y
Beau temps belle brise

Elle bat le beurre
Tes laitues naissent-elles
Matin d'été
Indo-européen
Et ta soeur
Les spécifs du clavier
Euréka youpi
Lucky Luke
Papa ira à la pêche

Il fait chaud
Martini on the rocks
la route est large
Dis-moi bonjour
Plies soles et bars
ZZZ...
Eaux douce d'Asie
Ah que j'aime
Demain dès l'aube

Aladin

Rhododendron
Si seulement ça cédait
Ty didrouz

Rapport d'anomalies Socrate
Toulouse
Sésame fermes-toi.

IV - SESSIONS AFFICHEES

SYSTEME D'AIDE AUX HANDICAPES AUDITIFS

D. BOURDACHE - M. LAMOTTE
Laboratoire d'Electricité et d'Automatique
C.O. 140 - 54037 NANCY Cédex

On a réalisé un système de visualisation de certains paramètres de la parole en vue de l'aide à l'éducation des sourds profonds.

Le dispositif comprend trois parties :

1. *L'analyseur*, composé de 20 filtres passe-bande quart d'octave, échelonnés de 200 à 6400 Hz, d'un passe-bas à 200 Hz et d'un passe-haut à 6400 Hz. Ce sont des filtres numériques récursifs du second ordre fonctionnant sur un signal vocal échantillonné à 16 kHz. Compte-tenu de l'arrangement des filtres, cela revient à faire travailler 20 filtres en 62 µs. La sortie des filtres est échantillonnée à 25 Hz.

2. *L'ensemble du traitement implanté sur microprocesseur TMS 9900* :

Il assure la prise en compte des commandes, l'acquisition des données et leur traitement en vue de synthétiser une image représentative d'un aspect de la parole selon l'exercice d'orthophonie choisi. Ces images sont la représentation stricte d'un phénomène, par exemple intensité en fonction du temps, ou une représentation symbolique, par exemple voyelle. Sur l'écran apparaît un tracé de référence fixe et un tracé représentatif de la prononciation instantanée de l'élève.

Comme il est nécessaire que les images soient générées et visualisées en temps réel, lès tracés sont composés de petits segments fournis par un générateur de caractères spéciaux. A tout instant une mémoire 64 x 38 contient le code de ces caractères.

3. *L'interface de visualisation* :

La mémoire 64 x 38 est explorée au rythme du balayage télévision, et le générateur de caractères construit le signal vidéo. La réception se fait sur un récepteur de télévision standard.

ASSISTANCE TO DEAF PEOPLE

D. BOURDACHE - M. LAMOTTE
Laboratoire d'Electricité et d'Automatique
C.O. 140 - 54037 NANCY Cédex

A visualisation system of some speech parameters has been carried out for the teaching assistance of deeply deaf people.

This system includes three parts :

1. *The analyzor* with twenty 1/4 octave band-pass filters form 200 Hz to 6400 Hz, with a 200 Hz low-pass filter and one 6400 Hz high-pass filter. These filters are second order recursive numerical ones, working on a vocal signal sampled to 16 Khz. According to the filters' arrangement, it follows that they work in 62 µs. The outputs of the filters are sampled to 25 Hz.

2. *The whole treatment* has been implanted in a microprocessor TMS 9900 giving orders, treating numerical data for finding out the drawing components typical of the orthophonic excercize. These pictures show one phenoma (for example the intensity in function of the time) or give a symbolic image of a vowel, for example.

A motionless reference picture and an illustration of the pupil's instantaneous pronounciation are drawed simultaneously on the T.V. screen.

As this device must work in real time, the plottings are composed of short segments given by a special character generator. At every moment, one memory 64 x 38 keeps in these character's codes.

3. *The visualisation interface* :

The 64 x 38 memory is explored with rate of the video scanning and the character generator builds up the video signal which is transfered on a standard video receptor.

Delgado Martins,Maria Raquel
Laboratoire de Phonétique.Faculté de Lettres.
Université de Lisbonne. Portugal

Perception de degrès d'accent dans la phrase

Résumé

Est-il possible aux sujets de percevoir auditivement plusieurs degrés d'accent dans une phrase de leur langue maternelle?

Nous avons testé,à l'écoute,dix phrases du portugais sur un groupe de 32 sujets portugais,naïfs, en leur demandant d'attribuer à chaque syllabe un degré d'accent de 1 à n -n étant le nombre de syllabes de la phrase et 1 l'accent principal,2 l'accent secondaire ect..

Nous avons analysé les résultats ainsi obtenus par une méthode statistique de régression multiple pour laquelle nous avons considéré comme variable dépendante le degré d'accent attribué à chaque syllabe et comme variables indépendantes les valeurs des paramètres acoustiques de la même syllabe,soit:la fréquence fondamentale,la durée,l'intensité,l'énergie et le pourcentage de durée et d'énergie de chaque syllabe par rapport à l'énoncé où elle s'insère.

Les valeurs de corrélation de régression multiple(R^2) correspondant à l'homogénéité des réponses attribuant un même degré d'accent à une même syllabe sont particulièrement significatifs pour les degrés d'accent 1,2 et 9,10,(respectivement 0,91;0,89;0,70;0,72).Ces valeures s'avèrent moins significatives pour les degrés d'accent intermédiaires ce qui suggére que la valeur des paramètres acoustiques n'est pas un indice systématique mais qu'elle peut contribuer à la reconnaissance de certains types d'accent.Une analyse plus approfondie de ces valeurs de corrélation permettra de calculer l'importance relative et la hiérarchisation de chaque paramètre par rapport à la corrélation globale.

Cette analyse permettra,en particulier,de tester au niveau de la phrase les hypothèses énoncées et verifiées au niveau de l'accent de mot (Delgado Martins,M.R.,1977),notamment l'importance de la valeur relative des paramètres acoustiques par rapport à l'énoncé.Dans une étape ultérieure de cette recherche nous essayerons de verifier si l'accent de mot s'organise dans la phrase en fonction de sa structure phonologique et syntaxique et si la "conscience" qu'a le sujet de la langue ont une fonction déterminante quant aux réponses obtenues par ce test.

Delgado Martins,M.R.(1977) Aspects de l'accent en Portugais.Thèse de Doctorat de 3ème. cycle-Université de Strasbourg(non-publiée)

Delgado Martins,Maria Raquel
Phonetics Laboratory
University of Lisbon-Portugal

Perception of stress level in the sentence

Abstract

Can subjects recognize several stress levels in a sentence of their native language?

We submitted 10 oral sentences to 32 native and naïve subjects who were asked to assign a level of stress to each syllable from 1 to n -n being the number of syllables and 1 the primary stress, 2 the secondary stress stc.

Multiple regression analysis was applied.For each syllable we choose as the dependent variable the level of stress assigned and as the independent variable the value of the following acoustic parameters:fundamental frequency,intensity,duration,energy for each syllable and the percentage of duration and energy of each syllable in relation to the sentence.

The values of multiple regression correlation(R^2) corresponding to the homogeneity of the responses assigning the same stress level to the same syllable are particularly significant for stress levels such as 1,2 and 9,10 (respectively 0,91;0,89;0,70;0,72).They are less significant for intermediate levels suggesting that the values of the acoustic parameters are not a systematic feature,but can contribute to the stress recognition.A more accurate analysis of the values of correlation will allow the evaluation of the importance and hierarchy of the acoustic parameters for the global correlation.

This analysis will permit to test the hypothesis formulated and verified in relation to the word stress level (Delgado Martins, M.R. 1977),in particular, the importance for the perception of the relative values of the acoustic parameters in relation to the whole sentence.Later on, we will verify if the word stress organize themselves in the sentence according to its phonological and syntactic structure and if the "conciousness" the subject has of the language is relevant on the responses obtained to the test.

Delgado Martins,M.R. (1977) Aspects de l'accent en Portugais.Thèse de Doctorat de 3ème. cycle-University of Strasbourg-(non publi).

Echantillonnages pour l'adaptation des systèmes de reconnaissance aux locuteurs

JEAN A. DREYFUS-GRAF

La reconnaissance automatique de la parole exige le recours à plusieurs niveaux de sources de connaissance, tels que : N1) acoustique, N2) phonétique, N3) phonologique, N4) lexical, N5) syntaxique, N6) sémantique, N7) contextuel.

La reconnaissance automatique de la parole exige généralement l'intervention de contraintes artificielles, telles que syntaxiques [1], lexicales [2], ou phonétiques [3], qui varient selon la forme de parole à reconnaître. Celle-ci peut être continue ou segmentée, large ou restreinte, quasi-naturelle ou phono-codée, nationale ou internationale, individuelle ou collective.

Au niveau de la source de connaissance phonétique ou phonologique (N2 ou N3), la machine doit subir un apprentissage préliminaire, portant sur un nombre "optimum" d'échantillons (modèles de référence) qu'on peut nommer "phono-types".

"Optimum" signifie : conciliant un minimum de temps d'apprentissage et d'encombrement de mémoire avec un maximum de fiabilité (ou minimum d'erreur).

Le Tableau 1. propose le principe général d'un système (réductible ou extensible) de 24 phono-types à 3 syllabes. Il comprend 14 mots internationaux et 10 mots nationaux, français dans le cas considéré. Il totalise 147 diphones (ou transitions), 72 syllabes et 30 phonèmes, dont 14 voyelles et 16 consonnes. Les 14 mots internationaux, basés sur 3 voyelles (o,a,i) et 14 consonnes (šfs jvz kpt gbd nm), sont communs à la plupart des langues indo-européennes. Le français y ajoute 2 consonnes (lr) et 11 voyelles. Au total, il y a les 14 voyelles (ûôo âaê eöü éi ō ā ī) et les 16 consonnes (šfs jvz kpt gbd nm lr), avec leurs transitions principales. Les autres diphones peuvent en être déduits par programmes de machine. La durée d'apprentissage, par la machine, des 24 phono-types serait d'environ 36 secondes pour la parole quasi-naturelle. Dans le cas de parole phono-codée, il suffirait d'un seul phono-type, durant 1,5 seconde, pour une adaptation individuelle.

Tableau 1. PHONO-TYPES: 24 mots (72 syllabes, 30 phonèmes (=14 voy.+ 16 cons.))

a) internationaux: 14 mots 3 voy.(oai), 14 cons.(âfs jvz kpt gbd mn)					b) nationaux: 10 mots, 2 cons.(lr) 14 voyelles (ûôo âaê eöü éi ō ā ī)			
âôšâšiš fôfâfif sôsâsis	jôjâjij vôvâviv zôzâziz	kôkâkik pôpâpip tôtâtit	gôgâgig bôbâbib dôdâdid	nônânin mômâmim	lôlâlil rôrârir	kûkôkok kâkakêk kekökük	kêkékik tōtātīt	klôklâklikl prôprâprêtr trôtrâtrêtr
symboles : ô=côte â=pâte i=gîte û=cou e=je ê=tête ō=ton Ī=cinq š=ch j=je o=cotte a=patte ü=fut ö=jeu é=blé ā=tant								

[1] BAHL, BAKER, COHEN, COLE, JELINEK, LEWIS, MERCER (I.B.M.), Automatic recognition of continously spoken sentences from a finite state grammar, IEEE, ICASSP, 1978

[2] RABINER, WILPON, Speaker Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary, IEEE, ASSP, December 1979

[3] DREYFUS-GRAF, Reconnaissance de mots quasi-naturels et phono-codés, GALF, 10èmes Journées d'Etude sur la Parole, Grenoble, 1979.

13.3.80

SUR L'UTILISATION DES PARAMETRES GLOBAUX DE LA VOIX POUR LE CLASSEMENT ET L'IDENTIFICATION DU LOCUTEUR.

G.IBBA - A.PAOLONI - Fondazione "U.BORDONI" - Roma (Italie)

Il est bien connu que le spectre du signal de parole contient même des caractéristiques liées au locuteur et, par conséquent, il peut être utilisé pour son identification. A ce but, nous avons effectué quelques determinations sur les spectres vocaux moyens, calculés à l'aide d'un analysateur spectral en temps réel et d'un calculateur électonique. Ces spectres étaient constitués par 400 bandes étroites de 12.5 Hz, correspondantes à une bande globale d'une largeur de 5000 Hz.

Tout d'abord, nous avons évalué la durée la plus convenable du message vocal en calculant les differences entre les spectres à l'aide d'une relation analitique (distance euclidée) appliquée au logaritme des valeurs d'amplitude de chaque ligne des spectres moyens normalisés. Nous avons assumé comme optimale la durée du signal vocal au-delà de laquelle la distance gardait une valeur à peu près constante. En adoptant une durée moyenne de 20 s pour l'emission vocale, on a calculé, avec le même procedé et pour un grand nombre de locuteurs, la distance entre le même locuteur (distance intra-locuteur) et entre locuteurs différents (distance inter-locuteurs) (fig.1).

Cette detérmination a été effectuée sur des signaux vocaux enregistrés sur un ruban magnetique d'une façon ortophonique et en limitant les spectres à bandes de fréquences de différentes amplitudes.

Les résultats de nos mesures sont resumés dans la fig.2 ; on voit que la position relative des deux courbes (intra et inter-locuteurs) permet de prévoir la possibilité d'une discrimination assez significative en ce qui concerne l'exclusion ou l'attribution de deux voix au même locuteur.

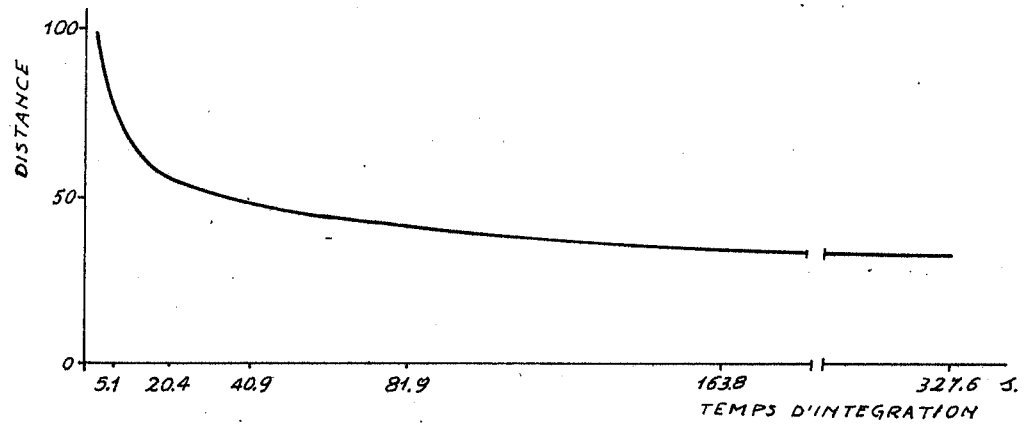


FIG. 1- VARIATION DES VALEURS DE DISTANCE INTRA-LOCUTEUR EN FONCTION DU TEMPS D'INTEGRATION.

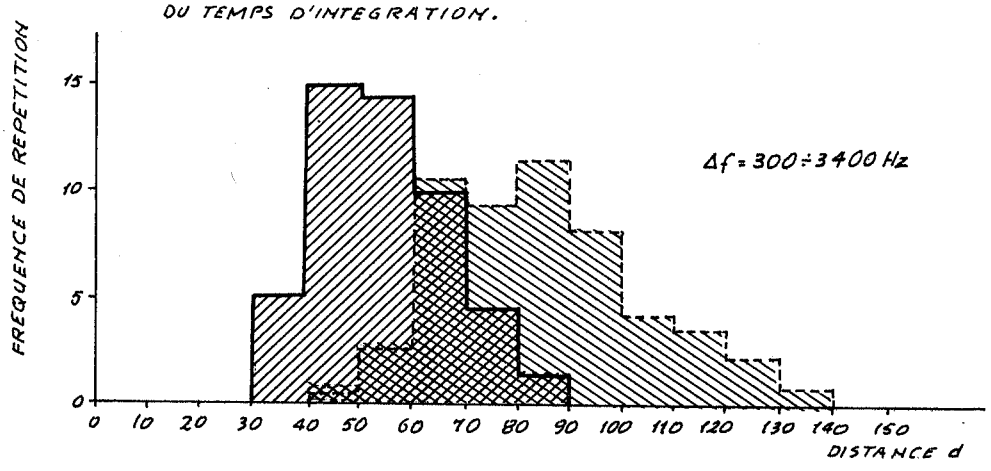


FIG.2 - HISTOGRAMMES RELATIFS AUX VALEURS DE DISTANCE INTRA-LOCUTEUR (—) ET INTER-LOCUTEURS (---).

RECONNAISSANCE DE MOTS ISOLES AVEC DETECTION DES FAUTES DE PRONONCIATION

M.T. GIORGETTI - M. LAMOTTE
Laboratoire d'Electricité et d'Automatique
C.O. 140 - 54037 NANCY Cédex

Il s'agit d'un appareil autonome permettant la reconnaissance de mots isolés parmi un nombre assez faible de mots de référence. Il est utilisé en enseignement programmé de langues. Plus précisément, au cours d'une leçon, l'élève est invité à prononcer un mot et le système doit reconnaître si ce mot a été bien prononcé et, dans le cas contraire, détecter la faute et la localiser.

On a donc, à tout moment, à reconnaître un mot parmi une liste qui contient la bonne occurence et les mauvaises prononciations possibles.

Le système comprend un analyseur à 8 canaux choisis d'après les données du triangle vocalique et échantillonnés à 50 Hz. Les mots de référence sont stockés sur cassette et appelés par blocs au fur et à mesure des besoins. La reconnaissance utilise une procédure de programmation dynamique qui détecte sur la fonction de retour la position de l'écart maximal par rapport à la forme de référence.

Le résultat peut présenter trois formes :

- bonne prononciation : le mot est écrit sur la console et on passe à l'exercice suivant ;
- erreur légère de prononciation : le mot correct est écrit en soulignant la partie du mot correspondant à la position de l'écart maximal et on donne l'ordre de répéter ;
- très mauvaise prononciation : le mot doit être répété.

Le traitement est effectué par un microprocesseur TMS 9900 et le programme occupe environ 2500 octets.

RECOGNITION OF ISOLATED WORDS WITH FINDING OF PRONOUNCIATION MISTAKES

M.T. GIORGETTI - M. LAMOTTE
Laboratoire d'Electricité et d'Automatique
C.O. 140 - 54037 NANCY Cédex

This work deals with an autonomous device which permits recognition of isolated words among a restricted number of reference words. This device can be used in programmated language teaching.

In the procedure, the student is asked to pronounce a word which will receive from the device either a good pronounciation label, or a localization indication of pronounciation mistake.

The recognition algorithm is then a comparison between the pronounced word and those of a list including the correct elocution and all the possible uncorrect ones.

One uses an eight channel analyzor which has been chosen according to the data of the vowel plot and a fifty hertz sampling. The reference words are tape-stored and called when needed. The recognition procedure is based on dynamic programming and finds out the place of the maximum difference between the return function and the reference pattern.

The result can be the following :

- if well pronounced, the word is written on the visualisation screen and one can go on ;
- if only slightly mistaken, the correct word is written with the indication of the mistaken part and an order is given for repeating ;
- if very badly pronounced, the word must be repeated.

A microprocessor TMS 9900 is used with a 2500 octets programm.

Reconnaissance vocale de séquences de mots adaptée au tri postal

Jean-Paul HATON et Olivier MOREL

Centre de Recherche en Informatique de Nancy
C.O. 140 54037 NANCY

Le but de ces travaux est la reconnaissance de séquences de chiffres prononcés continuement. Notre système s'insère dans une étude de faisabilité de commande vocale de machines automatiques de tri postal de paquets. Le nombre de chiffres est connu : soit deux (code de département) soit trois (code de ville).

Les chiffres sont analysés par un vocodeur à 15 canaux, puis traités par une méthode semi-globale comportant deux étapes :

- segmentation du signal, en utilisant des paramètres de voisement, d'énergie et de répartition spectrale.
- Reconnaissance des chiffres par un algorithme de programmation dynamique, que nous utilisons avec succès depuis plusieurs années pour la reconnaissance de mots isolés.(1)

Nous segmentons les séquences de deux chiffres selon diverses heuristiques utilisant des fonctions binaires (voisement, énergie, zones fricatives), et une fonction de variation spectrale.Le principe consiste à localiser les noyaux syllabiques, caractéristiques des chiffres, en recherchant les zones de forte énergie, puis à affiner le découpage en utilisant les autres paramètres. Le cas particulier du "zéro", qui comprend deux syllabes, est résolu en prononçant "O".

Notre système offre dès à présent des résultats très encourageants, pour les séries de 2 chiffres, malgré les effets de la coarticulation qui a tendance à déformer les chiffres. Ce problème est en voie de résolution par un apprentissage approprié, qui recherche la forme de référence moyenne entre les occurences d'un même chiffre dans des contextes différents. Nous utilisons aussi pour certains chiffres une forme différente selon que le chiffre est placé en début ou fin de prononciation.

Nous envisageons maintenant d'étendre notre système aux séries de 3 chiffres. De plus, pour tester ses performances, nous implantons actuellement un algorithme de localisation de mots. Celui-ci, qui compare localement une forme de référence à la séquence inconnue, est très largement inspiré de l'algorithme de recherche lexicale décrit dans (2).Nous pourrons ainsi disposer d'un point de comparaison, tant sur le plan des performances que des résultats.

Références :

(1) J.P.Haton " Reconnaissance analytique de la parole aux niveaux acoustiques, morphologique, lexical et syntaxique" RAIRO Informatique, 10, n° 9, pp. 57-75, Sept. 1976.
(2) J.M.Pierrel,J.F. Mari, J.P. Haton "Le niveau lexical dans le système MYRTILLE II" 10èmes J.E.P. GRENOBLE 30 mai 1979.

Automatic Recognition of Connected Digits Sequences for postal codes recognition

Jean-Paul HATON and Olivier MOREL

Centre de Recherche en Informatique de Nancy
C.O 140 54037 NANCY

This paper describes an experiment of recognition of sequences of digits pronounced continuously. The sequence represents the french ZIP codes, so the number of digits is known : two or three according to the operation.

The speech signal is analysed by a 15-channel spectral analyser. The digit sequence is then recognized using a semi-global procedure which consists of two successive steps.

- segmentation of the sequence into digits.
- recognition of the individual digits by a dynamic programming technique which was previously designed in our laboratory. (1)

We segment the two-digits sequences using heuristics based on pitch, energy and fricative zones, and a forth fonction of spectral variation. We first locate energy zones, which are expected to be the center of the digit, and then segment between these zones using the other parameters.

We have got encouraging results by now. Due to coarticulation effects, the global pattern of the digit can be affected by its position in the sequence. So we have developed a learning process which can take a mean pattern of a digit occuring in different contexts. We also use two different patterns for digits like "sept" /s ɛ t/ where the burst of the /t/ can disappear.

We shall now extend our system to three-digits sequences. We are also implementing a word spotting algorithm. (2) So we will soon be able to compare results and performances of both techniques.

References :

(1) J.P. Haton "Reconnaissance analytique de la parole aux niveaux acoustique, morphologique, lexical et syntaxique" RAIRO informatique, 10, n° 9, pp.57-75, Sept. 1976.

(2) J.M. Pierrel, J.F. Mari, J.P. Haton "Le niveau lexical dans le système MYRTILLE II" 10èmes J.E.P. Grenoble, 30 mai 1979.

Intelligibilité des mots codés chuchotés

Jens-Peter Köster, Monika Klaes et Herbert R. Masthoff
Université de Trèves

Depuis 1975, des travaux de recherche dans le domaine de l'élaboration des langages artificiels pour la communication verbale homme/machine sont en cours à l'Université de Trèves. Ces travaux ont abouti à l'établissement d'une série de lois déterminant la perception des phonocodes chez l'homme. La connaissance de telles lois est importante

a. pour la sélection des mots codés ayant des qualités requises pour un code qui doit assurer une perception parfaite dans des conditions bien déterminées

b. pour établir des hypothèses sur l'inventaire des sons parfaitement intelligibles pour des machines simples de reconnaissance de la parole codée.

Après avoir décrit les effets de filtrage et de masquage du signal naturel codé sur la perception, nous nous sommes intéressés à l'intelligibilité des mots codés chuchotés dans trois conditions particulières de présentation: non manipulés, filtrés et masqués. Voici quelques résultats d'une étude pilote récente:

- L'intelligibilité des mots codés chuchotés est soumise au même nombre et au même genre de lois générales que celles isolées pour la perception des mots codés normaux; ce qui change, cependant, c'est l'intelligibilité des éléments de base (sons chuchotés) qui réclame une structure particulière des codes définitifs (légère modification des codes développés à partir de mots codés normaux).

- La hiérarchie d'intelligibilité de mots codés chuchotés est fonction des conditions particulières de présentation (signal non manipulé → filtré → masqué).

- Un mot mal perçu est, en principe, remplacé par un mot typique, cette loi étant également en vigueur au niveau des éléments de base. Les règles de détail sont, pour les uns comme pour les autres, fonction des conditions particulières de présentation.

- L'intelligibilité de chaque élément de base atteint une valeur qui est typique pour le langage chuchoté en ce qui concerne la hiérarchie entre les éléments; le total des fautes observées, cependant, correspond exactement au taux de reconnaissance des éléments normaux:

	a	i	n	o	s	t	total
normal	0,1	0,34	0,54	0,26	0,34	0,57	4,6% de fautes sur l'ensemble
chuchoté	0,05	1,51	0,03	1,17	0,37	1,45	4,6% des éléments de base

- L'intelligibilité des éléments de base est fonction de leur position dans le mot codé chuchoté.

Intelligibility of Whispered Artificial Words

Jens-Peter Köster, Monika Klaes and Herbert R. Masthoff
Universität Trier

In 1975 a research program for the development of artificial languages for man/machine communication has been initiated at the Universität Trier. As a result, a series of rules has been established which govern the perception of phonocodes by man. Such rules are important for:

a. the selection of artificial words allowing the setup of codes which provide high perceptual performance under any well defined condition

b. hypothesizing sets of speech sounds perfectly intelligible for simple recognition devices.

After having described the effects of filtering and masking of the natural coded signal upon the perception, our interest has recently been centering on the problem of the intelligibility of whispered artificial words under three particular conditions: whispered signals only, + filtering, + masking. Here are some results of a pilot study conducted recently:

- The intelligibility of whispered artificial words is subject to the same number and kind of general rules as those stated with non-whispered artificial words; however, the intelligibility of the various whispered speech sounds changes requiring a different structure of the definitive code (slight modification of the codes developed with non-whispered artificial words).
- The hierarchy of intelligibility of whispered artificial words is a function of the particular conditions of presentation (no distortion → filtering → masking).
- In general, a wrongly perceived word will be substituted by another specific word. This rule equally applies to the level of the speech sounds. In detail the rules are - for both words and sounds - a function of the particular conditions of presentation.
- Concerning the hierarchy among the speech sounds, the intelligibility of each sound takes a specific value in whispered speech; the total of all mistakes observed, however, exactly corresponds to the recognition score of normal speech sounds:

	a	i	n	o	s	t	total	
normal	0,1	0,34	0,54	0,26	0,34	0,57	4,6%	percentage of mistakes
whispered	0,05	1,51	0,03	1,17	0,37	1,45	4,6%	related to the total of all occurring sounds

- The intelligibility of the sounds is a function of their position within the whispered artificial words.

LISSAGE DE FONCTIONS D'AIRES OBTENUES PAR METHODE 'ACOUSTIQUE

Jean-Paul Lefèvre, Bernard Tousignant, Michel Lecours

Département de Génie électrique
Université Laval, Québec

Pour évaluer la fonction d'aire du conduit vocal à partir de la mesure de la réponse impulsionnelle au niveau des lèvres [1], [2], on fait l'hypothèse d'une fonction d'aire finie, positive et dérivable deux fois et, on se place dans un contexte de propagation d'ondes planes, ce qui limite les signaux de mesure à environ 4 kHz. Ces contraintes conduisent à proposer des techniques de lissage des coefficients de réflexion au voisinage d'une discontinuité identifiée avant de procéder au calcul d'une fonction d'aire continue correspondante.

Nous avions précédemment utilisé un lissage de type sinusoïdal [3], [4]: les résultats ainsi obtenus montraient une meilleure convergence des calculs et avaient conduit à des fonctions d'aire représentées par quelques sections cylindriques raccordées de façon continue. Nous proposons ici d'utiliser pour le lissage des fonctions d'aire, une loi correspondant à une réponse impulsionnelle du type $ae^{-b|t|}$. Une telle loi de lissage, en plus de respecter les hypothèses de départ, ne présente pas de distorsion de phase. En termes de fonctions d'aire, il en résulte que la position d'une constriction n'est pas modifiée lors du lissage. Tout en respectant les contraintes de base sur la limitation de la largeur de bande, cette loi permet de présenter des variations plus abruptes de fonction d'aire que le lissage sinusoïdal et, donc, une représentation plus détaillée de la fonction d'aire, notamment lors d'une succession de variations de la section. Les premiers résultats obtenus depuis l'introduction de cette nouvelle loi indiquent notamment une meilleure maîtrise de la localisation de la glotte. Quelques sons obtenus par synthèse permettent d'apprécier les performances de la technique utilisée.

1 - M.M. Sondhi, B. Gopinath, "Determination of vocal-tract shape from impulse response at the lips", J. Acoust. Soc. Amer., vol. 49, no 6 (part 2), pp. 1867-1873, juin 1971.

2 - R. Descout, B. Tousignant, M. Lecours, "Deux méthodes de détermination de la fonction d'aire du conduit vocal dans le domaine temporel", 7èmes Journées d'Etude sur la Parole, Nancy, 19-21 mai 1976, pp. 307-318.

3 - R. Descout, B. Tousignant, J.P. Lefèvre, M. Lecours, "Détermination de la fonction d'aire du conduit vocal: quantification et interpolation", Actes du symposium Modèles articulatoires et Phonétique, Grenoble, 10-12 juillet 1977, pp. 31-40.

4 - B. Tousignant, J.P. Lefèvre, M. Lecours, "Speech synthesis from vocal tract area function acoustical measurements", Proc. IEEE ICASSP, Washington, D.C., 2-4 avril 1979, pp. 921-924.

WEIGHTING VOCAL TRACT AREA FUNCTIONS OBTAINED BY ACOUSTICAL MEASUREMENTS

Jean-Paul Lefèvre, Bernard Tousignant, Michel Lecours

Département de Génie électrique
Université Laval, Québec

For the evaluation of the vocal tract area function from the measurement of the impulse response at the lips [1], [2], one makes the assumption of a finite area function, positive and twice differentiable and of plane wave propagation, condition which limits the bandwith of the measuring signals to approximately 4 kHz. These contraints lead to weighting the reflection coefficients in the neighbourhood of a detected discontinuity before computing the values corresponding to the continuous area function.

We had used previously a sinusoidal weighting function: the results showed improved convergence in the computations and lead to the representation of area functions by cylindrical elements smoothly joined together [3], [4]. We propose here to adopt for weighting area functions a law corresponding to an impulse function of the type $ae^{-b|t|}$. In addition to respecting the above-mentioned assumptions, such a law avoids the introduction of phase distorsion, so that the weighting process does not change the position of a constriction in an area function. This law, while still respecting the basic assumptions on bandwidth limitation, allows more abrupt changes in the area function than the sinusoidal law: the more detailed representation of the area functions obtained appears particularly advantageous for modeling successive changes in the cross-section of the vocal tract. The first results obtained with this new weighting law show a better performance in glottis localization. A number of sounds have been synthesized and permit to evaluate the sound quality obtained with this kind of weighting.

1 - M.M. Sondhi, B. Gopinath, "Determination of vocal-tract shape from impulse response at the lips", J. Acoust. Soc. Amer., vol. 49, no 6 (part 2), pp. 1867-1873, juin 1971.

2 - R. Descout, B. Tousignant, M. Lecours, "Deux méthodes de détermination de la fonction d'aire du conduit vocal dans le domaine temporel", 7èmes Journées d'Etude sur la Parole, Nancy, 19-21 mai 1976, pp. 307-318.

3 - R. Descout, B. Tousignant, J.P. Lefèvre, M. Lecours, "Détermination de la fonction d'aire du conduit vocal: quantification et interpolation", Actes du symposium Modèles articulatoires et Phonétique, Grenoble, 10-12 juillet 1977, pp. 31-40.

4 - B. Tousignant, J.P. Lefèvre, M. Lecours, "Speech synthesis from vocal tract area function acoustical measurements", Proc. IEEE ICASSP, Washington, D.C., 2-4 avril 1979, pp. 921-924.

PHRASES FRANCAISES PHONETIQUEMENT EQUILIBREES

M. Lennig et P. Mermelstein
Recherches Bell-Northern, 3 Place du Commerce
Ile des Soeurs, Québec, Canada H3E 1H6

Il existe des phrases phonétiquement équilibrées pour l'anglais [1] mais non pas pour le français. Ces listes de dix phrases chacune ont des fréquences relatives pour chaque phonème qui reflètent celles qui existent dans la langue. Les listes de phrases phonétiquement équilibrées sont utiles pour toutes sortes d'expériences avec les systèmes de communication, tels que les tests d'intelligibilité, où il est important de ne pas favoriser certains phonèmes plus que d'autres. La présente communication offre deux listes de phrases françaises phonétiquement équilibrées. Nous avons utilisé une statistique χ^2 pour mesurer la différence entre une liste proposée et des données de fréquences relatives des phonèmes, calculées sur un grand corpus de théâtre français [2].

Pour créer chaque liste de phrases phonétiquement équilibrées, nous avons d'abord écrit dix phrases arbitraires. Nous avons ensuite calculé le χ^2 total entre les fréquences relatives des phonèmes dans la liste, et celles du grand corpus. Nous avons utilisé le critère de $\chi^2 < 4.5$ pour accepter une liste comme "phonétiquement équilibrée". Si la liste initiale ne satisfaisait pas le critère, nous en avons modifié les phrases jusqu'à ce que le critère soit satisfait.

[1] IEEE Recommended Practice for Speech Quality Measurements, IEEE Transactions on Audio and Electroacoustics, Vol. AU-17, No.3, September 1969.

[2] Szklarczyk, Lillian, 1961, Essai sur la structure phonologique du français. Thèse de doctorat inédite, University of Pennsylvania.

PHONETICALLY BALANCED SENTENCES IN FRENCH

M. Lennig and P. Mermelstein
Bell-Northern Research, 3 Place du Commerce
Nuns' Island, Quebec, Canada H3E 1H6

Phonetically balanced sentences exist for English [1] but not for French. These lists of 10 sentences each have relative frequencies for each phoneme which reflect phoneme frequencies in the language. Phonetically balanced sentences are useful in testing communication systems, where it is important not to favour one phoneme over another. This paper presents two lists of French phonetically balanced sentences. We have used an χ^2 statistic to measure the difference between a proposed list and relative frequency data on phonemes computed from a large corpus of French theatre [2].

To create each list of phonetically balanced sentences, we first wrote down ten arbitrary sentences. We then calculated the total χ^2 between the phoneme relative frequencies in the list and those in the large corpus. We used the criterion $\chi^2 < 4.5$ to accept a list of sentences as "phonetically balanced". If the initial list did not meet this criterion, we modified the sentences until the criterion was satisfied.

[1] IEEE Recommended Practice for Speech Quality Measurements, IEEE Transactions on Audio and Electroacoustics, Vol. AU-17, No.3, September 1969.

[2] Szklarczyk, Lillian, 1961, Essai sur la structure phonologique du français. Unpublished Ph.D. Thesis, University of Pennsylvania.

ETAT DES RECHERCHES SUR LA PAROLE EN AUSTRALIE

Cette revue des activités de recherche sur la parole en Australie fait suite à une visite de quelques laboratoires, effectuée courant février 80 à Sydney, Canberra et Melbourne. La situation particulière du pays (isolement géographique et pénurie d'ingénieurs) explique l'actuel développement "modéré" des programmes de recherche australiens sur le signal de parole.

Le laboratoire le plus avancé est le Speech and Language Research Center de l'Université MacQuarie àSydney où un programme de synthèse par régles a été mis au point (J.Clark). C'est un outil de travail particulièrement bien adapté à des études de perception grâce à la modification en temps réel des 12 paramètres de synthèse qui commandent un synthétiseur à formants. Des études sur la nasalité (Blair et Clark) et sur le chant (Bernard et connor) complètent, avec l'enseignement, les activités du laboratoire.

L'Australian National University de Canberra, créée dans les années 50, est dotée de moyens informatiques considérables : 1 PDP 11/45, 1 PDP 11/40 pour les seuls groupes mentionnés ci-après (3 ingénieurs+ 1 technicien).
L'Information Sciences Group du Department of Engineering Physics (A. Collins) s'est spécialisé dans l'analyse : LPC et FFT.L'application visée est l'identification du locuteur; la phase actuelle est celle de la constitution d'une base de données sur une centaine de locuteurs.
Le Vice-Chancellors Computing Research Group (B. Millar) poursuit une recherche parallèle depuis 1 1/2 an, en vue d'une modélisation paramétrique du signal de parole incluant l'information locuteur et l'information linguistique.

La section de Neuroaudiologie du N.A.L.(National Acoustic Laboratories) de Sydney (Ph. Dermody) étudie particulièrement les phénomènes d'acquisition du langage dans le cadre d'un programme d'aide aux enfants handicapés auditifs ou retardés dans leur apprentissage du langage. L'essentiel de l'activité passée de l'équipe concerne le domaine clinique; l'orientation récente vise à développer des tests de perception de parole couplés à la mesure des réponses évoquées au niveau central.

Aux Laboratoires de Recherches de Telecom Australia à Melbourne, les études en analyse-synthèse de parole sont à peine amorcées. Le principal résultat actuel consiste en l'amélioration du naturel de la parole produite par un synthétiseur à formants dont les paramètres sont déterminés par une analyse LPC. La mise en place d'un programme de recherche en synthèse et reconnaissance de la parole a été ajournée faute d'ingénieur.

Le Department of Otolaryngology de l'Université de Melbourne (équipe de chirurgiens et d'ingénieurs) vient de débuter une étude sur la perception de la parole au niveau central. Cet intêret va de pair avec le développement de la technique d'implantation cochléaire (5 opérés depuis 1 an). les expérimentations "post-opératoires" sont en cours.

Enfin signalons le Department of Electrical Engineering de l'Université d'Adelaïde -que je n'ai pas eu le temps de visiter- et dont l'axe principal de recherche est la reconnaissance de la parole aussi bien que du locuteur grâce à des données articulatoires.

PASCAL Dominique C.N.E.T. Lannion A Département Codage et modèle des Communicatio
Route de Trégastel 22301 LANNION Cedex.

STATE OF SPEECH RESEARCH ACTIVITIES IN AUSTRALIA

This survey of speech research activities in Australia follows a one week visit to some loboratories in Sydney, Canberra and Melbourne in february 80. The particular situation of the country (geographical isolation and lack of engineers) explains the present moderate development of australian speech research programs.

The MacQuarie University Speech and Language Research Center in Sydney is the most advanced laboratory. A computer program of synthesis-by-rule had been implemented (J. Clark). It is very well adapted to perceptual studies because it stimulates a formant synthesizer driven in real time by 12 parameters. This laboratory is also involved in studies on nasality (Blair and Clark) and singing (Bernard and Connor) and, of course, teaching.

The Australian National University in Canberra, created in the fifties, is overequipped in computer equipments : 1 PDP 11/45 and 1 PDP 11/40 for the only two following groups (3 engineers + 1 technician).
The Information Sciences Group from the Department Of Engineering Physics (A.Collins) is specialized in analysis : LPC and FFT for the main purpose of speaker identification; the department is actually constituting a data base using one hundred speakers.
The Vice-Chancellors Computing Research Group (B. Millar) has been doing similar research for 1 1/2 year in order to develop a parametric model of speech phenomena integrating speaker and message.

The Neuroaudiology Section of the National Acoustic Laboratories in Sydney : Specific interests include the investigation of the auditory processing factors underlying speech acquisition and development and their relation to language difficulties of learning disabled and hearing impaired children. The main point of this team past activities regards clinical area. The recent trend emphasis development of measures of speech perception together with their corresponding evocated responses.

At Telecom Australia Research Laboratories in Melbourne studies in Speech analysis -synthesis have just been initiated. The main actual result is on improving the naturalness of synthetic speech produced by an LPC formant synthesizer. The setting of a research program in speech synthesis and recognition had been delayed by the lack of engineer.

The University of Melbourne Department of Otolaryngology (surgeons and engineers) has just started a study on speech perception to learn more about the central system mechanisms. This interest is correlated with the development of cochlear implant technics (5 patients operated in the one past year). Post operative experiments are in progress.

The University of Adelaïde Electrical Engineering Department is mainly doing research on speech and speaker recognition using articulatory parameters.

Dominique PASCAL C.N.E.T. Lannion A
Departement Codage et Modèle des Communications
Route de Trégastel
22301 LANNION

PERCEPTION DE L'INTONATION - EXPERIENCES SUR LE DANOIS (PAROLE NATURELLE)

Nina Thorsen, Institut de Phonétique, Université de Copenhague,
96 Njalsgade, DK-2300 Copenhague, DANEMARK

Résumé

Dix sujets ont identifié 15 phrases (parole naturelle), identiques sauf en ce qui concerne la fréquence fondamentale (Fo), comme déclaratives, non-finales, ou interrogatives (choix forcé). Les réponses sont en corrélation très étroite avec le Fo: les pentes (déterminées par les syllabes accentuées) les plus descendantes sont identifiées comme déclaratives, les pentes horizontales comme interrogatives, et les pentes intermédiaires comme non-finales.

L'inspection des tracés du Fo a révélé des corrélations étroites entre la distribution des réponses et l'allure des tracés du Fo, avec prédominance des deux dernières voyelles accentuées et de la voyelle inaccentuée finale (mais ni la montée finale allant de la voyelle accentuée à la voyelle finale inaccentuée, ni le mouvement du Fo dans la dernière voyelle accentuée ne montrent de corrélations étroites avec les réponses). Tous les points mesurés au cours des deux derniers groupes accentuels étant en corrélation/concordance étroite, il est impossible de déterminer quel point ou quelle combinaison de points constitue l'indice/les indices d'identification des contours. Toutefois les analyses statistiques suggèrent que le Fo de la dernière voyelle accentuée et celui de la voyelle inaccentuée suivante sont des paramètres indépendants pour l'identification des contours. Cela ne veut pas dire que le Fo de ce qui précède ne soit pas pertinente pour l'identification des contours de l'intonation, au contraire: à cause de sa concordance avec le groupe accentuel final, elle doit nous renseigner sur l'identité du contour, et cette information peut être utilisée, comme cela a été démontré par des expériences où un nombre (plus ou moins grand) de syllabes finales avaient été tronquées. L'identification des stimuli n'est sérieusement perturbée que quand il ne reste que le premier groupe accentuel. Par contre, les stimuli sont très bien identifiés quand on ne présente que le dernier groupe accentuel.

Dans un second test sept sujets ont identifié les mêmes stimulis comme déclaratifs ou non-déclaratifs. La majorité des phrases précédemment jugées non-finales et non la moitié ont été catégorisées comme non-déclaratives. Cela semble confirmer l'hypothèse suivant laquelle la catégorisation par les sujets dans le premier test a été linguistique, plutôt que purement phonétique.

[Résumé d'un article publié dans Journal of the Acoustical Society of America, 67 (1980).]

PERCEPTION OF INTONATION CONTOURS - EXPERIMENTS WITH NATURAL DANISH STIMULI

Nina Thorsen, Institute of Phonetics, University of Copenhagen,
96 Njalsgade, DK-2300 Copenhagen, DENMARK

Abstract

Ten subjects identified 15 natural Danish utterances, differing only in their fundamental frequency (Fo) course, as being either declarative, non-final, or interrogative (forced choice). Responses are closely correlated with Fo: the most steeply falling intonation contours are identified as being declarative, the least falling ones as being interrogative, and contours in the middle of the continuum as being non-final.

Close inspection of the course of Fo in the stimuli revealed rather high correlations between response distributions, on the one hand, and the course of Fo on the other, with a predominance of the last two stressed vowels and the final post-tonic vowel (whereas neither the final rise from stressed to unstressed vowel nor the Fo movement within the last stressed vowel yielded high correlations with responses). As all points of measurement during the second and third stress groups were highly correlated/concordant, there is no way to decide which one of these points or how many of them together cue identification of the contours, but statistical analyses suggest that Fo in the last stressed and post-tonic vowels are the independent parameters, as far as correlation with responses is concerned. This does not mean that the preceding Fo course is irrelevant for the identification of intonation contours, quite the contrary: because of its concordance with the last stress group it does carry information about the contour, and this information may be turned into account as was demonstrated by the identification of stimuli where a greater or lesser number of syllables had been cut away from the end. Only when just the first stress group remains is identification seriously affected. On the other hand, utterances are identified very well indeed when only the last stress group is presented.

In a subsequent experiment, seven subjects identified the same utterances as being either declarative or non-declarative. The majority of the (formerly) non-final sentences were now labelled non-declarative, rather than being equally distributed among the declarative and non-declarative categories. This may be taken as an indication that the categorization performed by the listeners in the first experiment is linguistic, rather than purely phonetic.

[Summary of a paper published in Journal of the Acoustical Society of America 67 (1980).]

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

UN MODELE DE DETECTION AUTOMATIQUE DES FRONTIERES INTONATIVES ET SYNTAXIQUES

MARIO ROSSI ET ALBERT DI CRISTO INSTITUT DE PHONÉTIQUE D'AIX

1.- INTRODUCTION

Nous présentions, dans "Propositions pour un modèle d'analyse de l'intonation"(ROSSI et DI CRISTO, 1977), une méthode d'interprétation prosodique susceptible d'être appliquée à la détection automatique des frontières des groupes intonatifs. L'analyse de type pluriparamétrique que nous préconisions procédait à l'effacement des variations microprosodiques et au transcodage perceptif des données et visait à extraire les traits liés de façon univoque aux instructions linguistiques (LIEBERMAN, 1967; ATKINSON, 1973). Des règles dites intonosyntaxiques (DI CRISTO, 1975), dont on ne donnait qu'un échantillon dans un exemple d'application,devaient permettre d'emettre des hypothèses sur l'organisation en constituants à partir des frontières prosodiques reconnues et de leur hiérarchie. Ce modèle, dont les axiomes demeurent parfaitement fondés, présente toutefois deux inconvénients majeurs.

1) L'interprétation des données brutes supposait la reconnaissance préalable de certains traits qui, dans un système de reconnaissance automatique, ne sont pas toujours identifiés de façon sûre : il s'agit en particulier du trait /Nasal/ qui était utilisé pour apporter une correction à la durée objective des voyelles. D'autre part, la détection des maximums de hauteur, points d'ancrage éventuels des frontières intonatives, procédait selon une méthode de proéminence locale qui risquait d'introduire des pics parasites en correspondance fortuite avec un maximum de durée, correspondance d'autant plus plausible que nous maîtrisons encore mal les effets des variations micro-temporelles. La détection des maximums, par ailleurs, ne faisait aucune place aux intonèmes progrédients inversés réalisés comme une chute ou une rupture négative (MARTIN, 1977., VAISSIERE, 1975).

Enfin les facteurs de pondération utilisés pour la fréquence et la durée dérivaient de l'analyse d'un corpus trop étroit. Une étude systématique sur la microprosodie du français nous fournit aujourd'hui des valeurs de coefficients de correction fiables (DI CRISTO, 1978).

2) La conception des règles intonosyntaxiques laissait supposer à tort qu'il était toujours possible de déduire directement l'organisation syntaxique de l'énoncé, ou du moins d'émettre une hypothèse sur cette organisation, à partir de la structuration prosodique. Nous revenions implicitement à une conception mécaniste des relations entre l'intonation et la syntaxe (STOCKWELL, 1960, 1972) qui a été suffisamment critiquée pour que nous ayons à y revenir (BRESNAN, 1972).

Nous faisons remarquer ailleurs (ROSSI, 1979) que s'il est légitime de penser que la hiérarchie intonative en structure sous-jacente reflète la hiérarchie syntaxique, dans l'énoncé l'organisation prosodique peut être bouleversée.

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

UN MODELE DE DETECTION AUTOMATIQUE DES FRONTIERES INTONATIVES ET SYNTAXIQUES

MARIO ROSSI ET ALBERT DI CRISTO INSTITUT DE PHONÉTIQUE D'AIX

RÉSUMÉ

Nous présentons un modèle de détection automatique des frontières intonatives et syntaxiques qui prolonge, en apportant des solutions nouvelles, la méthode exposée dans (ROSSI et DI CRISTO, 1977).

Dans la première partie, nous définissons des règles d'énonciation que nous mettons en correspondance avec la syntaxe. Cette première étape nous conduit à formuler trois types de règles:

i.) des règles génératives d'énonciation;
ii.) des règles transformatives d'énonciation;
iii.) des règles de correspondance intonation-syntaxe.

L'algorithme de reconnaissance des frontières intonatives et syntaxiques, développé dans la seconde partie, représente un modèle de perception qui est fondé sur l'effacement des variations microprosodiques intrinsèques et co-intrinsèques et sur le transcodage perceptif des données acoustiques. Cet algorithme s'articule suivant trois étapes:

a) Calcul des configurations objectives de Fo.
b) Stylisation des paramètres prosodiques.
c) Recherche des frontières intonatives et de leur hiérarchie.

L'ordre des corrections apportées dans le transcodage des données acoustiques n'est pas indifférent.Il est lié aux hypothèses que nous pouvons faire sur la hiérarchie des mécanismes d'intégration mis en oeuvre dans le processus de reconnaissance de la parole.

SUMMARY

We present a model for the automatic detection of intonative and syntactic boundaries which is a development, with a number of new solutions, of the method outlined in (ROSSI et DI CRISTO, 1977).

In the first part we define rules of enunciation linked to the syntax. We are led, in this first stage, to formulate three types of rules:

i.) generative enunciation rules;
ii.) transformative enunciation rules;
iii.) Enunciation-syntax correspondance rules.

The algorithm for detecting intonative and syntactic boundaries is developped in the second part. It represents a perceptual model based on the cancelling of microprosodic intrinsic and co-intrinsic variations and on the perceptual recoding of acoustic data. The algorithm comprises the following three stages:

a) Calculation of objective Fo patterns.
b) Stylising of prosodic parameters.
c) Identification of intonative boundaries and their hierarchy.

The order in which corrections are applied in the recoding of acoustic data is not indifferent, but it is linked to the hypothesis which can be formulated concerning the mechanisms of perceptual integration which are brought to bear during the process of speech understanding.

De la sorte, deux hiérarchies intonatives différentes peuvent représenter une seule et même structure syntaxique. Inversement, le même énoncé, et la même structure intonative, par exemple, "Des abris | côtiers↑ ouverts au public↓ * , peuvent renvoyer à deux organisations syntaxiques dérivées différentes :

1- $SN^{(GN \mid GA)_{SN}\uparrow SA\downarrow}$

2- $SN^{(\ GN \mid GA)\uparrow GA)_{SN}\downarrow}$

La possibilité de cette double interprétation est évidemment conditionnée par la nature de la présupposition. Dans le premier exemple, la phrase répond à la question : De quels abris côtiers s'agit-il ? et "ouverts au public" est un syntagme adjectival à valeur attributive. Dans le deuxième exemple, la phrase répond à la question : Qu'est-ce que c'est ?. Dans ce cas, "ouverts au public" est un groupe adjectival inclus dans le syntagme nominal. L'intonation est, semble-t-il, liée plus directement à la structure énonciative, c'est-à-dire à l'organisation en thème et rhème (le donné et l'apport d'information) ; la structure énonciative à laquelle on a d'abord accès devrait ensuite être mise en relation avec la syntaxe. Sans ce détour, il sera difficile de trouver une correspondance constante entre la hiérarchie syntaxique et la hiérarchie intonative. Dans l'exemple "Ce sont des abris ↑côtiers ↓", l'intonème progrédient majeur /↑/ ne sépare pas deux constituants de même niveau, puisque le deuxième (GA) est un élément du SN inclus dans SV ; pire, /↑/ sépare un GA, constituant complet, d'une suite GV + GN, c'est-à-dire d'un constituant mal formé et indéfinissable. En revanche, le recours à la structure énonciative, permet de mettre en évidence la présence de deux constituants énonciatifs bien formés, le thème (ou support) et le rhème (ou apport d'information) séparés par une frontière majeure. C'est grâce à ce maillon intermédiaire, penson-nous, qu'il sera possible d'établir les liens complexes entre énonciation, syntaxe et sémantique.

Nous présentons un modèle de détection automatique des frontières intonatives et syntaxiques qui tient compte des critiques et des réflexions précédentes et des apports de nos recherches récentes. Dans une première partie nous tentons de définir des règles d'énonciation que nous mettons en correspondance avec la syntaxe ; dans une deuxième partie, nous exposons l'algorithme d'interprétation prosodique des données ; en conclusion, nous fournirons des exemples d'analyse intonative et syntaxique.

2.- PRESENTATION DU MODELE

Nous envisageons trois types de règles :
1) des règles génératives d'énonciation,
2) des règles transformatives d'énonciation,
3) des règles de correspondance énonciation-syntaxe.

---

*/↑/ = intonème progrédient majeur ; /|/ = intonème progrédient mineur ; /↓/ = intonème terminal. Pour signification des abréviations contenues dans les règles, voir, en appendice, le Lexique des abréviations.

1 - Règles génératives d'énonciation :

Ce sont des règles de réécriture récursives, indépendantes du contexte : (*) :

Rg.1 PE→ (T1 ↑) R1↓

Rg.2 T1→T2 | R2

Rg.3 R1→R1 ↑ R1n

Rg.4 R1→T2 ↑ R2

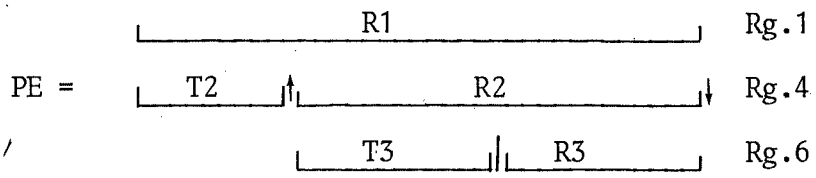
Rg.5 T2→T3 (|) R3

Rg.6 R2→T3 | R3

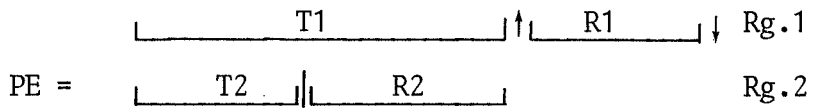
Rg.7 XX→X ^{v} X, v = et, ou.

Dans Rg.7, la disjonction "v" signifie que deux constituants adjacents de même nature et de même niveau sont coordonnés par et, ou. La reécriture de PE est conditionnée par la présupposition, c'est-à-dire le contenu de la question sous-jacente.

Exemple : (Que fait-il ?) Il dévore des poissons multicolores



(Quels poissons dévore-t-il ?) Il dévore des poissons multicolores



2 - Règles transformatives d'énonciation :

Les règles transformatives (Rt) comprennent a) des règles sémantiques (intono-sémantiques) : les règles de rhématisation ($R^{T}RH$) et d'emphase ($R^{T}EMPH$) ; b) des règles d'ajustement prosodique (intonotactiques).

On ne donne ici que des exemples de ce que peut être le résultat de cette catégorie de règles. Cette partie, pour être représentative de la compétence et de la performance des locuteurs, doit être complétée par des recherches ultérieures. Mais on trouvera d'autres exemples de règles intonotactiques dans (DI CRISTO, 1975) et dans (ROSSI, 1979)

---

* PE = phrase énonciative, marquée par la présence d'un intonème terminal ; T1 = thème de niveau 1 ; R1 = rhème de niveau 1 ; T2, R2 = respectivement thème et rhème de niveau 2, c'est-à-dire d'un rang inférieur au niveau 1. Les parenthèses enserrent un élément facultatif. Sauf spécification contraire, on utilise le formalisme de la grammaire de CHOMSKY.

(règles de dominance et de proximité). La $R^{T}RH$ entraîne une permutation qui place le rhème en tête de la phrase énonciative et transforme l'intonème progrédient du thème en une intonation parenthétique (⟷):

Ex. Il vient (T1) ↑ demain (R1) ↓ → demain (R1) ↓ il vient (T1) ⟷

La transformation d'emphase provoque généralement le relèvement d'un ou plusieurs niveaux; pour l'intonème terminal, le niveau de réalisation est le plus souvent le suraigu (LEON, 1971).

Ex.

Il dévore | des poissons ↑ multicolores ↓ → Il dévore | des poissons ↑ multicolores^{SA} ↓

3- Règles de correspondance énonciation-syntaxe (Rs):

Rs.1 T1 → (SN, SV, ADv)

Rs.2 R1 → αR1 /α R1—

Rs.3 R1 → αSV / SN (-α X), ADv (-α X), Ø—

Rs.4 R1 → αADv /SV (-α X), Ø—

Rs.5 R1 → αSN /SV (-α X), Ø—

Rs.6 R1 → ά^{SA} /SN (-α X), SV (-α X), Ø—

Rs.7 T2,3→ (GN, SN, GA, GV)

Rs.8 R2,3→ (GA/SN (X), GN (X)—

Rs.9 R2,3→ SN/GV —

La Rs.2 signifie que les suites de R1 représentent des constituants syntaxiques de même nature. (-α X) indique que toute réécriture d'un constituant est bloquée par un constituant adjacent de même nature; la présence de (X) dans Rs.8 indique que cette contrainte n'existe pas dans ce contexte.

Ex.

(Quels poissons dévore-t-il?)

Il dévore des poissons multicolores

Rg.1 T1 — R1

Rg.2 T2 — R2

Rs.1 SV

Rs.6 SA

Rs.7 GV

Rs.9 SN

Structure énonciative: $_{T1}$ (T2 | R2) $_{T1}$ ↑ R1 ↓

Structure syntaxique : $_{SV}$ (GV | SN) $_{SV}$ ↑ SA ↓ (*)

Comme on le voit incidemment par cet exemple, le recours à l'analyse énonciative permet de résoudre une des difficultées que nous avions évoquée dans l'introduction: la présupposition confère à R1 une fonction attributive; de ce fait, le constituant (SV) qui précède le syntagme adjectival (SA) est complet et / ↑ / sépare deux constituants bien formés.

3.- ALGORITHME D'INTERPRETATION PROSODIQUE

Cet algorithme comprend deux modules. Le premier est destiné à normaliser les valeurs de Fo et à identifier les configurations objectives de Fo(Figure 1). La normalisation vise à effacer partiellement les effets microprosodiques aux frontières des voyelles et à lisser la courbe mélodique; A la suite de ces deux opérations, on calcule les configurations de Fo: tons montants, tons descendants, tons complexes (convexes ou concaves).

Dans le deuxième module, on cherche à styliser les paramètres et à détecter les frontières intonatives et leur hiérarchie. La logique de ce module repose sur deux principes fondamentaux:

a) Le caractère pluriparamétrique de l'intonation;

b) La nécessité d'une transformation perceptive des données objectives, afin de retrouver l'invariant qui structure l'énoncé.

COLLIER (1974) critique à juste titre la plupart des études sur l'intonation dont les auteurs se contentent d'un relevé des mesures objectives et de leur interprétation directe. D'autre part, les modèles de reconnaissance automatique de la parole qui prennent en compte la prosodie se fondent exclusivement sur des données acoustiques. Or les paramètres prosodiques (Fo, intensité, durée), soumis qu'ils sont à des variations de tous ordres (contextuelles, intrinsèques, idiosyncrasiques, etc.) n'entretiennent pas des correspondances biunivoques avec les unités syntaxiques ou énonciatives qui structurent la phrase. Le continuum prosodique est interprété, chez l'auditeur, par des mécanismes de perception qui opèrent une sorte de filtrage pour retrouver dans le signal les traits liés aux instructions linguistiques.

Il est donc nécessaire, pensons-nous, de simuler ces processus de décodage, non seulement dans la perspective d'une analyse linguistique, mais aussi dans celle de la reconnaissance automatique. Les echecs relatifs de l'application de la prosodie à la détection des frontières linguistiques (LEA, 1973) sont dus, nous semble-t-il, à l'inadéquation du modèle de la performance,représenté par l'analyse objective, à la reconnaissance des unités syntaxiques, liées à la compétence du locuteur.

L'analyse objective, qui constitue la première étape, doit être pluriparamétrique et faire intervenir les différents traits qui organisent la mélodie (relief, pente, niveaux, configurations), la durée et l'intensité.

La présence d'un relief mélodique et d'un glissando n'indique

---

(*) Nous considérons avec KAYNE (1977) que le pronom personnel est un élément du syntagme verbal (SV).

pas toujours une limite de groupe intonatif; ces deux traits peuvent être liés à la mécanique respiratoire ou à la présence d'un accent expressif. Dans ce cas précis, le relief mélodique se réalise généralement sur la première syllabe du mot et conserve, de ce fait, une fonction démarcative, mais à un niveau inférieur.

Si,en revanche, le relief mélodique et le glissando sont accompagnés d'une augmentation de durée, ils constituent alors deux traits qui permettent d'identifier l'intonème progrédient et qui assument une fonction démarcative à un niveau supérieur, celui du syntagme.

Les corrections apportées aux données objectives, pour simuler le filtrage opéré par l'auditeur sont de deux ordres:

a) Les corrections contextuelles (microprosodie co-intrinsèque, effets du voisement sur Fo et la durée, par exemple: DI CRISTO, 1978).

b) Les corrections perceptives en fonction des caractéristiques spécifiques (DI CRISTO, 1978; ROSSI, 1971a), du seuil de glissando (ROSSI, 1971b; 1978), de l'intégration temporelle de l'intensité (ROSSI, 1970) et de l'interaction de Fo et de l'intensité (ROSSI, 1979b). L'ordre de ces ajustements n'est pas indifférent, il est lié aux hypothèses que nous pouvons faire sur la hiérarchie des mécanismes de perception.

Le voisement avant la voyelle a pour effet d'abaisser Fo; dans ce contexte, la variation mélodique est importante et perceptible (DI CRISTO, 1978). Il est indéniable qu'elle joue un rôle dans le décodage du trait de voisement (MASSARO and COHEN, 1976), mais elle n'a pas de fonction prosodique. Autrement dit, l'auditeur opère dans ce cas un filtrage au niveau central et pour l'interprétation prosodique ignore ce glissando contextuellement provoqué. Il convient donc, pour que le glissando calculé ait une valeur intonative, de corriger en priorité l'effet du voisement. Le degré de la correction apportée varie selon que la mélodie sur la partie non contiguë à la voyelle est statique ou montante(DI CRISTO, 1978).

La fréquence fondamentale intrinsèque est liée à une contrainte articulatoire; de ce fait, elle est vraisemblablement corrigée selon un processus d'analysé par synthèse (LIEBERMAN, 1967). On peut penser que la variation mélodique détectée à la périphérie du système auditif est réinterprétée au plus haut niveau. La correction en fonction de Fo intrinsèque doit donc précéder le calcul du glissando, si on tient à identifier la fonction intonative de ce dernier. Toutefois, pour Fo, l'ordre des corrections intrinsèques et co-intrinsèques n'est pas décisif.

En ce qui concerne la durée, l'ordre des corrections intrinsèques et co-intrinsèques est impératif. Contrairement à ce qu'on a pu penser, la sourdité de la consonne n'a pas d'effet de troncation, du moins dans la réalisation de l'intonème progrédient (DI CRISTO, 1980), mais le voisement a un effet allongeant bien connu qui accroît la perceptibilité du glissando. Comme on l'a vu pour Fo, on doit supprimer l'allongement dû au voisement afin de dégager de la variation mélodique la part qui revient à l'intonation et pour éviter d'attribuer à une variation paramétrique conditionnée un rôle fonctionnel qu'elle n'assume pas. En revanche, la correction en fonction de la durée spécifique n'est opérée, comme on peut s'en rendre compte dans l'algorithme (figure 1), qu'après le calcul du glissando. En effet, on a remarqué que la variation intrinsèque de la durée n'a pas d'incidence sur la per-

ceptibilité du glissando (Di CRISTO, 1978, p. 1029); en d'autres termes, intervient un phénomène de compensation qui justifie notre démarche. Pour cette même raison, si on corrige la durée intrinsèque avant le calcul du glissando, on s'aperçoit que le point de hauteur se réalise en dehors de son niveau normal: soit au delà si on allonge la voyelle haute, soit en deçà si on abrège la voyelle basse.

La correction de la durée spécifique doit s'effectuer après le calcul du glissando, pour l'identification correcte des maxima de durée. Mais quelles voyelles corriger? Les voyelles hautes ou les voyelles basses? En l'absence d'une hypothèse sérieuse, nous avons adopté une solution pragmatique. Nous avons remarqué empiriquement que la détection des maxima de durée était plus exacte si on allongeait les voyelles hautes. C'est donc cette solution que nous avons adoptée. Elle présente, en outre, l'avantage d'éviter le recours à une règle "d'incompressibilité" (KLATT, 1976) qui entrainerait inévitablement une complication de l'algorithme.

Parallèlement à cet ajustement, nous avons allongé la durée de la voyelle qui est affectée d'un glissando perceptible. On sait, en effet (DURAND, 1936; LEHISTE, 1976; PISONI, 1976), qu'une variation mélodique donne l'illusion d'une durée supplémentaire. Etant donné, d'autre part, qu'un glissando perceptible à toutes les chances de se réaliser à la fin d'un groupe intonatif, la probabilité de le voir coïncider avec un maximum de durée est ainsi plus grande.

Après avoir évalué l'action des variations d'intensité sur la mélodie et avoir calculé les glissandos:

a) on recherche les maxima locaux de durée;
b) on corrige l'intensité afin d'obtenir des valeurs approchées de la sonie;
c) on calcule les points de hauteur (PH) qui correspondent à la valeur moyenne d'une mélodie statique et au point situé aux 2/3 d'une mélodie montante ou descendante.

Les corrections perceptives et contextuelles ainsi apportées sont nécessairement schématiques. Ce schématisme dérive des contraintes inhérentes à la reconnaissance automatique. Afin de minimiser l'erreur, nous sommes obligés d'opérer avec une information sûre et de nous contenter d'un nombre réduit de traits, en l'occurrence: haut/bas et voisé/non voisé. Il est évident que si nous n'avons pas accès, par exemple, au trait /nasal/, l'évaluation de la durée subjective reste très approximative.

Toutefois, nous possédons maintenant l'essentiel de l'information qui nous permettra, dans la dernière partie de l'algorithme, de détecter et de hiérarchiser les frontières. On recherche les points de hauteur minimum (MIN PH) et maximum (MAX PH). Connaissant MIN PH, on peut placer la frontière terminale et savoir si la phrase est une déclarative neutre ou non. Si MAX PH coïncide avec un maximum local de durée (MAX D) et de sonie (MAX SO), il précède une frontière continuative majeure; s'il ne coïncide pas avec un MAX D, il réalise un ictus mélodique, c'est à dire un sommet mélodique sans valeur accentuelle. L'ictus (') frappe la syllabe initiale d'un mot lexical ou d'un mot phonétique (article + nom, possessif + nom, adjectif + nom, etc.); on le fait donc précéder de la frontière de mot (=).

Par interpolation linéaire entre PH initial et MAX PH, entre MAX PH et MIN PH, on calcule sur chaque voyelle les valeurs HD de la

ligne de déclinaison (notion différente de celle développée par BRECKENRIDGE et LIBERMAN, 1977). Tout PH supérieur à un HD respectif indique un relief de hauteur. S'il coïncide avec un MAX D, il précède une frontière prosodique, sinon il est la manifestation d'un ictus (').
Lorsque l'ictus est suivi, dans une limite définie, d'un MAX D, MAX D précède une frontière intonative de niveau 3 (|3). L'ictus et la frontière de niveau 3 sont les signes démarcatifs d'un mot lexical ou d'un mot phonétique..

Ex.

Les 'ballons | 3 percés ↑ défilaient.↓

Les 'petites filles | 3 de mon frère ↑ sont jolies.↓

Le 'gros chat noir | 3 de mon voisin ↑ s'est enfui.↓

La hiérarchisation des frontières prosodiques est essentiellement fonction du degré de relief des points de hauteur et de la présence d'un maximum de sonie et/ou d'un glissando.

4.- EXEMPLE D'APPLICATION

-Ce sont des conques ombragées

a) Valeurs des paramètres objectifs

	Ce	sont	des	conques	om	bra	gées.
Fo	243	259	238-217	245-286-250	258-222	215-205	168
Durée	54	159	73	136	126	261	183
Int.	60	61	61	59	60	56	51

b) Valeurs des paramètres corrigés à la sortie de l'algorithme

Haut.	243	253	229	279	242	210	168
Durée	54	123	91	136	107	174	229
Sonie	45	74	52	54	50	50	46
	Ce	sont	des	conques	om	bra	gées.↓
		|		↑			

- Interprétation énonciative

Etant donné la structure énonciative, les règles énonciatives fournissent trois hypothèses:

a) PE = $_{T1}$ (T2 | R2) $_{T1}$ ↑ R1↓

b) $PE = {}_{R1}\left({}_{T2}\ (T3\ \ (\ |\ )\ R3)\ {}_{T2}\uparrow R2\right)_{R1}\downarrow$

c) $PE = {}_{R1}\left({}_{R1}\ (T2\ |\ R2\ )\ {}_{R1}\uparrow R1\right)_{R1}\downarrow$

Le choix entre ces trois structures dépend du contenu de la question sous-jacente. Chaque structure énonciative renvoie à plusieurs structures syntaxiques. La connaissance de la présupposition permettrait de choisir entre ces dernières et réduirait ainsi le nombre des hypothèses.

-Hypothèses syntaxiques

1) T1 Présupposé

a) $_{SN}(GN\ |\ GA)_{SN}\uparrow \left\{\begin{matrix}SV\\SA\end{matrix}\right\}\downarrow$

Exemples de phrases

Les hommes |blancs ↑ bouffent ↓

Des abris | côtiers ↑ ouverts au public ↓

T1 R1

b) $_{SV}(GV\ |\ SN)_{SV}\uparrow \left\{\begin{matrix}SA\\ADv\end{matrix}\right\}\downarrow$

Exemples de phrases

Il mange |des poissons ↑ multicolores ↓

Il mange |des poissons ↑ le vendredi ↓

c) $_{ADv}(GN\ |\ GA\ )_{ADv}\uparrow SV\downarrow$

Exemple de phrase

Le vendredi |après midi ↑ il travaille ↓

2) Aucune présupposition; pas de coordination.

a) $_{SN}$ (GN | GA ↑ GA)$_{SN}$↓ De grands | enfants ↑ bien sages ↓

R1

b) $_{SV}$ (GV | GN ↑ GA)$_{SV}$↓ Ce sont | des abris ↑ côtiers ↓

3) Aucune présupposition; coordination

a) $_{SN}\left(_{SN}\ (GN\ |GA\ )_{SN}\ \uparrow\ ^{v}\ SN\right)_{SN}\downarrow$

Exemple de phrase

Des structures | compliquées ↑ et des phrases ↓

R1 R1

b) $_{SV}\left(_{SV}\ (GV\ |SN\ )_{SV}\ \uparrow\ ^{v}\ SV\ (SN)_{SV}\right)_{SV}\downarrow$

Exemple de phrase

Il faut | des structures ↑ et des phrases ↓

c) $_{ADv}\left(_{ADv}\ (GN\ |\ GA)_{ADv}\ \uparrow\ ^{v}\ ADv\right)_{ADv}\downarrow$

Exemple de phrase

Au temps | des tzars ↑ et après ↓

5.- CONCLUSION

L'algorithme que nous présentons vise essentiellement à neutraliser les effets intrinsèques et co-intrinsèques des segments phonétiques sur les paramètres prosodiques, à convertir les données objectives en valeurs perceptives et à rechercher l'invariant qui structure l'énoncé. La détection des frontières des groupes intonatifs nécessite l'application préliminaire d'une méthode de stylisation prosodique (DI CRISTO et al, 1979; NISHINUMA et ROSSI, 1979). Cette démarche est justifiée si l'on admet que c'est précisément sur des formes stylisées que l'auditeur opère pour identifier les indices et les traits et pour reconnaître les structures.

Cet algorithme est donc fondé sur la connaissance des règles transformatives, des règles phonotactiques, en particulier, dont il

prend la réciproque.

Les règles phonotactiques d'ajustement perceptif utilisées n'ont pas été présentées dans le cadre des règles transformatives dont nous n'avons proposé qu'un échantillon provisoire. Les règles phonotactiques sont des règles qui relèvent soit de la compétence soit de la performance du sujet.Nous n'avons exploité ici que celles qui relèvent de la performance (règles d'ajustement contextuel ou perceptif liées à des contraintes chez le locuteur et l'auditeur) ; les règles du premier type sont celles que le locuteur-auditeur met en oeuvre pour compenser, par exemple, la variation involontaire d'un paramètre (DI CRISTO, 1980). Cette dernière classe de règles revêt une importance de premier plan et devrait être introduite dans l'algorithme définitif.

Du point de vue de la reconnaissance automatique l'algorithme présente un avantage certain par rapport au précédent car il ne suppose connue que l'information relative aux limites des voyelles, et aux traits /haut /, /bas /, /+ voisé /et /+ silence /.

En ce qui concerne la sortie pour l'interprétation énonciative et syntaxique, il produit, grâce à l'analyse pluriparamétrique, 4 niveaux de frontières. Il est peu probable qu'ils soient tous pertinents : il semble que les 3 intonèmes /↑/, /|/ et /|3/ suffisent à rendre compte de la hiérarchie intonative. (/|3/ est un intonème progrédient mineur de rang inférieur, inversé), vraisemblablement parce qu'il est difficile de percevoir plus de deux niveaux de relief différents réalisés par une montée.

La technique de détection de cet intonème progrédient descendant par l'ictus et le maximum local de durée est une méthode originale en relation avec l'une des règles phonétiques qui spécifient la réalisation de TERM + (voir règles transformatives).

L'algorithme de l'interprétation prosodique fournit par conséquent des frontières intonatives hiérarchisées. Ensuite les règles énonciatives permettent d'inférer des hypothèses sur l'organisation des groupes de sens de l'énoncé.

Dans un système de reconnaissance où les questions posées par la machine ont un contenu et une structure connus, le nombre des hypothèses syntaxiques dérivées des règles de correspondance est plus réduit.

On pourrait penser que les règles énonciatives introduisent un niveau inutile qui multiplie indûment les structures syntaxiques. En effet, apparemment entre les phrases "Des abris côtiers ouverts au public" et "De grands enfants bien sages", il n'existe par de différence syntaxique. En réalité, l'organisation énonciative se projette sur la structure syntaxique, permettant ainsi de situer dans la phrase le support et l'apport d'information. La conjonction de ces 2 types de règles conduit à une connaissance plus riche de la chaîne puisqu'elle donne accès non seulement à l'organisation syntaxique mais aussi à une partie appréciable du contenu sémantique.

LEXIQUE DES ABREVIATIONS

PE	: Phrase énonciative
T1	: Thème de niveau 1
T2	: Thème de niveau 2
T3	: Thème de niveau 3
R1	: Rhème de niveau 1
R2	: Rhème de niveau 2
R3	: Rhème de niveau 3
R1n	: Rhème de niveau 1, n définissant le nombre
SA	: Syntagme adjectival
SN	: Syntagme nominal
SV	: Syntagme verbal
GA	: Groupe adjectival
GN	: Groupe nominal
GV	: Groupe verbal
Adj	: Adjectif
Adv	: Adverbe
Rs	: Règles de correspondance intonation-syntaxe
Rt	: Règles transformatives d'énonciation
TRM	: Transformation de rhématixtion
TEMPH	: Transformation d'emphase
PER	: Période
PER 1	: Première période
PER fin	: Dernière période
Pi	: Période(s) initiale (s)
Pf	: Période(s) finale(s)
v	: valeur de
d	: pente ou dérivée de Fo
INTEM	: sous-programme de traitement des phrases interrogatives et emphatiques.
↓	: intonème terminal
↑ 1	: intonème progrédient majeur de niveau 1
| 1	: intonème progrédient mineur de niveau 1
V	: Voyelle
V + Haute	: Voyelle haute ou fermée

V + Basse : Voyelle basse ou ouverte
V INI : Voyelle initiale
V FIN : Voyelle finale
MAX : Valeur maximale
MIN : Valeur minimale
MAX.PH : Valeur maximale de hauteur
MIN.PH : Valeur minimale de hauteur
MAX.D : Maximum local de durée
MAX.SO : Maximum local de sonie.

REFERENCES BIBLIOGRAPHIQUES

ATKINSON J. (1973) : Aspects of Intonation in speech : Implications from an Experimental Study of fundamental frequency, Ph.D., The University of Connecticut.

BRECKENRIDGE J. and LIBERMAN M. (1977) : The declination effect in perception, Bell Laboratories Publ., mineographed.

BRESNAN J. (1972) : Stress and syntax : a reply ; Language, 48 : 326-34.

COLLIER, R. (1974) : Intonation from a structural viewpoint : a criticism, Linguistics, 129, 5-28.

DI CRISTO A. (1975) : Recherches sur la structuration prosodique de la phrase française, Actes VIe Journées d'Etudes sur la Parole (Toulouse), 1 : 96-116.

DI CRISTO A. (1978) : De la Microprosodie à l'Intonosyntaxe, thèse de Doctorat d'Etat (Université de Provence), 1274 p.

DI CRISTO A. (1980) : Variabilité acoustique et intégration perceptive des cibles prosodiques, Communication soumise pour être présentée aux XIè Journées d'Etudes sur la Parole (Strasbourg).

DI CRISTO A. , ESPESSER R. , NISHINUMA Y. (1979) : Présentation d'une méthode de stylisation prosodique, Communication présentée au IXe Congrès International des Sciences Phonétiques (Copenhague).

DURAND M. (1936) : Voyelles Longues et Voyelles Brèves, Paris, 195 p.

KAYNE, J. (1977): French Syntax, M.I.T. Press.

KLATT, D. (1976): Linguistic uses of segmental duration in English, J.A.S.A, 59 (5): 1208-21.

LEHISTE, I. (1976): Influence of fundamental frequency pattern on the perception of duration, J. of Phonetics, 4: 113-17.

LEON, P. (1971): Essais de Phonostylistique, Stud. Phonetica 4, Didier.

LIEBERMAN, P. (1967): Intonation, Perception and Language, M.I.T. Press.

MARTIN, P. (1977): Syntax and Intonation: An Integrated Theory, Toronto Semiotic Circle, Pub. nr 2.

MASSARO, D. and COHEN, H.M. (1976): The contribution of fundamental frequency and voice onset time to the /zi/ - /si/ distinction, J.A.S.A., 60 (3): 704-17.

NISHINUMA, Y. and ROSSI, M. (1979): Essai d'automatisation de l'analyse prosodique du français, Communication Présentée au IX e Congrès des Sciences Phonétiques (Copenhague).

PISONI, D. (1976): Fundamental frequency and perceived vowel duration: Res. On Speech Percept. Prog. Rept 3: 145-54.

RIETVELD, A.C. and BOVES, L. (1979): Automatic detection of prominence in Dutch Language, Proc. Inst. Phon. Catholic Univ. Nijmegus, 3: 72-78.

ROSSI, M. (1970): Au sujet des paramètres de l'accent, Actes du VI e Cong. des Sciences Phonétiques (Prague): 779-86.

ROSSI, M. (1971a): l'Intensité Spécifique des voyelles, Phonetica , 24: 129-61.

ROSSI, M. (1971b): Le seuil de perception des glissandos, Phonetica, 23: 1-33.

ROSSI, M. (1973) l'Intonation prédicative en français dans les phrases transformées par permutation, Linguistics, 103: 64-94.

ROSSI, M. (1978): La perception des glissandos descendants dans les contours prosodiques, Phonetica, 35: 11-40.

ROSSI, M. (1979a): Le français, langue sans accent?, Studia Phonetica 13 (sous presse).

ROSSI, M. (1979b): Interaction of intensity glides and frequency glissando, In Honor to D.B. Fry, Lang. & Speech (sous presse).

ROSSI, M. et DI CRISTO, A. (1977): Propositions pour un modèle d'analyse de l'intonation, Actes VIII e Journées d'Etudes sur la Parole (Aix-en-Provence): 323-29.

STOCKWELL, R.P. (1960): The place of intonation in a generative grammar of English, Language, 36: 360-67.

STOCKWELL, R.P. (1972): The role of intonation: reconsiderations and other considerations, in: Penguin Books: 87-109.

VAISSIERE, J. (1975): On french prosody, M.I.T.- Q.P.R., 114, 212-23.

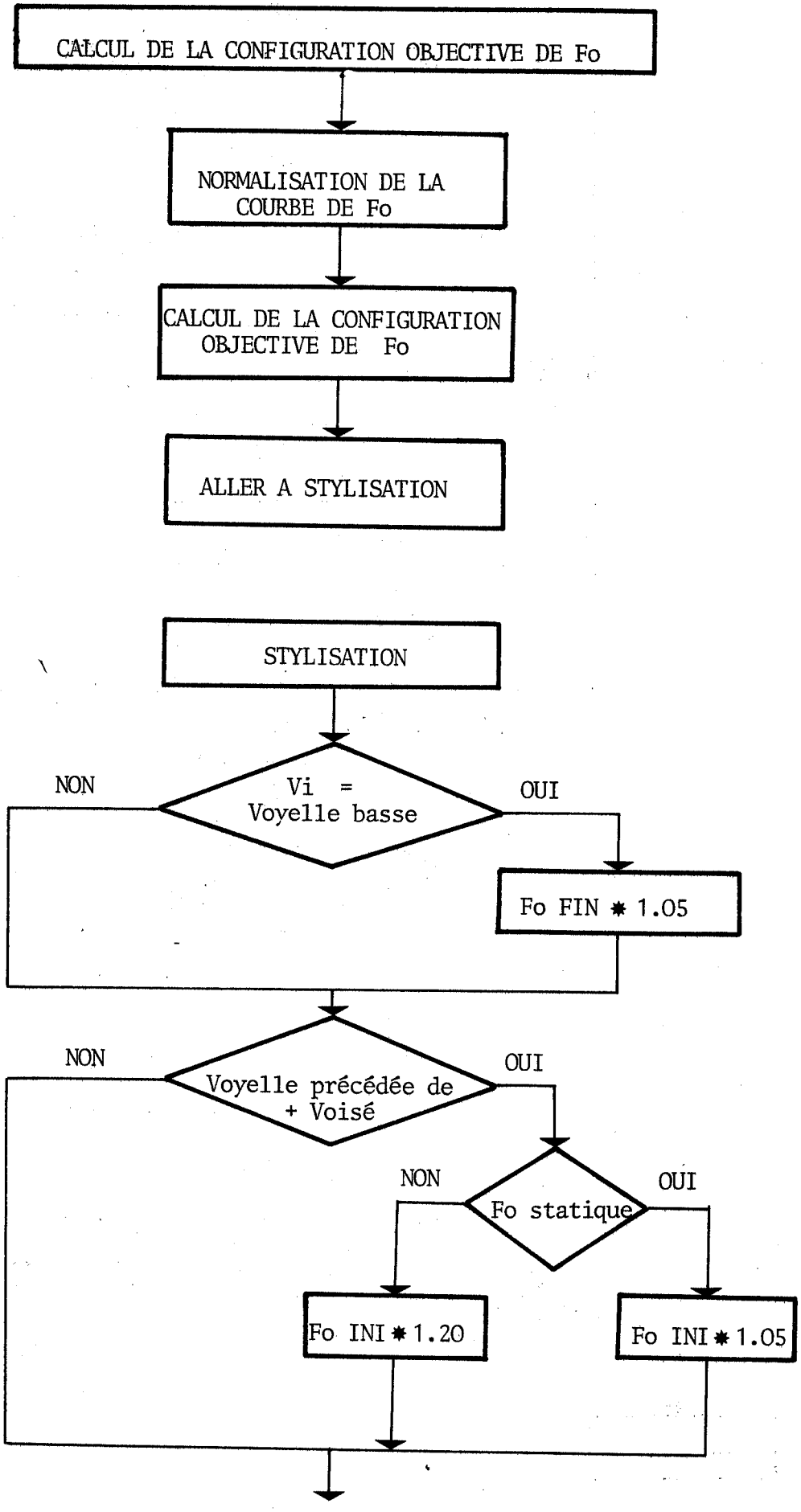
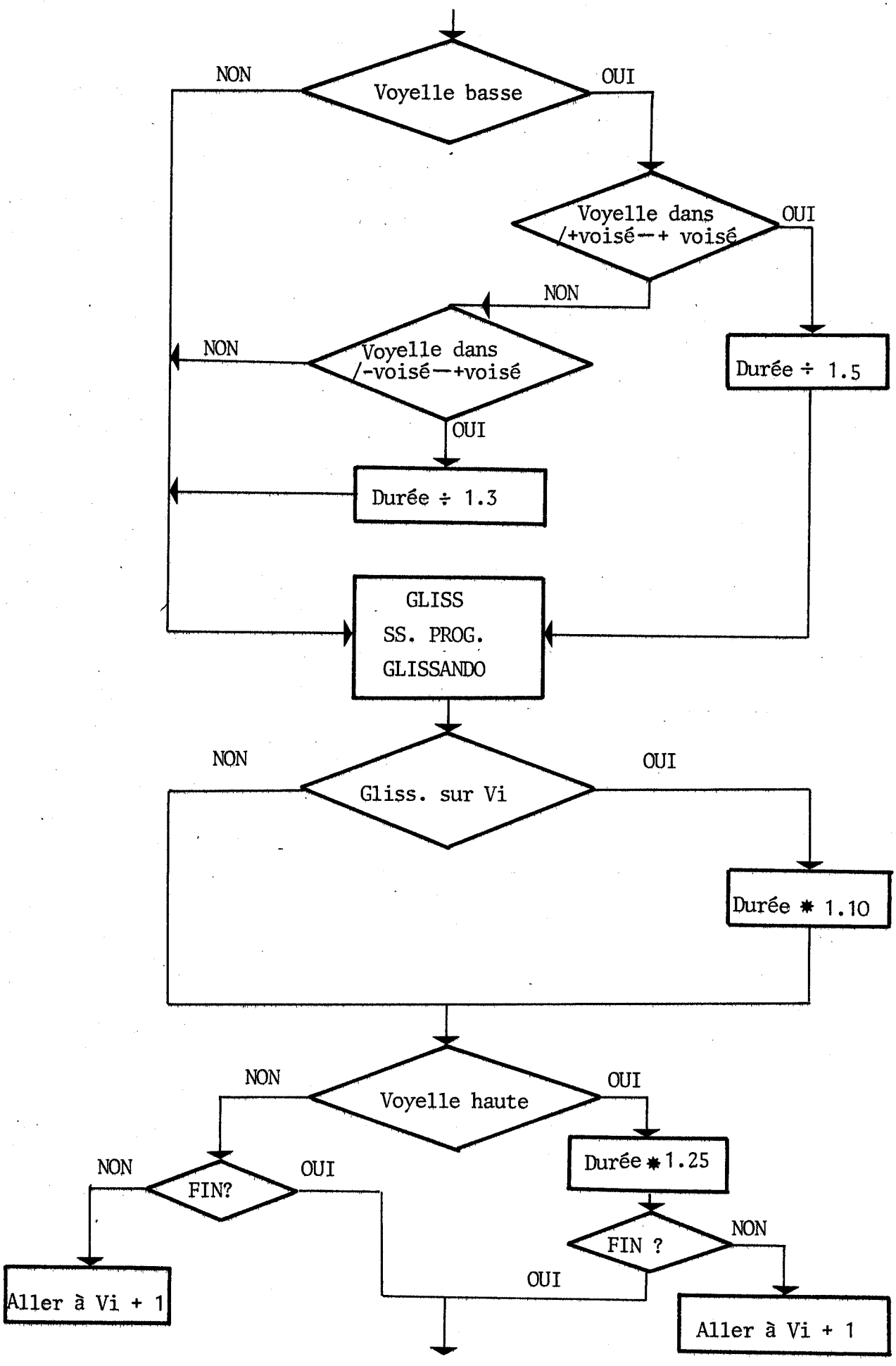
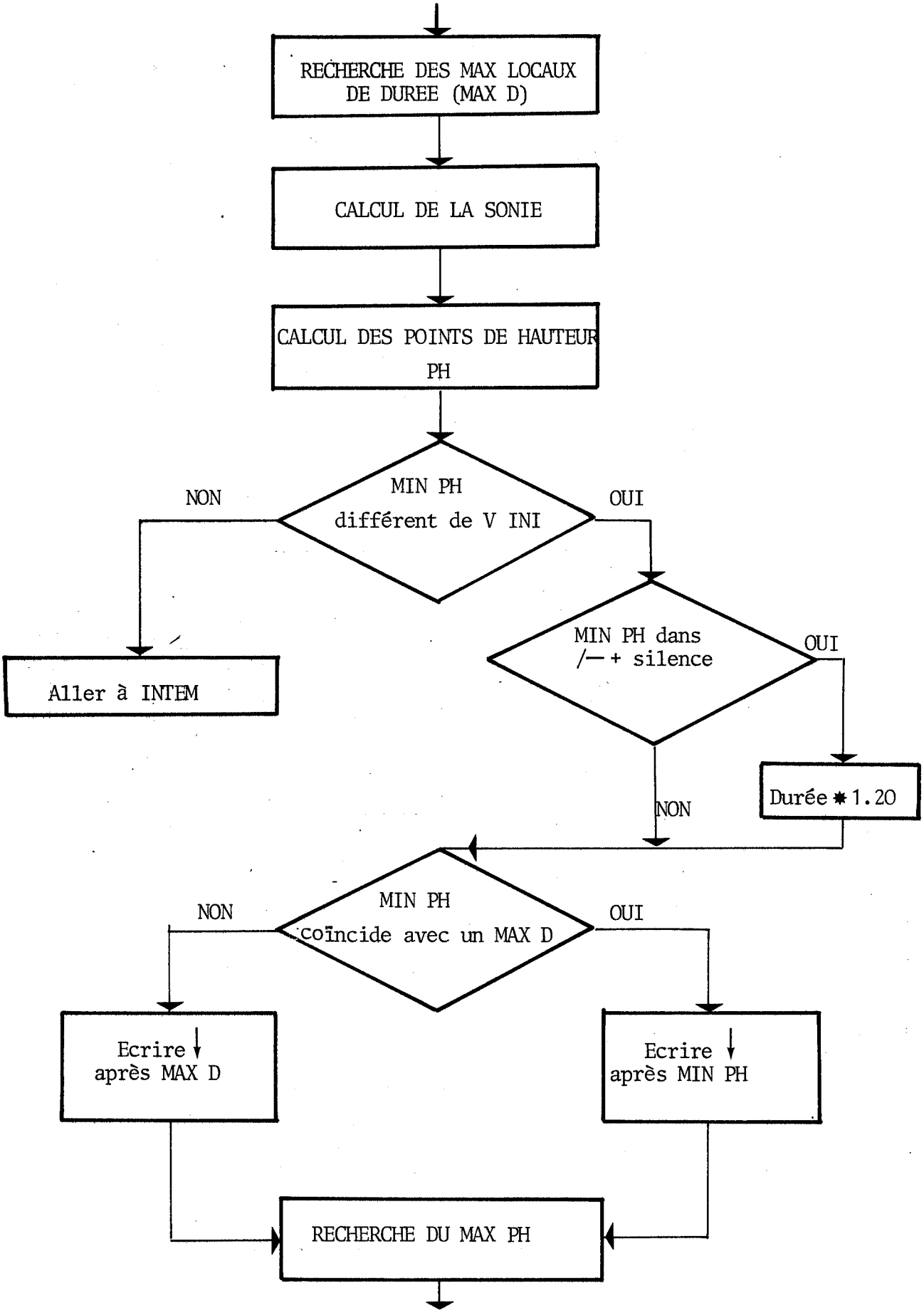
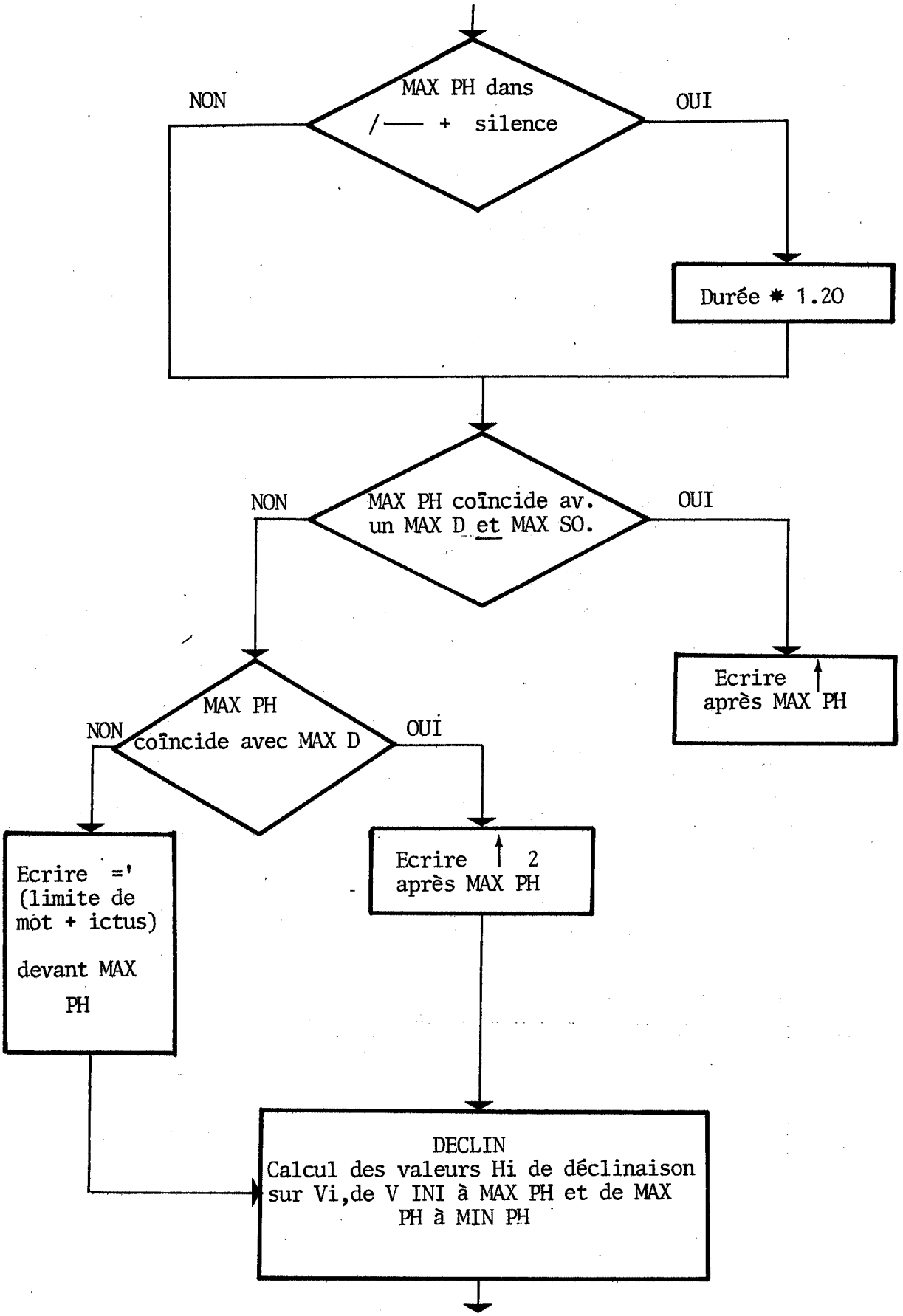
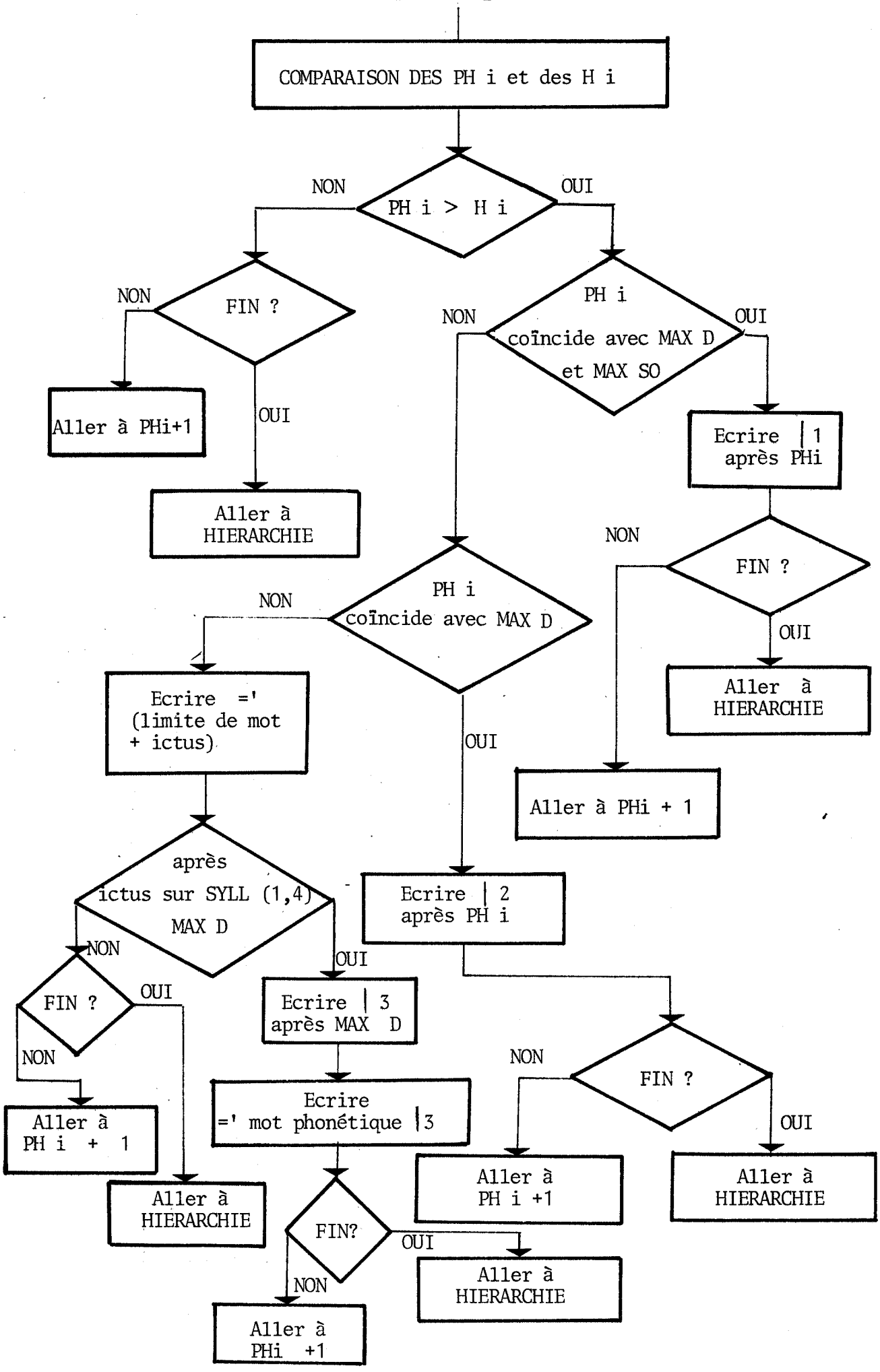


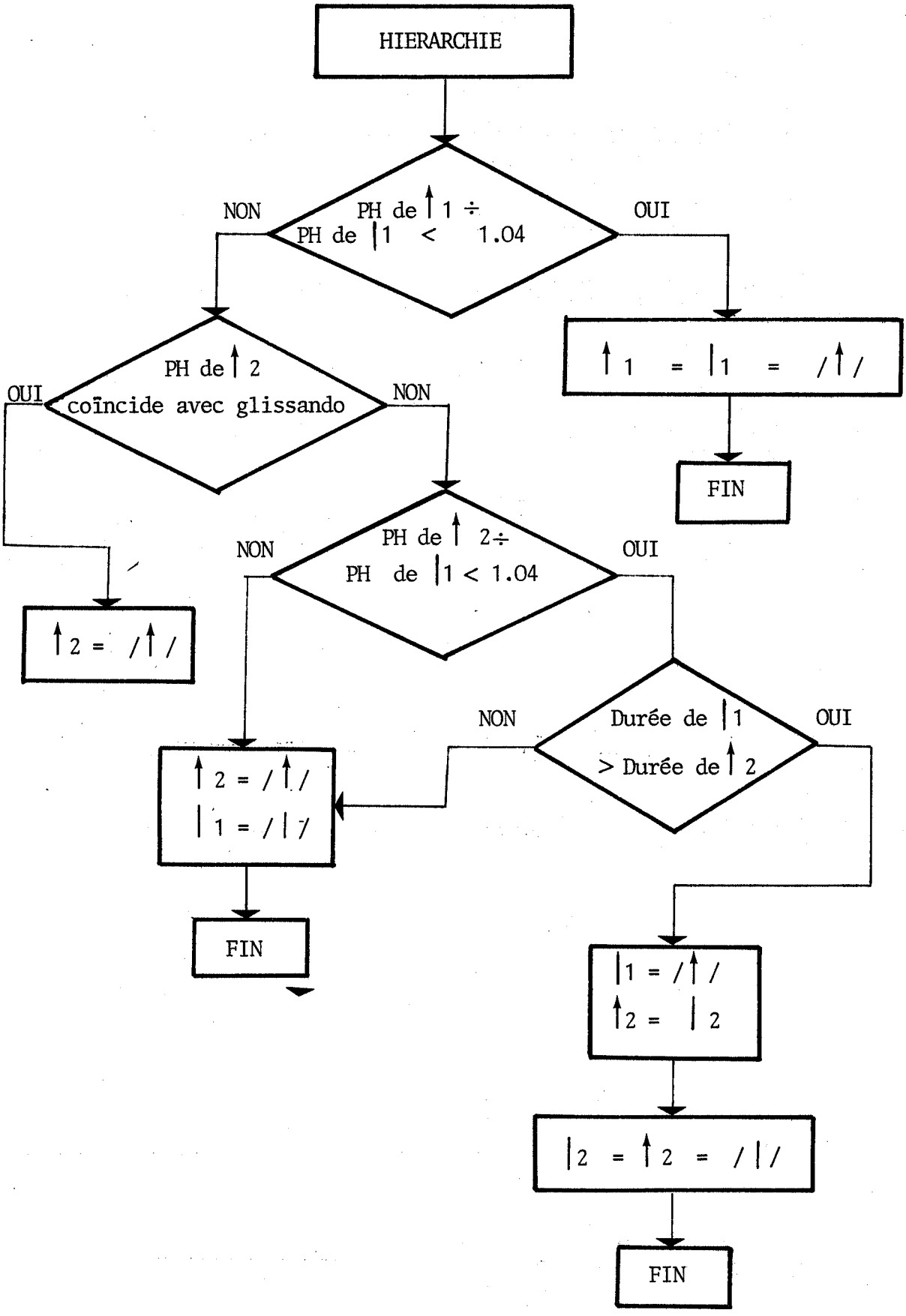
Figure 1. Présentation de l'algorithme de détection des frontières intonatives


Voyelle basse
NON
OUI
Voyelle dans /+voisé—+ voisé
OUI
NON
Voyelle dans /-voisé—+voisé
NON
OUI
Durée ÷ 1.5
Durée ÷ 1.3
GLISS
SS. PROG.
GLISSANDO
Gliss. sur Vi
NON
OUI
Durée * 1.10
Voyelle haute
NON
OUI
Durée * 1.25
FIN?
NON
OUI
FIN ?
NON
OUI
Aller à Vi + 1
Aller à Vi + 1


RECHERCHE DES MAX LOCAUX DE DUREE (MAX D)
CALCUL DE LA SONIE
CALCUL DES POINTS DE HAUTEUR PH
MIN PH différent de V INI
NON
OUI
Aller à INTEM
MIN PH dans /— + silence
OUI
Durée * 1.20
NON
MIN PH coïncide avec un MAX D
NON
OUI
Ecrire ↓ après MAX D
Ecrire ↓ après MIN PH
RECHERCHE DU MAX PH


MAX PH dans
/—— + silence
NON
OUI
Durée * 1.20
MAX PH coïncide av.
un MAX D et MAX SO.
NON
OUI
Ecrire ↑
après MAX PH
MAX PH
coïncide avec MAX D
NON
OUI
Ecrire ='
(limite de
mot + ictus)
devant MAX
PH
Ecrire ↑ 2
après MAX PH
DECLIN
Calcul des valeurs Hi de déclinaison
sur Vi,de V INI à MAX PH et de MAX
PH à MIN PH


COMPARAISON DES PH i et des H i
PH i > H i
NON
OUI
FIN ?
NON
OUI
Aller à PHi+1
Aller à HIERARCHIE
PH i coïncide avec MAX D et MAX SO
NON
OUI
Ecrire | 1 après PHi
FIN ?
NON
OUI
Aller à HIERARCHIE
PH i coïncide avec MAX D
NON
OUI
Ecrire =' (limite de mot + ictus)
Aller à PHi + 1
après ictus sur SYLL (1,4) MAX D
NON
OUI
Ecrire | 2 après PH i
FIN ?
NON
OUI
Ecrire | 3 après MAX D
Aller à PH i + 1
Aller à HIERARCHIE
Ecrire =' mot phonétique | 3
FIN ?
NON
OUI
Aller à PH i +1
Aller à HIERARCHIE
FIN?
OUI
NON
Aller à HIERARCHIE
Aller à PHi +1


HIERARCHIE
NON
PH de ↑1 ÷ PH de |1 < 1.04
OUI
↑1 = |1 = /↑/
FIN
PH de ↑2 coïncide avec glissando
OUI
NON
↑2 = /↑/
NON
PH de ↑2÷ PH de |1 < 1.04
OUI
NON
Durée de |1 > Durée de ↑2
OUI
↑2 = /↑/ |1 = /|/
FIN
|1 = /↑/ ↑2 = |2
|2 = ↑2 = /|/
FIN

XIemes JOURNÉES D'ÉTUDE SUR LA PAROLE

28, 29 et 30 mai 1980

STRASBOURG

APPLICATION D'UN MODÈLE AUDITIF À L'ÉTUDE DES CONFUSIONS DES FRICATIVES NON-VOISÉES

LACERDA* Francisco Paulo ORGANISME CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DE LISBOA
LIEU Av. 5 de Outubro,
85 - 5º/6º 1000 LISBOA
PORTUGAL

(*Cette étude a été realisée au Departement de Phonétique de l'Institut de Linguistique de l'Université de Stockholm, graces à des bourses d'études de la Junta Nacional de Investigação Cientifica e Tecnológica, de l'Instituto Nacional de Investigação Cientifica et de la F.Gulbenkian).

0 - INTRODUCTION

L'objectif à long terme de cette étude est une approche des principes d'organisation des systèmes de consonnes fricatives dans les différentes langues du monde.

L'hypothèse selon laquelle les systèmes de fricatives sont organisés de façon à maximaliser les distances auditives inter-classes est inspirée par les récents travaux de Bjorn Lindblom. Tester cette hypothèse implique un étude topologique de l'espace auditif pour la représentation des sons fricatifs et une estimation des distances auditives en corrélation avec les jugements perceptifs de ressemblance.

L'étude des espaces auditifs pour les voyelles et le test de performance du modèle auditif qui ont été developpés par Liljencrants et Lindblom (1972), Lindblom (1975), Lindblom, Lubker et Pauli (1977) et Bladon et Lindblom (1979) est poursuivie et appliquée ici à l'étude des fricatives non-voisées, extraites du discours naturel. Le sujet peut être appelé un "locuteur de phonétique universelle" ce qui signifie qu'il est capable de produire les fricatives d'accord avec les normes de l'Alphabet Phonétique International".

Bien que l'on puisse contester l'emploi de cette désignation, l'important est que, un ensemble de stimuli dynamiques, contenant toutes les particularités imposés par le locuteur, est ici utilisé pour étudier la corrélation entre la perception de ces stimuli et leur représentation auditive.

Ce rapport rend compte de l'étude préliminaire sur l'espace auditif pour les fricatives. Dans cette étude les stimuli, bien que dynamiques, seront traités par le modèle auditif.

Ces stimuli, variant en fonction du temps, seront dans une phase postérieure traités par des techniques de de groupement et d'estimation de trajectoires dans l'espace-temps des représentations auditives, ce qui nous permettra de nous appercevoir des principaux traits qui caractérisent les fricatives dans l'espace auditif et, à long terme, de mieux comprendre la perception de la parole.

1 - LE MODÈLE AUDITIF

Le modèle auditif utilisé dans cette étude est inspiré par le travail de Zwicker et Scharf (1965) et utilise un certain nombre d'expressions mathématiques proposées par Schroeder, Atal et Hall (1978).

L'information spectrale est obtenue avec un programme en ILS qui calcule des FFT pour des segments de 512 points et donne une représentation sous forme d'un spectre de puissance avec 256 points au long de l'échelle de fréquence. Cette information spectrale est alors traitée par MONA, un programme en basic, qui exécute les pas suivants:

1) La conversion de l'échelle de fréquence en échelle bark.
2) Le calcul de la sortie du filtre auditif.
3) La génération de la représentation en PHONE-BARK.
4) La génération de la représentation SONE/BARK vs BARK.

Pour la conversion de l'échelle de fréquence en échelle bark (1er pas) nous avons utilisé la formule de Schroeder (Schroeder et el, 1978) mais nous l'avons étendue, en introduisant une courbe parabolique de façon à rendre valable la conversion jusqu'à 8500Hz.

Les résultats de cette conversion correspondent de très près à la courbe de Zwicker (Zwicker, 1961).

Pour le calcul de la sortie du filtre auditif,MONA utilise des fenêtres de dispersion en fréquence, avec des pentes de +25 et -10 dB par bande critique (Schroeder et al, 1978), dérivées par Schroeder à partir des travaux de Zwicker sur l'effet de masque d'un bruit de bande étroite (Zwicker, 1958).

La représentation phone vs bark - est dérivé à partir de la sortie du filtre auditif au moyen d'un ensemble de courbes isophoniques selon la recommendation ISO nº226.

La représentation sone/bark utilise le modèle proposé par Zwicker et Scharf (1965) - "model of loudness summation".

2 - MESURE DES DISTANCES AUDITIVES

Selon Plomp (1970) le timbre des sons complexes peut être considéré comme un attribut multi-dimmensionnel et "since loudness of a complex tone can be considered successefully as the sum of the contributions of the different frequency bands (Zwicker and Scharf, 1965) we may suppose similarly that the total difference between two frequency spectra can be considered as the sum of the differences in each band".

Ce concept de distance auditive utilisé par Plomp (1970) pour un certain nombre de sons complexes, dont les voyelles, et appliqué à l'étude de l'espace auditif pour les voyelles par Liljencrants et Lindblom,(1972),Lindblom (1975),Lindblom, Lubker et Pauli(1977)et Lindblom et Bladon (1979)sera maintenant testé pour le cas des fricatives non-voisées extraites du discours naturel.

Les distances auditives utilisées dans ce travail sont des distances de Hamming ("city-block distances") calculées pour chaque paire de représentations auditives générées par le programme MONA.

3 - LES TESTS DE PERCEPTION

Comme nous l'avons déjà mentionné ci-dessus, les stimuli utilisés dans cette étude ne sont pas des stimuli synthétiques, ils ont été extraits d'une liste de logatomes de structure VCV produite par un locuteur.

Les consonnes sont des fricatives de l'ensemble
C = {ɸ, f, θ, s, ʂ, ʃ, ç, x, χ, ħ, h}
produites en contexte vocalique symétrique iCi, uCu, aCa, ce qui donne un total de 33 stimuli différents.

Nous avons élaboré deux différents tests de perception avec ces stimuli:

1) Un test d'identification de fricatives où chaque stimulus est répété 5 fois. Les stimuli ont été présentés par ordre aléatoire, en 11 groupes de 15 stimuli. L'intervalle de temps entre chaque stimulus à l'intérieur d'un groupe est de 5 sec et l'intervalle entre deux groupes de stimuli de 15 sec.

2) Un test d'identification du contexte vocalique qui est en tout égal au premier test sauf en ce qui concerne l'ordre de présentation des stimuli, qui est différente, et l'intervalle entre les stimuli qui a été réduit à 3 sec.

Les feuilles de réponses contenaient l'ensemble de réponses possibles. Les deux tests, de choix forcé, ont été présentés avec des écouteurs calibrés. Nous avions besoin de controler le niveau de pression sonore, pendant la présentation des stimuli pour pouvoir calibrer aussi le programme MONA.

Après la mise au point du niveau d'intensité (la référence étant un signal de 1 KHZ enregistré dans les mêmes conditions des stimuli du test) les sujets du test étaient soumis à une période d'entraînement de façon à apprendre à bien utiliser les symboles phonétiques dans les feuilles de réponses et à minimiser les effets d'une situation de test.

Pendant la période d'entraînement les sujets ont écouté les logatomes VCV (dans un ordre connu de façon à pouvoir mettre en rapport les fricatives et les symboles phonétiques correspondants) et aussi des stimuli utilisés dans le test. Ces derniers cependant, ont été présentés par ordre aléatoire, de façon à eviter l'apprentissage. Tous les stimuli ont été présentés dans les mêmes conditions du test.

Le nombre de fricatives utilisés comme stimuli étant supérieur au nombre de fricatives existantes dans les langues maternelles des sujets il était possible de prévoir des différences dans l'identification des sons avec ou sans signification linguistique.

Il est bien connu que le processement des sons avec et sans signification linguistique est fondamentalement différent. Cependant, en ce qui concerne le niveau auditif le processement doit être le même malgré la nature des sons (Stevens and House, 1972). Ce n'est qu'après le processement periphérique que les stratégies linguistiques doivent être tenues en ligne de compte. Une fois qu'il est bien probable que les fricatives isolées ne soient pas considérées comme linguistiquement significatives, il nous semble que

les résultats des tests de perception seront valables en ce qui concerne le niveau auditif, le seul qui nous concerne ici.

4 - RÉSULTATS DES TESTS DE PERCEPTION

Les tests de perception ont été appliqués à 6 sujets suédois, 2 anglais, 2 finlandais et un allemand.

Les résultats auraient dû être organisés en fonction de la langue maternelle des sujets de façon à vérifier l'influence du système de fricatives existant dans la langue des sujets sur la distribution des réponses.

Cependant, comme le nombre de sujets finlandais, anglais et allemands est trop petit pour que l'on puisse considérer les résultats comme significatifs, nous présentons seulement une matrice de confusion pour les sujets suédois et une autre pour l'ensemble des sujets.

Les matrices de confusion présentées ici ont été dérivées des matrices originales par la méthode de symétrisation de Klein, Plomp et Pols (1970). Cette méthode utilise un concept de ressemblance selon lequel, deux stimuli sont d'autant plus semblables que la distribution des réponses est semblable.

1) Le test d'identification de fricatives

Les tableaux 1 à 3 présentent les matrices de confusion des fricatives independemment du contexte vocalique pour les sujets suédois seulement et en contexte vocalique de [i] pour les sujets suédois et pour l'ensemble des sujets.

	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	
ɸ												ɸ
f	44											f
θ	48	60										θ
s	7	4	16									s
ʂ	8	4	12	29								ʂ
ʃ	4	0	11	28	60							ʃ
ç	8	1	11	22	36	72						ç
×	26	4	9	11	20	19	18					×
χ	19	7	10	13	21	19	18	77				χ
ħ	18	7	6	2	6	4	3	43	54			ħ
h	21	10	9	4	8	4	3	42	51	73		h
	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	

Tableau 1 - Indices de ressemblance (%) pour les sujets suédois, sans discrimination du contexte vocalique.

Les cases de ces matrices contiennent les valeurs en pourcentages des indices de ressemblance obtenus par:

$$Sij = \frac{Sij}{Sii} \times 100\%$$

où i représente le stimulus présenté et j une réponse particulière.

	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	
♠												ɸ
f	47											f
θ	73	53										θ
s	10	3	10									s
ʂ	7	3	7	28								ʂ
ʃ	0	0	0	17	13							ʃ
ç	0	0	0	24	20	90						ç
×	7	3	3	28	27	33	33					×
χ	10	0	7	24	50	13	17	50				χ
ħ	7	0	3	3	7	3	0	10	30			ħ
h	3	0	0	0	7	0	0	20	30	72		h
	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	

Tableau 2 - Indices de ressemblance (%) en contexte vocalique de [i] pour les sujets suédois.

	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	
ɸ												ɸ
f	51											f
θ	60	60										θ
s	16	11	13									s
ʂ	13	5	6	31								ʂ
ʃ	5	2	2	30	44							ʃ
ç	5	2	2	26	44	93						ç
×	20	9	9	26	31	39	38					×
χ	25	13	17	22	38	24	27	65				χ
ħ	13	5	6	6	11	0	0	36	47			ħ
h	11	4	4	4	9	0	0	33	38	78		h
	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	

Tableau 3 - Indices de ressemblance (%) en contexte vocalique de [i] pour l'ensemble des sujets.

2) Discrimination du contexte vocalique

Avec ce test, qui présente toujours des fricatives, nous prétendons vérifier si les sujets peuvent trouver les voyelles adjacentes à partir des variations spectrales dues aux effets de coarticulation.

Une fois que toutes les voyelles utilisées comme contextes des fricatives, appartiennent aux langues maternelles des sujets il nous semble raisonnable d'additionner toutes les matrices de confusion et d'analyser les résultats globaux.

Ce test n'est pas un test de reconnaissance. Il s'agît de choisir à partir d'un son donné, le contexte vocalique le plus probable, ce qui implique, une sorte de stratégie d'estimation de la distance auditive la plus courte.

Nous n'établissons pas ici de rapport entre les résultats de ce test de perception et l'analyse auditive, ce qui nous semble prématuré, mais nous pouvons cependant essayer de poser des hypothèses d'explication de nos résultats.

Il est évident que dans le cas de [h] il y a un bon score d'identifications correctes du contexte vocalique.Ceci peut être expliqué par le fait que le bruit d'aspiration excite fondamentalement la même configuration du conduit vocal que les voyeles adjacentes, ce qui résulte, au premier abord dans une même enveloppe spectrale pour la fricative et pour les voyelles adjacentes.

Une observation des spectrogrammes des réalisations VhV montre une très faible énergie de F1, pendant la production du [h] surtout en contexte de voyelle haute. Cette différence entre l'enveloppe "idéale" et l'enveloppe observée est due à une plus grande ouverture de la glotte qui, à son tour, est due à un plus grand accouplement des cavités sous-glottales et qui résulte dans une atténuation des basses fréquences. C'est pourquoi, le premier formant des fricatives est pratiquement absent en contexte de [i] et de [u]. C'est une situation semblable à celle de la production des voyelles chuchotées et on peut penser que l'identification de la voyelle est basée sur F2 et F3.

Dans les cas de [ɸ]et [x] nous avons aussi observé un bon score de discrimination. L'analyse spectrographique montre pour ces deux sons de grandes variations spectrales qui dépendent du contexte vocalique.

Cependant, pour toutes les autres fricatives les matrices de confusion suggèrent l'impossibilité d'interpréter les contextes d'après les variations spectrales. L'information fournie par les donnés brutes, montre que la plupart des réponses ont été de [i] quand il s'agissait de fricatives palatales et palato-alvéolaires et de [u] quand il s'agissait de fricatives uvulaires ou pharyngales.

Si nous retournons aux cas de [ɸ] et de [x] , il y a des faits articulatoires qui sont probablemment en relation avec les bons scores de discrimination.

La fricative bilabiale peut être traitée de façon semblable à [h] . Nous pouvons considérer que la langue n'a pas de rôle actif dans leur articulation et que, par consé-

quence, est libre de créer des cavités tout à fait différentes derrière la source sonore, selon les voyelles adjacentes ce qui a comme conséquence des spectres résultants différents.

En ce qui concerne la fricative vélaire, son point d'articulation peut expliquer la sensibilité au contexte vocalique. En effet, des modèles acoustico-articulatoires comme ceux de Fant,(1960), Stevens,(1968),Klatt et Stevens, (1969), et Lindblom et Sundberg,(1971), suggèrent en ce qui concerne la région d'articulation vélaire, d'importantes variations de F2 selon le point d'articulation. Ceci peut expliquer, pourquoi les sujets réussissent aussi bien à détecter la voyelle adjacente.

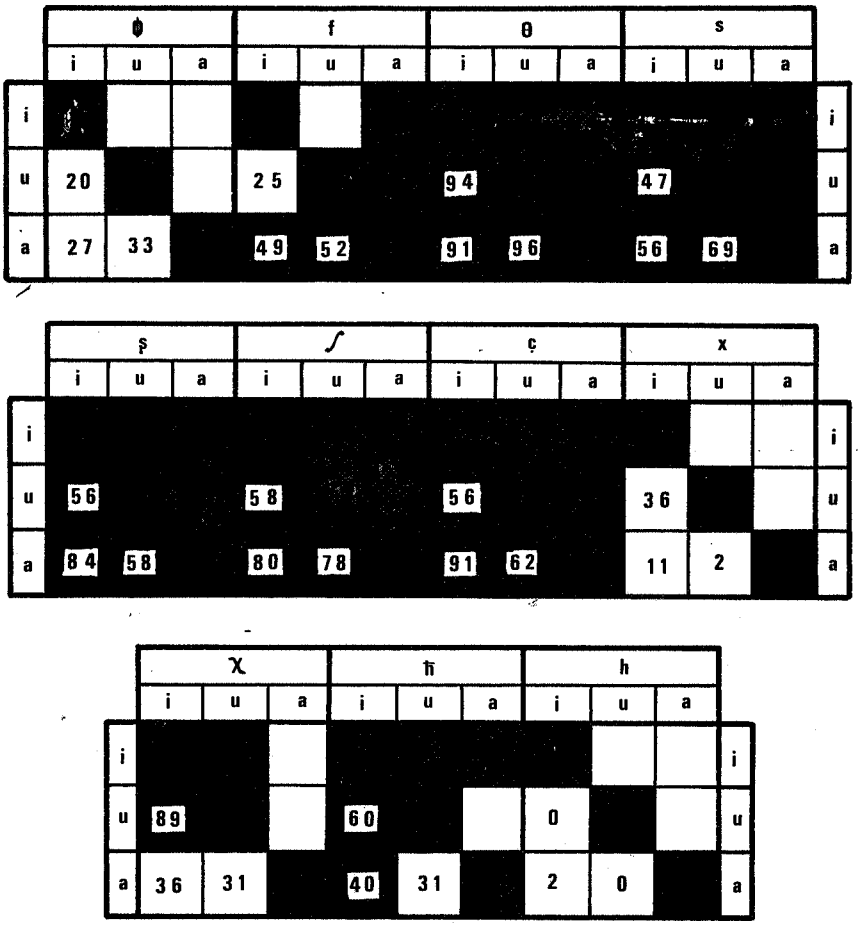
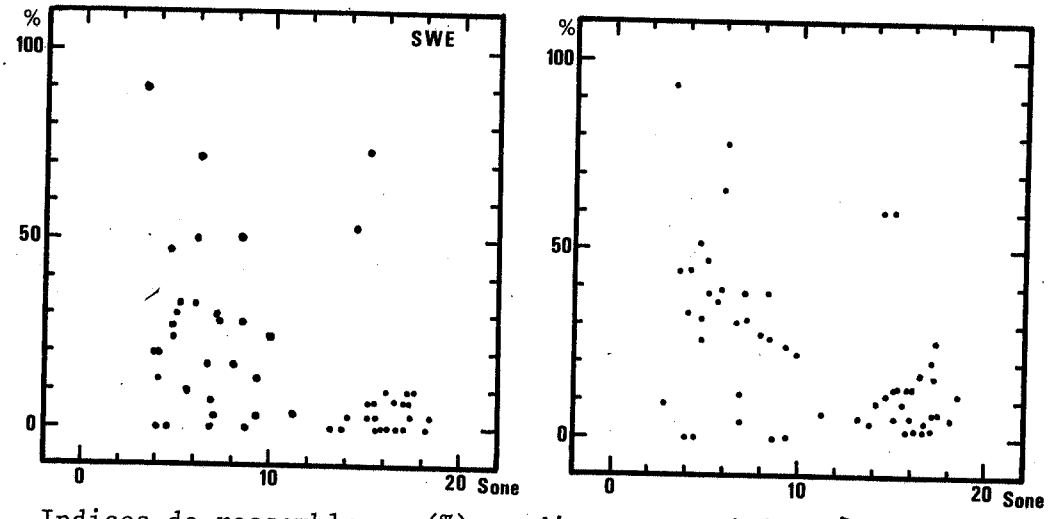


Tableau 3 - Indices de ressemblance (%) des contextes vocaliques des consonnes fricatives pour l'ensemble des sujets. Les cases noires correspondent à des indices plus grands ou égaux à 40%.

5 - LES DISTANCES AUDITIVES ET LES CONFUSIONS DE FRICATIVES

Pour le moment, les relations entre les distances auditives et les confusions des fricatives seront analysées seulement pour le contexte de [i] (avec les limitations auxquelles nous avons déjà fait référence ci-dessus: échantillons de 25ms par représenter les stimuli dynamiques).

Les graphiques suivants représentent l'indice de ressemblance (%) versus la distance auditive correspondante, calculés pour chaque paire de fricatives. Les distances auditives ont été obtenues à partir des représentations sone/ /bark vs bark avec une distance de Hamming(5).



Indices de ressemblance (%) vs distance auditive. À gauche pour les sujets suédois et à droite pour l'ensemble des sujets.

	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	
ɸ	■	■	■									ɸ
f	■	■	■									f
θ	■	■	■									θ
s				■								s
ʂ					■	■	■					ʂ
ʃ					■	■	■					ʃ
ç					■	■	■					ç
×								■			■	×
χ									■	■		χ
ħ									■	■		ħ
h								■			■	h
	ɸ	f	θ	s	ʂ	ʃ	ç	×	χ	ħ	h	

Tableau 4 - Résultats de la technique de groupement de Jarvis et Patrick (Nombre de voisins considéres=4. Nombre de voisins partagés=3)

Ces graphiques correspondent aux réponses des sujets suédois et aux réponses de l'ensemble des sujets.

La corrélation entre les confusions observées et la distance auditive n'est pas aussi claire que celle que l'on peut trouver dans Lindblom et al (1977). En gros, ces graphiques montrent qu'une augmentation des distances auditives est en relation avec une plus petite ressemblance perceptive des fricatives. Cependant ses relations ont besoin d'être étudiées plus en profondeur, notamment en tenant en ligne de compte les aspects dynamiques des stimuli.

Si l'hypothèse, selon laquelle les confusions perceptives sont en rapport inverse avec les distances entre les représentations auditives, est correcte, il y aura des groupements de représentations auditives correspondant aux fricatives souvent confondues.

Dans ce sens, la topologie de l'espace auditif est encore plus importante que les distances entre les représentations auditives.

Nous avons appliqué à l'ensemble des représentations auditives la technique de groupement fondée sur le nombre de voisins les plus proches partagés (Jarvis et Patrick, 1973). Nous pouvons remarquer l'accord entre les résultats de la méthode de groupement et les principales confusions(*) faites par les sujets.

Évidemment que, une fois que ces résultats sont spéculatifs, ils sont très sensibles à des changements des paramêtres de groupement et il est préférable de ne pas pousser plus loin la discussion sans connaître les nouvelles données.

Un jour, un modèle auditif comme celui-ci permettra de clarifier des questions d'organisation et d'acquisition du langage. En effet il nous semble qu'il est hors de question que quelle que soient les stratégies utilisées, elles ne feront que travailler sur les résultats de l'analyse auditive.

Selon la théorie des Quanta de Stevens, les langues tendent à utiliser des sons pour lesquels il n'y a pas de stabilité acoustique critique en fonction de petits changements articulatoires.

Il sera intéressant de vérifier si, en effet, parmi les sons possibles, qui satisfont cette contrainte acoustique-articulatoire les langues choisissent,en effet, ceux qui présentent le maximum de distance dans l'espace auditif.

(*) - Celles qui correspondent à des indices plus grands ou égaux à 40% (voire tableau 3).

Références bibliographiques

FANT, G. - Acoustic Theory of Speech Production ,The Hague, Mouton et Co., 2e. édition, (1970).

HALLE, M. et STEVENS, K. - "Some Reflections on the Theoretical Bases of Phonetics", in B. Lindblom and S. Ohman (eds),Frontiers of Speech Communication Research, Academic Press,Londres,(1979).

HEINZ, J. et STEVENS, K. - "On the properties of voiceless fricative consonants", JASA, Vol. 33, nº 5, 589--596 Mai,(1961).

JARVIS, R. et PATRICK, E. - "Clustering Using a Similarity Measure Based on Shared Near Neighbors", IEEE Transactions on Computers, Vol. C-22, nº 11, Novembre ,(1973).

JASSEM, W. - "Acoustical description of voiceless fricatives in terms of spectral parameters" in Speech Analysis and Synthesis, Warsaw.

KLATT, D. et STEVENS K. - "Pharyngeal consonants", Quarterly Progress Report, nº 93, Research Laboratory of Electronics, M.I.T., pp. 207-216,(1969).

KLEIN, W., PLOMP, R., et POLS, L. - "Vowel spectra, vowel spaces and vowel identification", JASA, 48, pp. 999-1009, (1970).

LILJENCRANTS, J. et LINDBLOM, B. - "Numerical Simulation of vowel quality systems: The role of perceptual contrast", Language, Vol. 48, nº 4, Déc., (1972).

LINDBLOM, B., LUBKER, J. et PAULI, S. - "An acoustic-perceptual method for the quantitive evaluation of hypernasality", Journal of Speech and Hearing Research, Vol. 20, nº 3, Septembre, (1977).

LINDBLOM, B. et SUNBERG, J. - "Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement", JASA, Vol. 50, nº 4, (1971).

PLOMP, R. - "Timbre as a multidimentional attribute of complex tones", In R. Plomp et G. F. Smoorenburg (eds.), Frequency Analysis and Periodicity in Detection in Hearing. Leider: Sijthoff, pp. 397--414, (1970).

POLS, L., TROMP, H. et PLOMP, R. - "Frequency analysis of Dutch vowels from 50 male speakers", JASA, Vol. 53, pp. 1093-1101, (1973).

SCHROEDER, M., ATAL, B. et HALL, J. - "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception", (1978).

STEVENS, K. - "Acoustic Correlates of Place of Articulation for Stop and Fricative Consonants", QRP, nº 89, Research Laboratory of Electronics, M.I.T., pp. 119-205, (1968).

STEVENS, K. - "The Quantal Nature of Speech:Evidence from Articulatory - Acoustic Data", in E.E. David,Jr et P.B. Denes (eds.), Human Communication:A Unified View (Mc Graw-Hill Publ. Co.,New York,1972).

STEVENS, K. et HOUSE, A. - "Speech Perception", in Foundations of Modern, Auditory Theory, Vol. II, Chap. 1, 1-57, Jerry V. Tobias (ed.), Academic Press, New York, Londres, (1972).

ZWICKER, E. - "Subdivision of the audible frequency range into critical bands (frequenz-gruppen)", JASA, Vol. 33, 248, (1961).

ZWICKER, E. et SCHARF, B. - "A model of loudness summation", Psychology Rev. 72, pp. 3-26, (1965).

TABLE DES MATIERES

Imprimerie de l'Université des Sciences Humaines de STRASBOURG

Dépôt légal : 2ème trimestre 1980